PhD Dissertation

International Doctoral School in Information and Communication Technology

DISI - University of Trento

Machine Learning for Tract Segmentation in dMRI data

Nguyen Thien Bao

Advisor:

Prof. Paolo Avesani

Co-Advisor:

Dr. Emanuele Olivetti

Università degli Studi di Trento

April 2016

Abstract

Diffusion MRI (dMRI) data allows to reconstruct the 3D pathways of axons within the white matter of the brain as a set of streamlines, called tractography. A streamline is a vectorial representation of thousands of neuronal axons expressing structural connectivity. An important task is to group the same functional streamlines into one tract segmentation. *This work is extremely helpful for neuro surgery or diagnosing brain disease. However, the segmentation process is difficult and time consuming due to the large number of streamlines (about* 3×10^5 *in a normal brain) and the variability of the brain anatomy among different subjects. In our project, the goal is to design an effective method for tract segmentation task based on* machine learning *techniques and to develop an interactive tool to assist medical practitioners to perform this task more precisely, more easily, and faster. First, we propose a design of interactive segmentation process by presenting the user a clustered version of the tractography in which user selects some of the clusters to identify a superset of the streamlines of interest. This superset is then re-clustered at a finer scale and again the user is requested to select the relevant clusters. The process of re-clustering and manual selection is iterated until the remaining streamlines faithfully represent the desired anatomical structure of interest. To solve the computational issue of clustering a large number of streamlines under the strict time constraints requested by the interactive use, we present a solution which consists in embedding the streamlines into a Euclidean space (call dissimilarity representation), and then in adopting a state-of-the art scalable implementation* *of the k-means algorithm. The dissimilarity representation is defined by selecting a set of streamlines called prototypes and then mapping any new streamline to the vector of distances from prototypes. Second, an algorithm is proposed to find the correspondence/mapping between streamlines in tractographies among two different samples, without requiring any transformation as the traditional tractography registration usually does. In other words, we try to find a* mapping *between the tractographies. Mapping is very useful for studying tractography data across subjects. Last but not least, by exploring the mapping in the context of dissimilarity representation, we also propose the algorithmic solution to build the common vectorial representation of streamlines across subject. The core of the proposed solution combines two state-of-the-art elements: first using the recently proposed tractography mapping approach to align the prototypes across subjects; then applying the dissimilarity representation to build the common vectorial representation for streamline. Preliminary results of applying our methods in clinical use-cases show evidence that our proposed algorithm is greatly beneficial (in terms of time efficiency, easiness.etc.) for the study of white matter tractography in clinical applications.*

Keywords

Machine Learning, Tract Segmentation, Brain Connectivity, dMRI Data, Neuro Imaging

Acknowledgements

I wish to express my sincere appreciation and thanks to my advisor professor Dr. Paolo Avesani, for the excellent guidance and advice during the entire course of my PhD. His wise academic advice and ideas have played an extremely important role in the work presented in this thesis. Without Prof. Paolo's support, this thesis would not have been possible. I also would like to be indebted the help of my co-advisor Dr. Emanuele Olivetti. His research suggestions and encouragement have been endless source of inspiration and support for my work.

I am also grateful to professor Dr. Lauren O'Donnell at Harvard University, for being so welcoming during my stay in the United States. Your advice on both research as well as daily life have been priceless. I would furthermore like to thank Dr. Eleftherios Garyfallidis for all his patience when explaining the basics of dMRI data, tractography, etc, at my very early step to work in this field. His constant availability for questions both in general theory and detail techique has been a great support for finishing this dissertation.

A big thank you to Nivedita Agarwal, Neuroradiologist at S. Maria del Carmine Hospital, and Assistant Professor of Neuroradiology, School of Medicine, who helped me with designing and collecting the dMRI data sets used in this work.

I would like to thank my friends and colleagues in NiLab - in particular, Dr. Sandro Vega Pons, and Dr. Vittorio Lacovella for their inspirations and discussions.

A special thank to ICT, Doctoral School and CiMec of Trento University, especially two secretaries of ICT, for their assistance during the past three years I stay in Italy.

My great acknowledgement to my master Ching Hai, for her uncon-

ditional love and spirit guidance at the most difficult moments of my life.

Finally, I would like to thank my family for their everlasting love and support. I want to give this dissertation as a present to my Father who always looks after my steps in life.

Nguyen Thien Bao
Trento, 20 Feb. 2016

Publications

[1] Emanuele Olivetti, **Thien B. Nguyen**, and Eleftherios Garyfallidis. *The approximation of the dissimilarity projection*. The 2nd IEEE International Workshop on Pattern Recognition in NeuroImaging, 0:8588, 2012.

[2] Emanuele Olivetti, **Thien B. Nguyen**, and Paolo Avesani. *Fast Clustering for Interactive Tractography Segmentation*. The 3rd IEEE International Workshop on Pattern Recognition in NeuroImaging, 2013.

[3] **Thien B. Nguyen**, Emanuele Olivetti, and Paolo Avesani. *Multiple-Scale visualization of large data based on hierarchical clustering*. International Journal of Computer and Electrical Engineering, 6(2):7782, 2014.

[4] **Thien B. Nguyen**, Emanuele Olivetti, and Paolo Avesani. *Mapping Tractography Across Subjects*. NIPS Workshop on Machine Learning and Interpretation in Neuroimaging, MLINI 2014.

[5] Diana Porro-Munoz, Emanuele Olivetti, Nusrat Sharmin, **Thien Bao Nguyen**, Eleftherios Garyfallidis, and Paolo Avesani. *Tractome: A Visual Data Mining Too for Brain Connectivity Analysis*. International Journal of Data Mining and Knowledge Discovery (DAMI), 2014.

[6] Paolo Avesani, **Thien Bao Nguyen**, Nivedita Agarwal, Mark Bromberg, Lubdha Shah, and Emanuele Olivetti. *Tractography Mapping for Dissimilarity Space Across Subjects*. The 5th IEEE International Workshop on Pattern Recognition in NeuroImaging (PRNI 2015), June 10-12 2015, Stanford University, CA, USA.

Posters

[1] Eleftherios Garyfallidis, Stephan Gerhard, Paolo Avesani, **Thien B. Nguyen**, Vassilis Tsiaras, Ian Nimmo-Smith, and Emanuele Olivetti. *A software application for real-time, clustering-based exploration of tractographies*. In 18th Annual Meeting of the Organization for Human Brain Mapping (OHBM), 2012.

[2] **Thien B. Nguyen**, Paolo Avesani, and Lauren O'Donnell. *Comparison of voxel-based and tract-based affine registration for tract segmentation*. In 20th Annual Meeting of the Organization for Human Brain Mapping (OHBM) 2014, pages 1858, June 2014.

[3] Paolo Avesani, Emanuele Olivetti, **Thien Bao Nguyen**, Nusrat Sharmin, and Nivedita Agarwal. *White-Matter Alignment Across Subjects by Tractography Mapping*. In 21th Annual Meeting of the Organization for Human Brain Mapping (HBM 2015).

Software

Tractome: A Visual Data Mining Tool for Brain Connectivity Analysis

http://tractome.org/

https://github.com/FBK-NILab/tractome

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 The Context

The brain, the central part of the nervous system, consists of the grey matter, known as cerebral cortex, and the white matter. Its core components are the neurons (nerve cells), that are in-charge of all the communication and processing within the brain. Neurons are divided into three main parts: cell body, dendrites and axons. The grey matter is composed of dense concentrations of the cell bodies and dendrites of these neurons and all the processing of the brain takes place here. On the other hand, the white matter works as the brain's connective cabling. It is composed of billions of myelinated axons that connect, i.e. transmit signals between neurons in different regions of the brain [25]. The patterns and structures of these anatomical links between regions in the brain are known as anatomical connectivity [55] [17] of white matter. Anatomical connectivity can vary among people if, for example, they have mental disorders, neurologic or neuropsychiatric diseases. Therefore, research about the anatomical connectivity of the white matter hence becomes essential in neuroscience and is also the main focus of this work.

Currently, diffusion magnetic resonance imaging (dMRI), a non-invasive technique, is popularly used to find the anatomical connectivity in brain [6,

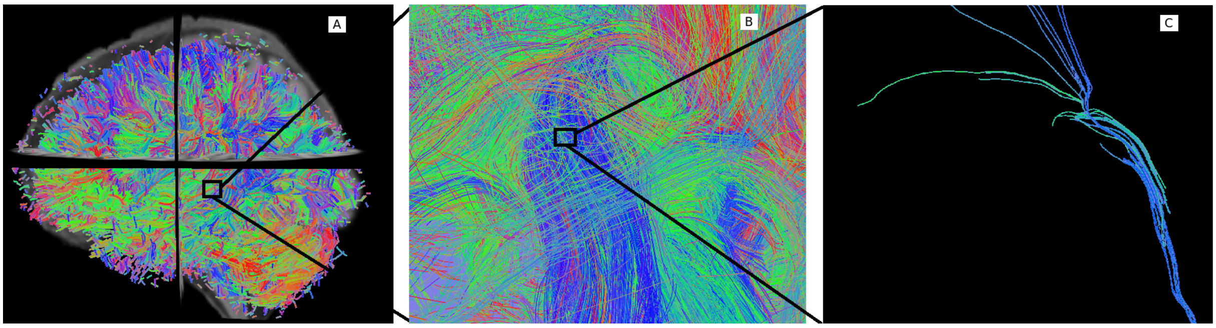


Figure 1.1: (A) Tractography overlaid with the structural image (only 10% of the streamlines are shown). The colour encodes the orientation of the mid-segment of every streamlines using a colour map based on [23]. (B) Amplifying an area of the tractography. (C) Small subset of streamlines.

99]. It measures the displacement distribution of water molecules in the brain tissue, that is mechanically constrained by the myelinated axons. Thereby, it provides information about the local orientation of white matter axons. The data obtained with this technique, can be used to extract the anatomical connectivity information by using deterministic tractography algorithms [69, 56, 30]. These algorithms reconstruct the approximate trajectories of the axons as polylines, so they resemble the white matter anatomical structures (see Figure 1.1). A polyline in this context is called *streamline*, and the full brain streamlines are called *brain tractography*. It is worth to notice that one streamline represents approximately 10^4 neural axons sharing the same structural connectivity path. In some cases, *streamline* is also called ***track*** or ***fiber***. The whole set of *streamlines* of a brain is called ***tractography***. And given that the resolution of modern MRI scanners is in the order of $1mm^3$, a full brain tractography consists of $\approx 3 \times 10^5$ streamlines (see Figure 1.1). ***Bundle*** is a set of streamlines with similar spatial and shape characteristics e.g. they are close to each other according to a streamline distance, while ***tract*** is the real anatomical group of neuronal axons.

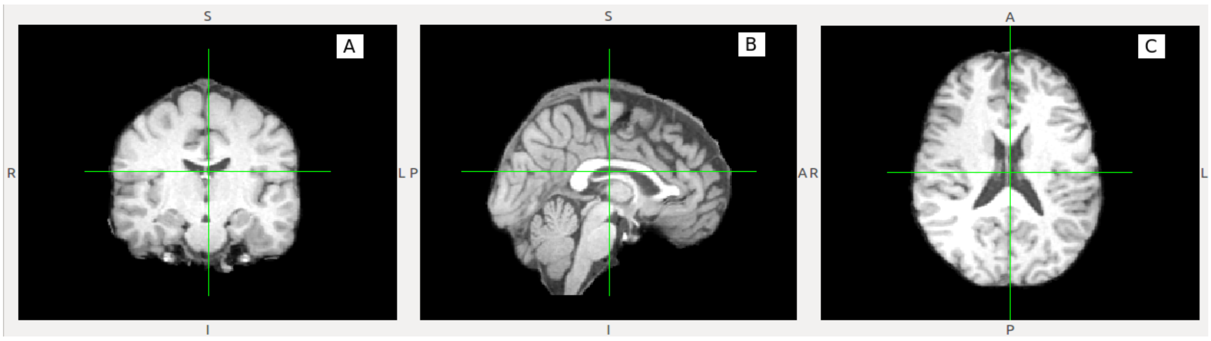


Figure 1.2: The structural image of the brain with different type of views. The 2D views: (A) coronal, (B) sagittal, (C) axial

The exploration of tractography data sets has hence become very useful to neuroanatomists. Information like the shape of streamlines, their spatial location and the relation with each other, allows to identify and study the subsets of streamlines related to specific function(s). From there, it can be also determined if there is (or the status of) an ongoing neurodegenerative process.

With these data, there are two main approaches for the study of anatomical connectivity: automatic and manual. The automatic analysis has gained popularity over the last few years. It is based on machine learning and data mining algorithms. It is mainly aimed at a fast segmentation of the white matter into sets of streamlines that follow similar trajectories [105, 39, 91]. Nevertheless, the automatic segmentation of the tractography is not always in agreement with the real anatomical structures of the white matter. Therefore, neuroanatomists still strongly rely on their manually guided visual exploration. This manual task though, is complex and slow. The manual exploration of the streamlines is usually supported by the overlaid structural image, such that experts can orient themselves into specific regions of the brain they are focusing their analysis (Figure 1.1). Moreover, the number of streamlines can be really large, usually in the order of hundreds of thousands, making the exploration i.e. shape recognition, spatial localization, quite difficult. See for example, in Figure 1.1.A, where only a 10% of the total amount of streamlines is shown, it is still difficult to visually understand the data.

1.2 The Problem

Recently, the literature about machine learning techniques to apply for analyzing and studying the white matter tractography is increasing. Although it has gained some encouraging results, but these results are still under the satisfactory of medical practitioners. In this work, we want improve the support of machine learning techniques for studying the white matter tractography. We want to help the medical practitioners to analyse the white matter tractography data more easily and more accurately based on machine learning techniques. The things that we want to investigate in this project are :

- **Brain tractography segmentation:** Traditionally the segmentation task is done by neuroanatomists and it consumes a lot of time and effort due to the large number of streamlines (about 3×10^5 in a normal brain). Moreover, the variability of the brain anatomy among different subjects makes the segmentation a difficult task [12]. Up to now, there are two machine learning approaches for tractography segmentation: supervised [15] and unsupervised [34] learning. The unsupervised techniques often rely on expert-crafted streamline-streamline distance functions [26, 111] encoding informative relationships for the segmentation task, then followed by a clustering algorithm (agglomerative, k-means, Gaussian mixture model, etc. see [105] for a recent brief review). Supervised tract segmentation [15, 75] instead aims at learning how to segment the tractography from expert-made examples provided as input. Although both supervised and unsupervised techniques get some encouraging results, but they are below the expectation of medical practitioners. Unsupervised techniques usually work on the whole tractography while medical practitioner often focus on a specific tract. In the case of supervised learning, the lack of ground truth data makes the results not good and need the refinement from experts. Although both supervised and unsupervised learning have gain some encouraging results, but they are still under the satisfactory of medical practitioners. In this work, we try to assist the medical practitioners to do the tract segmentation task more accurately, more easily in order to improve the quality of the segmentation.

- **Vectorial representation for tractography streamline:** Most of the machine learning and patern recognition techniques used for tractography analyses (such as supervised and unsupervised learning for tractography segmentation, clustering for tractography visulaization, ...) require the input to be from a vectorial space. This requirement contrasts with the intrinsic nature of the tractography because streamlines have different lengths and different number of points and for this reason they cannot be directly represented in a common vectorial space. This lack of the vectorial representation avoids the use of some of those algorithms and of computationally efficient implementations. In this thesis, we try to define a new representation for tractography streamline that can be fed to the most machine learning technigues.

- **Tractography registration:** Initially, tractography is originally in the space of scanner with different coordinate, it is necessary to align tractography from different brains together or combine different tractographies of same brain using different image registration techniques, for further analysis.

- **A common vectorial representation for streamline across subjects:** Current neuroscientific analyses of white matter tractography data are limited to qualitative intra-subject comparisons. It is then quite difficult to use the information for direct inter-subject comparisons [37, 7]. Thus, when applying machine learning techniques for inter-subject tractography analyses, it leads to the need of defining a common vectorial representation for tractography streamlines not only intra-subject but also across subjects.

1.3 The Solution

In this part, we shortly describe the solution for the problem mentioned before.

- **BOI (Bundle of Interest) based tractography segmentation:** The drawback of all the current tractography segmentation approaches is that they work on a large number of tracks and most of them are not interested to medical practitioners; or they focus on a target tract but the variance between brains makes it difficult to generate well. The results from both case are neccessary to be refined by experts. In this work, we want to overcome these disadvantage by proposing a framework using **BOI** (Bundles of Interest) contrasting with ROI (Region of Interest). While ROI concerns about which streamlines go through some interesting regions, BOI focuses only on streamlines inside some specific bundles. Because all of the current approaches only work on the tracks without caring the anatomy [102], it makes difficult to validate the result. Using BOI would make medical practitioners concentrate on which tracts they are working on, and of course these tracts also correlate to the anatomy.

- **Interactive visualization tractography:** In order to help medical practitioners to do the segmentation task more easily and quickly, we provide an interactive tool for visualization tractography data in $3D$ space. While all the current methods are off-line and medical practitioners can not interact or modify the result of segmentation, our tool is able to support them instantly to refine the segmentation result manually. This tool has to adapt to the real time responses of the user. This also differentiates our method from most of the current state of art approaches that do not adjust to user feedback.

- **Dissimilarity approximation:** The dissimilarity space representation could be the way to provide a vectorial representation for streamline, and for this reason it is crucial to assess the current machine learning techniques that require the input to be from a vectorial space. Actually, the dissimilarity representation is an Euclidean embedding technique defined by selecting a set of objects (e.g. a set of streamlines) called *prototypes*, and then by mapping any new object (e.g. any new streamline) to the vector of distances from the prototypes. This representation [88, 5, 16] is usually presented in the context of classification.

- **Tractography mapping:** Tractography registration is most often performed by applying the transformation resulting from the registration of other images, such as $T1$ or fractional anisotropy (FA), to tractography [38, 106, 37, 113]. In contrast to all current tractography registration methods based on rigid or non-rigid shape transformation of one tractography into another, we suggest to find which streamline of one tractography corresponds to which streamline in the other tractography, without transformations. This correspondence is a *mapping* from one tractography to the other. We propose to solve the problem of finding the mapping between two tractographies through a graph-based approach similar to that of the well-known graph matching problem [18, 109] in pattern recognition literature .

- **A common vectorial representation for streamline across subjects:** By exploring the tractography mapping idea in the context of dissimilarity representation, we propose a new common vectorial representation for streamlines across subjects. This representation, as far as we know, is the first approach that create a common space for representing streamlines from multiple subjects without requirement of co-registering subjects in the same space.

1.4 Innovative Aspects

This research is motivated to support medical practitioners to analyse and study the brain white matter tractography more easily and accurately. Results of tractography studying are immediately applicable to surgical intervention, and to the treatment of psychological and psychiatric disorders. The main contributions of this thesis are the following:

- First, we design an effective method for tract segmentation task using *machine learning* based on BOI approach.

- Second, we present a solution to solve the computational issue of clustering a large number of streamlines under the strict time constraints requested by the interactive use. The solution consists in embedding the streamlines into a Euclidean space using dissimilarity representation technique, and then in adopting a state-of-the art scalable implementation of the k-means algorithm.

- Third, we propose a methodology to map the tractography from one subject to another subject, i.e to find the correspondence of streamlines between two different tractographies without co-registering tractographies together.

- Fourth, based on exploring the dissimilarity representation idea in the context of tractography mapping, we are able to build up a common vectorial representation for streamline across subjects with high accuracy and low computational cost.

- Fifth, we develop a scientific interactive visualization tool, the implementation of the framework that we propose for tract segmentation task, to help medical practitioners to perform this segmentation task more precisely and easily based on BOI approach.

1.5 Structure of the Thesis

Chapter 2 presents the state of the art of the current white matter tractography analysis. The first part introduces the dMRI technique and how to reconstruct the tractography from dMRI data. The analysis of tractography is subdivided into two parts: tractography segmentation and tractography registration. In tractography segmentation section, we present the overview of the current segmentation methods, and point out some limitation of these methods. The later part describes the registration approaches for tractography including voxel-based and tract-based method.

Chapter 3 introduces the first main contribution, the dissimilarity approximation for tractography. The proposed approach solves the problem of how to represent streamlines, which have different number of 3D points and differ in sizes, in an Euclidean space. This work is motivated by practical applications about executing common algorithms, like spatial queries, clustering or classification, on large collections of objects that do not have a natural vectorial space representation (i.e streamlines in our case). The lack of the vectorial representation of streamlines avoids the use of some of those algorithms and of computationally efficient implementations. The dissimilarity space representation could be the way to provide such a vectorial representation. The dissimilarity representation is an Euclidean embedding technique defined by selecting a set of objects (e.g. a set of streamlines) called *prototypes*, and then by mapping any new object (e.g. any new streamline) to the vector of distances from the prototypes. The use of a stochastic approximation of an optimal algorithm for prototype selection is also discussed in this chapter. Finally, we provide practical examples both from simulated data and human brain tractographies, and confirm that dissimilarity approximation is able to provide a fast and accurate vectorial representation for tractography.

Chapter 4 proposes an alternative way of the traditional tractography registration. Tractography registration is most often performed by applying the transformation resulting from other images (T1, FA, DTI) to tractography data, or to register tractographies themselves. However, the above methods can not deal with a new coming tractograhy, except for running the whole registration process again with all data plus the new comer. In contrast with the registration methods, instead of finding the transformation between tractographies, in this work, we want to directly map the *source* tractography to the *target* tractography. We believe to be the first to recast the problem as mapping problem rather than registration problem. By taking advantage of more than thirty year graph-matching research, we propose a graph-based solution for tractography mapping problem and explain similarities and differences with the well-known graph matching problem. We define the loss function based on the pairwise streamline distance, and reformulate the mapping problem as the problem of minimizing that loss function. To our knowledge, this is also the first graph-matching-based objective function applied to tractography. Moreover, we propose an algorithm for building the common vectorial representation for streamlines across subject. The core idea is to combine the dissimilarity representation with tractography mapping. Tractography mapping allows to find the correspondence between streamlines across subjects, while dissimilarity representation is able to build an Euclidean representation for streamline. We apply the proposed algorithm in the context of tractography segmentation. Experiments using real dMRI data demonstrate the potential of the proposed method for medical or neuroscientific analyses of white matter tractography data.

Chapter 5 describes a novel interactive system for human brain tractography segmentation to assist neuroanatomists in identifying white matter anatomical structures of interest from dMRI data. The difficulty in segmenting and navigating tractographies lies in the very large number of reconstructed neuronal pathways, i.e. the streamlines, which are in the order of hundreds of thousands with modern dMRI techniques. The novelty of our system resides in presenting the user a clustered version of the tractography in which he/she selects some of the clusters to identify a superset of the streamlines of interest. This superset is then re-clustered at a finer scale and again the user is requested to select the relevant clusters. The process of re-clustering and manual selection is iterated until the remaining streamlines faithfully represent the desired anatomical structure of interest. The computational issue of clustering a large number of streamlines under the strict time constraints requested by the interactive use, is solved by embedding the streamlines into a Euclidean space and then in adopting a state-of-the art scalable implementation of the k-means algorithm. We tested the proposed system on tractographies from amyotrophic lateral sclerosis (ALS) patients and healthy subjects that we collected for a forthcoming study about the systematic differences between their corticospinal tracts. The latter part of this chapter contains the demonstration of the usefulness of our proposed interactive visualization tractography segmentation software tool in the neuroscientific analyses activities. The first one is to study the characterisation of the amiotrophic lateral sclerosis (ALS) disease through the corticospinal tract. The second one uses the result of tract segmentation for validation two tractography registration methods, voxel-based and tract-based method.

Chapter 6 concludes the thesis work.

Chapter 2

State of the Art

Neuroimaging techniques allow researchers and clinicians to gain insights of unprecedented quality on the cerebral anatomy. Three main focuses of neuroimaging include *brain decoding*, *brain mapping* and *brain connectivity*. Both brain mapping and brain decoding concern about the prediction or detection of a cognitive stimulus given a recording brain activity and reverse. Contrast with them, brain connectivity tries to build a model of the connections between different brain areas. *Functional connectivity* focuses on the correlation between the brain activity of anatomically remote areas. *Effective connectivity* investigates on finding a causal link between different brain structures. *Anatomical connectivity* refers to the structural links between different areas that develops in the white matter of the brain, and it also the main focus of this work. In this chapter, we review the state of the art of some activities that study the anatomical connectivity using the diffusion magnetic resonance imaging (dMRI) data.

2.1 Diffusion magnetic resonance imaging (dMRI) data and deterministic tractography

DMRI data allow to reconstruct the 3D pathways of axons within the white matter of the brain as a set of streamlines, called tractography. A streamline is a vectorial representation of thousands of neuronal axons expressing structural connectivity. In this part, we will discuss more detail of the pipeline to reconstruct the tractography from raw dMRI data.

2.1.1 From raw data to NIfTI format

Most of the dMRI scanner produces data in DICOM format (.dcm - Digital Imaging and Communications in Medicine). DICOM is the most common standard for receiving scans from a hospital [10, 67, 70]. The DICOM standard was created by the National Electrical Manufacturers Association (NEMA)[1] to aid the distribution and viewing of medical images, such as CT scans, MRIs, and ultrasound. A single DICOM file contains both a header (which stores information about the patient's name, the type of scan, image dimensions, etc), as well as all of the image data (which can contain information in three dimensions) [45].

Figure 2.1 shows an example of the hypothetical DICOM image file. In this one, DICOM header uses the first 794 bytes to describe the image dimensions and retain other text information about the scan. The size of this header varies depending on how much header information is stored. For example, in the Figure 2.1, the header defines an image which has the dimensions $109 \times 91 \times 2$ voxels, with a data resolution of 1 byte per voxel, and the total image size will be 19838. Following the header is the image data, and both the header and the image data are

[1] `http://dicom.nema.org/`

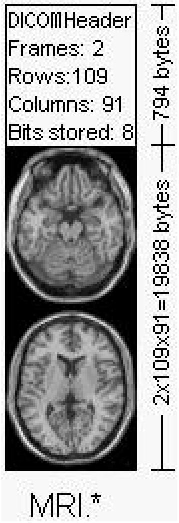


Figure 2.1: An example of the DICOM image file. The image is reproduced from `http://www.cabiatl.com/mricro/dicom/`

stored in the same file. More information about DICOM format can be found on the official webpage of DICOM [2].

Although DICOM is the most common standard for receiving scans, it is quite complex format, and difficultly to be understood. DICOM data, thus, needs to be converted in the format of NIfTI (Neuroimaging Informatics Technology Initiative) [108]. NIfTI is a modern incarnation of the Analyze format, but includes important information like the orientation of the image [22]. It was for scientific analysis of brain images [3]. The images can be stored as a pair of files (hdr/img, compliant with most Analyze format viewers), or a single file (nii). Many tools like FSL [49], NiBabel[4], MRIcron[5], ...are also able to read compressed (nii.gz) images. NIfTI format attempts to keep spatial orientation information, therefore, it should reduce the chance of making left-right errors.

[2] `http://medical.nema.org/`

[3] `http://nifti.nimh.nih.gov/`

[4] `http://nipy.org/nibabel`

[5] `http://www.nitrc.org/projects/mricron`

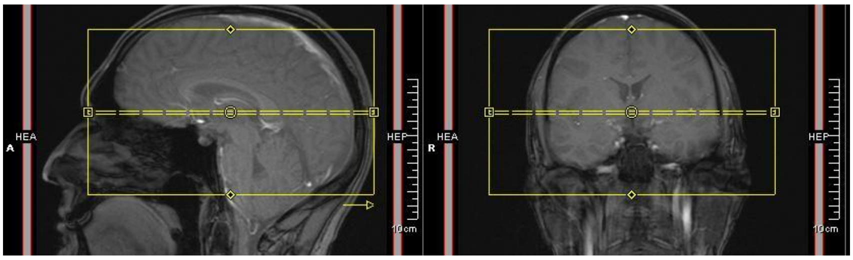


Figure 2.2: Image acquired orthogonal to scanner bore

When converting image from DICOM format to NIfTI format, beside the NIfTI file image, most of DICOM image conversion tools also generate (.bvec) and b-value (.bval) text files (contains diffusion gradient vector and the b-value). These files are very important to reconstruct diffusion properties. Because in diffusion tensor imaging (DTI) method, we construct tensors by collecting a series of direction-sensitive diffusion images [6]. Therefore, in addition to recording the images, the scanner also saves these directions. A potential concern is that the scanner manufacturers can choose to either report the vectors with reference to the scanner bore, or with reference to the imaging plane (i.e., imaging grid). This is not a problem if the images are always acquired precisely orthogonal to the scanner bore (Figure 2.2), as the image and scanner have the same frame of reference. However, problems can arise when the image plane is not aligned with the scanner bore (i.e., oblique acquisitions). In this situation, it is important to ensure that these vectors are in the same frame of reference as the image. Moreover, the eigenvectors of the tensor, and consequently tractography programs are sensitive to proper interpretation of the bvecs relative to the imaging plane.

2.1.2 Reconstruction

From the dMRI data, tractography is created in two steps: reconstruction and tracking. Reconstruction is about computing the information about the spatial distribution of the diffusion signal within each voxel. While tracking tries to connect many signals to form a tractography based on orientation signal of each voxel.

It is usually to extract brain image only from the actual dMRI data in NIfTI images before doing reconstruction, because the result of scanner contains not only brain but also other things close to brain which can distract the processing of tracking. Brain extraction is the process of removing the skull and the rest of the head from the brain (see Figure 2.3). The resulting file only contains a representation of the brain's anatomy.

Brain extraction can be done with the FSL [6] program BET (Brain Extraction Tool) [96]. BET takes an image of a head and removes all non-brain parts of the image. It uses a deformable model that evolves to fit the brain's surface by the application of a set of locally adaptive model forces. This method is fast and requires no preregistration or other pre-processing before being applied. Result of BET is a file saved with a brain extension at its end. An example of BET result can be seen in the bottom line of the Figure 2.3, while original image data is at the top.

After brain extraction, we can do reconstruction step. The main purpose of this is to estimate the orientation information from the diffusion signal within each voxel which is adequate for accurate tractography generation. In the last few years, there has been an increasing number of techniques which are proposed to recover the signal directions inside the voxel from dMRI data, and the most simple one is Diffusion Tensor Model [6]. But in many cases this model is not sufficiently [2], because most of voxels inside brain contain multiple streamline bundles crossings while this model is only working with single tensor. Many other reconstruction methods have been proposed to overcome the limitations of this Diffusion Tensor model, such as Diffusion Spectrum Imaging [14]

[6] `http://www.fmrib.ox.ac.uk/fsl/index.html`

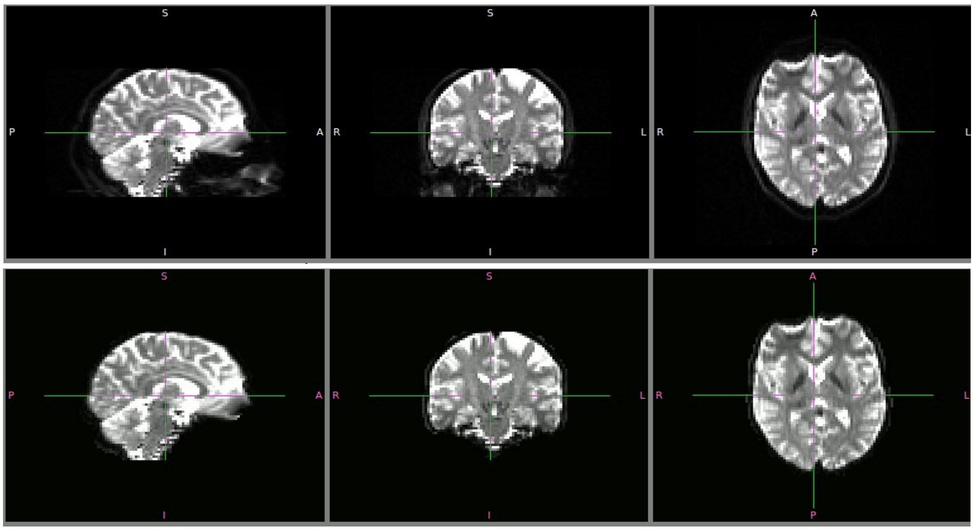


Figure 2.3: An example of the dMRI data after doing brain extraction. Top: the original NIfTI images. Bottom: the result of doing brain extraction

or Higher Order Tensors [80]. The overview of these model can be found more detail in [42].

To visualize the 3D diffusion data, [42] proposed two approaches. The first one is to replace the displacement distribution with an isosurface, which is a surface that passes through all points of equal probability density value. And the second one is to compute the orientation distribution function (ODF) from the displacement distribution. Figure 2.4 represents these two approaches.

2.1.3 Tracking

In the previous part 2.1.2, information about orientation of streamlines at every voxel has been gained. Based on these information, tractography algorithms (or tracking algorithms) can be used to join these directions up to reconstruct complete tracks and hence approximate anatomical tracts. This processing is known as tracking, which connect voxels in order to create *tracks* (or *fibers*, *streamlines*), using the spatial information computed during the reconstruction step. Basically, tracking proce-

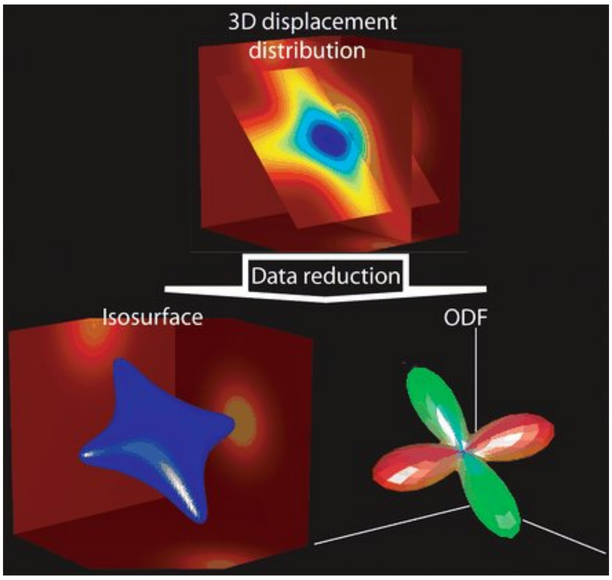


Figure 2.4: Two approaches maybe used to simplify the visual representation of 3D diffusion data. Top: the reconstruction of the 3D displacement probability distribution from the diffusion signal. Left: the replacement of the displacement distribution with an isosurface. Right: the computation of the commonly used Orientation Distribution Function (ODF). This displacement distribution simulates the crossing of two fibres. In general, the ODF is used essentially to identify the primary directions of the underlying fibres. Picture is reproduced from [42]

dure consists of starting at a seed location and following the preferred direction until we reach a new voxel. Then, we can change to this voxels referred direction and continue until an entire track is propagated. An example about creating track from orientation of streamline signal within voxels is presented in Figure 2.5

Tracking algorithms can be grouped in three categories: local, global and simulated. The local approaches tries to propagate a strealine from a starting (seed) point using locally greedy criteria, i.e. tracking sequentially through orientation estimates in adjacent voxels [24]. The global ones find the best path between two points of interest, based on some optimization criterion, rather than identifying paths arising from a single point [57, 46]. The simulated approaches simulate the diffusion process or solve the diffusion equation to reconstruct tracks [40, 51]. Due to

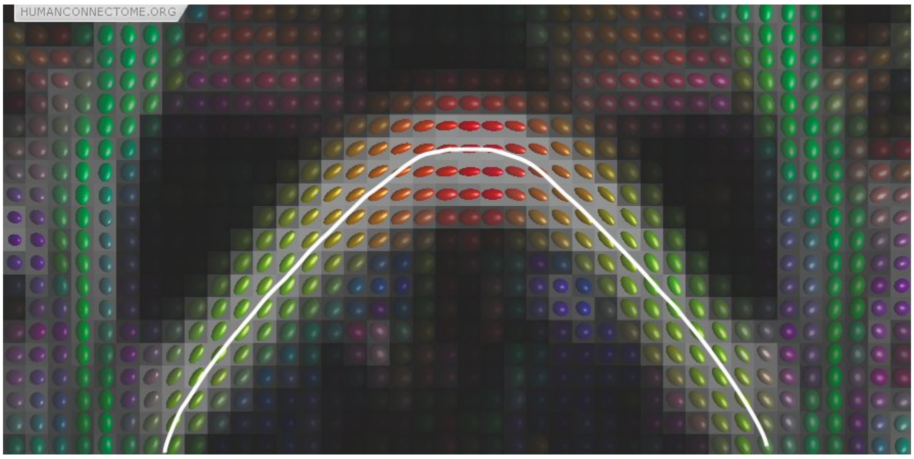


Figure 2.5: Tracking from tensor direction information. The white line shows the streamline obtained by joining a set of voxels based on their diffusion direction information. The color is a complementary way of coding the preferred direction where red denotes left-right, green denotes back-front and blue denotes up-down. Picture is produced from [30]

the need of creating the whole brain tractography, only local techniques are described in this part.

In local techniques, deterministic and probabilistic tracking algorithm are the two best known families [24]. Deterministic fiber tracking uses the principal direction of diffusion to integrate trajectories over the image, and to make a series of discrete locally optimum decisions [69]. It is fast, simple and easy to interpret. The disadvantages of deterministic algorithms are that a pathway either exists or not (no uncertainty) and that they do not explore the entire space of possible white matter tracts. Probabilistic procedure is to calculate a spatial distribution of tracks arising from a single seed rather than a single track. It considers the tensor as a probability distribution of fiber orientation [8, 41, 83, 9]. An example about the difference between deterministic and probabilistic is shown in the Figure 2.6.

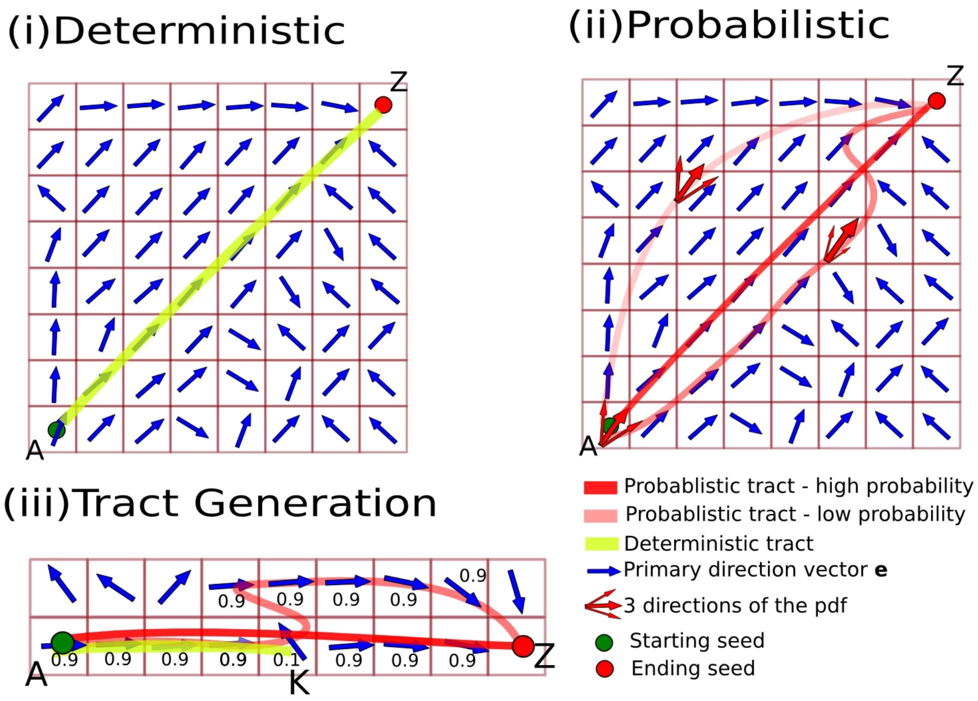


Figure 2.6: A simplified example showing in (i) and (ii) the same data set. (i) The yellow line shows the result of deterministic tractography which is given by a single trajectory and in (ii) is given by connectivity matrix depicting in red the probability of different pathways throughout the hole slice. For the ease of understanding, only 3 possible pathways are depicted. Finally, in (iii) an example is given where it is shown that probabilistic tractography weights more closer connections. However, it can track further deep than deterministic tractography. Picture is produced from [30]

2.2 Tractography segmentation

From the dMRI data, by using *fiber-tracking algorithms*, we can extract the structural connectivity information, called *tractography*, of the brain. However, the resulting tractography datasets are highly complex and include thousands of fibers (about $\approx 3 \times 10^5$), which requires techniques or method to create the exact anatomic brain before doing further studying. The **tractography segmentation** aims at grouping some fiber tracts belonging to a common anatomical area, into one segmentation, and it is a task of interest in neurological studies [12], for example for the study of

Alzheimer disease. Traditionally, the segmentation task is done by neuroanatomists, and it consumes a lot of time and effort due to the large number of streamlines (about 3×10^5 in a normal brain). Moreover, the variability of the brain anatomy among different subjects makes this task more difficult. Furthermore, clinical studies often use the segmentation of white matter bundles in order to perform comparisons between populations, and thus, it is also an press on the accuracy of the segmentation task. Therefore, . Recently, there is a rise of applying pattern recognition techniques to solve this problem [73, 111, 75], however the segmentation of tractography is still not completely solved problem. In the following part, the brief survey about currently trends in segmentation tractography is presented.

Atlas approach: Atlas are the models of white mater structure in brain. Firstly, atlas are created from experience of experts without being driven from data. After that, atlas are used as model of clusters for tractography segmentation. All streamlines would be grouped into the closest cluster in atlas. O'Donnell and Westin [73] generated a tractographic atlas using spectral embedding and expert anatomical labeling. They then automatically segmented the new tractography using again spectral clustering and embedding the tracks as points in the embedded space, to the closest existing atlas clusters. The true affinity matrix was too big to compute therefore they used the Nystrom approximation: working on a subset and avoid generating the complete distance matrix. However, the important information from the full data set may be lost after sub-sampling.

ROI - region of interest: One of the first idea for segmentation is to use the region of interest (ROI) [102]. This approach tried to reconstruct tracts passing through ROI by exploiting existing anatomical knowledge of tract trajectories. First, some target tracts must be defined. It also requires to specified manually some regions where tracts start, end or pass through. Then streamlines would be filtered based on the constraint of passing through ROIs. ROI approach needs a prior knowledge about the trajectory and is used only for well-characterized white matter tracts. In order to refine the segmentation, multi-ROIs were used to include or exclude tracks.

Unsupervised learning: From the point of view of algorithmic approaches, the segmentation task has traditionally been addressed with unsupervised techniques over only diffusion data [111]. This typical framework first defines a pairwise distance between fibers and inputs the similarity matrix to standard clustering algorithms. Various distance functions between fibers have been proposed: the Euclidean distances between fiber shape descriptors [13]; the similarity between two fibers based on the number of points sharing the same voxel [50]; distance from the $B - spline$ representation [63]; closest point distance, mean of closest distances and Hausdorff distance [33]. Then, following is a clustering algorithm such as agglomerative, k-means, Gaussian mixture model, etc (see [105] for a recent brief review of applying these algorithm for tractography).

The disadvantage of these clustering algorithms is that they require manually specifying the number of clusters or a threshold to stop merging or splitting clusters. The different numbers of chosen clusters vary significantly the performance of clustering [68]. Recently, there are some approaches try to solve this problem by auto choosing the number of clusters. In [73], a large cluster number for spectral clustering is chosen, and then these clusters are manually merged to obtain models for white matter structures. Zvitia et al. [115] and Wassermannet et al. [107] decide the number of clusters based on mean-shift. By adding a penalty to a larger cluster number, Neji et al. [71] solved the optimization using linear programming to chose the number of clusters. Recently, Garyfallidis et al. [32] proposed a very quick clustering algorithm, called QuickBundles. It took one random streamlines as initial cluster, and calculated the distance from all the un-clustered streamlines to the representatives of clusters. Only the streamline with the minimum distance was grouped into the closest cluster if the distance was less than a given threshold, other while, that streamline became a new cluster.

Although these approaches avoid manually choosing number of cluster, the drawback is the high space and time complexity of computing pairwise distances between fibers. Whole brain tractography produces $\approx 3 \times 10^5$ streamlines fibers per subject, the pairwise distance between fibers is difficult to compute. And it becomes more serious when clustering fibers of multiple subjects. To avoid computing pairwise distances between fibers, Savadjiev et. al. [93] clustered diffusion orientation distribution functions maxima instead of clustering fiber tracts directly. This algorithm based on the geometric coherence of fiber orientations. Maddah et al. in [64] proposed a probabilistic approach to cluster fibers. It used a Dirichlet distribution as a prior to incorporate anatomical information. However, this algorithm also required establishing point correspondence which was difficult to define.

Supervised learning: The most disadvantage of unsupervised approach is that it works on the whole tractogrpahy and tries to cluster tractography into many tracts, while the requirement of medical practitioners only focuses on some specific tracts. Supervised segmentation is the method of partitioning according to some provided tract examples, therefore, it only focuses on a specific tract [63, 73] . Firstly, the target tracts should be specific, such as corticol spinal tracts (see Figure 5.13). Then, a repository of samples must be collected. A sample is an expert-made assignment of streamlines to the target tracts. These

samples are used to train a classify model, which is used to cluster a new streamline. In this setting, each streamline can be class-labelled as being member of the fiber tract of interest or not. For this reason the supervised segmentation problem becomes a binary classification problem. Maddah et al. [63] used the B-spline representation of the streamlines, and classified by the nearest-neighbor algorithm with respect to an atlas. Wang et al. [105] proposed a non-parametric Bayesian framework using a hierarchical Dirichlet processes mixture (HDPM) model, and the models of bundles were learned from how voxels are connected by fibers in training data instead of comparing fiber distances. Olivetti [77] combined both structural and functional connectivity to study jointly in a pairwise approach with the goal of assessing the contributions of structural information and functional information when segmenting the tracts. Recently, [75] solved this classification problem basing on the dissimilarity representation. After projecting all streamlines into some prototypes, one streamline-streamline distance function is computed in this new representation space, and it is used for classifying.
Although supervised approaches focus on a specific tract as requirement of medical practitioners. However, because the number of data for training and testing is very small due to the vague time for collecting enough the truth background data of manual segmentation tractography, the results usually are bellow the expectation of medical practitioners, and they need to be refined to use in clinical applications.

Most of these above methods often require the data to lie in a vectorial space, which is not the case for streamlines. Streamlines are polylines in 3D space and have different lengths and numbers of points. The lack of the vectorial representation avoids the use of some of those algorithms and of computationally efficient implementations. The dissimilarity space representation [88, 5, 16] could be the way to provide such a vectorial representation. The dissimilarity representation, a specific Euclidean embedding technique, is usually used in the context of classification and clustering problems. It is defined by selecting a set of objects (e.g. a set of streamlines) called *prototypes*, and then by mapping any new object (e.g. anynew streamline) to the vector of distances from the prototypes. It is a *lossy* transformation in the sense that some information is lost when projecting the data into the dissimilarity space. To the best of our knowledge this loss, i.e. the degree of approximation, has received little attention in the literature. In [86] the approximation was studied to decide among competing prototype selection policies only for classification tasks. In this work we are interested in assessing and controlling this loss without restriction to the classification scenario.

2.3 Tractography registration

Current neuroscientific analyses of white matter tractography data are limited to qualitative intra-subject comparisons. Thus, it is quite difficult to use the information for direct inter-subject comparisons [37, 7]. This leads to the need of initial alignment, or registration, of tractographies together via some methods before doing further study.

Registration is the problem of identifying the process of geometric transforming the coordinate system of an input image to be as spatially aligned to a reference image, more generally establishing a homology among the input images [43]. In this scenario, a group of transformations needs to be established to put all the inputs into correspondence [114]. Inter-modal registration allows precise spatial localization across images of the same subject but under different modalities, while intra-modal one tries to specify spatial localization across multiple subjects [37]. The most important transformation is the affine transformation, as in Equation 2.1, which has 12 degrees of freedom (DOF) in $3D$.

$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & l_{13} & t_x \\ l_{21} & l_{22} & l_{23} & t_y \\ l_{31} & l_{32} & l_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{2.1}$$

where l_{ij} are the nine parameters of a linear transformation, the t_i are the translation parameters in $3D$; x, y, and z are the input coordinates and x', y', and z' are the transformed coordinates. Any registration technique can typically be described by three components: a transformation, a similarity measure and an optimization. The transformation is applied to an input image to increase its similarity with the reference image. The similarity measure measures the similarity between the reference image and the input image after transforming. And the optimization algorithm iteratively determines the optimal transformation parameters as a function of the similarity measure. Image registration plays an important role in medical image analysis, group analysis and statistical parametric mapping. Because of its importance in both research and medical applications, medical image registration has been intensively investigated for almost three decades and numerous algorithms have been proposed. More detail can be found in the recent survey of medical image registration in [65].

Specific to the tractography registration, different authors classified tractography registration in different ways. Some classifications are based on what kind of registration techniques are used [59] and some are based on what kind of diffusion data are used [58]. If we consider from the point view of registration technique, it is basically based on different similarity measures: rigid, non-rigid registration; linear, non-linear registration and the feature based registration. According to the data type registration, there are three alternative approaches: scalar or vectorial registration, tensor registration and fiber or streamline registration. We choose to keep the registration based on data type to elaborate.

From scalar image based registration point of view, mutual information is used to measure the similarity between images. Affine co-registration along with mutual information is performed with diffusion weighted images [62]. Orientation information of the diffusion tensor preserves after affine transformation in order to align anatomical structure. Scalar registration are used at early stage of dMRI registration with the scaler images; without considering the directional images. More details can be found in [82, 35].

In tensor based method [92], FA (Fractional Anisotrophy) mapping or affine registration is applied on tractography along with tensorial value of the images. We can distinguished tensor based registration with the scalar registration by additional deformation model which keep the tensor orientation consistent according to the anatomical structure of the image. Direct and feature based methods are discussed in [92], where direct approach is based on Diffusion Tensor Constancy Constraint (DTCC) along with finite strain reorientation schema, and feature based method is based on singular value decompositions (SVD). Above described image based and tensor based model are voxel based registration and by considering the anatomical images, not on 3D reconstructed tractography.

Actually for the tractographiy data registration, no transformation is meaningful since it is not possible to reconcile two different anatomies by means of rigid (ornon-rigid) transformations. Also recent methods are voxel based and computationally very expensive due to voxel to vexel similarity measuring cost. And as it calculates the spatial transformation iteratively, it could suffer for local optima. That is the rea-

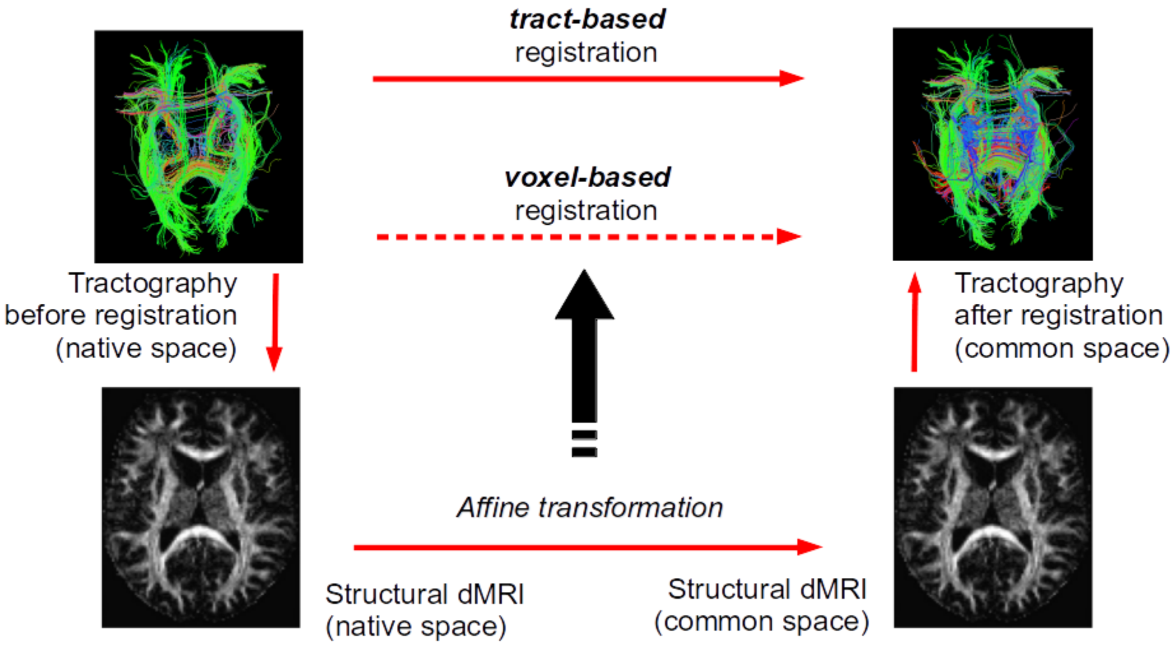


Figure 2.7: Tractography registration: voxel based method and tract-based method

son why the concept of streamline-streamline registration comes on the mind of researchers as the quality being optimized during registration are closely related to final goal of streamline registration. When we use the voxel based method, we have information about one voxel, but a streamline (i.e, its a set of voxel information) could be used for registration techniques. Hence, the problem is how to use the streamline information to register the whole brain tractrography. In fiber or streamline registration, the concept is to register the tractography in native space directly, without any prior knowledge of structural images and any kinds of transformation. Figure 2.7 demonstrates the overall concept of tractography registration both in image and streamline based. The solid lines show how images are converted from native to common space. Afterward, the affine transformation is used to wrap tractography in common space that is illustrated in the right upper corner of the figure.

In streamline registration, streamline is usually considered as a set of points [66, 59, 115]. Points are represented in high dimensional point space. In [66], Efficient Interactive Closest Feature point (ICF) is used to register different tractographies. Computational complexity on high dimensional search is handled by implementing approximate nearest neighbors techniques. In [115], streamlines are projected on high dimensional feature space with a 3D coordinates sequence. Fiber model is extracted by adaptive mean shift (AMS) clustering. Gaussian Mixture Model (GMM) is represented by assigning weight to each fiber model. The registration is performed as the alignment of two GMMs by maximizing the correlation ratio.

Recently another unbiased multi-subject registration is proposed in [72]. In that paper registrations are done by minimizing the entropy based objective function. Distance between the streamlines are calculated and represented by the Gaussian kernel distribution. This registration technique works with the whole brain with group wise registration.

In this work, instead of finding the shape transformation of one tractography into another, we try to find which streamline in one tractography correspond to which streamline in the other tractography, without any transformation. In other words, we try to find a mapping between the tractographies. The tractography mapping is similar to the well-known graph matching problem [18, 109] in pattern recognition literature. During the last decade, *graph matching* has paid a huge attention due to the application of it in modern scientific discipline and applied field [61]. The graph matching problem can be described as follows. An undirected weighted graph $G = (V, E)$ of size N is a finite set of vertices $V = \{1, \ldots, N\}$ and edges $E \subset V \times V$. Given two graphs G_A to G_B with the *same* number of vertices N, the problem of matching G_A and G_B is to find the correspondence between vertices of G_A and vertices of

G_B, which allows to align, or register, G_A and G_B in some optimal way. The correspondence between vertices of G_A and of G_B is defined as a *permutation* P of the N vertices, i.e. there a one-to-one correspondence between the two set of vertices. P is usually represented as a binary $N \times N$ matrix where P_{ij} is equal to 1, if the ith vertex of G_A is matched to the jth vertex of G_B, otherwise 0.

In literature, efficient algorithms for finding the matching matrix P can be either optimal or approximate methods [18, 36, 109, 110]. Optimal matching algorithms always find an exact solution if it exists, and have exponential time complexity in the worst case, which makes them unattractive for many applications. In contrast, approximate or suboptimal matching algorithms find the local minima of the matching cost with the polynomial time complexity respect to the number of nodes. Generally, there are no guarantees to reach the global minimum, but often the approximation is not very far from the global one [18]. Almohamad et. al. [3] solved the quadratic problem by using the simplex algorithm, while [90] used a method based on Lagrangian relaxation network. In [36], Gold et. al. proposed the graduated non-convexity assignment to avoid poor local optima. The relaxation of the discrete optimization problem to be continuous one for the graph-matching was introduced in [109, 110]. Recently, a new graph matching algorithm has proposed with the exploration of factorizing affinity matrix in [112].

By considering each streamline as a vertex and the edge connecting vertex s_i and s_j as the distance between the two streamlines, $d(s_i, s_j)$ (the concept of distance between two streamline is presented in the next Section 2.4). Then, intuitively, the problem of tractography mapping becomes very similar to that of graph matching, but with some key differences. Firstly, the size of the two tractographies is in general not the same. Global differences in the anatomy of the brains, e.g. different volume, motivates this difference. Secondly, in general there is not a one-to-one correspondence between the streamlines but a many-to-one correspondence. This is anatomically likely if we consider that a given anatomical structure (*tract*), e.g. the cortico-spinal tract (CST), whose streamlines should have direct correspondence across subjects, may have different number of streamlines. In this case, for example, multiple streamlines of one CST would correspond to a single streamline in the other CST. Because of these differences, it is generally not possible to directly apply efficient graph matching algorithms to the problem of mapping tractographies.

2.4 Notation

Let the polyline $s = \{\vec{x_1}, \ldots, \vec{x}_{n_s}\}$, where $\vec{x} \in \mathbb{R}^3$, be a *streamline* (or fiber, track) reconstructed from dMRI data by deterministic tractography algorithms [69]. Note that each streamline has a different number of 3D points with other streamlines. Let the *tractography* $\mathbb{T} = \{s_1, \ldots, s_n\}$ be defined as a set of n streamlines. Current dMRI techniques operated on adult humans generate tractography of size in the order of 3×10^5 streamlines. Let τ be an anatomical fiber tract of interest, e.g. the cortical spinal tract (see figure 5.13), and let $\mathsf{T} \subset \mathbb{T}$ be its corresponding streamline-based approximation within given the tractography.

In the literature of tract segmentation or registration, it usually refers to distances between pair of streamlines as a leading way to incorporate domain specific information. The recent survey about streamline distance can be found more detial in [111]. A popular group of distances is the modified Hausdorff distances [26] and among the most popular [111] are

- $\boldsymbol{d}_1(s_A, s_B) = \frac{1}{n_{s_A}} \sum_{i=1}^{n_{s_A}} d(\boldsymbol{x}_i^A, s_B)$

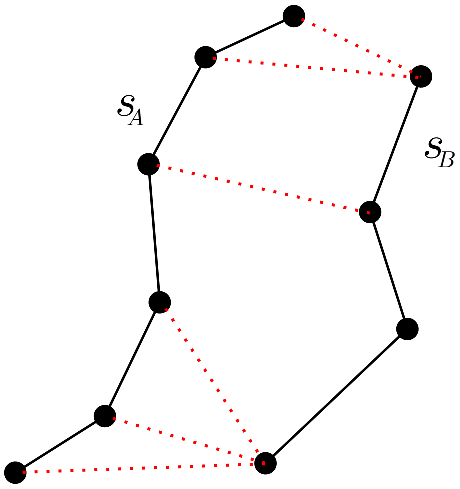


Figure 2.8: Many distances between two streamlines, s_A and s_B (solid line), that are proposed in the literature are based on the set of minimum distances between each point of s_A to s_B. The set of minimal distances is represented here as dotted lines.

- $\boldsymbol{d}_2(s_A, s_B) = \min_{i=1,\ldots,n_{s_A}} d(\boldsymbol{x}_i^A, s_B)$
- $\boldsymbol{d}_3(s_A, s_B) = max_{i=1,\ldots,n_{s_A}} d(\boldsymbol{x}_i^A, s_B)$

where (see Figure 2.8)

$$d(\boldsymbol{x}_i^A, s_B) = \min_{j=1,\ldots,n_{s_B}} ||\boldsymbol{x}_i^A - \boldsymbol{x}_j^B||_2 \tag{2.2}$$

which can be combined in order to get the symmetric versions:

- $\boldsymbol{h}_a(\boldsymbol{d}, s_A, s_B) = \frac{\boldsymbol{d}(s_A,s_B)+\boldsymbol{d}(s_B,s_A)}{2}$
- $\boldsymbol{h}_b(\boldsymbol{d}, s_A, s_B) = \min(\boldsymbol{d}(s_A, s_B), \boldsymbol{d}(s_B, s_A))$
- $\boldsymbol{h}_c(\boldsymbol{d}, s_A, s_B) = \max(\boldsymbol{d}(s_A, s_B), \boldsymbol{d}(s_B, s_A))$

Note that all distances defined above are not metric [26] because $\boldsymbol{d}(s_A, s_B) = 0$ does not imply that $s_A = s_B$.

Chapter 3

Dissimilarity Representation for Tractography

Diffusion magnetic resonance imaging (dMRI) data allow to reconstruct the 3D pathways of axons within the white matter of the brain as a tractography. The analysis of tractographies has drawn attention from the machine learning and pattern recognition communities. Many of the current learning algorithms require the input to be from a vectorial space. This requirement contrasts with the intrinsic nature of the tractography because its basic elements, called streamlines or tracks, have different lengths and different number of points and for this reason they cannot be directly represented in a common vectorial space. In this work we propose the adoption of the dissimilarity representation which is an Euclidean embedding technique defined by selecting a set of streamlines called prototypes and then mapping any new streamline to the vector of distances from prototypes. We investigate the degree of approximation of this projection under different prototype selection policies and prototype set sizes in order to characterise its use on tractography data. Additionally we propose the use of a scalable approximation of the most effective prototype selection policy that provides fast and accurate dissimilarity approximations of complete tractographies.

3.1 Introduction

Deterministic tractography algorithms [69] can reconstruct white matter fiber tracts as a set of *streamlines*, also known as *tracks*, from diffusion Magnetic Resonance Imaging (dMRI) [6] data. A streamline is a mathematical approximation of thousands of neuronal axons expressing anatomical connectivity between different areas of the brain, see Figure 3.1. Recently there has been an increase of attention in analysing tractography data by means of machine learning and pattern recognition methods, e.g. [111, 105]. These methods often require the data to lie in a vectorial space, which is not the case for streamlines. Streamlines are polylines in 3D space and have different lengths and numbers of points. The goal of this work is to investigate the features and limits of a specific Euclidean embedding, i.e. the dissimilarity representation, that was recently applied to the analysis of tractography data [75].

The dissimilarity representation is an Euclidean embedding technique defined by selecting a set of objects (e.g. a set of streamlines) called *prototypes*, and then by mapping any new object (e.g. any new streamline) to the vector of distances from the prototypes. This representation [88, 5, 16] is usually presented in the context of classification and clustering problems. It is a *lossy* transformation in the sense that some information is lost when projecting the data into the dissimilarity space. To the best of our knowledge this loss, i.e. the degree of approximation, has received little attention in the literature. In [86] the approximation was studied to decide among competing prototype selection policies only for classification tasks. In this work we are interested in assessing and controlling this loss without restriction to the classification scenario.

This work is motivated by practical applications about executing common algorithms, like spatial queries, clustering or classification, on large

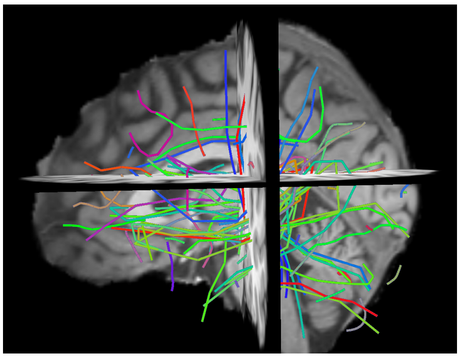


Figure 3.1: A set of 100 streamlines, i.e. an example of prototypes, from a full tractography

collections of objects that do not have a natural vectorial space representation. The lack of the vectorial representation avoids the use of some of those algorithms and of computationally efficient implementations. The dissimilarity space representation could be the way to provide such a vectorial representation and for this reason it is crucial to assess the degree of approximation introduced. Besides this characterisation we propose the use of a stochastic approximation of an optimal algorithm for prototype selection that scales well on large datasets. This scalability issue is of primary importance for tractographies given that a full brain tractography is a large collection of streamlines, usually $\approx 3 \times 10^5$, a size for which algorithms may become impractical. We provide practical examples both from simulated data and human brain tractographies.

3.2 Methods

In the following we present a concise formal description of the dissimilarity projection together with a notion of approximation to quantify how accurate this representation is. Additionally we introduce three strategies for prototype selection that will be compared in Section 3.3.

3.2.1 The dissimilarity projection

Let $\mathcal{X}$ be the space of the objects of interest, e.g. streamlines, and let $X \in \mathcal{X}$. Let $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$ be a distance function between objects in $\mathcal{X}$. Note that d is not assumed to be necessarily metric. Let $\Pi = \{\tilde{X}_1, \ldots, \tilde{X}_p\}$, where $\forall i\ \tilde{X}_i \in \mathcal{X}$ and p is finite. We call each $\tilde{X}_i$ as *prototype* or *landmark*. The *dissimilarity representation/projection* is defined as $\phi_\Pi^d(X) : \mathcal{X} \mapsto \mathbb{R}^p$ s.t.

$$\phi_\Pi^d(X) = [d(X, \tilde{X}_1), \ldots, d(X, \tilde{X}_p)] \tag{3.1}$$

and maps an object X from its original space $\mathcal{X}$ to a vector of $\mathbb{R}^p$.

Note that this representation is a *lossy* one in the sense that in general it is not possible to exactly reconstruct X from $\phi_\Pi^d(X)$ because some information is lost during the projection.

We define the distance between projected objects as the Euclidean distance between them: $\Delta_\Pi^d(X, X') = ||\phi_\Pi^d(X) - \phi_\Pi^d(X')||_2$, i.e. $\Delta_\Pi^d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$. It is intuitive that Δ_Π^d and d should be strongly related. In the following sections we will present more details and explanations about this relation.

3.2.2 A measure of approximation

We investigate the relationship between the distribution of distances among objects in $\mathcal{X}$ through d and the corresponding distances in the dissimilarity representation space through Δ_Π^d. We claim that a good dissimilarity representation must be able to accurately preserve the partial order of the distances, i.e. if $d(X, X') \leq d(X, X'')$ then $\Delta_\Pi^d(X, X') \leq \Delta_\Pi^d(X, X'')$ for each $X, X', X'' \in \mathcal{X}$ almost always. As a measure of the degree of approximation of the dissimilarity representation we define the Pearson correlation coefficient ρ between the two distances over all possible pairs of objects in $\mathcal{X}$:

$$\boldsymbol{\rho} = \frac{\mathrm{Cov}(d(X,X'), \Delta_{\Pi}^{d}(X,X'))}{\sigma_{d(X,X')}\sigma_{\Delta_{\Pi}^{d}(X,X')}} \tag{3.2}$$

where $X, X' \sim P_X$. In practical cases P_X is unknown and only a finite sample S is available. We can approximate $\boldsymbol{\rho}$ as the *sample* correlation $\boldsymbol{r}$ where $X, X' \in S$. An accurate approximation of the relative distances between objects in $\mathcal{X}$ results in values of $\boldsymbol{\rho}$ far from zero and close to 1[1].

In the literature of the Euclidean embeddings of metric spaces, the term of *distortion* is used for representing the relation between the distances in the original space and the corresponding ones in the projected space. The embedding is said to have *distortion*$\leq c$ if for every $x, x' \in \mathcal{X}$:

$$d(x,x') \geq \Delta_{\Pi}^{d}(x,x') \geq \frac{1}{c}d(x,x'). \tag{3.3}$$

An interesting embedding of metric spaces is described in [60]. It is based on ideas similar to the dissimilarity representation and has the advantage of providing a theoretical bound on the distortion. Unfortunately this embedding is computationally too expensive to be used in practice.

We claim that correlation and distortion target are slightly different aspects of the embedding quality, the first focuses on the *averaged* differences between the original and projected space, and the second on the worst case scenario. For this reason we claim that, in the context of machine learning and pattern recognition applications, correlation is a more appropriate measure.

[1] Note that negative correlation is not considered as accurate approximation. Moreover it never occurred during experiments

3.2.3 Strategies for prototype selection

The definition of the set of prototypes with the goal of minimising the loss of the dissimilarity projection is an open issue in the dissimilarity space representation literature. In the context of classification problems the policy of random selection of the prototypes was proved to be useful under certain assumptions [5]. In the following we address the issue of choosing the prototypes in order to achieve the desired degree of approximation but we do not restrict to the classification case only. We define and discuss the following policies for prototype selection: random selection, farthest first traversal (FFT) and subset farthest first (SFF). All these policies are parametric with respect to p, i.e. the number of prototypes.

Random Selection

In practical cases we have a sample of objects $S = \{X_1, \ldots, X_N\} \subset \mathcal{X}$. This selection policy draws uniformly at random from S, i.e. $\Pi \subseteq S$ and $|\Pi| = p$. Note that sampling is *without replacement* because identical prototypes provide redundant, i.e. useless, information. This policy was first proposed in [29] for seeding clustering algorithms. This policy has the lowest computational complexity $O(1)$.

Farthest First Traversal (FFT)

This policy selects an initial prototype at random from S and then each new one is defined as the farthest element of S from all previously chosen prototypes. The FFT policy is related to the *k-center* problem [44]: given a set S and an integer k, what is the smallest ϵ for which you can find an ϵ-cover[2] of S of size k? [3]. The k-center problem is known to be an NP-hard [44], i.e. no efficient algorithm can be devised that always returns the optimal answer. Nevertheless FFT is known to be close to the optimal solution, in the following sense: If T is the solution returned by FFT and T^* is the optimal solution, then $\max_{x \in S} d(x, T) \leq 2 \max_{x \in S} d(x, T^*)$. Moreover, in metric spaces, any algorithm having a better ratio must be NP-hard [44]. FFT has $O(p|S|)$ complexity. Unfortunately when $|S|$ becomes very large this prototype selection policy becomes impractical.

Subset Farthest First (SFF)

In the context of radial basis function networks initialisation, a scalable approximation of the FFT algorithm, called *subset farthest first* (SFF), was proposed in [100]. This approximation is also claimed to reduce the chances to select outliers that can lead to a poor representation of large datasets. The SFF policy samples $m = \lceil cp \log p \rceil$ points from S uniformly at random and then applies FFT on this sample in order to select the p prototypes. In [100] it was proved that under the hypothesis of p clusters in S, the probability of not having a representative of some clusters in the sample is at most $pe^{-m/p}$. The computational complexity of SFF is $O(p^2 \log p)$. Note that for large datasets and small p this prototype selection policy has a much lower computational cost than FFT.

3.3 Experiments

In the following we describe the assessment of the degree of approximation of the dissimilarity representation across different prototype se-

[2] Given a metric space $(\mathcal{X}, d)$, for any $\epsilon > 0$, an ϵ-cover of a set $S \subset \mathcal{X}$ is defined to be any set $T \subset X$ such that $d(x, T) \leq \epsilon, \forall x \in S$. Here $d(x, T)$ is the distance from point x to the closest point in set T.

[3] Note that in our problem k is called p.

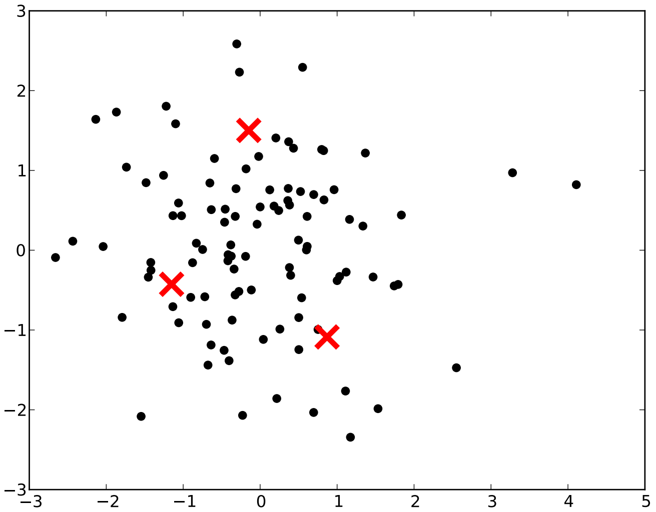


Figure 3.2: A 2-dimensional example of 50 points (black circles) drawn from $\mathcal{N}(\mathbf{0}, I)$ and 3 prototypes (red stars) drawn from the same pdf.

lection policies and different numbers of prototypes. The aim is to investigate the trade-off between accuracy and computational cost. The experiments are carried out on 2D simulated data and on real tractographies reconstructed from dMRI recordings of the human brain.

3.3.1 Simulated data

Let $\mathcal{X} = \mathbb{R}^2$, $P_X = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{\mu} = [0, 0]$, $\Sigma = I$, $d(X, X') = ||X - X'||_2$, $p = 3$ and $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3 \sim P_X$. Then $\phi^d_\Pi(X) = \left[||X - \tilde{X}_1||_2, ||X - \tilde{X}_2||_2, ||X - \tilde{X}_3||_2\right] \in \mathbb{R}^3$. Figure 3.2 shows a sample of 50 points drawn from P_X together with the 3 prototypes $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3$. Figure 3.3 shows the sample projected into the dissimilarity space together with the prototypes.

The selection of the prototypes according to different policies is explained in Section 3.2.3. For SFF we chose $c = 3$ in order to have high probability (> 0.95) of accurately representing S through the subset. Each dataset was projected in the dissimilarity space. The correlation $\boldsymbol{\rho}$

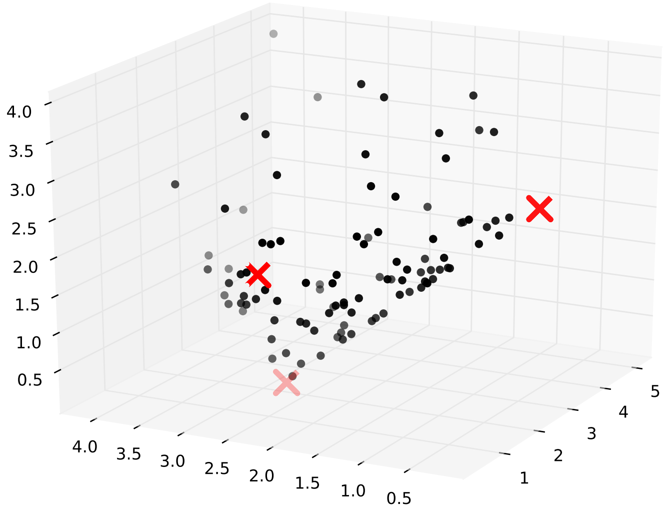


Figure 3.3: The dissimilarity projection of the dataset and prototypes of Figure 3.2.

between distances in the original space and the corresponding distances in the projected space was estimated by computing 50 repetitions of the simulated dataset. The average correlation and one standard deviation for each prototype selection strategy are shown in Figure 3.5.

In this simulated dataset both SFF and FFT performed significantly better than the random selection, on average. FFT showed a small advantage over SFF when $p < 10$.

3.3.2 Tractography data

We estimated the dissimilarity representation over tractography data from dMRI recordings of the MRI facility at the MRC Cognition and Brain Sciences Unit, Cambridge UK. The dataset consisted of 12 healthy subjects; 101 $(+1$, i.e. $b = 0)$ gradients; b-values from 0 to 4000; voxel size: $2.5 \times 2.5 \times 2.5mm^3$. In order to get the tractography we computed the single tensor reconstruction (DTI) and created the streamlines using

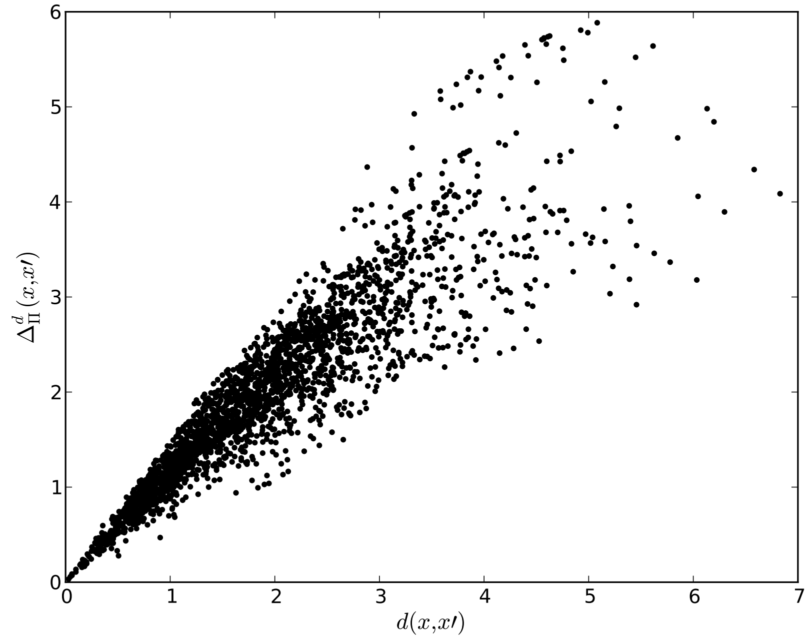


Figure 3.4: The correlation distances of 50 points (2D dimention) in original space (d x-axis) and in the projection space (Δ_{Π}^{d} y-asix) with random prototype selection policcy.

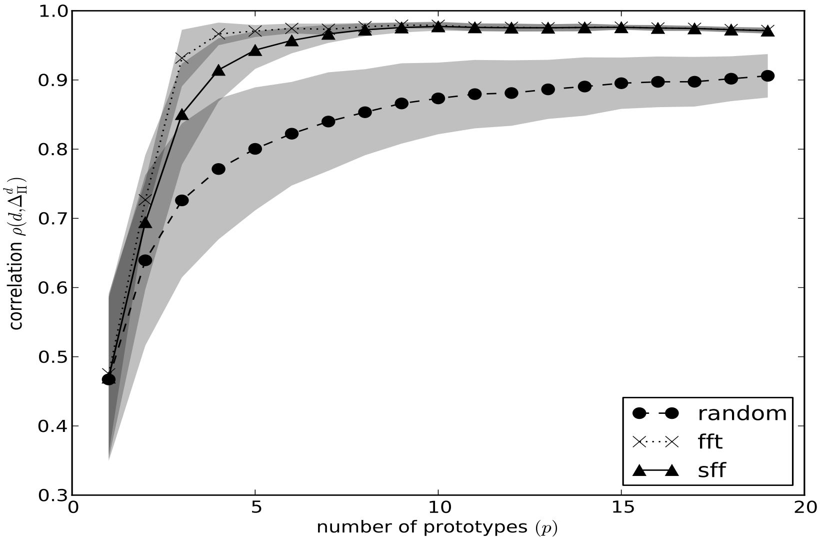


Figure 3.5: Average correlation between d and Δ_{Π}^{d} across different prototype selection policies and different numbers of prototypes.

EuDX, a deterministic tracking algorithm [30] from the DiPy library [4]. We obtained two tractographies using 10^4 and 3×10^6 random seed respectively. The first tractography consisted of approximately 10^3 streamlines and the second one of 3×10^5 streamlines. An example of a set of prototypes from the largest tractography is shown in Figure 3.1.

As the distance between streamlines we chose one of the most common, i.e. the symmetric minimum average distance from [111] defined as $d(X_a, X_b) = \frac{1}{2}(\delta(X_a, X_b) + \delta(X_b, X_a))$ where

$$\delta(X_a, X_b) = \frac{1}{|X_a|} \sum_{\mathbf{x}_i \in X_a} \min_{\mathbf{y} \in X_b} ||\mathbf{x}_i - \mathbf{y}||_2. \tag{3.4}$$

As it is shown in Figure 3.6 for the case of a tractography of 10^3 streamlines both FFT and SFF($c = 3$) had significantly higher correlation than the random sampling for all numbers of prototypes considered. We confirmed that the SFF selection policy is an accurate approximation of the FFT policy for tractographies. Moreover we noted that after $15 - 20$ prototypes the correlation reaches approximately 0.95 on average (50 repetitions) and then slightly decreases indicating that a little number of prototypes is sufficient to reach a very accurate dissimilarity representation.

Figure 3.7 shows the correlation for SFF and the random policy when the tractography has 3×10^5 streamlines, i.e. the standard size of a tractography from current dMRI recording techniques. In this case FFT is impractical to be computed because it requires approximately 15 minutes on a standard desktop computer for a single repetition when $p = 50$. The cost of computing SFF is instead the same of the case of 10^3 streamlines, as its computational cost depends only on the number of prototypes. It took ≈ 2 seconds on standard desktop computer when $p = 50$

[4] `http://www.dipy.org`

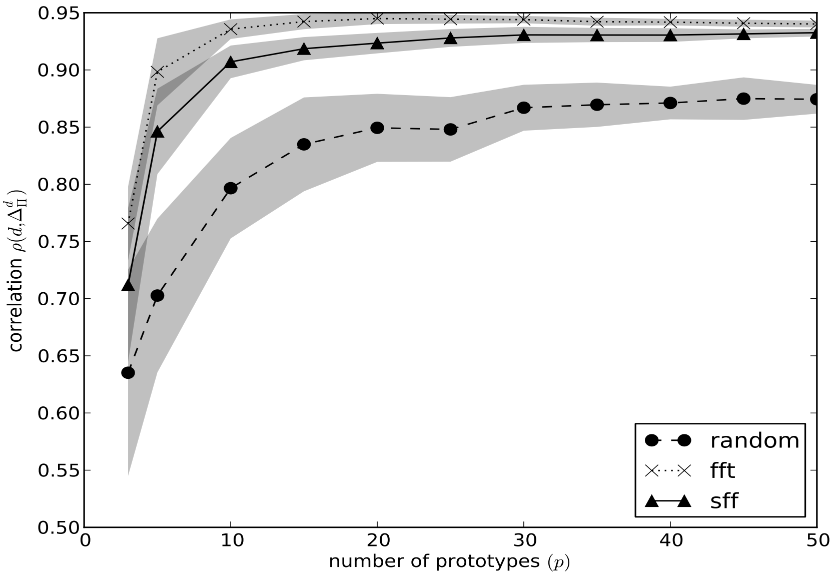


Figure 3.6: The correlation between of d and Δ_Π^d over a 10^3 streamlines tractography for different prototype selection policies.

to compute one repetition. We observed that for 3×10^5 streamlines SFF significantly outperformed the random policy and reached the highest correlation of 0.96 on average (50 repetitions) for $15 - 25$ prototypes.

Note that the figures presented in this section refers to data from subject 1 of the dMRI dataset. We conducted the same experiments on other subjects obtaining equivalent results.

3.3.3 Dissimilarity for fast clustering tractography

In this part, we explain how to explore the dissimilarity in the clinical application for using dMRI data. Our work is motivated by a clinical research hypothesis about the characterisation of the amiotrophic lateral sclerosis (ALS) disease. The ALS disease is known to be affected by the corticospinal tract (CST) [21], an anatomical structure that connects cortical motor areas to the spine and the body. For this reason, the first task

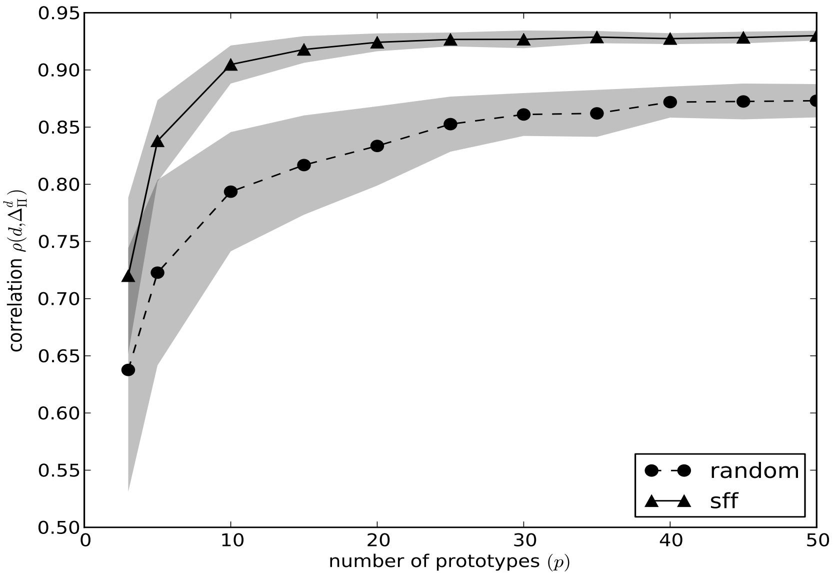


Figure 3.7: The correlation between of d and Δ_Π^d for a full tractography of 3×10^5 streamlines with random, and SFFprototype selection policies.

in the endeavour of characterisation of the ALS disease is to segment the CST from the full brain tractography of each subject.

Tractography segmentation can be performed manually or automatically. Despite an increasing literature in automatic segmentation (see a brief review in [105]), the application in the clinical domain usually still rely on manual segmentation. The manual segmentation process usually consists in selecting the subset of the streamlines connecting a few manually located regions of interest[5]. This task is a lengthy and complex one due to a very large set of streamlines in tractography, in the order of 3×10^5, which makes it intrinsically difficult both to inspect and to unfold anatomical structures.

In this part, we conceived a novel computer-assisted interactive process based on clustering algorithms with the help of dissimilarity representation, and aimed at greatly reducing the time required to manu-

[5] See for example `http://www.trackvis.org`.

ally segment a given anatomical white matter structure of interest. Our approach is based on a fast-clustering technique based on dissimilarity representation, by means of which the expert is presented with a summary of the streamlines, i.e. the clusters represented by their medoids [6]. The expert manually selects the medoids/clusters of interest in order to remove most of the streamlines not related to the anatomical structure of interest. Interacting with the summary, instead of the actual streamlines, is much simpler for the user. In this part, we only mention the fast clustering based on dissimilarity, the more detail of the interactive segmentation procedure can be found in the Chapter 5.

The core of the problem is to cluster a large number of streamlines in no more than a few seconds, to allow a comfortable interactive user experience to the expert. The proposed solution combines two state-of-the-art elements: first a recently proposed Euclidean embedding algorithm for streamlines, i.e. the dissimilarity representation with the scalable *subset farthest first* (SFF) prototype selection policy [76]. This embedding provides fast and accurate vectorial representation of streamlines. Second, a recently proposed improvement of the k-means clustering algorithm called *mini-batch* k-means [94] (MBKM). This algorithm, which require the data to lie in a vector space, drastically reduces the convergence time to the actual clusters in case of large and very-large sets of objects. We claim that the dissimilarity embedding together with the MBKM algorithm provides a viable solution to problem of fast clustering of streamlines.

Mini-Batch k-means

The k-means clustering problem is a cornerstone of the clustering literature. Given k, the number of clusters, the problem is to find k clus-

[6] A medoid is the element of a cluster closest to its centre.

ter centres $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$, $\mathbf{c} \in \mathbb{R}^p$, and to assign each element of the vectorial dataset $\Phi(T) = \{\phi(X_1), \ldots, \phi(X_M)\} \subset \mathbb{R}^p$ to the closest cluster[7]. The k-means problem is then to compute centres C such as to minimise the loss function $f(C) = \sum_{\phi(X)\in\Phi(T)} D(\phi(X), C)^2$, where $D(\phi(X), C) = \min_{\mathbf{c}\in C} ||\phi(X) - \mathbf{c}||_2$ is the distance between $\phi(X)$ and the closest centre. The exact solution of the k-means problem is NP-hard and the computational complexity of the standard algorithm, the Lloyd's algorithm, has been proved to be $O(M^{34})$ in the general case [4], even though much less in practical applications.

The *mini-batch k-means* (MBKM) algorithm [94] is a recently proposed modification of the standard algorithm that is able to reduce the computational costs by orders of magnitude. The intuitive idea is to use a stochastic gradient descent approach to find the centres C starting from a random initialisation. This idea was introduced in [11] where the points of the dataset were given one at a time in an online fashion.

Instead of updating the centers with one streamline at a time, the MBKM algorithm proposes to use multiple random subsets of the dataset, i.e. the *mini batches*, to update the cluster centres and to estimate the per-centre learning rates. As soon as the objective function $f(C)$ converges the process stops. The pseudocode algorithm of MBKM is shown Algorithm 1.

The computational complexity of the MBKM algorithm is not known in the general case but empirical results in [94] show a reduction of two orders of magnitude in computation time with respect to the standard k-means.

[7] From now on, we denote $\phi^d_\Pi(X)$ as $\phi(X)$ to simplify the notation without introducing ambiguity.

Algorithm 1 Mini-batch k-Means algorithm

```
Given: $k$, mini-batch size $b$, iteration $t$, dataset $X$
Initialize each $c \in C$ with an x picked randomly from $X$
$v \leftarrow 0$
for i = 1 to t do
    $M \leftarrow b$ examples picked randomly from $X$
    for $x \in M$ do
        $d[x] \leftarrow f(C, x)$ //Cache the center nearest to x
    for $x \in M$ do
        $c \leftarrow d[x]$ //Get cached center for this x
        $v[c] \leftarrow v[c] + 1$ //Update per-center counts
        $\eta \leftarrow \frac{1}{v[c]}$ //Get per-center learning rate
        $[c] \leftarrow (1 - \eta)c + \eta x$ //Take gradient step
```

From Centroids to Medoids

In order to visually present the clusters of streamlines to the user one representative streamline of each cluster needs to be selected. In the general case the dissimilarity representation is not invertible, i.e. given a vector $\mathbf{c} \in \mathbb{R}^p$ it is not possible to construct the streamline $X_\mathbf{c}$ such that $\mathbf{c} = \phi(X_\mathbf{c})$. This means that the centroids C obtained with the k-means or the MBKM cannot be shown to the user as streamlines. For this reason we decided to display the *medoid* of each cluster, i.e. the streamline of the tractography closest to each centroid $\mathbf{c}$. The exhaustive search of the medoids requires the computation of kM distances, which is too slow for interactive use. For this reason we adopted a data structure for efficient computation of the nearest neighbour in high-dimensional spaces: the *Ball Tree*. We refer the reader to [78] for additional details. We present empirical results of the time required for the computation of the medoids in Table 3.1.

ALS Dataset

The data was recorded with a $3T$ scanner at the Brain Institute, University of Utah. It consisted of 12 ALS patients and 12 healthy controls (64 gradients; b-value$=$ $1000.$; anatomical scan ($1 \times 1 \times 1mm^3$)). We reconstructed the streamlines using EuDX, a deterministic tracking algorithm [30] from the DiPy library [8]. The tractography was then embedded in $\mathbb{R}^p$ using the dissimilarity representation presented in Section 3.2 with $p = 40$ and the SFF prototype selection procedure ($c = 3$) as suggested in [76]. The prototype selection and the actual embedding of $\approx 3 \times 10^5$ streamlines required $\approx$ 180s. The resulting matrix $\phi(T) \in \mathbb{R}^{300K \times 40}$ was computed once and stored, so that the time to compute the projection did not affect the interactive segmentation.

The average timings of the clustering algorithms of are reported in Table 3.1. In the first column (size) are reported the size of the subset of streamlines that were clustered. The second column (k) reports the number of clusters, according to the notes expressed above. The third (k-means) and the fourth (MBKM) report the time for clustering. Note that the clustering of the whole tractography can be computed once and stored, so its time does not affect the interactive use. The Fifth column reports the size (b) of the mini-batches for the MBKM, which was always 100 except for the full tractography for which we observed a significant gain in time when increasing it to 1000. The sixth column reports the time to compute the medoids from the centroids provided by k-means and MBKM. Each medoid was computed with simple exhaustive search within each cluster. The time to compute all medoids was always negligible with respect to the clustering time. All computations were per-

[8] `http://www.dipy.org`

Table 3.1: For a given number of streamlines (1st column, size) and a given number of clusters (2nd column, k) the time to compute the clustering with k-means and MBKM is reported in the 3rd and 4th columns, respectively. The size (b) of the mini-batches for MBKM is in the 5th column. The time to compute the medoids from the centroids is in the 6th column.

size	k	k-means	MBKM	b	medoids
500	50	$0.3s$	$\mathbf{0.2}s$	100	$0.003s$
1000	50	$0.6s$	$\mathbf{0.2}s$	100	$0.004s$
5000	50	$6.1s$	$\mathbf{0.4}s$	100	$0.009s$
10000	50	$14.4s$	$\mathbf{0.6}s$	100	$0.018s$
15000	50	$29.9s$	$\mathbf{0.7}s$	100	$0.026s$
250000	150	$> 1000s$	$\mathbf{13.3}s$	1000	$0.72s$

formed on a standard desktop computer.

In this part, in order to handle the computational burden of clustering a large number of streamline under strong time constraints, we proposed a solution based on the dissimilarity representation and the MBKM algorithm. As shown in Table 3.1 (4th column) the time required to cluster the streamlines with the proposed solution was always the lowest and always $<$ 1s, thus meeting the requirements for a comfortable user experience. Conversely, the time required by the standard k-means algorithm was inadequate (see the 3rd column in Table 3.1).

3.4 Discussion

In this document we investigated the degree of approximation of the dissimilarity representation for the goal of preserving the relative distances between streamlines within tractographies. Empirical assessment has been conducted on two different datasets and through various prototype selection methods. All of the results from both simulated data and real tractography data reached correlation ≥ 0.95 with respect to the distances in the original space. This fact proved that the dissimilarity representation works well for preserving the relative distances. Moreover on tractography data the maximum correlation was reached with just approximately $20 - 25$ prototypes proving that the dissimilarity representation can produce compact feature spaces for this kind of data.

When comparing the different prototype selection policies we found that FFT had a small advantage over SFF but only when the number of prototypes was very low ($p < 10$). Both FFT and SFF always outperformed the random policy. Moreover, since the computational cost of SFF does not increase with the size of the dataset but only with the number of prototypes, we observed that the SFF policy can be easily computed on a standard computer even in the case of a tractography of 3×10^5 streamlines. This is different from FFT which is several orders of magnitude slower than SFF, thus computationally less practical.

We advocate the use of the dissimilarity approximation for the Euclidean embedding of tractography data in machine learning and pattern recognition applications. Moreover we strongly suggest the use of the SFF policy to obtain an efficient and effective selection of the prototypes. We also applied dissimilarity representation to mini-batch k-mean clustering algorithm for supporting segmentation tractography. Experiments showed that combination of dissimilarity and MNBKM could meet the requirements for a comfortable interactive use. It showed a potential of using dissimilarity for tractography in real medical applications.

Chapter 4

Mapping Tractography Across Subjects

Diffusion tensor imaging (DTI) and tractography provide mean to study the anatomical structures within the white matter. When studying the tractography data across subjects, it requires to align or register tractographies together. This registration is most often performed by applying the transformation resulting from other images (T1, FA, DTI) to tractography data, or to register tractographies themselves. However, the above methods can not deal with a new coming tractograhy, except for running the whole registration process again with all data plus the new comer. In contrast with these registration methods, instead of finding the transformation between tractographies, in this work, we try to find which streamline in one tractography corresponding to which streamline in the other tractography without any transformation, or to directly map the *source* tractography to the *target* tractography. We present what we believe to be the first recasting the problem as mapping problem rather than registration problem. Moreover, by taking advantage of more than thirty year graph-matching research, we propose a graph-based solution for tractography mapping problem and explain similarities and differences with the well-known graph matching problem. We define the loss function based on the pairwise streamline distance, and reformulate the mapping problem as the problem of minimizing that loss function. To our knowledge, this is also the first graph-matching-based objective function applied to tractography. Moreover, we also try to explore the dissimilarity representation idea with the help of mapping in the context of finding the correspondences between streamlines across different tractographies. Experiments using real dMRI data demonstrate the potential of the proposed method for medical or neuroscientific analyses of white matter tractography data.

4.1 Introduction

Diffusion magnetic resonance imaging (dMRI) [6] is a modality that provides non-invasive images of the brain white matter. It captures the diffusion process of the water molecules in each voxel which represents important structural information of the axons of the neurons. From dMRI data, tracking algorithms [69, 111] allow to reconstruct the $3D$ pathways of axons within the white matter of the brain as a set of streamlines, called tractography. A *streamline* is a vectorial representation of thousands of neuronal axons expressing structural connectivity, and *tractography* is a set of N streamlines ($N \sim 3 \times 10^5$ usually).

Current neuroscientific analyses of white matter tractography data are limited to qualitative intra-subject comparisons. Thus, it is quite difficult to use the information for direct inter-subject comparisons [37, 7]. This leads to the need of initial alignment, or registration, of tractographies via some methods before doing further study. Registration is most often performed by applying the transformation resulting from the registration of other images, such as $T1$ or fractional anisotropy (FA), to tractography [38, 106, 37, 113]. Recently, [72] proposed group-wise registration using the trajectory data of the streamlines. The idea to work on tractography rather than other images is quite innovative. And, it may be advantageous to directly align the streamlines because the result would be closely related to the final goal of registration.

Similar to [72], in this work, we explore the idea of working on tractography rather than other images. However, in contrast to all current tractography registration methods, which are based on rigid or non-rigid shape transformation of one tractography into another, our approach tries to find which streamline of one tractography corresponds to which streamline in the other tractography, without transformations. This correspondence is a *mapping* from one tractography to the other.

We propose to solve the problem of finding the mapping between two tractographies through a graph-based approach similar to that of the well-known graph matching problem [18, 109]. In the graph matching problem the aim is to find which node of one graph corresponds to which node of another graph, under the assumption that graphs have the same number of nodes and that the correspondence is one-to-one.

Given a tractography of N streamlines $T = \{s_1, \ldots, s_N\}$ and a distance function d between streamlines, we can create an undirected weighted graph by considering each streamline as a vertex and the edge connecting vertex s_i and s_j as the distance between streamline s_i and s_j, $d(s_i, s_j)$. Then, intuitively, the problem of tractography mapping becomes very similar to that of graph matching, but with some key differences. Firstly, the size of the two tractographies/graphs is in general not the same. Global differences in the anatomy of the brains, e.g. different volume, motivates this difference. Secondly, in general there is not a one-to-one correspondence between the streamlines/vertexs but a many-to-one correspondence. This is anatomically likely if we consider that a given anatomical structure (*tract*), e.g. the cortico-spinal tract (CST), whose streamlines should have direct correspondence across subjects, may have different number of streamlines. In this case, for example, multiple streamlines of one CST would correspond to a single streamline in the other CST. Because of these differences, it is generally not possible to directly apply efficient graph matching algorithms to the problem of mapping tractographies.

In the following we formally describe the tractography mapping problem starting from the graph matching problem and define the details of the optimization problem to solve. We provide a preliminary algorithmic solution, based on simulated annealing, to minimize the proposed loss function. Then, we apply our proposed solution to a tractography segmentation task in order to compare a standard registration-based method to our proposed method on a fair ground.

Moreover, as discussed in the Chapter 3, dissimilarity representation for tractography provides a fast and accurate vectorial representation of the streamlines. However, its use is limited to intra-subject analysis only, because the choice of prototypes is subject-specific. This fact prevents the use of dissimilarity representation in the context inter-subject studies, because they require inter-subject comparisons [37, 7]. The best practice for studying structural connectivity across subjects recommends the alignment of all tractographies to one common space before the quantitative assessment. In this part, we combine dissimilarity representation and mapping to build a common vectorial representation of streamlines across subjects. First, given two tractographies, we compute the prototypes of one and, by mapping them, we obtain those of the other tractography. Second, we build a common vectorial space by simply merging the two dissimilarity representations, now aligned because prototypes were mapped. With such common space, we are able to align tractographies of two subjects and we claim that the quality of such alignment is superior to that of affine-based registration.

In Section 4.2, the algorithmic elements of the proposed method are formally described. In Section 4.3, we describe some experiments of only mapping and the combination of mapping and dissimilarity representation. We report the details of the actual use of the proposed solution in the context of the cortico-spinal tract (CST) segmentation. We quantitatively describe the result of mapping and provide figures to evaluate the viability of the proposed solution. In Section 4.4 we discuss the results and show that the proposed solution is quite challenging, but also promises many benefit in the field of brain connectivity. We conclude with a summary of our contribution and open challenges that needs to be solved in future work.

4.2 Methods

Our basic approach is to consider a streamline as a point in the space of co-relation with other streamlines in tractography. The co-relation between two streamlines can be defined as the distance between them. Streamline s_j^B in the target tractography T_B is known as the correspondence of streamline s_i^A in the source tractography T_A if it reserves the co-relation with other mapped streamlines in the mapped source in target. It means that the co-relation of mapped streamlines in mapped source to target, must similar to the co-relation of the source streamlines in the source T_A. For that reason, we then choose the mapping of T_A into T_B on the collection of permutation T_B by maximizing the similarity, or minimizing the difference between mapped T_A and T_A.

4.2.1 Tractography mapping

An undirected weighted graph $G = (V, E)$ of size N is a finite set of vertices $V = \{1, \ldots, N\}$ and edges $E \subset V \times V$. The graph matching problem can be described as follows. Given two graphs G_A to G_B with the *same* number of vertices N, the problem of matching G_A and G_B is to find the correspondence between vertices of G_A and vertices of G_B, which allows to align, or register, G_A and G_B in some optimal way. The correspondence between vertices of G_A and of G_B is defined as a *permutation* P of the N vertices, i.e. there a one-to-one correspondence between the two set of vertices. P is usually represented as a binary $N \times N$ matrix where P_{ij} is equal to 1, if the ith vertex of G_A is matched to the jth vertex of G_B, otherwise 0. Given A and B, i.e. the $N \times N$ adjacency matrices of the two graphs, the quality of the matching is assessed by the discrepancy, or loss, between the graphs after matching as:

$$L(P) = \|A - PBP^\top\|_2 \tag{4.1}$$

where $\|A\|_2 = \sqrt{\sum_{ij}^N A_{ij}^2}$ is the Frobenius norm. Therefore, the graph matching problem becomes the problem of finding P^* that minimize L over the set of permutation matrices $\mathcal{P}$:

$$P^* = \operatorname*{argmin}_{P \in \mathcal{P}} \|A - PBP^\top\|_2 \tag{4.2}$$

which is a combinatorial optimization problem. The exact solution to this problem has extremely high complexity and only approximate solutions are available in practical cases [18, 109].

A tractography can be encoded as a fully-connected undirected weighted graph [1]. Let $T_A = \{s_1^A, \ldots, s_N^A\}$ and $T_B = \{s_1^B, \ldots, s_M^B\}$, where $s = \{x_1, \ldots, x_{n_s}\}$ is a streamline and $x \in \mathbb{R}^3$, be the tractographies of two subjects. With a given distance function d between streamlines, we define two graphs G_A and G_B, where their nodes are respectively T_A and T_B, and their adjacency matrix is $A \in \mathbb{R}^{N \times N}$ and $B \in \mathbb{R}^{M \times M}$, with $A_{ij} = d(s_i^A, s_j^A)$ and $B_{kl} = d(s_k^B, s_l^B)$. Our current choice of d is discussed in Section 3.3, however any common streamline distance from the literature can be used.

Given two graphs G_A and G_B, the problem of mapping T_A to T_B becomes that of finding the correspondence between vertices of G_A and vertices of G_B. Such correspondence, called *mapping* [1], can be represented as a binary matrix Q, where $Q_{ij} = 1$, if the ith vertex of G_A is mapped to the jth vertex of G_B or, according to tractography notation, $s_i^A \in T_A$ is mapped to $s_j^B \in T_B$, otherwise 0. With mapping, $\sum_k Q_{ij} = 1$, which means that the correspondence can be many-to-one. Given A and B, the quality of the mapping is measured by the discrepancy, or loss, between the two graphs after the application of Q:

$$L(Q) = \|A - QBQ^\top\|_2 \tag{4.3}$$

Note that, in general, Q is not a permutation matrix. In order to find the optimal mapping Q^*, we minimize L so that T_B is most similar to T_A:

$$Q^* = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \|A - QBQ^\top\|_2 \tag{4.4}$$

where $\mathcal{Q}$ is the set of all possible mappings. Notice that $|\mathcal{Q}| = M^N$ which, given the typical size of tractographies, or even small part of them, is prohibitively high as combinatorial optimisation problem. For this reason only approximate solutions can be found in practical cases. In general $N \neq M$ and Q is a mapping and not just a permutation, therefore the tractography mapping problem is more general than the graph matching problem, i.e. the size of the search space $\mathcal{Q}$, i.e. M^N, is much larger than $\mathcal{P}$. As a consequence, the efficient solutions available in the literature of graph matching, e.g. [109], are not applicable, because they heavily rely on the assumptions that we violate here. In Section 3.3 we implemented a simple preliminary solution to the combinatorial optimization problem by means of the Simulated Annealing meta-heuristic [54].

Moreover, the distance between streamlines $d(s, s')$ in Equation 4.3, is computed when both s and s' belong to tractography of one subject, not across subjects. Thus, it does not require to put both tractographies T_A, T_B in the same space. It, as far as we know, is prominent comparing to other tractography registration.

In order to compare the proposed method against a standard registration procedure on a fair ground, we cannot rely on the value of the loss function L as in Equation4.3, because it is defined only in the case of mapping. As the evaluation criterion, we measured the overlap between the aligned T_A from subject A to B, called $T_{A(B)}$, and the expert-segmented tract T_B of subject B, in terms of voxels, as proposed in [37]. The idea is that the more voxel-overlap between $T_{A(B)}$ and T_B, the better the mapping is. Our hypothesis is that reducing L leads to better overlap between tractographies, which is important for practical applications like segmentation. In Section 3.3 we describe experiments to test this hypothesis and provide the necessary details. Here we introduce the metric that we use for comparing registration and mapping. As proposed in [37], we compare the set of voxels crossed by the streamlines of each tractography after mapping or after registration. We considered two scores: the sensitivity, or True Positive Rate, and the False Discovery Rate. TPR and FDR are computed as following:

$$TPR = \frac{|T_{A(B)} \cap T_B|}{|T_B|} \tag{4.5}$$

$$FDR = \frac{|T_{A(B)} \setminus T_B|}{|T_{A(B)}|} \tag{4.6}$$

Note that in the above equations, $|T|$ is the volume computed as number of voxels that any streamline $s \in T$ goes through, and $|T_1 \cap T_2|$ indicates the number of voxels in common between T_1 and T_2.

4.2.2 Common vectorial representation across subjects

The *dissimilarity representation* [88] is a lossy Euclidean embedding algorithm that maps general objects into $\mathbb{R}^p$. Dissimilarity representation for tractography was previously introduced in [76], and it is also claimed that this embedding provided fast and accurate vectorial representation of streamlines. Our assumption is that when exploring the dissimilarity representation idea based on mapping, we can get the better representation for streamlines, and it thus would provide a better correspondences between streamlines across different tractographies.

Dissimilarity representation

The *dissimilarity representation* [88] is a Euclidean embedding technique for generic spaces, originally proposed in the context of classification and clustering problems [16]. Given a set of objects, e.g., a tractography, the dissimilarity representation is defined by two elements: a distance function and a subset of objects, i.e., a subset of streamlines, called *prototypes*. Then, the dissimilarity representation maps every new object, i.e. every other streamline, into $\mathbb{R}^p$ [76] through $\phi^d_\Pi(s) : \mathcal{S} \mapsto \mathbb{R}^p$ s.t.

$$\phi^d_\Pi(s) = [d(s, \tilde{s}_1), \ldots, d(s, \tilde{s}_p)] \tag{4.7}$$

where d is a given distance function between streamlines, and $\Pi = \{\tilde{s}_1, \ldots, \tilde{s}_p\}$ is a given set of p streamlines used as prototypes. The quality of the resulting Euclidean embedding is strongly dependent on the choice of d and on the selection and number of prototypes (see [86, 76]).

An efficient procedure to select effective prototypes in the case of tractography data was presented in [76]: the *subset farthest first* (SFF) algorithm. This procedure is a scalable approximation of the well known farthest first traversal (FFT) algorithm which is a standard greedy solution to the well known k centre problem. This problem, put in our context, entails selecting a set Π of p streamlines [1] such that the sum of the distances of each streamline of the tractography to closest streamline in Π is minimised. Intuitively the streamlines in Π are designed to be a *representative* sample of the whole tractography. The FFT algorithm selects one streamline at random from the tractography as the first prototype $\tilde{s}_1$ and then iteratively adds a new prototype as the streamline maximising the distance to the already selected prototypes. The SFF algorithms is a stochastic scalable version of FFT, which subsamples $m = \lceil cp \log p \rceil$ streamlines from the whole tractography, and then applies FFT to the subsample. For the case of tractography data, when $c >= 3$ the SFF algorithm is comparable to the FFT algorithm with high probability, following the proof in [100] and the empirical results in [76].

Common vectorial representation

By combining mapping and dissimilarity representation, we propose to build a common vectorial representation of streamlines across subjects, as described in Algorithm 2.

After finding corresponding prototypes in the two sets, the respective embeddings of the streamlines are aligned. The quality of this alignment mainly depends on three approximations:

- The one introduced by the dissimilarity representation of T_A.
- The one introduced by the dissimilarity representation of T_B.

[1] Note that here we use p to denote the size of Π instead of the k of the "k centre problem". This is to avoid confusion with the notation we adopt in this paper.

Algorithm 2 Common vectorial representation

```
Step 1: Select prototypes $\Pi_A$ from tractography $T_A$ (SFF)
Step 2: Find prototypes $\Pi_B$ by mapping $\Pi_A$ to $T_B$
Step 3: Compute the dissimilarity representation of $T_A$ based on $\Pi_A$
Step 4: Compute the dissimilarity representation of $T_B$ based on $\Pi_B$
Step 5: Align dissimilarities according to map of $\Pi_A$ to $\Pi_B$
```

- The one introduced by the mismatch of mapping Π_A to Π_B.

Ideally, if the error introduced by each of these approximations were zero, then the corresponding streamline $s_B \in T_B$ of $s_A \in T_A$ would be the nearest neighbour of $\phi^d_{\Pi_A}(s)$ in the embedding of T_B. In this work we provide only indirect experimental evidence of this observation and leave its detailed investigation to future work.

By exploring the mapping idea in the context of dissimilarity representation, we propose a new common vectorial representation for streamlines across subjects, no matter what subjects are in the same space or not. This representation, as far as we know, is the first approach that can create a common space for representing streamlines from multiple subjects without requirement of co-registering subjects in the same space, and it can be free to to be used for further research purpose.

4.3 Experiments

We designed two experiments to evaluate the mapping method and the common vectorial representations. The first experiment aimed at providing empirical evidence that reducing the loss in Equation 4.3 is related to an increase of the overlap between tractographies. The second one was conducted to afford that the combination of DR and tractography mapping provides an accurate alignment of the streamlines of two subjects, even better than traditional affine-based registration. We considered the scenario of tract segmentation. There, the task was the identification of a desired tract in the tractography of a subject, given the equivalent tract segmented from the tractography of an other subject. We applied our proposed tractography mapping framework on the task of Cortical Spinal Tract (CST) segmentation. CST is a set of streamlines projecting from the lateral medial cortex associated with the motor homunculus, and is known to affect the characterisation of the Amyotrophic Lateral Sclerosis (ALS) disease. From a pre-defined CST in one tractography, we tried to infer the CST in another tractography. The goals were, first to investigate the behavior of the TPR and FDR indices, when minimizing the loss function in Equation 4.3; and second to understand whether the combination of mapping and dissimilarity representation in Algorithm 2 could provide a better correspondence between source and target tractography or not.

4.3.1 Data and preprocessing

The dataset used for the experiment is based on dMRI data recorded with a $3T$ scanner at Utah Brain Institute[2], 65 gradients (64 + $b0$); b-value = 1000; anatomical scan ($2 \times 2 \times 2mm^3$). The tractography was reconstructed with the EuDX algorithm [30] using the dipy[3] toolbox. We considered 4 healthy subjects and focused the analysis on the corticospinal tract (CST). CST is a set of streamlines projecting from the lateral medial cortex associated with the motor homunculus. This tract is of main interest for the characterization of neurodegenerative diseases, like the

[2] The authors are grateful to Prof. Mark B.Bromberg, Prof. Lubdha Shah and Prof. Perry Renshaw of the Department of Neurology and the Department of Radiology, University of Utah (US), for their assistance in acquiring MR data

[3] `http://www.dipy.org`

amyotrophic lateral sclerosis (ALS). The CST tracts were segmented by the expert neuroanatomists using a toolbox [74] that supports an interactive selection of streamlines. The size of the segmented tracts is reported in Table 4.2 (see column *size*).

4.3.2 Design experiments

We considered four alternative methods to align tractographies in a common space. The first, as baseline, was the affine registration of the tractographies in a common MNI space using the voxel-based FLIRT method [47]. FLIRT is an affine FA-image based registration, with 6 DOF (degrees of freedom), and uses correlation ratio as the cost function. The registration is defined as follows: first, FA images were registered to the MNI-FMRIB-58 FA template, then the affine transformation was applied to the tractographies. The TPR and FDR index computed between the CST_A and CST_B in common space is reported in Table 4.2 (see column FLIRT). The second (ODON) [72] and the third (GARY) [30] methods computed the affine transformations by minimizing a loss function based on distances among streamlines. The fourth method was the proposed approach based on mapping/combination of mapping and dissimilarity representation as described in Section 4.2 .

To encode the tractography as graph, we used the common distance between streamlines, Mean Average Minimum distance (MAM) [111], based on the Hausdorff distance. Given two streamlines $s = \{x_1, \ldots, x_K\}$ and $s' = \{x'_1, \ldots, x'_{K'}\}$, the distance metric d_{MAM} is calculated as:

$$d_{MAM}(s, s') = \frac{1}{2}(D(s, s') + D(s', s)) \tag{4.8}$$

where $D(s, s') = \frac{1}{K}\sum_{i=1}^{K} d(x_i, s')$, and $d(x, s') = \min |x - x'_j|$, $j = 1, ..., K'$. We calculated d_{MAM} as the mean of the average of the minimum distance between pairs of points along the streamlines. This distance computation is a symmetric one, and thus can take advantage of matrix subtractions that we need to calculate in the loss function Equation 4.3.

The four algorithms were applied to all 8 pairs of tracts (4 pairs left, 4 right), in the following way. Given the segmented CST_A of one subject, as source tract, and the tractography of another subject, T_B, the task was to find the corresponding $CST_B \subset T_B$.

Mapping a tract such as CST, which usually comprises 10^2 streamlines, to an entire tractography T_B, which usually consist of 10^7 streamlines, is computationally extremely expensive because the space of all possible mappings $\mathcal{Q}$ has size $|T_B|^{|CST|}$. For this reason, we introduced a heuristic to retain some of the streamlines in T_B. The intuitive idea was to define a superset of streamlines of the CST for subject B, denoted CST_B^+. The heuristic is in two steps: first, we computed the medoid s_m of CST_B, and the radius $r = \max\{d(s_m, s_i), \forall s_i \in CST_B\}$. Second, we filtered the streamlines in T_B such that $CST_B^+ = \{s_j \in T_B | d(s_m, s_j) \leq \alpha \cdot r\}$, where $\alpha = 3$. An example of CST_A and the extension CST_B^+ can be found in Figure 4.3-B (the CST is in the middle with green colour, while the extension is red). The segmentation task was then to identify CST_B in CST_B^+. The sizes of CST and CST^+ of all subjects are reported in Table 4.1, both as number of streamlines and voxels. The adjacency matrices of both source and target tractography CST_A, CST_B^+ were precomputed for computational speed.

Computing the optimal mapping Q^* requires to solve, even in an approximate way, the minimization problem of Equation 4.4. As a preliminary strategy to approximate the optimal mapping Q^*, we implemented the simulated annealing (SA) [54] meta-heuristic, a reference method for combinatorial optimization. SA requires the definition of a function to move from the current state, i.e. the current mapping Q, to a (potentially

Table 4.1: Data description: for each subject, the size of CST and CST^+ are reported, both as number of streamlines (the third and fourth column), and number of voxels (the last column).

	subject ID	#stream. CST	#stream. CST^+	#voxel CST
Left	202	156	858	453
	204	163	897	465
	209	95	523	322
	212	74	407	349
Right	204	124	682	426
	205	60	330	221
	206	100	550	346
	212	68	374	365

better) neighbouring one. As transition function we used a stochastic greedy one where, given the current mapping Q, one streamline of CST_A is selected at random and then it is greedily re-mapped to the streamline in CST_B^+ providing the greatest reduction in the loss. As starting point of the annealing process, we used the 1-nearest neighbour of CST_A with respect to CST_B^+ after the registration of T_A and T_B. We ran the simulated annealing for 1000 iterations, which required a few minutes on a standard computer[4].

The SFF policy was used to select 50 prototypes for dissimilarity representation as it was suggested in [76]. More detail can be found in the Chapter 3.

4.3.3 Results

The proposed pipeline was applied to map the CST_A to CST_B^+, for all subjects in the ALS dataset (4 CST-Left $+$ 4 CST-Right). We mapped one CST-Left to the other 3 CST-Left, and did the same for CST-Right. Thus,

[4] We are aware that this method of combinatorial optimization can be significantly improved, but we claim that the it was sufficient to do a preliminary investigation of the relation between the loss L and the overlap between tractographies, by means of the Jaccard and BFF index.

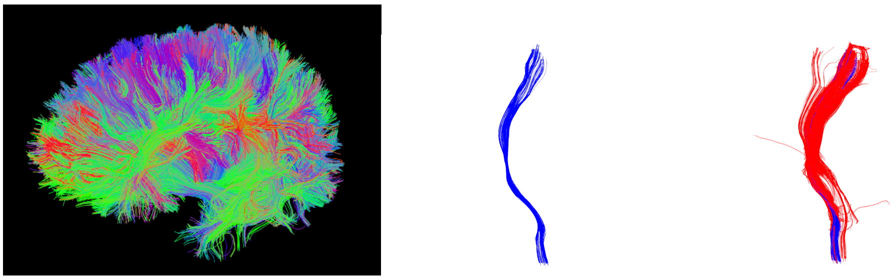


Figure 4.1: The visualization of CST and CST extension. Whole tractography with about 3×10^5 streamlines is on the left. The original CST Left of subject 204 from ALS dataset is in the middle, and the extension of CST Left, with $\alpha = 3$, is on the right.

the pipeline was run 24 times ($4 \times 3 + 4 \times 3$) for each experiments.

Experiment 1: Mapping CST_A to CST_B^+

With each mapping the CST_A to CST_B^+, we used 06 different max-iteration thresholds in SA optimization process: 100, 200, 400, 600, 800, and 1000. These parameters were set empirically as a compromise between fast optimization and good convergence. Two examples of the optimization process of the loss function under different max-iteration thresholds, when mapping subject ID 209-Left and subject ID 205-Right to other subjects, were presented in Figure 4.2. With other mapping, we also got the similar plots. The results showed that the approximation of the minimum value of loss function could be reached from 400 to 800 iterations with simulated annealing.

Due to the limit of paper space, we just present one detail example from all 24 results in Table 4.2, which presents values of TPR and FDR index when mapping CST-Right of subject ID 205 to other subjects. Table 4.3 shows the mean of TPR and FDR index from all 12 mapping results of CST-Left, and 12 mapping results of CST-Right.

The results reported in Figure 4.2 show the behaviour of the loss during the optimization process for the mapping of CST_A (subject ID 205),

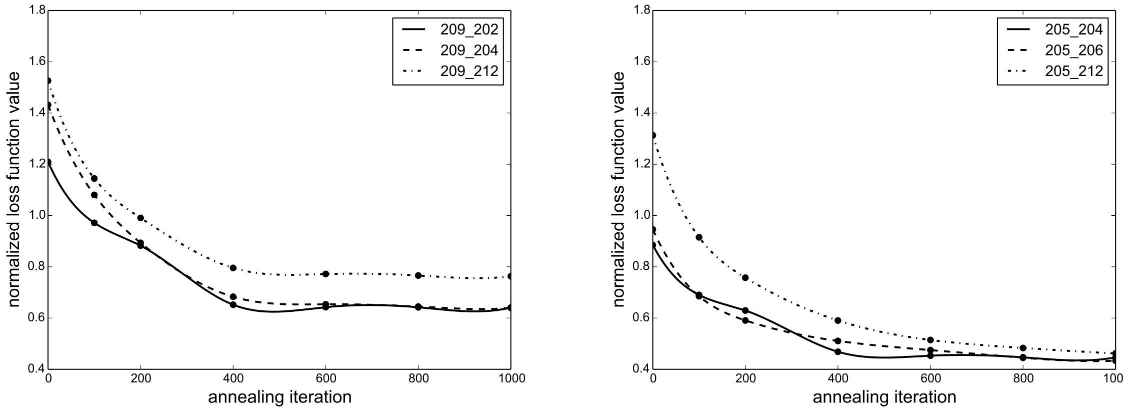


Figure 4.2: Plots of the normalized loss ($L_{norm} = \frac{L}{|CST_A|}$) as a function of number of iterations with simulated annealing, when mapping the CST-Left of subject 209 to those of subjects 202, 204 and 212 (left), and the CST-Right of subject 205 to those of subjects 204, 206 and 212 (right). under different number of iterations in simulated annealing algorithm. The initial state is 1-nearest neighbor (1-NN) of CST_A to CST_B^+ when both of them are in MNI space.

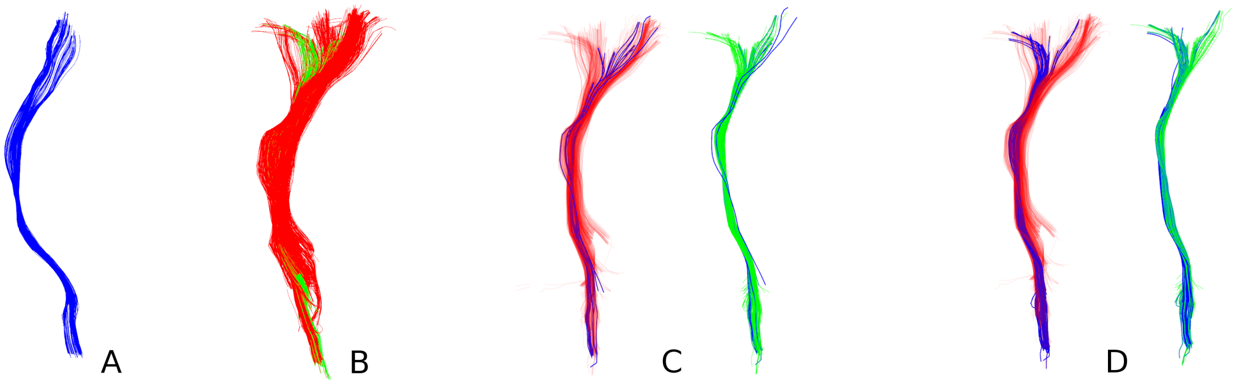


Figure 4.3: (A) - The CST-Left of the subject ID 204 used as source for mapping, CST_A; (B) - Red colour: the subset of the whole tractography of subject ID 202 used as target, CST_B^+; Green colour: the ground-truth CST of the target, CST_B; (C) - Blue colour (on both left and right): 1-NN from source to target after registering to MNI space; Red-left: the target CST_B^+; Green-right: the ground-truth CST_B; (D) - Blue colour (on both left and right): the result of mapping source to target using our proposed method; Red-left: the target CST_B^+; Green-right: the ground-truth CST^T

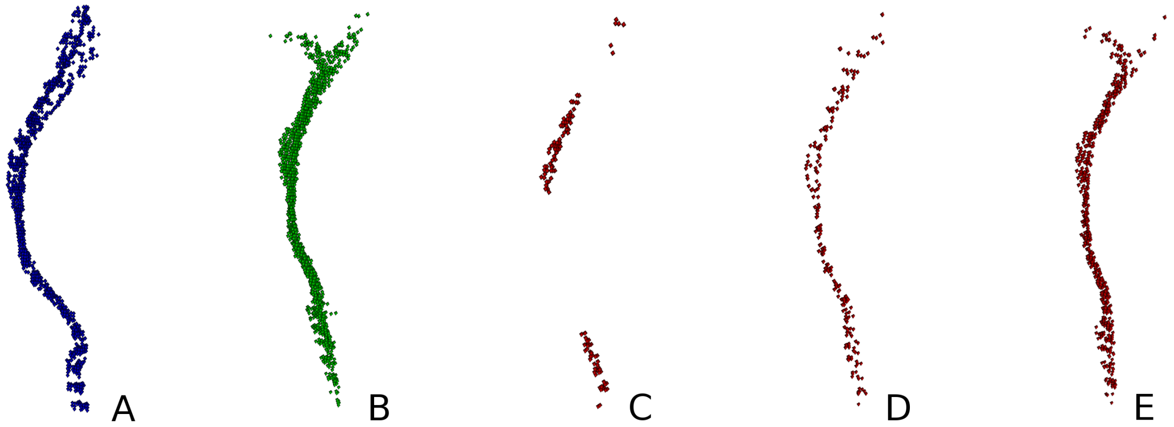


Figure 4.4: The visualization of the volume (voxel unit) of CST. (A) - the CST-Left of subject ID 204 used as source. (B) - the CST-Left of subject ID 202 used as the ground-truth in target. (C) - the overlap of source and ground-truth in target after co-registering both of them to MNI space. (D) - the overlap of 1-nearest neighbour (1-NN) of source in the target, with the ground-truth target, when both of them are in MNI space. (E) - the overlap of mapped source and the ground-truth target using our normal mapping method with 1000 iterations for SA.

with respect to the tractography of three other subjects (subject IDs 204, 206 and 212). In all cases, as the number of iterations increases, the value of loss function decreases. In Figure 4.3 we show an example of experiment with the outcome of FLIRT registration and mapping which refers to subjects 204 and 202. In subfigure A, the source tract CST_A is shown in blue, in subfigure B the target tract CST_B is shown in green and the related superset of streamlines CST_B^+ in red. In subfigure C, the result of FLIRT registration is presented, both with respect to the superset CST_B^+ on the left and with respect to the target tract CST_B on the right. On the right side, it is illustrated the set of streamlines (blue) from the source tract CST_A associated to streamlines of target tract CST_B. The association between streamlines of CST_A and CST_B is computed as nearest neighbour after the FLIRT registration. The ratio between blue and green streamlines represents the portion of target

Table 4.2: Comparison of registration vs. mapping. The subject IDs of CST_A and CST_B are reported in the first two columns (CST Right). Their sizes in number of streamlines together with that of CST_B^+ are in columns three to five. The last nine columns report the overlap between CST_A and CST_B in terms of true positive rate (TPR) and false discovery rate (FDR) for the four compared methods: FLIRT, ODON [72], GARY [30] and our proposed MAPP (mapping) method, respectively.

A	B	size			TPR (True Positive Rate)			
ID	ID	$\lvert CST_A \rvert$	$\lvert CST_B \rvert$	$\lvert CST_B^+ \rvert$	*FLIRT*	*ODON*	*GARY*	*MAPP*
205	206	60	100	550	0.20	0.28	0.27	0.46
	204	60	124	682	0.19	0.20	0.24	0.30
	212	60	68	374	0.16	0.16	0.14	0.47

A	B	size			FDR (False Discovery Rate)			
ID	ID	$\lvert CST_A \rvert$	$\lvert CST_B \rvert$	$\lvert CST_B^+ \rvert$	*FLIRT*	*ODON*	*GARY*	*MAPP*
205	206	60	100	550	0.68	0.89	0.48	0.07
	204	60	124	682	0.64	0.63	0.55	0.37
	212	60	68	374	0.74	0.76	0.74	0.08

tract correctly detected. On the left side of subfigure C, blue streamlines represents the portion of source tract CST_A not associated to target tract CST_B. In subfigure D, the result of mapping is presented, with the same strategy of presentation of subfigure C. On the right side the visualization shows a greater amount of (blue) streamlines correctly mapped into target tract. Even on the left side the amount of (blue) streamlines erroneously mapped is greater. The sum of blue streamlines on the left and right side represents the portion of streamlines projected from the source to the target. The registration based on FLIRT does not preserve after the alignment the same amount of streamlines from the source tract.

The overlap between CST_A and CST_B provided by FLIRT registration is generally quite poor. This is partly expected because even after the registration of T_A and T_B, CST_A and CST_B may have a systematic displacement due to the variability of anatomy across subjects. The results of mapping at different iterations of the optimization process shows a remarkable global increase in the overlap and a general trend of improved alignment when more iterations are computed.

Table 4.3: The average mean and standard deviation of True Positive Rate (TPR) and False Discovery Rate (FDR) for the four compared methods: FLIRT, ODON, GARY and MAPP. The *FLIRT* column is calculated when both CST_A and CST_B are in the MNI space after alignment using FLIRT registration method. The fifth and sixth columns are calculated in the same way but using the tract-based registration method proposed in [72] and [30], respectively. And *MAPP* indicates the results using our mapping method.

	size		TPR (True Positive Rate)			
	$\|CST_A\|$	$\|CST_B^+\|$	*FLIRT*	*ODON*	*GARY*	*MAPP*
Left	122 ± 40	671 ± 220	0.21 ± 0.06	0.14 ± 0.07	0.30 ± 0.05	$\mathbf{0.53} \pm 0.31$
Right	88 ± 27	484 ± 147	0.27 ± 0.07	0.23 ± 0.12	0.26 ± 0.08	$\mathbf{0.55} \pm 0.14$
All	105 ± 38	578 ± 206	0.24 ± 0.06	0.19 ± 0.10	0.27 ± 0.06	$\mathbf{0.52} \pm 0.14$

	size		FDR (False Discovery Rate)			
	$\|CST_A\|$	$\|CST_B^+\|$	*FLIRT*	*ODON*	*GARY*	*MAPP*
Left	122 ± 40	671 ± 220	0.80 ± 0.06	0.87 ± 0.04	0.58 ± 0.07	$\mathbf{0.31} \pm 0.19$
Right	88 ± 27	484 ± 147	0.74 ± 0.07	0.79 ± 0.08	0.65 ± 0.08	$\mathbf{0.32} \pm 0.17$
All	105 ± 38	578 ± 206	0.76 ± 0.07	0.82 ± 0.07	0.62 ± 0.08	$\mathbf{0.31} \pm 0.15$

Experiment 2: Exploring the dissimilarity representation based on mapping

The proposed Algorithm 2 was applied to the full ALS dataset including 4 CST-Left plus 4 CST-Right. The pipeline used 50 prototypes selected by SFF policy as in [76]. Note that, in this experiment, for a fair comparison, we used the 1-NN to choose the correspondences of the source in the target when both source and target were aligned together after registration. We just presented in the Table 4.4 the detail of TPR and FDR for finding the correspondence between CST-Right of subject ID 205 to 3 other subjects. Table 4.5 shows the comparison of our method with other state-of-

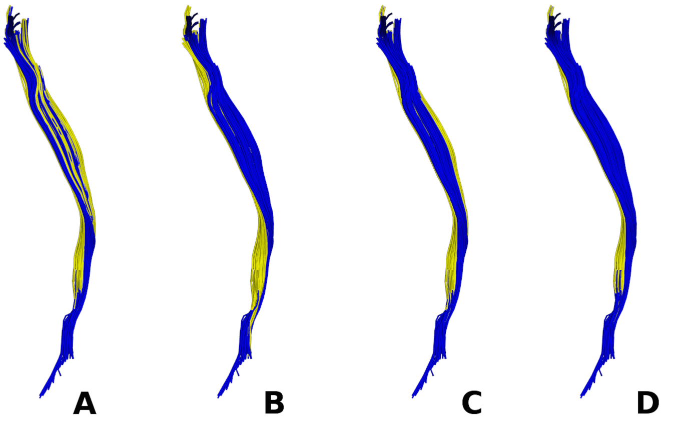


Figure 4.5: Visualization of the result when the left CST of subject 204 is obtained from the left CST from subject 202: (A) FLIRT registration, (B) ODON [72], (C) GARY [30], (D) DMAP, dissimilarity and mapping. Blue color denotes the correctly aligned streamlines, while the yellow color the incorrect ones.

the-art registration methods. It strongly confirms that our method is outstanding with the others in finding the correct correspondences, but not too much better in removing incorrect ones. The reason is that the CST is not symmetric, therefore it is difficult to map correctly. After dividing the dataset according to the size of the CST, we get the better result. Table 4.6 reports the mean and the standard deviation of TPR and FDR for all four methods, over all pairs of subjects. The results are split in two groups because the task is not symmetric: $|CST_A| > |CST_B|$ (source > target) and $|CST_A| < |CST_B|$. The results demonstrate successful segmentation of the CST of the brains when combining the dissimilarity representation with mapping method. Additionally, Figure 4.5 shows an example of CST alignment of the four methods, where blue streamlines indicate the ones contributing to TPR and the yellow ones to FDR.

Table 4.4: The performance of exploring the dissimilarity based on mapping for finding the correspondences of CST-Right of subject ID 205 to other subjects are represented by the TPR and FDR index. The subject IDs of CST_A and CST_B are reported in the first two columns. Their sizes together with that of CST_B^+ are in columns three to five. The last four columns report the overlap between CST_A and CST_B in terms of TPR index (higher is better) or FDR index (lower is better), with FLIRT registration method (6th column), tract-based ODON [72] method (7th column), tract-based GARY [30] method (8th column), and with exploring the dissimilarity representation based on mapping (DMAP). Note that, for a fair comparison, the results of column six, seven and eight are calculated by using the 1-nearest neighbour of CST_A in the CST_B^+ when both are in the common space after registration with FLIRT, ODON and GARY method respectively.

A	B	size			TPR (True Positive Rate)			
ID	ID	$\|CST_A\|$	$\|CST_B\|$	$\|CST_B^+\|$	*FLIRT*	*ODON*	*GARY*	*DMAP*
205	206	60	100	550	0.36	0.32	0.27	0.42
	204	60	124	682	0.22	0.24	0.23	0.29
	212	60	68	374	0.31	0.35	0.22	0.45

A	B	size			FDR (False Discovery Rate)			
ID	ID	$\|CST_A\|$	$\|CST_B\|$	$\|CST_B^+\|$	*FLIRT*	*ODON*	*GARY*	*DMAP*
205	206	60	100	550	0.14	0.22	0.21	0.11
	204	60	124	682	0.39	0.25	0.32	0.40
	212	60	68	374	0.10	0.17	0.41	0.10

4.4 Discussion and Conclusion

With tractography mapping, we have addressed an alternative way for doing tractography registration without affine transformation information. We also have linked the tractography mapping problem to the graph matching problem, and proposed a graph-based solution for tractography mapping problem by optimizing the loss function. We have explored the mapping as a solution for the task of tract segmentation. Experiments have shown that mapping by minimizing the loss function can successfully map one tractography to another tractography. The comparisons of the results from our proposed method and other tradi-

Table 4.5: The average performance of segment CST by combining mapping and dissimilarity approximation is represented by TPR and FDR index. The *FLIRT* column is calculated with the 1-nearest neighbour of CST_A in the CST_B^+ when both CST_A and CST_B are in the MNI space after alignment using FLIRT registration method. The fifth and sixth columns are calculated in the same way but using the tract-based ODON [72] and GARY [30] registration method. And *DMAP* indicates the results of our proposed method in Algorithm 2.

	size		TPR (True Positive Rate)			
	$\|CST_A\|$	$\|CST_B^+\|$	*FLIRT*	*ODON*	*GARY*	*DMAP*
Left	122 ± 40	671 ± 220	0.30 ± 0.09	0.36 ± 0.15	0.36 ± 0.09	$\mathbf{0.41} \pm 0.13$
Right	88 ± 27	484 ± 147	0.40 ± 0.09	0.41 ± 0.14	0.31 ± 0.08	$\mathbf{0.48} \pm 0.11$
All	105 ± 38	578 ± 206	0.35 ± 0.10	0.39 ± 0.15	0.33 ± 0.09	$\mathbf{0.44} \pm 0.13$

	size		FDR (False Discovery Rate)			
	$\|CST_A\|$	$\|CST_B^+\|$	*FLIRT*	*ODON*	*GARY*	*DMAP*
Left	122 ± 40	671 ± 220	0.37 ± 0.16	0.34 ± 0.14	$\mathbf{0.24} \pm 0.16$	0.34 ± 0.17
Right	88 ± 27	484 ± 147	$\mathbf{0.28} \pm 0.14$	0.35 ± 0.13	0.36 ± 0.13	0.31 ± 0.14
All	105 ± 38	578 ± 206	0.32 ± 0.15	0.35 ± 0.13	$\mathbf{0.30} \pm 0.15$	0.32 ± 0.15

tional registration methods (FLIRT, ODON, GARY) have presented that our method is more prominent and demonstrates the potential of mapping for medical or neuroscientific analyses of tractography.

By combining mapping with dissimilarity approximation, we have proposed a new method for constructing the common vectorial representation of streamlines across subjects. The aggregated results for TPR reported in Table 4.6 show that the proposed method, based on dissimilarity and tractography mapping, outperforms the methods based on registration, while keeping a comparable FDR. This occurs both when the source tract is larger or smaller than the target tract. In our opinion, the lower performance of registration methods is a tendency to underestimate the target tract: when the affine transformation results in poor alignment of the two tractographies, the nearest neighbour computation may return the same streamline multiple times. In other words, many

Table 4.6: Average mean and standard deviation of true positive rate (TPR) and false discovery rate (FDR) for the four compared methods: FLIRT, ODON [72], GARY [30] and DMAP (dissimilarity + mapping). Results are aggregated in two groups, according to the size of source and target tracts.

	source < target		source > target	
Method	TPR	FDR	TPR	FDR
FLIRT	0.30 ± 0.09	0.32 ± 0.17	0.39 ± 0.10	0.33 ± 0.14
ODON	0.30 ± 0.08	0.34 ± 0.15	0.46 ± 0.15	0.36 ± 0.11
GARY	0.29 ± 0.08	0.33 ± 0.12	0.37 ± 0.08	$\mathbf{0.27} \pm 0.18$
DMAP	$\mathbf{0.40} \pm 0.10$	$\mathbf{0.30} \pm 0.15$	$\mathbf{0.48} \pm 0.13$	0.35 ± 0.16

streamlines from the source tract are projected to the same streamline of the target tract. This issues does not happen for the proposed method, that better preserves the proportion of streamlines between source and target tracts.

The proposed method presents some limitations too. As mentioned in Section 4.2, the quality of the provided alignment has no theoretical guarantees. There, we speculated that such approximation can be decomposed in three separate parts, i.e., two due to the dissimilarity representation and the one due to mapping; but more work is needed to identify their cumulative effects and interactions. Additionally, the mapping of prototypes is currently limited to a number of streamlines much smaller than a full tractography. For this reason improvements to the scalability issue of mapping algorithms are necessary in order to extend the proposed solution set of streamlines larger than single tracts.

Chapter 5

An Interactive Visual Tool for Tractography Segmentation

Diffusion magnetic resonance imaging data allows reconstructing the neural pathways of the white matter of the brain as a set of 3D polylines. This kind of data sets provides a means of study of the anatomical structures within the white matter, in order to detect neurologic diseases and understand the anatomical connectivity of the brain. To the best of our knowledge, there is still not an effective or satisfactory method for automatic processing of these data. Therefore, a manually guided visual exploration of experts is crucial for the purpose. In order to make the use of the advantages of both manual and automatic analysis, we have developed a new visual data mining tool for the analysis of human brain anatomical connectivity. With such tool, humans and automatic algorithms capabilities are integrated in an interactive data exploration, analysis process and identifying white matter anatomical structures of interest from diffusion magnetic resonance imaging (dMRI) data. However, because of the large size of these data sets, visual exploration and analysis has also become intractable. The difficulty in visual exploration, navigating and analysis segmenting tractographies lies in the very large number of reconstructed neuronal pathways, i.e. the streamlines, which are in the order of hundreds of thousands with modern dMRI techniques. The novelty of our system resides in presenting the user a clustered version of the tractography in which user selects some of the clusters to identify a superset of the streamlines of interest. This superset is then re-clustered at a finer scale and again the user is requested to select the relevant clusters. The process of re-clustering and manual selection is iterated until the remaining streamlines faithfully represent the desired anatomical structure of interest. In this work, we present a solution to solve the computational issue of clustering a large number of streamlines under the strict time constraints requested by the interactive use. The solution consists in embedding the streamlines into a Euclidean space and then in adopting a state-of-the-art scalable implementation of the k-means algorithm. We tested the proposed system on tractographies from real dMRI data set that we collected for a forthcoming study about the systematic differences between the corticospinal tracts.

5.1 Introduction

Brain connectivity analysis is the field dedicated to investigating aspects of the organization and dynamics of the brain. There are three different but related forms of brain connectivity: anatomical, functional and effective [55]. Our work focuses on the anatomical connectivity. Diffusion magnetic resonance imaging (dMRI) [6] is a non-invasive technique, well established in the neuroimaging community for this purpose. It measures the translational displacement (diffusion) of water molecules in the brain tissue, which is mechanically constrained by the myelinated axons. Thereby, it provides information about the local orientation of white matter axons. The data obtained with this technique, can be used to extract the anatomical connectivity information by using

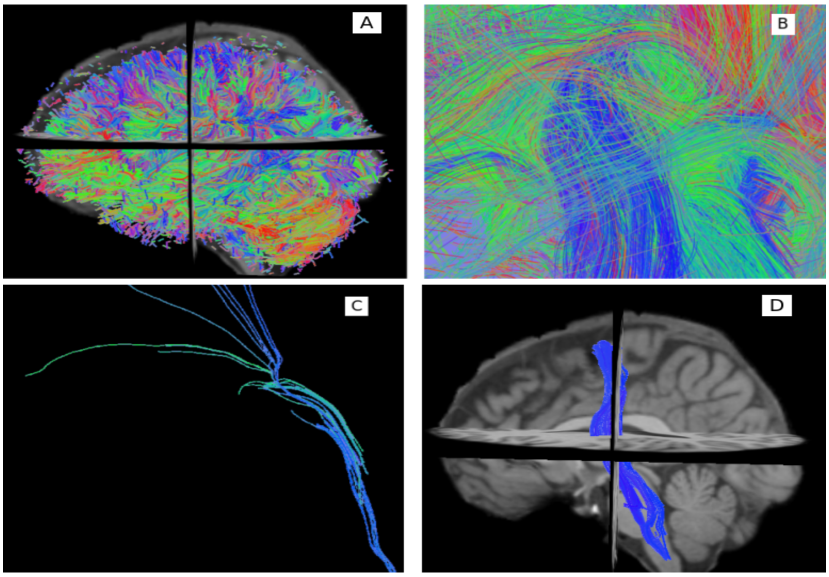


Figure 5.1: (A) Tractography overlaid with the structural image(only 10% of the streamlines are shown), (B) Amplifying an area of the tractography (C) Small subset of streamlines (D) Corticospinal tract right of a healthy subject with the 3D view of the structural image.

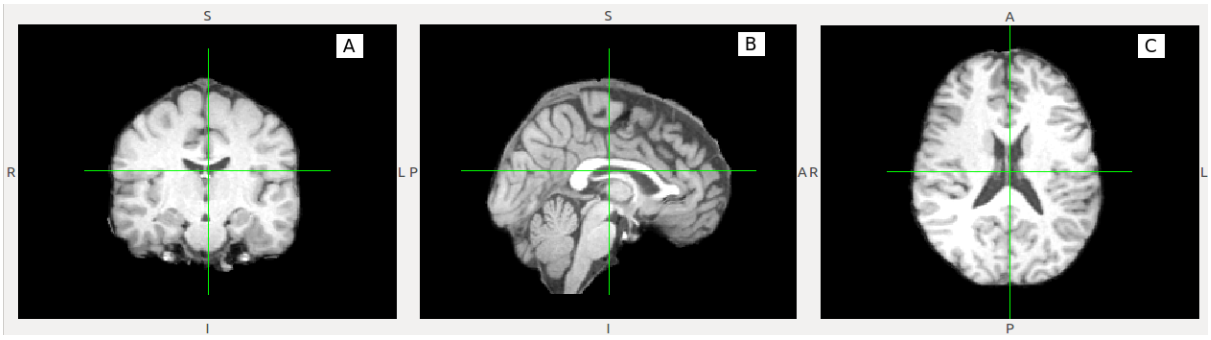


Figure 5.2: The structural image of the brain with different type of views. The 2D views: (A) coronal, (B) sagittal, (C) axial

deterministic tractography algorithms [69], [56], [30]. These algorithms reconstruct the approximate trajectories of the axons as polylines, so they resemble the white matter anatomical structures (see Figure 5.1). A polyline in this context is called *streamline*, and the full brain streamlines are called *brain tractography*. It is worth to notice that one streamline represents 10^4 axons approximately.

The exploration of tractography data sets has become then very useful to neuroanatomists. Information like the shape of streamlines, their spatial location and relation with each other, allows for the identification and study of anatomical structures of interest within the white matter, i.e. locating subsets of streamlines which are related to specific function(s), and from which it can be also determined if there is (or the status of) an ongoing neurodegenerative process (see Figure 5.1.C). For this purpose, besides the tractography, a high-resolution structural magnetic resonance image is typically available. This image shows a good contrast between gray matter and white matter, thus it is commonly used as reference for visualizing and studying brain anatomy (see Figures 5.2 and 5.1.D), e.g. if a neuroanatomist wants to explore the fornix structure, she knows that it is the bundle of fibers (axons) that carries signals from the hippocampus to the hypothalamus. With these data, there are two main approaches for the study of anatomical connectivity: automatic and manual. The automatic analysis has gained popularity over the last years, and it is based on machine learning and data mining algorithms, mainly for clustering (for more details see Section 5.2). It is mainly aimed at a fast segmentation of the white matter into sets of streamlines that follow similar trajectories [105], [39], [91]. Nevertheless, the automatic segmentation of the tractography is not always in agreement with the real anatomical structures of the white matter. Therefore, neuroanatomists still strongly rely on their manually guided visual exploration. This manual task though, is complex and slow. The manual exploration of the streamlines is usually supported by the overlaid structural image, such that experts can orient themselves into which regions of the brain they are focusing their analysis (Figure 5.1). Moreover, the number of streamlines can be really large, usually in the order of hundreds of thousands, making the exploration i.e. shape recognition, spatial localization, quite difficult. See for example, in Figure 5.1.A, where only a 10% of the total amount of streamlines is shown, it is still difficult to visually understand the data. For the previous presented reasons we believe that combining the advantages of automatic and manual exploration can be a solution to improve the current practice.

Visual data mining is precisely the field that aims at integrating humans in the data-mining process by interaction with visual representations of abstract data, thus applying humans perceptual abilities and their domain expertise for the analysis of large datasets [95], [52]. Visual data mining tools allow viable data exploration and often provide effective results [52], [95], [98]. This field strongly relies on the visualization and interaction, therefore a new requirement is added to the standard data mining process. Scalability becomes a major problem not only in the automatic analysis of the large data sets, but also in the interactive visualization [28], [53], [95]. The strategy is then to develop a solution which integrates high-performance analysis algorithms with appropriate and efficient visualization and interaction techniques.

In this paper, we propose a visual data mining tool, Tractome, for the analysis of human brain anatomical connectivity. This work focuses on computer-assisted tractography segmentation and describes the solution we developed to build a software system to support neuroanatomists and medical doctors in studying the white matter. With such tool, both the automatic algorithms and neuroanatomists capabilities are integrated in an interactive tractography exploration process. To the best of our knowledge, this mixed approach for tractography exploration has not yet been proposed in the literature.

In Tractome, we address the scalability issue as it demands. For the interactive visualization, we follow the visual scalability analysis made in [28]. With this purpose, we define a *visual metaphor* that allows presenting the experts with a comprehensible summary of the tractography. It makes also possible the concurrent reading of the anatomical connectivity when they are presented with an overlapped view of the structural image and the tractography. Moreover, interactive techniques [52] are exploited in the tool, such that the neuroanatomist can browse through the streamlines and see the details of any anatomical structure of interest on demand.

To obtain the abstract or compact version of the tractography, the system relies on automatic clustering of the data. If we want to provide a comfortable user experience, the clustering algorithm has to be fast and scalable. However, the standard representation of tractography data prevents the achievement of this goal. Streamlines are 3D polylines that can even have different lengths and different number of points (see Figure 5.1 for a visual representation of streamlines). This representation may hinder the use of many machine learning and data mining algorithms, as most of them are based on a vectorial representation of the data. Even though some algorithms do not strictly require a vectorial representation of data, it frequently allows fast and scalable implementations. In the designed tool, we use what is called the Dissimilarity Representation (DR) approach [87], in order to embed the tractography data into an Euclidean space. The DR maps each streamline in the space of its dissimilarities with respect to a selected set of representative streamlines, called prototypes. This leads to a vectorial representation of the streamlines, where the domain information is taken into account, provided that an informative streamline-to-streamline distance function is available. A crucial step in the DR approach is the selection of the prototypes [87], [76]. Thus, given the large number of streamlines, a fast prototype selection algorithm is also needed. For this tool, we propose a domain-based heuristic that improves the general prototype selection policy that was proposed in [76]. This policy is a stochastic approximation of an effective algorithm for prototype selection that scales well with the large collection of streamlines.

With respect to the clustering task, we recently proposed the application of the mini-batch k-means algorithm [94] for segmenting tractography data [74]. Mini-batch k-means is an approximation of the k-means clustering algorithm, which dramatically reduces the computation of clusters in very large data sets.

In Section 5.2, we give more details about tractography data and the available tools for their analysis. The designed visual data mining tool is presented in Section 5.3, where we explain in detail how we approach the scalability issue in every stage of the whole process. The detail of Tractome software architect is described in Section 5.4. Section 5.5 is dedicated to the experimental part, where we show in practice how our tool performs on a real data set. Some case studies by using our software Tractome for different neuroscientific analyses purpose are described in Section 5.6. Finally, the conclusions of this study are drawn in Section 5.7.

5.2 Basic Concepts and Related Works

When reconstructing white matter axons, the result is a set of streamlines $T = \{s_1, s_2, \ldots, s_m\}$, which is called *brain tractography*. Each streamline s_i is defined as a polyline $s_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_{s_i}}\}$, with $\mathbf{x}_{ij} \in \mathbb{R}^3$. We use n_{s_i} to denominate the number of points of streamline s_i because this number usually differs from one streamline to another. A set of streamlines with anatomical meaning is called a *tract* (see Figure 5.1.C).

Tractography data can be analysed automatically and manually. In fact, the literature on methods for automatic extraction of specific tracts has increased over the years. Both supervised and unsupervised learning have been used for this purpose [91], [39], [105]. However, it can be difficult to model the different structures into groups over different subjects, hence unsupervised approaches are most common (see [105] for a larger review). The idea is to partition the whole tractography into clusters, such that streamlines that are spatially closer and have similar shape are grouped together. Thereby, the obtained clusters are supposed to represent discernible tracts. Still, the drawback of just applying a clustering algorithm is that the grouping of streamlines is usually not optimal for the detection of anatomical structures. False merges of streamlines into the same cluster or false splits into different clusters may take place, leading to a segmentation of the tractography that may not have an anatomical meaning. The most common approach to deal with this issue is to incorporate background anatomical knowledge from experts in the process. The main idea has been to create a template or atlas of one or several tracts, which can be obtained by a manual extraction of known white matter tracts from a set of subjects [73], [39], [91], [103]. This approach is unsatisfactory as it introduces the problem of corregistration of brain tractographies of the different subjects. Other works propose to learn a model from the data as well as the clustering parameters without supervision [105]. In any case, even if a reasonable segmentation of the tractography is obtained, this is useful for analysing the main (most distinguishable) tracts only.

Even though the automatic tract extraction is quite popular, in the clinical domain preference is usually given to manually exploring and analysing the data. Trackvis [1][104] for example, is a software tool commonly used by the neuroanatomists and doctors for this purpose. Anyhow, the manual approach of this kind of tools can make the task lengthy

[1] http://trackvis.org

and complex. Moreover, the exploration of the entire brain tractography is not possible. Thus, they provide sub-sampling and filtering tools to reduce the amount of objects visualized on the screen, such that users can explore and locate specific tracts[104]. Filtering in this case is usually based on manually placing one or more regions of interest (ROIs), such that the set of streamlines passing through these ROIs are selected. However, a problem of this approach is that, to the best of our knowledge, the definition of ROIs in this kind of tools is indirect, as they are defined on the structural image and not on the tractography data. Hence, in order for the extraction of the tracts to be accurate, it strictly requires for the structural image and tractography to be correctly aligned. Another important aspect is that this kind of approach is strongly sensitive to the quality of the reconstructed streamlines. Due to noise in the measurement process, or to the performance of the tractography algorithm, the reconstruction of a streamline may fail, resulting instead into multiple disconnected polylines. As a consequence, these streamlines would not fulfil the condition of connecting the ROIs defined by the experts. When using this approach, it is a common experience to get a reduced or non-existent set of streamlines, instead of the actual tract.

5.3 Tractome: A Visual Data Mining tool for tractography analysis

Following the concepts and general scheme of visual data mining as a human-centred discovery process, we define an interactive visual data mining tool, Tractome, for the exploration and analysis of tractography data. It is based on the Visual Exploration Paradigm (Visual Analyt-

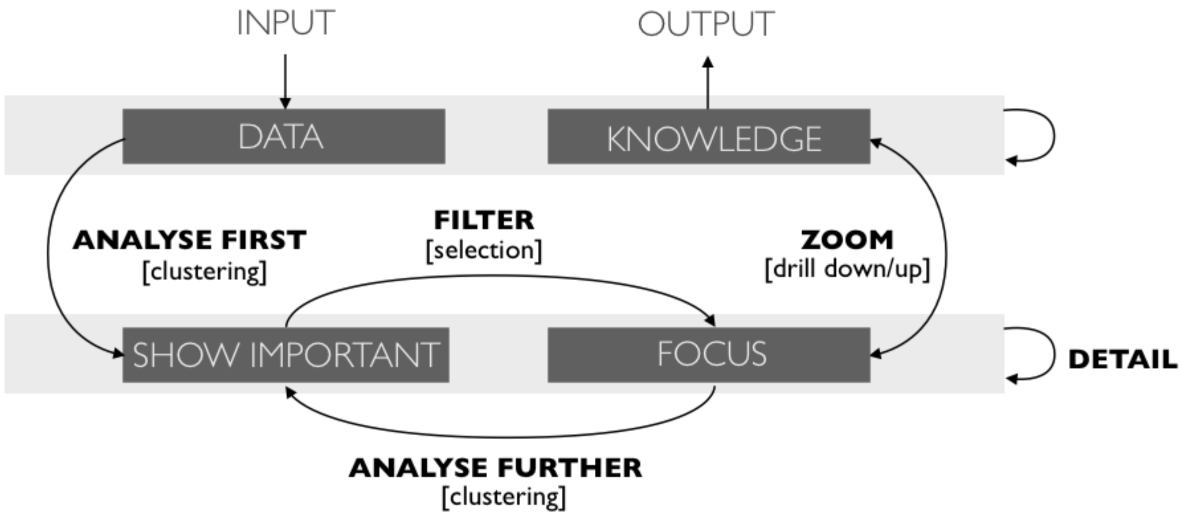


Figure 5.3: Workflow of the Visual Data Mining tool for analysis of tractography data.

ics Mantra): *"Analyse First, Show the Important, Zoom*[2]*, Filter and Analyse Further, Details on Demand"* [53] and follows visual scalability requirements [28]. Thereby, experts are provided with a tool that allows them for fast and easy exploration of tractography data. Moreover, they are involved in the knowledge discovery process, by guiding the exploration with their knowledge on the domain. See Figure 5.3 for the workflow diagram.

Given that tractography algorithms can generate an extremely large number of densely packed streamlines, it is difficult to interact with these data or to do any visual interpretation (see Figure 5.1). Therefore, the initial stage of the system consists in *Analysing First* the data, by segmenting the tractography with a fast-clustering technique. More technical details about the implementation of this step will be given at Subsections 5.3.1 and 5.3.2. As result, the tool *Shows the Important* by presenting a simplified version of the data set, i.e. the clusters represented

[2] Zoom is referred here to the definition in [52], which means that the data representation changes to present more details at higher zoom levels.

by one streamline each, together with a 3D perspective view (the 3D slicer, see Figure 5.1.D) of the structural image. Thereby, the expert can start the exploration with an overview of the whole tractography, and the interaction and concurrent reading of the data with the structural image is much simplified. A clearer view will be now available, instead of the crowded picture with thousands of streamlines. In case the user wants to see the whole data, she can *Zoom* on the clusters, and a view of all the streamlines is presented. Afterwards, the expert can interact with the system by doing some *Filtering*, such that she can *focus* only on the clusters related to the anatomical structure of interest. Based on the selection, the user can *Zoom* again on the chosen representatives and will be able to see the streamlines belonging to the corresponding clusters. From this first cycle, the expert can refine the exploratory process by specifying the parameter of the clustering algorithm, i.e desired number of clusters, re-cluster again the streamlines belonging to the focused area, and drill-down in order to inspect the details about the data. In this way, she can explore and *Analyse further* until the anatomical structure(s) of interest is found. At any iteration of this procedure, the expert can manipulate and explore the data in more *Detail* by using common visual interaction tools like panning, rotation, translation, traditional zooming. Subsection 5.3.3 is dedicated to defining the visualization and interactivity techniques that are used in the tool.

We have continuously stated that, in order for the interactivity feature of the tool to be successful, we need to tackle the scalability problem. Hence, in the following subsections, it will also be explained how our choices, for the different steps of the system, are meant to deal with this issue.

5.3.1 Dissimilarity Representation

Despite that a vectorial representation of data is not an absolute requirement for machine learning and data mining, many algorithms rely on this type of representation. It facilitates the development of efficient algorithms, as it allows performing fast queries, as well as algorithms based on simple linear algebra operations.

Given that the size and complex representation of the brain tractography data is in contrast with the scalability requirement of our interactive tool, we apply a previously proposed vectorial representation for streamlines [76].

The Dissimilarity Representation (DR) [87] approach was mainly introduced for classification purposes. However, it has also been used in the context of unsupervised learning [87], [79]. In this approach, new features are defined for the objects, such that they are represented by their (dis)similarities to a set of objects that are representative of the problem at hand, i.e. the prototypes. In such way, every object is then represented by a vector of dissimilarities, instead of the attributes from the original feature space.

So, let us define the DR approach based on the tractography application. Given $\mathcal{X}$ our space of objects i.e. the streamlines, and the brain tractography $T \subseteq \mathcal{X}$ (as defined in Section 5.3), let $\Pi = \{\hat{s}_1, \hat{s}_2, \dots, \hat{s}_p\}$, $\Pi \subset \mathcal{X}$ be a set of prototypes of size p, and let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ be a dissimilarity measure between streamlines. A mapping $\phi(\cdot, \Pi) : T \to \mathbb{R}^p$ is done, such that every object is associated with its dissimilarities to all prototypes in Π, $\phi(s_i, \Pi) = [d(s_i, \hat{s}_1), d(s_i, \hat{s}_2), \dots, d(s_i, \hat{s}_p)]$. A way of handling the DR is to interpret the dissimilarity vectors as features. The obtained mapping to $\mathbb{R}^p$ is equipped with the traditional inner product and Euclidean metric, and we have the so-called dissimilarity space

(DS). In this way, the dissimilarity matrix $\Phi(T, \Pi)$ is used as input for the classification or clustering algorithms [87], [27].

The prototypes, are usually selected as the most representative objects of the data set, i.e. $\Pi \subseteq T$, or Π might be even the equal to T. However, in our case, due to the large number of streamlines, computing the whole dissimilarity matrix would be computationally infeasible. Hence, we need to find a set of prototypes, such that the computational cost is reduced. Moreover, note that this Euclidean embedding is a lossy one, in the sense that in general it is not possible to reconstruct s_i from $\phi(s_i, \Pi)$. The quality of the representation is strongly dependent on the choice of d and the selection of the prototypes. Therefore, it is important that we use a suitable distance and an efficient method to select effective prototypes for our application. The following two subsections are dedicated to these topics.

Dissimilarity Measure

One topic of research in the literature about tractography data analysis, is the selection of a suitable distance between streamlines [111]. It should allow the incorporation of domain specic information when clustering streamlines. Moreover, if we map the tractography into the DS, the geometry and structure of the data will be determined by the measure. Therefore, it is important to choose a measure that fits for the problem at hand. Most of the studied distances for streamlines are modified versions of the Hausdorff distance [26]. They are based on the set of minimum distances between each of the points of the compared streamlines(see Figure 5.4). In our work we use the symmetric minimum average distance [19], which is defined as:

$$d(s_A, s_B) = \frac{1}{2}(\delta(s_A, s_B) + \delta(s_B, s_A)) \tag{5.1}$$

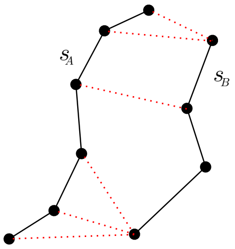


Figure 5.4: Set of minimum distances (dotted lines) between each point of two streamlines s_A and s_B (solid lines).

where

$$\delta(s_A, s_B) = \frac{1}{n_{s_A}} \sum_{x_j \in s_A} \min_{y \in s_B} ||x_j - y||_2 \tag{5.2}$$

given that n_{s_A} is the number of points of streamline s_A.

This measure is symmetric, which is a desirable property in order to remove inconsistencies when changing the order of the streamlines in the comparison. Unfortunately, it is non-metric, as $d(s_A, s_B) = 0$ does not necessarily imply that $s_A = s_B$. However, the metric requirements are not essential for the DR approach. It has been shown that this modified version of Hausdorff [19] gives better results when clustering streamlines [111]. Therefore, it is one of the most common measures used in this domain [19], [111], [39].

Prototype selection Method

Due to the size of tractography data sets, the computation of the whole dissimilarity matrix, i.e $\Pi = T$, for the DR approach has two main disadvantages: excessively large storage requirements and very high computational complexity. Therefore, it is required to compute a small set of prototypes, in order to tackle these problems. Nevertheless, we cannot use just any prototype selection method. In this case, we need an efficient procedure that can scale well on large data sets.

Random Selection for example, is a very common algorithm and with the lowest computational complexity $\mathcal{O}(1)$. In this case, given our set of streamlines or tractography T, a subset Π of p number of streamlines is randomly selected. This algorithm has shown to work reasonably well for the DR approach [86], [87].

In [76], the authors proposed the use of the Subset Farthest First (SFF) [100] algorithm for selecting effective prototypes from tractography data. This procedure is a stochastic scalable approximation of the well known Farthest First Traversal (FFT) algorithm, which has a computational complexity of $\mathcal{O}(p|T|)$. It is also claimed that it reduces the chances to select outliers [100]. The SFF first samples $m = [cp \log p]$ streamlines from T uniformly at random, where $c = 3$. Afterwards, FFT is applied on this subsample i.e. one streamline is randomly selected as prototype $\hat{s}_1$ and a new prototype is iteratively added such that it is the streamline maximising the distance to the already selected prototypes. It was proved that under the assumption of p clusters in T, the probability of not having a representative of some clusters in the sample is at most $pe^{\frac{-m}{p}}$ [100], i.e. a sample of size m is a meaningful summary of T. The computational complexity of SFF is $\mathcal{O}(p^2 \log p)$. For large data sets and small p, this prototype selection policy has a much lower computational cost that FFT.

Even though the SFF policy has shown to be a suitable option for prototype selection in tractography data sets, it is still based on an initial random selection of streamlines. Therefore, the final representation set will depend on whether the random selection is a poor subsample of the tractography or not.

In this paper we propose a new algorithm of prototype selection for tractography data sets, named Spatial SFF (S+SFF). It is basically the SFF algorithm with a new heuristic, which is based on background knowledge information about the application. When looking at a full brain tractography, it can be noticed that there are many short streamlines. It is very likely that these are just noise artifacts that have, of course, no anatomical meaning. Therefore, the first step of our algorithm is to exclude these streamlines for our selection. Thus, given z_{min}, a minimum allowed size for a streamline, we will just use the set of streamlines $T_{min} = \{s_1, s_2, \ldots, s_l\}$, such that $z_s > z_{min}$. Moreover, longer streamlines are believed to have greater potential to be useful landmarks, as they are more likely to be present in most subjects [32]. Hence, before proceeding with the selection, streamlines will be organized in descendent order according to their size. The goal of the selection step is to obtain a subsample of spatially distributed streamlines, such that streamlines from all brain areas are considered. The procedure is as follows: the longest streamline is selected and the next longest streamline is iteratively added, if it is not intersecting with any of the previously selected streamlines. Afterwards, the FFT is applied on the obtained subsample, as for the SFF algorithm. The computational cost of S+SFF is $\mathcal{O}(|T| + l \log l + pl)$. This is more computationally expensive than SFF, but in practice $l << |T|$, therefore it is still feasible.

In Section 5.5, we investigate the trade-off between accuracy and computational cost across the different prototype selection policies and different numbers of prototypes. The policy used in this tool was selected based on this study.

5.3.2 Clustering

As mentioned in the beginning of this section, an important aspect for the design of this tool was the selection of the clustering algorithm. This clustering action has two main functions in this tool. The first one is to use the obtained clusters as a way to present the user with a summary of the data, thus favouring the visual scalability of the tool (see subsection 5.3.3 for more details). Second, to provide the user with a segmentation of the data with a "possible" anatomical meaning, such that she has a guide for the visual data mining process. However, even when the later gives an added value, it is not our aim to find the most meaningful or accurate partitioning of a tractography, as it is the case of the traditional automatic analysis of tractography data sets (as referred in Section 5.2). Our main goal is to use the clustering results to provide the user with a comfortable visualization and interactivity. Therefore, the core problem is to use an algorithm such that a large number of streamlines can be clustered in no more than a few seconds.

The k-means clustering algorithm is very popular in machine learning and data mining, due to its simplicity and efficiency. Given k, the number of clusters, the clustering problem is to find k cluster centres $C = \{c_1, c_2, \ldots, c_k\}$, $c \in \mathbb{R}^p$, and to assign each element of the vectorial data set $\Phi(T) = \{\phi(s_1), \phi(s_2), \ldots, \phi(s_m) \in \mathbb{R}^p\}$ to the closest cluster. The k-means is then based on computing centres C such as to minimise the loss function $f(C) = \sum_{\phi(s)\in\Phi(T)} D(\phi(s), C)^2$, where $D(\phi(s), C) = \min_{c\in C} ||\phi(s) - c||_2$ is the distance between $\phi(s)$ and its closest centre. Even though it has been shown that the computational complexity of the standard implementation for k-means is much less than the theoretical bound of $\mathcal{O}(m^{34})$ in practical applications [4], it is impractical when clustering tractography data in an interactive setting. For this purpose, we recently introduced the use of a scalable k-means implementation, known as mini-batch k-means (MBKM) [94], for clustering streamlines [74].

The MBKM algorithm is a modification of the standard k-means algorithm, that is able to reduce the computational costs by orders of magnitude. Instead of updating centres with one streamline at a time, the

MBKM uses multiple random subsets of the data sets, termed as *mini-batches b*, in order to update the cluster centres and to estimate the per-centre learning rates. The stopping criterion is also the convergence of the loss function $f(C)$. The computational complexity of the MBKM algorithm is not known in the general case, but empirical results in [94] show a reduction of two orders of magnitude in computation time with respect to the standard k-means. We show similar results for tractography data in Section 5.5.

5.3.3 Data Visualization and Interaction

Up to know, we have described the requirements of the proposed tool from the point of view of the automatic algorithms. However, the success of a visual data mining tool does not only depend on the computational complexity and scalability of the algorithms. The visualization and the interactivity feature are also essential factors in its workflow. In this subsection, we define the visualization and interaction techniques used in this tool. These were selected by following the literature about visual scalability [28], [89].

- **Visual Metaphor:** According to the literature, one of the most important factors affecting the visual scalability is the selection of a proper visual metaphor [28]. These are the visual objects by which the original data are encoded and displayed. They should be related with the domain and be easy for the user to understand. Thus, in order to present the user with an overview of the data and to *Show the Important*, we have defined a visual metaphor [28], which is defined as: "the medoid streamline of a cluster". Thus, given the tractography T and a partition of the data P, the summary of the tractography data will be visualized as $M = \{\mu_1, \mu_2, \ldots, \mu_k\}$, where μ_i is the medoid streamline of cluster $p_i \in P$ (Figure 5.5 B). We use medoids instead of centroids because we want to show real representatives of the data. Averaging objects, i.e. streamlines, does not make sense in this application, it has no anatomical meaning. Moreover, in order to compute centroids we would have to do it in the DR and we have no way to invert the DR in order to create the corresponding streamline.

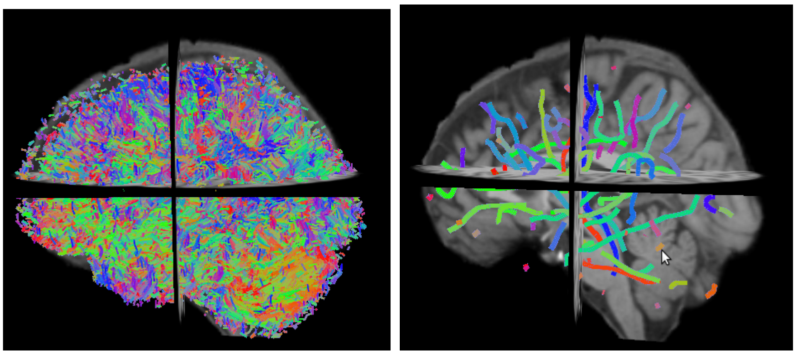


Figure 5.5: Analyse First and Show Important. (A) Whole tractography. Streamlines are colored by following the Directionally Encoded Color convention [81]. (B) Visualization of summary of the data by medoids of clusters.

- **Interaction techniques:** As explained above, from the summary of the data, the neuroanatomists can explore the tractography by interactively filtering and visualizing the streamlines that belong to the specified clusters. In this tool, we use *Selection* and *Zooming* as interaction techniques [52], [53]. Other common interactivity options are also available.
  - Selection: The user can iteratively select and focus on a set of medoid(s)/cluster(s) of interest, by just pointing and clicking it (them). The unselected clusters (those of no interest) can be hidden. Notice that the selected medoids change color (white) in the view, because this color is not used in the color convention for painting streamlines. See Figure 5.6.

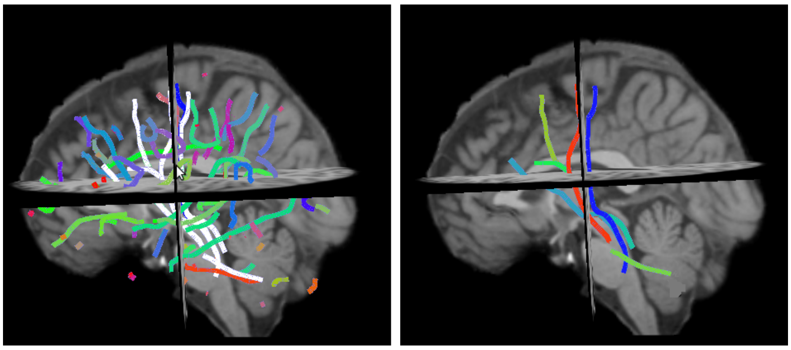


Figure 5.6: Filtering. (A) Selected clusters (in white) (B) Only selected clusters are shown. The rest were removed from the view.

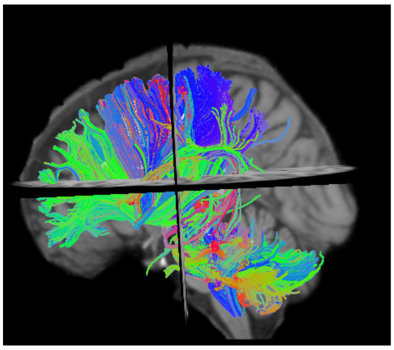


Figure 5.7: Zooming. Streamlines belonging to the selected clusters are shown in the view, with a different line thickness and colors.

– Zooming: The selection can be expanded, such that all streamlines belonging to the selected clusters are also visualized, thus providing the expert with a way to drill-down for a higher level of detail. It is worth to remark that the detailed composition of each cluster, i.e. the set of streamlines, is displayed in a different line thickness, such that the expert can differentiate between the different levels of detail (see Figure 5.7).

– Other interactivity tools: The user can interact with the data by panning on the view and also by using the standard zooming to amplify or decrease the size of the displayed objects. Moreover, the scene can be modified by rotating it, or by translating the 3 orthogonal planar views (slices). Another tool is to hide/show the representatives of the clusters on demand.

In Section 5.5 we describe in more detail how the tractography exploration works by means of these techniques, on a real example.

5.3.4 Limitations and potentialities of Tractome

In general, our current version of Tractome has some issues that can be considered as limitations. However, many of them have potential for future research works and new functionalities of the tool. These issues are:

1. *Effective abstraction:* The main goal of the clustering step(s) in Tractome is to support an abstraction of the data that enables a more comfortable visualization/interaction. However, it is to some extent also desirable that the obtained partition is "meaningful" in order to favour the exploration process. When this is not the case, i.e. streamlines of similar shape fall in different clusters or those with different shape fall in the same one, it is currently not possible to readjust the composition of the clusters by selection. As there is no direct selection of streamlines, but selection of representatives (clusters), the neuroanatomist may find two problems: 1- She cannot remove specific unwanted streamlines from the selected clusters, therefore she will have to carry them to further clustering steps until they can be separated and ignored. 2- Some streamlines from other clusters are desired in the obtained bundle, but they are removed in the selection process because the whole cluster is not of interest and therefore the corresponding representative is not selected. The solution to these problems can be content for future research works.

2. *Preprocessing computation:* In order to obtain the initial clustering of the data the first time the system runs, a number of pre-computations are needed. However, these pre-computations can be included in the pipeline for obtaining the tractography from the dMRI data, such that they are computed once and later used by the system.

3. *Mixed strategy:* The approach of Tractome, based on the selection of a set of streamlines (bundle) of interest, is in contrast to that of ROIs (see Section 5.2). However, combining the two approaches may bring additional benefits in the process of finding the anatomical structure of interest. Part of our future work for the new version of Tractome is to support this mixed strategy for the exploration of tractography.

4. *Segmentation assessment:* We have verified with the neuroanatomists and also shown in the example of Subsection 5.5.2, that the expert should be able to arrive to the desired set of streamlines in a few steps. However, this is only based on practical experience. There is no information available on a "correct" segmentation of the tractography, therefore it is difficult to do a proper evaluation on how accurate the tool is.

Tractome was created with the specific purpose of solving a practical problem of great importance, i.e. the analysis of human brain natomical connectivity. However, we believe that the methodology proposed for this tool, i.e. DR + Clustering + Visualization and Interaction, can be used as a template for designing Visual Data Mining tools for other kinds of data which have e.g. complex non-vectorial representations. In this approach, the most determinant step for generalization is the one of the DR. Its success depends on finding a dissimilarity measure that allows taking into account relevant characteristics of the analysed data.

Moreover, it should be ensured that the relative distances between objects in the original feature space, is preserved in the dissimilarity space. The main advantage of going to this representation is that now the transformed data will be in a Euclidean space, which allows performing linear algebra operations that facilitate the development or application of efficient algorithms. Such is the case for example, of the MBKM which can only be applied on the proposed dissimilarity space. With respect to the visualization and interaction techniques, they can also be applied, if the Visual Metaphor based on the medoids and the Selection and Zooming techniques fit the purposes of the specific application. In this case, it is necessary to find an appropriate visual object that is related to the domain and easy for the user to understand.

5.4 Software Architecture

Tractome is organized as a three-layer model [3] which is often used in software development and known as a well-established software architecture pattern. Three-layer architecture describes the separation of functionality of a software into different layers including presentation (or user interface) layer, application (or business) layer, and data (or low level) layer; therefore, it allows to change or upgrade any of the three layers independently in response to changes in requirements. The general view of Tractome in three layer pattern is showed in Figure 5.4. In this part, we will describe each layer of the Tractome software. The all functions of each class in Tractome is showed in the Figure 5.4.3.

[3] It was developed by John J. Donovan in Open Environment Corporation (OEC), a tools company he founded in Cambridge, Massachusetts.

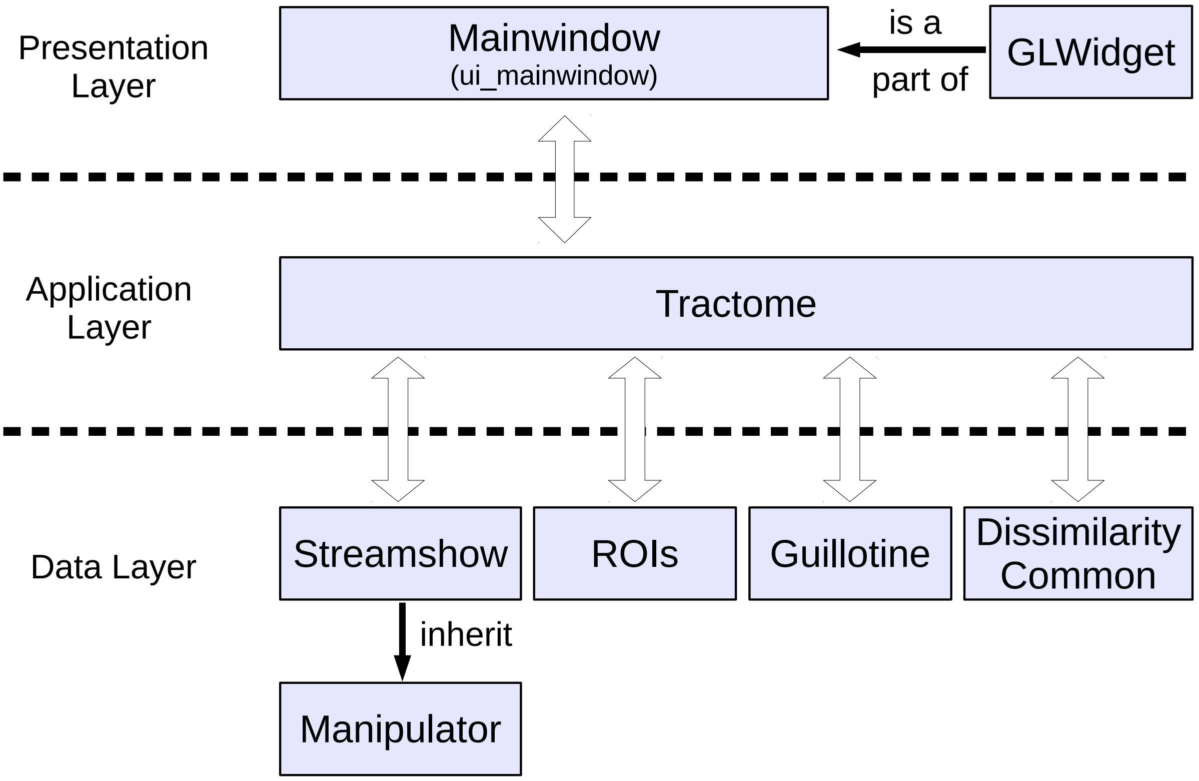


Figure 5.8: The architecture of Tractome software in three-layer pattern.

5.4.1 Presentation layer

This is the topmost level of the application. The presentation layer provides the application user interface (UI). Typically, this layer involves the use of Graphical User Interface (GUI) for user interaction, such as load structural image, load tractography, load segmentation, create ROI, apply ROIs, cluster tractography etc. The presentation tier also displays information related to the current state of the segmentation such as name of object, number of clusters, number of streamlines, voxel-size, volume, etc (the left of the interface in Figure 5.4). The presentation layer communicates with the application tiers by outputting the request from users and getting back the changes for visualization from the application tier.

In this presentation layer, *GLWidget* class (in the file glwidget.py) is inherited from the common QGLWidget class of QtOpenGL [4]. It is a widget for rendering and displaying OpenGL graphics. Some basic properties are the width/height of the widget, the back ground colore (bgcolor), and the orthogonal projection mode or not (ortho). GLWidget provides three convenient functions to perform the typical OpenGL tasks: initializeGL (sets up the OpenGL rendering context, defines display lists, etc), paintGL (renders the OpenGL scene), and resizeGL (sets up the OpenGL viewport, projection). Main actions of GLWidget class are mousePressEvent (defines what to do when mouse is pressed), mouseMoveEvent (what happens when mouse is move across), wheelEvent (changes the zoom), and keyPressEvent (handle all key press events).

On the top of GLWidget class, there are other two main files: mainwindow.py and ui_mainwindow.py. The ui_mainwindow.py defines all the properties (position, size, color, default value, etc) of GUI objects which would appear for user to interact. The GUI of Tractome is showed in Figure 5.4.1. The other file, mainwindow.py, specifies all actions or events that user can interact with each object in GUI.

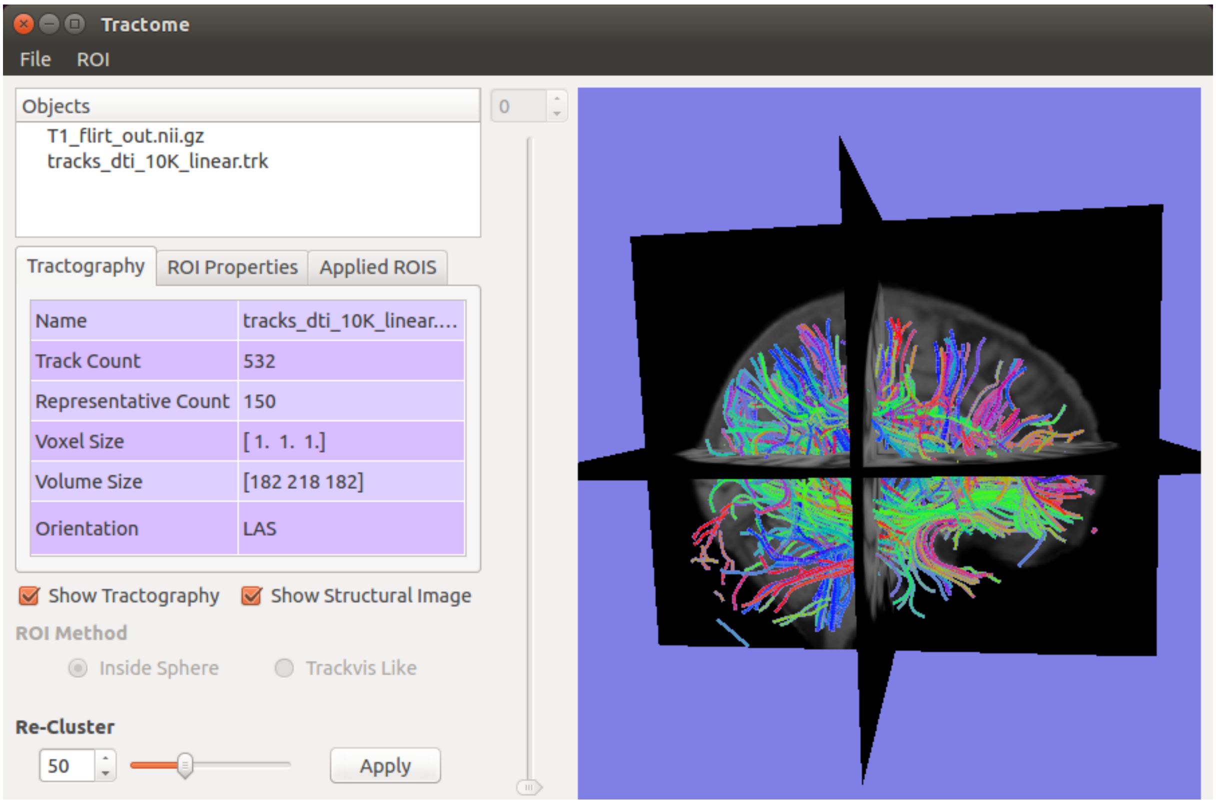
We used Eric4 [5], a full featured Python and Ruby editor and IDE, written in python, to design GUI; and then exported it to python language to generate these two mainwindown.py and ui_mainwindow.py files.

5.4.2 Application layer

Application layer is pulled from the presentation layer, and controls all application functionality by performing detailed processing. It is accessed on occasion by the user services layer. It receives the request from the presentation layer, divides the request to many sub-missions and sends these sub-missions to each components of the data layer.

[4] http://qt-project.org/doc/qt-4.8/qtopengl.html

[5] http://eric-ide.python-projects.org/



T1_flirt_out.nii.gz
tracks_dti_10K_linear.trk

Name	tracks_dti_10K_linear....
Track Count	532
Representative Count	150
Voxel Size	[1. 1. 1.]
Volume Size	[182 218 182]
Orientation	LAS

☑ Show Tractography ☑ Show Structural Image

ROI Method

● Inside Sphere ○ Trackvis Like

Re-Cluster

50

Figure 5.9: The graphic user interface of Tractome software. On right: the structural image and tractography; On left: the information of the current state of segmentation.

In our software, the application layer is coded in the source file of tractome.py. We create a class named tractome, and each function of tractome class corresponds to an event or action happening in the presentation layer. In general, it would be divided into three categories of functions: ones involving to streamlines, ones taking care of structural images, and ones working with the ROIs. Functions involving to streamlines are load tractography, save current working streamlines, save current log of segmentation, load a log of segmentation, remove streamlines, select streamlines, re-cluster the current working streamlines, etc. Structural image actions include hide or show a specific image (superior/inferior or coronal plane, anterior/posterior or horizontal plane, and left/right or sagittal plane), move a plane. ROIs functions can be listed as create a sphere, apply spheres on streamlines, update ROIs information, etc.

5.4.3 Data layer

The tractome data layer includes programs to manage actor for drawing streamlines, structural images in a very detail way. This layer is accessed through the business services layer. This layer keeps all the action of drawing and displaying actor independent from both application and presentation layer. The list of files in this layer is as following:

- *Manipulator.py* file defines the Manipulator class, which implements the logic of the operations for selecting, unselecting, expanding, hiding, showing etc. of streamline clusters. It basically provides set operations for streamline representatives.

- *Streamshow.py* file contains all the action that users can interact with streamlines. The main class in this file is StreamlineLabeler class which is inherited from the class Manipulator. Therefore, it includes all function from the mother class of Manipulator such as select/unselect the representative track, expand/collapse the selected streamlines, recluster the selected streamlines, invert selected streamlines to unselected, hide/show all representative streamlines, etc.

- *ROIs.py* file contains the class SphereTractome, which defines properties and functions for interacting between users and ROIs. Some basic properties of ROIs are coordinates (x,y,z) of the center, radius of sphere, color, and dimension. The main function of SphereTractome class is tractome_inside, which finds streamlines that are inside the sphere defined by the center and radius.

- *Guillotine.py* file describes the Guillotine class, inherited from the Slicer class of fos.actor [6]. It is a volume slicer actor to visualize a 3D volumetric image of the head as slices. It provides functions for showing/hiding and moving three planes (coronal plane, horizontal plane, and sagittal plane).

- *Dissimilarity.py* is a module that implements the computation of the dissimilarity representation of a set of objects from a set of prototypes given a distance function. In this module, various prototype selection policies are available such as FFT (furthest first traversal), SFF (subset furthest first).

5.5 Experiments

In this section we will describe in detail how the interactive exploration of tractography data is performed with the proposed tool. Moreover, we will show numerical experiments that provide evidence of the accuracy and efficiency of the automatic algorithms involved in the process.

For these experiments we used the dMRI recordings of 10 healthy subjects (100307, 124422, 161731, 245333, 528446, 556766, 201111, 199655, 239944 and 366446) from the Human Connectome Project [101], [97]. They were acquired on a Siemens Skyra 3T scanner (90 gradients; b-value= $1000 s/mm^2$; anatomical scan $(1.25 \times 1.25 \times 1.25 mm^3)$). From these data we reconstructed the streamlines using EuDX, a deterministic tracking algorithm [30] from the DiPy library [31]. The obtained tractography for each subject consists of approximately of 5×10^5 streamlines.

The proposed tool was implemented in Python code on top of the

[6] https://fos.readthedocs.org/en/latest/actors.html

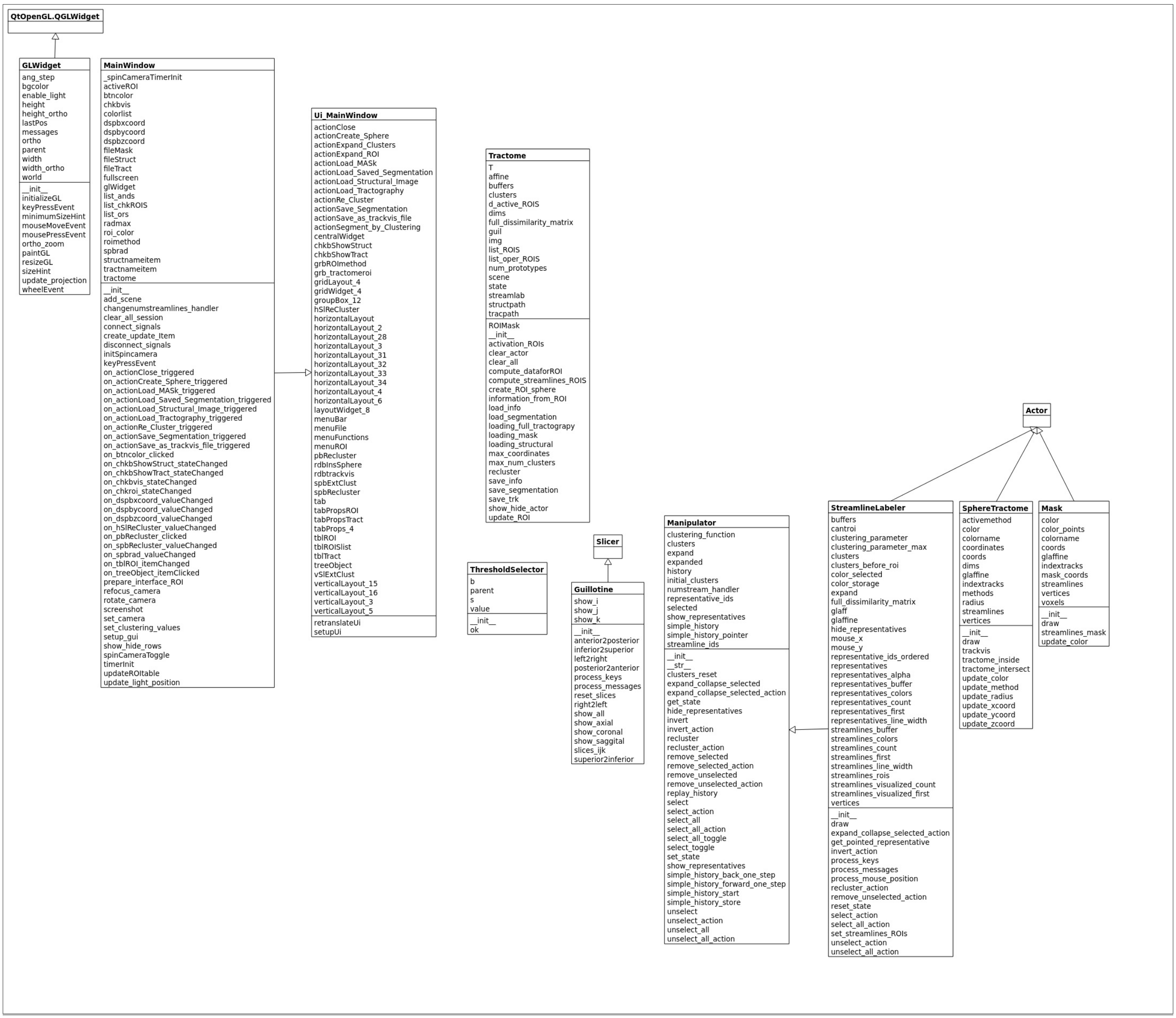


Diagram: /home/nusrat/tractome, Page 1

Figure 5.10: The list of functions of each class in Tractome software.

DiPy [7], Fos [8] and OpenGL [9]. The software project is distributed under a Free/Open Source license [10]. The code of the k-means and the MBKM are from scikit-learn [85] [11]. The code used to generate the tractographies

[7] http://nipy.org/dipy

[8] https://github.com/fos/fos

[9] http://opengl.org

[10] http://www.tractome.org

[11] http://scikit-learn.org

has also been published in the softwares' web page.

5.5.1 Dissimilarity Representation and Prototype Selection

As it was mentioned in Section 5.3, the first step of the workflow (see Subsection 5.3.1) is to project the data into the dissimilarity space $\mathbb{R}^p$. With this purpose, a set of prototypes has to be selected in advance. Next, we present a study of the degree of approximation of the dissimilarity representation across different prototype selection policies i.e. Random, SFF and S+SFF, and different numbers of prototypes. The aim is to investigate the trade-off between accuracy and computational cost. For SFF we chose $c = 3$ in order to have high probability (> 0.95) of accurately representing T through the subset. For S+SFF, only streamlines with size $z > 10$ will be used. By using this value, the set of streamlines to be analysed is drastically reduced in the order of 10^5 times.

In order to evaluate how accurate our dissimilarity representation is, we investigate the relationship between the distribution of distance among objects in $\mathcal{X}$ through d and the corresponding distances in the dissimilarity space through Δ^d_Π, where $\Delta^d_\Pi : \phi(s) \times \phi(s') \to \mathbb{R}^+$ and $\Delta^d_\Pi = ||\phi(s) - \phi(s')||_2$. It was claimed in [76], that a good dissimilarity representation must be able to accurately preserve the partial order of the distances, i.e. if $d(s, s') \leq d(s, s'')$ then $\Delta^d_\Pi(s, s') \leq \Delta^d_\Pi(s, s'')$ for each s, s' , $s'' \in \mathcal{X}$ almost always. As a measure of the degree of approximation of the dissimilarity representation we define the Pearson correlation coefficient ρ between the two distances over all possible pairs of objects in $\mathcal{X}$:

$$\boldsymbol{\rho} = \frac{\mathrm{Cov}(d(s, s'), \Delta^d_\Pi(s, s'))}{\sigma_{d(s,s')}\sigma_{\Delta^d_\Pi(s,s')}} \tag{5.3}$$

where given P_s a probability distribution over $\mathcal{X}$, $s, s' \sim P_s$. In practical cases P_s is unknown and only a finite sample F is available. We can approximate $\boldsymbol{\rho}$ as the *sample* correlation $\boldsymbol{r}$ where $s, s' \in F$. An accurate approximation of the relative distances between objects in $\mathcal{X}$ results in values of $\boldsymbol{\rho}$ far from zero and close to 1[12].

The correlation and standard deviation for each prototype selection strategy are shown in Figure 5.11 [13]. The correlation ρ between distances in the original space and the corresponding distances in the projected space was estimated by computing 50 repetitions for each subject.

We can observe that, as an overall behaviour for all subjects, SFF significantly outperformed the random policy, in agreement to what was reported in [76]. Moreover, it can be seen that S+SFF, outperformed the other two policies. In particular, we conducted a one-tailed t-test comparing the correlation values of the 50 repetitions of S+SFF against those of SFF for each number of prototypes p (11 values) and for each of the 10 subjects. Of these $11 \times 10 = 110$ tests, when $p \geq 5$, the obtained p-values were always lower than 5.2×10^{-13}. Even considering an overly conservative Bonferroni correction, the p-values were sufficiently low to reject the null hypothesis of S+SFF equal to SFF for $p \geq 5$.

We observe that S+SFF reached the highest correlation of 0.95 on average (50 repetitions) with respect to the distances in the original space, using only $15 - 25$ prototypes, which is half of the amount used in the previous paper with the SFF policy [76].

It is also worth to notice that the standard deviation with this last approach, $\hat{\sigma}_{\text{S+SFF}}$, is smaller than that of SFF, $\hat{\sigma}_{\text{SFF}}$, thus implying that the proposed heuristic is more stable. To support this point, we conducted a one-tailed t-test of the standard deviations of correlation for the 10 subjects, comparing S+SFF vs. SFF for each number of prototypes p (11

[12] Note that negative correlation is not considered as accurate approximation. Moreover it never occurred during experiments.

[13] The figure is restricted to 6 of the 10 subjects for lack of space. The graphs of all subjects showed an equivalent behaviour.

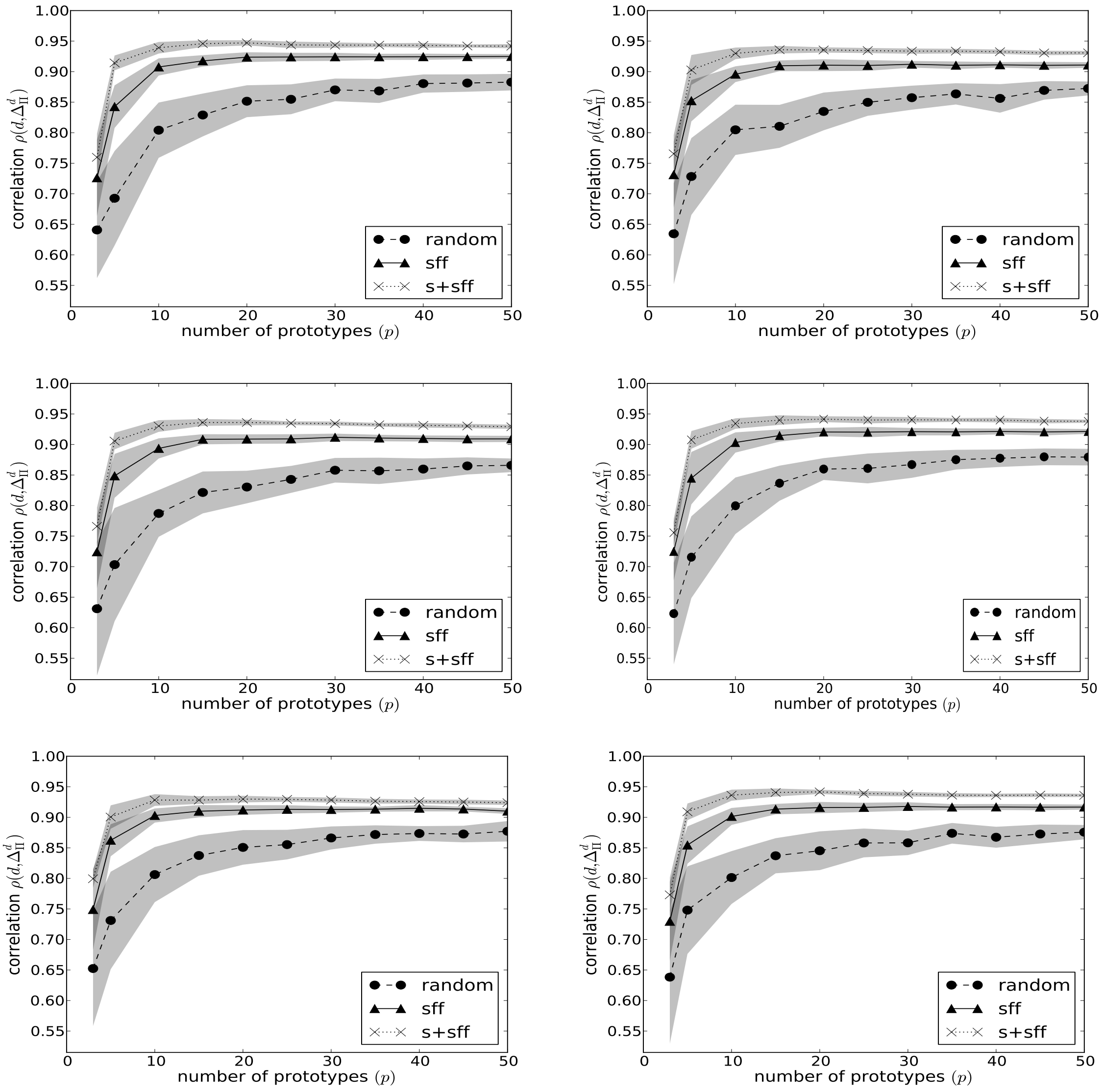


Figure 5.11: Average correlation between d and Δ^d_Π across the different prototype selection policies and different numbers of prototypes. Each figure corresponds to a different subject.

values). In all these 11 tests, the p-values were always lower than 1.2×10^{-5}. Even considering a very restrictive Bonferroni correction, the p-values were sufficiently low to reject the null hypothesis of $\hat{\sigma}_{\text{S+SFF}} = \hat{\sigma}_{\text{SFF}}$.

From these results we can conclude that the dissimilarity representation works well for preserving the relative distances. Given the number of prototypes with which the maximum correlation was reached, it is proven that this approach can produce compact feature spaces for this kind of data. Moreover, we observed that the S+SFF policy can be easily computed on a standard computer even in the case of a large tractographies. Therefore, we use the S+SFF to obtain an efficient and effective selection of the prototypes in our tool.

5.5.2 Tractography exploration example

In the following, we describe the steps for the tractography exploration process with our tool, based on the workflow presented in Section 5.3. For the example, we use subject 100307 from the HCP data set, and the goal is to explore the Corticospinal Tract (CST). For the embedding of the tractography in the dissimilarity space, and based on the study from the previous subsection, we use $p = 20$ prototypes and the new S+SFF prototype selection policy, according to the results in Section 5.5.1. The prototype selection and the actual embedding of $\approx 5 \times 10^5$ streamlines required ≈ 630s. The resulting matrix $\phi(T) \in \mathbb{R}^{500K \times 20}$ was computed once and stored, so that the time to compute the projection did not affect the interactive segmentation.

The full brain tractography (Figure 5.12.A) of ≈ 470000 streamlines was initially clustered in $k = 150$ clusters and the medoids were presented to the user (5.12.B). We have observed from previous trials with experts, that $k = 150$ is approximately the highest number of medoids the users could comfortably interact with in the 3D scene when the whole tractography is presented. Afterwards, the expert selected 11 clusters by clicking on the corresponding medoids (5.12.C, in white). These medoid streamlines are those approximating the structure, or positioned in the area (provided by the structural image) that correspond to the CST. In order to know if she is going in the proper direction and the right medoids were selected, the expert usually explores the composition of the clusters. With this purpose, the expert expands on the selected clusters (Zooming), which in this case represent a set of ≈ 35000 streamlines (5.12.D), that are shown in the same screen with a different line thickness. In order to be able to continue and do a further analysis, the expert re-clusters the streamlines of the initially selected clusters, into new $k = 50$ clusters (5.12.E) and selects 15 of them (5.12.F, in white). In this case, and from now on, the number of clusters is interactively specified by the expert. We have also observed that 50 medoids are approximately the highest number she can comfortably interact with after the initial selection from the full tractography. The 15 selected clusters corresponds to ≈ 9600 streamlines (5.12.G) that are then re-clustered into $k = 50$ clusters (5.12.H). For the selection and exploration of the clusters, the user may need to interact with the 3D slicer by rotating it, such that she can explore the different areas that streamlines are passing through. Moreover, the expert may need to do panning, zoom-in or zoom-out on a set of streamlines, such that she can explore and study the conditions of the structures of the streamlines. In two further steps the user reduced the selected streamlines to ≈ 2000 (5.12.J) and then to ≈ 650 (5.12.M), until he finally reached the desired anatomical structure, i.e. CST.

5.5.3 Clustering analysis

The average timings of the MBKM algorithm in each step of the exploration process are reported in Table 5.1. These are also compared to those of the k-means algorithm, in order to prove its scalability with respect to the former algorithm. In the first column (size) are reported the size of the subset of streamlines that were clustered. The second column (k) reports the number of clusters, according to the notes expressed above. The third (k-means) and the fourth (MBKM) report the time for clus-

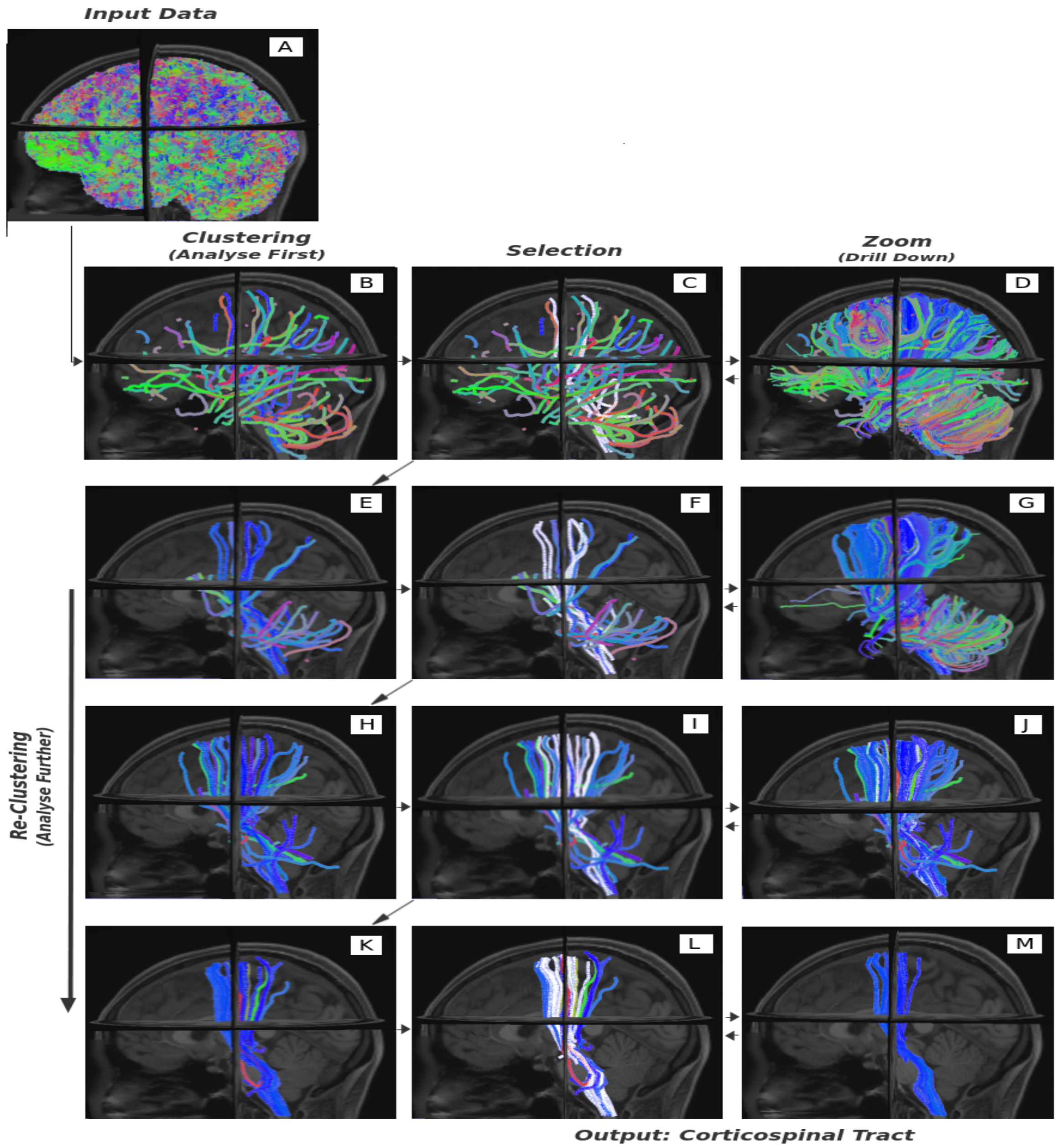


Figure 5.12: The segmentation process. (A) Full tractography $\approx 5 \times 10^5$ streamlines; (B) Computation of 150 clusters (C) Selection of 11 clusters (in white); (D) ≈ 35000 streamlines corresponding to previous selection; (E) Computation of 50 clusters (F) Selection of 15 clusters; (G) ≈ 9600 streamlines corresponding to the previous selection; (H) Computation of 50 clusters (I) Selection of 15 clusters; (J) ≈ 2000 streamlines corresponding to the previous selection; (K) Computation of 50 clusters (L) Selection of 25 clusters; (M) ≈ 650 streamlines corresponding to previous selection and representing the segmented CST.

size	k	k-means	MBKM	b	medoids
500	50	$0.3s$	$\mathbf{0.2}s$	100	$0.006s$
1000	50	$0.6s$	$\mathbf{0.2}s$	100	$0.007s$
5000	50	$7.2s$	$\mathbf{0.4}s$	100	$0.011s$
10000	50	$19.8s$	$\mathbf{0.6}s$	100	$0.015s$
15000	50	$32.4s$	$\mathbf{1}s$	100	$0.021s$
470000	150	$> 2000s$	$\mathbf{42}s$	1000	$0.28s$

Table 5.1: For a given number of streamlines (1st column, size) and a given number of clusters (2nd column, k) the time to compute the clustering with k-means and MBKM is reported in the 3rd and 4th columns, respectively. The size (b) of the mini-batches for MBKM is in the 5th column. The time to compute the medoids from the centroids is in the 6th column.

tering[14]. The Fifth column reports the size (b) of the mini-batches for the MBKM, which was always 100 except for the full tractography for which we observed a significant gain in time when increasing it to 1000. The sixth column reports the time to compute the medoids from the centroids provided by k-means and MBKM. Each medoid was computed with simple exhaustive search within each cluster. The time to compute all medoids was always negligible with respect to the clustering time. All computations were performed on a standard desktop computer.

As it is shown in the Table, the time required to cluster the streamlines with the MBKM was always the lowest one, i.e $<$ 1s during interactive use, thus meeting the requirements for a comfortable user experience.

5.6 Case Study

In this part, we demonstrate the usefulness of our proposed interactive visualization tractography segmentation software tool in the neuroscientific analyses activities. The first one is to study the characterisation of the amiotrophic lateral sclerosis (ALS) disease through the corticospinal

[14] The clustering of the whole tractography can be computed once and stored, so its time does not affect the interactive use.

tract. The second one uses the result of tract segmentation for validation the tractography registration method.

5.6.1 Corticospincal Tract segmentation for ALS disease analysis

In this part, we demonstrate the usefulness of tract segmentation using our Tractome software, in clinical study. Our work is motivated by a clinical research hypothesis about the characterisation of the amiotrophic lateral sclerosis (ALS) disease. Amyotrophic Lateral Sclerosi (***ALS***), also known as motor neurone disease or Lou Gehrigs disease, is a progressive neurodegenerative disease that affects nerve cells in the brain and in the spinal cord controlling voluntary movement. Motor neurons reach from the brain to the spinal cord and from the spinal cord to the muscles throughout the body. As motor neurons degenerate, they can no longer send impulses to the muscle fibers that normally result in muscle movement. The result is wasting and atrophy of muscles, leading to difficulties in speaking, swallowing, stumbling, permanent fatigue and cramping, amongst other symptoms. The ALS is known to be affected by CST (Corticol Spinal Tract) [21] and for this reason, the long term goal is to characterise these effects through tractography data.

Usually, CST starts from the cerebral cortex, and terminates in the spinal cord. Note that fibers after crossing over from one side to the other in the medulla, continue downward in the lateral corticospinal tract on the opposite side and go to muscles (see Figure 5.13-left). Each crossed corticospinal tract, therefore, conducts motor impulses from one side of the brain on interneurons or anterior horn motoneurons on the opposite site of the cord. That is the reason why impulses from one side of the cerebrum cause movements of the opposite side of the body. An example of CST can be found in the Figure 5.13.

From the prior knowledge of neuroscientists and doctors, there is

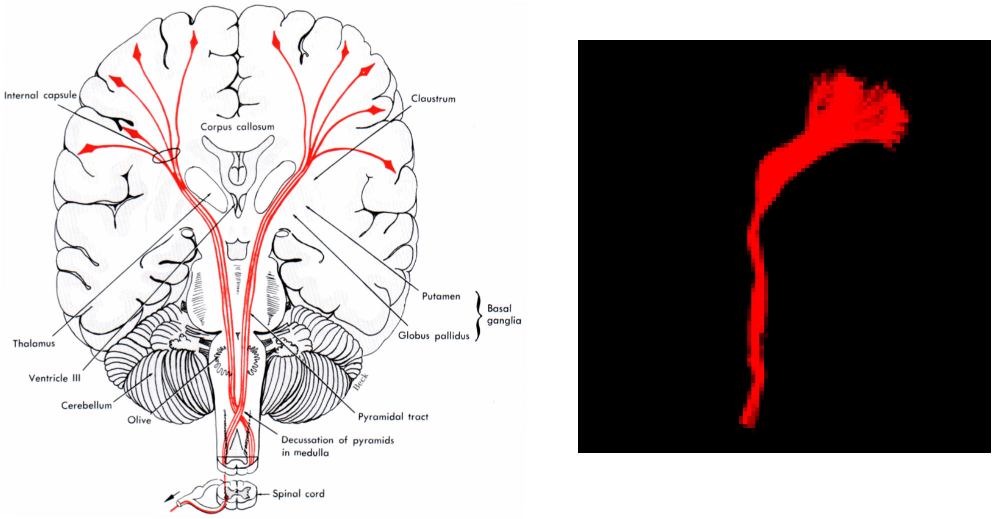


Figure 5.13: Left - the Cortico Spinal Tracts in general. Right - the left CST segmentation of the control ID 201 in the dataset ALS with 487 streamlines

an evidence about the reducing of the number of fibers in CST of ALS patients compared with control people [21]. It is also the same situation with the volume of CST. Beside the number of fibers and the volume, fractional anisotropy (FA) and mean diffusion (MD) also play an important role for recognizing the ALS disease. Following are some quantitative features which may effectively affect on ALS patients: *fiber count* - the number of streamlines belonging to CST; *fiber length* in mm (min, max, mean length); *fiber volume* - number of voxels occupied by all streamlines or the bounding geometry cylinder of CST; *fiber density* - ratio between fiber count and voxel number; *fragmentation* - be quantified by the ratio between fiber count and the volume; *fractional anisotropy (FA)* - defined as mean value of the standard deviation in the three eigenvalues and in the range 0 to 1

$$FA = \frac{1}{\sqrt{2}} \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{5.4}$$

and *mean diffusion (MD)* - the average diffusion rate in all directions

$$MD = \frac{trace(DT)}{3} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \tag{5.5}$$

where $(\lambda_1, \lambda_2, \lambda_3)$ is eigenvalues of diffusion at a given voxel.

The dMRI data from ALS patients and healthy controls collected were collected with the aim of studying the effects of the ALS disease on the corticospinal tract (CST) (see Figure 5.13), an anatomical structure that connects cortical motor areas to the spine and the body. Traditionally, diagnostic decision making has involved using evidence provided by patient data coupled with priori experience of physician. Up to now, it is still very much an art for many physicians due to a lack of quantitative tools and measurements. In this work, we overcome this drawback by exloring our proposed software tool for segmenting CST from the full brain tractography of each subject.

The dataset in our experiments was recorded with a $3T$ scanner at Utah Brain Institute[15]. This dataset consisted of 12 healthy controls and 12 subjects; 64 $(+1$, i.e. $b = 0)$ gradients; b-values 1000; anatomical scan $(2 \times 2 \times 2mm^3)$. We reconstructed the streamlines using EuDX tracking algorithm [30] from dipy[16] with 3×10^5 random seeds. CST segmentation was done by doctors[17] using our interactive visualization tool. An example of CST segmentation from ALS dataset is in figure 5.13 (left). As the result, we had 48 segmentations (24 of patients including 12 left CST and 12 right CST; and similar for controls).

After segmentation, we computed the value of quantitative features:

[15] The authors are grateful to Prof. Mark B.Bromberg, Prof. Lubdha Shah and Prof. Perry Renshaw of the Department of Neurology and the Department of Radiology, University of Utah (US), for their assistance in acquiring MR data

[16] `http://www.dipy.org`

[17] We thank Nivedita Agarwal, Department of Neuroradiology, S.Maria del Carmine Hospital, Rovereto, Italy and Francesca Maule, University of Trento, for the segmentation of the corticospinal tracts.

fiber count, fiber length (min, max, mean), fiber volume, fiber density, fragmentation, FA and MD. We, then, did a $t - test$ on each set of left and right. It showed that, in the right CST, fiber number significantly decreased ($p = 0.00042$) in ALS patients ($mean_{fiber-number} = 294$) compared with controls ($mean_{fiber-number} = 640$). In contrast, patients had the fiber min length slightly higher then controls ($p = 0.07$, patients: $mean_{min-length} = 74.25$ and controls: $mean_{min-length} = 53.9$). Moreover, the volumn of the left CST dramatically diminished ($p = 0.0034$) between patients ($mean_{volumn} = 6038$) and healthy peoples ($mean_{volumn} = 4230$). These are just some preliminary results and it needs more investigation to confirm the difference between healthy and ALS-diseased brain. But it also shows an strong evidence that the tract segmenation has a bright capability for applying in clinical diagnosing application.

5.6.2 Comparison between voxel-based and tract-based registration

Traditionally the segmentation task is done by neuroanatomists, and it consumes a lot of time and effort due to the large number of streamlines (about 3×10^5 in a normal brain). Moreover, the variability of the brain anatomy among different subjects makes the segmentation become a difficult task [12]. The first task in this endeavour is to align or register tractographies from different subject together. Registration is the problem of identifying the process of geometric transforming the coordinate system of an image to be as spatially aligned to a reference image, more generally establishing a homology among the input images [43]. In this scenario, a group of transformations needs to be established to put all the inputs into correspondence [114]. Specific to the tractography registration, it is most often performed by applying the transformation resulting from an image-based fractional anisotropy (FA) or diffusion tensor imaging (DTI) [38, 106, 37, 113]. Recently, O'Donnell et. al. proposed the unbiased multiple subject registration using the trajectory data produced by streamline tractography [72]. The idea to work on deterministic tractography rather than dMRI images or FA images is quite innovative, and it may be advantageous to register the tracts themselves as the quantity being optimized would be closely related to the final goal. An open point of [72] is how to collect evidence that the proposed approach is effective in practice.

In this work, we attempt to evaluate the result of [72] in the context of tractography segmentation. The transformed tractography after registration will be fed into a classifier to do the segmentation of tracts. We conceive an experiment of applying the whole process on the clinical case study of dMRI dataset. The results are used to empirically analyze the usefulness of [72] in the context of supervised segmentation of tracts.

Data acquisition and tractography

The dataset and tractography in this experiment is the same as in the previous one (see more detail in Section 5.6.1).The dataset consisted of the brain tractography from 12 ALS patients and 12 healthy controls.

Tractography registration

Method 1*: affine FA to MNI Atlas.*

FLIRT [48] is a linear (affine) image registration algorithm in FSL [18], and is specifically developed for brain imaging. It is a robust and accurate automated affine registration tool based around a multi-start, multi-resolution global optimization method [47]. It is available as part of the FSL software package. It provides different cost functions, including the within-modality functions Least Squares and Normalised Correlation,

[18] `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`

as well as the between-modality functions Correlation Ratio, Mutual Information and Normalised Mutual Information. Additionally, it can be run with a number of different transformation model (degrees of freedom - DOF) and it generates a global cost function as weighting of all cost functions. In our experiment, we ran this algorithm with the default parameter as recommendation from FSL (6 ODF and correlation ratio as cost function). All registrations were performed on the FA volumes, which are computed from the corresponding diffusion weighted images, with the FMRIB-58 FA template [19] in MNI atlas space as the reference image. Bellow, we refer to this registration as MNI-registration method.

Method 2: unbiased groupwise tractography registration, tract-based affine.

Instead of doing registration by applying the transformation resulting from an image-based fractional anisotropy (FA) as in method 1, recently O'Donnell et. al. proposed the unbiased multiple subject registration method using the trajectory data produced by streamline tractography [72]. It presents a brain tractographyd as a probability distribution on trajectories. The brain distribution is constructed as a kernel density estimate from the tractography, and an atlas distribution is constructed as a mixture of the constituent brain distributions. From this atlas, the entropy of all fibers is calculated. By minimizing this entropy the registration can successfully align the tractography in multiple subjects. The most advantages of this method is that it does not require any other prior information from outside of the set of fibers needed to be registered. Moreover, because it is based on only tractography data, the aligning the tracts themselves as the quantity being optimized would be closely related to the final goal. In this project, we used the current

[19] the average of 58 well-aligned FA images from healthy male and female subjects aged between $20-50$

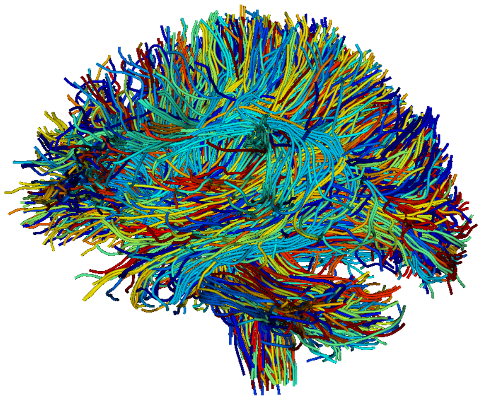


Figure 5.14: An example of the atlas after registration using tract-based method. Note that each color is a different subject

library available at whitematteranalysis [20]. Each tractography was subsampled with 750 fibers for creating the atlas, and each fiber was down sampled to 5 points.

Expert selection of CST

The groundtruth data was manually segmented by experts from both CiMec center [23] and Azienda Provinciale Sanitari, Trento, Italy [24] using our computer aided tractography segmentation software tool, Spaghetti [25]. First, the full tractography is initially clustered in k clusters/medoids (k is around 150 to 200). The expert then manually selects the medoids/-

[20] `https://github.com/ljod/whitematteranalysis`

[23] `http://www.unitn.it/en/cimec`

[24] `http://www.apss.tn.it/`

[25] `https://github.com/dporro/spaghetti`

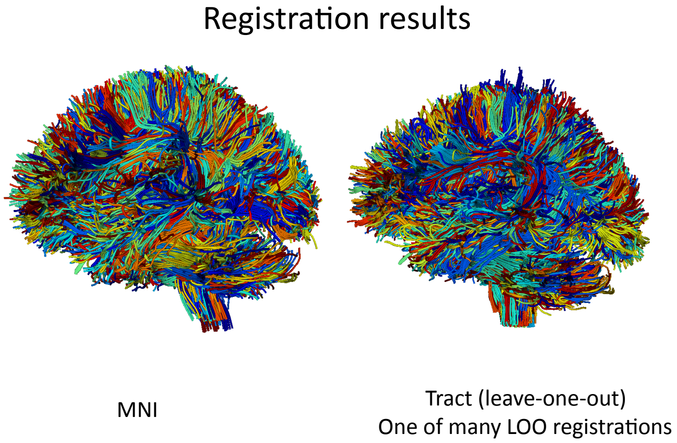


Figure 5.15: At left, the visualization of tracts from alll subjects to MNI space, performed by affine FA image registration to FA−FMRI58 atlas, then applying the transform to the tractography. Each subject is presented by a unique color. At right, the visualization of results of unbiased tract-based registration performed with the software package whitematteranalysis [22].

clusters of interest in order to remove most of the streamlines not related to the anatomical structure of interest. The process of reclustering the selected streamlines and of manual selection by the expert is iterated until the expert is confident of having segmented the structure of interest (in our case it is corticospinal tract). More detail about Spaghetti can be found in [74]. As the result, each subject/control has two CSTs, one in the left brain and another in the right one. In total, we have 12 left CSTs of patients, 12 right CSTs of patients, and the same quantity for healthy controls.

Note that, the tractography used in this step is only the MNI-registration result, because the current version of Spaghetti does only support in the MNI space not in native space. Moreover, instead of saving the CSTs as the real streamlines belonging to CSTs, we only saved CST as indices of the streamlines, which is the ordered number of each streamline in the whole tractography ranging from 0 to $N-1$, where N is the size of tractography in the number of streamlines. With these indices, it is easy to extract the real CSTs both in MNI-registered tractography or in tract-based registered tractography.

Training of SVM classifier for CST segmentation

Dissimilarity representation: In this experiment, we used SVM (support vector machine) classifier[20] to automatically segment the CST. Due to the fact that each streamline has different length and different number of points, while SVM requires the data to lie in a vectorial space, it is necessary to find a representation ϕ of streamline in a vectorial space, by mapping a streamline s from its original space $\mathbb{T}$ to a vector of $\mathbb{R}^p$ - $\phi : \mathbb{T} \mapsto \mathbb{R}^p$, where p is the dimension of the new space. In [72], authors downsample each streamline into 5-point length to calculate the distance between two streamlines. Downsampling is somehow simple and easy to calculate, but it also has some limitations(see [76] for more detail). Here, we proposed to use the *dissimilarity approximation* [88] instead of downsampling. This replacement promises a better result of registration comparing with the downsampling method.

The *dissimilarity representation* [88] is defined as $\phi_{\Pi}^{d}(X) : \mathcal{X} \mapsto \mathbb{R}^p$ s.t. $\phi_{\Pi}^{d}(X) = [d(X,\tilde{X}_1),\ldots,d(X,\tilde{X}_p)]$, where $d(\cdot,\cdot)$ is a distance function between two streamlines, and $\Pi = \{\tilde{X}_1,\ldots,\tilde{X}_p\} \subseteq \mathcal{X}$ is a set of p streamlines called *prototypes*. The quality of the Euclidean embedding is strongly dependent on the selection of the prototypes (see [86, 76]). The dissimilarity representation for streamlines was previously proposed in [76].

One-class-SVM classifier: Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. In this experiment, we used the current version of SVM classifier available at scikit-learn[84]. For the purpose of evaluation the correctness of the classifier, we used leave-one-out fashion to separate the initial tractography dataset into two sub-dataset, one for training and the other for testing.

At first, for the training step, all $n-1$ subjects $S = \{s_1, s_2, ..., s_{n-1}\}$ in the training dataset would be fed in the groupwise registration process, and it resulted in $n-1$-registered subjects $S^{'} = \{s^{'}_1, s^{'}_2, ..., s^{'}_{n-1}\}$. Beside the set of $n-1$ registered subjects, we also created the *atlas* A of n-1 training subjects. An example of *atlas* can be found in the Figure 5.14.

Let $T = \{T_1, T_2, ..., T_{n-1}\}$ be $n-1$ corticospinal tracts of $S^{'}$. The dissimilarity approximation was then used to represent each streamline $t \in T_i, \forall i \in [1, n-1]$ as a p-dimension vector: $t^{'} = (d(t, x_1), d(t, x_2), ..., d(t, x_p))$, where $d(a, b)$ was the distance between two streamlines a and b; and $\pi = (x_1, x_2, ..., x_p)$ was a set of p prototypes chosen from T. From the representation, the Euclidean distance was calculated between each pair of streamline representation to create the dissimilarity matrix. After that, the dissimilarity matrix was fed into the SVM classifier for training. Note that, in this experiment, the purpose was to know which streamline would belong to the CST, so instead of using the multiple class SVM, the one-class-SVM classifier was then used.

Testing of SVM classifier for CST segmentation

The testing subject was first registered to the atlas A (created during the tractography registration step when prepairing for training SVM) separately with the registration process of training set. The aim was to make sure that the testing subject was completely new with the training set, and thus it did not provide any information which could effect on the final classifier. After registered, the testing subject was also re-represented using the same prototype set p, and this representation was put into one-

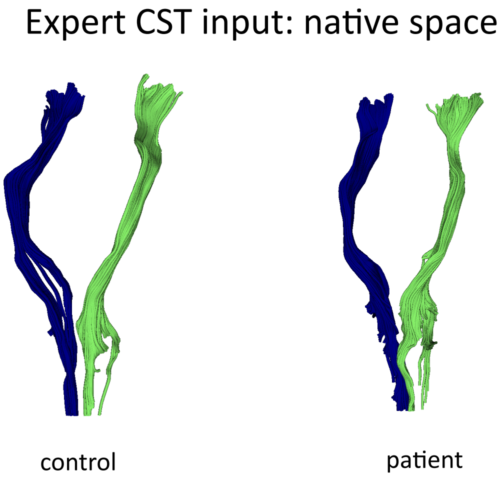


Figure 5.16: An example of the input CST from controls and patients. To create this test dataset, 46 expert segmentations were produced using our interactive segmentation software.

class-SVM classifier for automatically segmenting the CST.

Experiments

The entire pipeline consists of the following steps: 1. tractography registration to a common coordinate system; 2. extraction of CST according to expertly selected tracts (represented as indices into the tract file); 3. creation of a CST segmenter by training an SVM classifier using expert labeled training data; and 4. testing of the classifier. We performed this pipeline experiment in a leave-one-out fashion, and for each type of registration. Thus, the whole pipe line was run 24×2 times. Tract-based registration used 750 fibers of length $> 120mm$ per subject to form the atlas, and 2000 for registration to the atlas.

Cross-validation: from the N-objects dataset, we trained the one-class

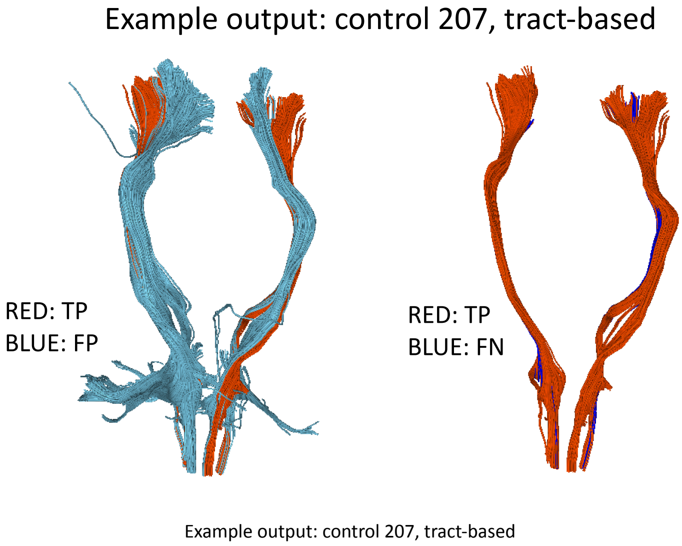

Example output: control 207, tract-based
RED: TP
BLUE: FP
RED: TP
BLUE: FN
Example output: control 207, tract-based

Figure 5.17: An example of the result of CST segmentation in one subject. At left, result calculated in tract-atlas space, in the middle, result calculated in MNI space. For comparison, at right the input expert segmentation.

SVM classifier on $N-1$ objects, and d the test on the remain object. We measured true positive (TP), false negative (FN), false positive (FP), true negative (TN) in units of numbers of tractography trajectories, by comparing the output of the SVM classifier to the known ground truth labels. The precision and recall were also calculated. We evaluated the effect of the registration methods using a paired t-test on the leave-one-out results. Example results of both styles of registration, applied to the tractography from multiple subjects are in Figure5.15. Examples of the input and output of SVM CST classifier are in Figures 5.16 and 5.17. The average and standard deviation of the performance of the SVM classifier over all subjects, are presented in Table I. It is possible to see that there is no significant difference between two registration methods.

This work represents an evaluation of a novel method for white matter registration. It demonstrates the potential of tractography registra-

Table 5.2: Performance of CST SVM classifier over 20 leave-one-out trials, in two atlas coordinate systems. On the left, unbiased tract-based registration atlas space vs on the right MNI FMRIB58−FA atlas space. The t-test does not detect a different in these methods.

	tract-based		MNI		
	mean	std	mean	std	t-test
training-error	0.10	0.01	0.10	0.01	0.18
training-correct	0.90	0.01	0.90	0.01	0.18
precision	0.45	0.15	0.46	0.17	0.23
recall	0.93	0.06	0.92	0.08	0.48
TP	566.71	252.02	563.43	257.82	0.66
FN	44.95	44.47	48.24	48.10	0.66
FP	733.33	372.44	703.29	400.39	0.39
TN	17856.24	3620.64	17886.29	3569.35	0.39

tion, which is sufficient to align the group data comparably to a popular affine registration method with respect to automatic tract segmentation. However, it is also clear that future work is needed, for example to increase robustness to tractography variability across subjects; to optimize parameter settings, to increase the appeal of such a tractography registration method; and to tune the parameter of classifier to increase specificity of the SVM itself with reducing the false positive.

5.7 Conclusions

We created a visual data mining tool for the exploration and analysis of tractography data sets. It provides the expert with a meaningful summary of the data and allows an iterative visual exploration of these large data sets, by integrating it with automatic clustering. In order to provide the user with a comfortable interaction with the tool, we used a scalable automatic algorithm. Moreover, a more effective representation for brain tractography data is used, such that the use of efficient implementations of the clustering algorithms are possible. These also allow for the use of advanced visualization and interaction techniques, such that the visual scalability requirement is accomplished. Considering the observed timings and the scalability of the interactive visualization, we observed that a trained neuroanatomist could find the studied tract in approximately 5 iterations. Thus, based on the performed studies, we believe that this software is compliant with the requisites of a scalable visual data mining tool. Consequently, it should provide neuroanatomists with a more useful system for the exploration of tractography data sets.

Chapter 6

Conclusion

The aims of this PhD thesis were to develop methods using machine learning techniques to enhance the robust segmentation of a specific tract from a whole brain white matter tractography. In the final chapter, we present the summary of our main original contributions, and discuss the extent to the works already described in previous chapters, that needs to be done.

6.1 Summary

In this document we investigated the using of machine learning techniques in neuroimaging for tractogarphy segmentation task. First, we propose a design of interactive segmentation process based on BOI (bundle of interest) approach instead of traditional ROI (region of interest) one. While ROI concerns about which streamlines go through some interesting regions, BOI focuses only on streamlines inside some specific bundles. Using BOI would make medical practitioners concentrate on which tracts they are working on, and thus, the final target tract would be easy to get.

Second, we suggested to use dissimilarity representation as a novel way to present streamlines in a vectorial space, which is required for most of the current machine learning algorithms. We investigated the degree of approximation of the dissimilarity representation for the goal of preserving the relative distances between streamlines within tractographies. Empirical assessment has been conducted on both simulated and real dMRI datasets, and through various prototype selection methods. The results from real tractography data reached correlation ≥ 0.95 with respect to the distances in the original space. This fact proved that, the dissimilarity representation works well for preserving the relative distances. Moreover on tractography data the maximum correlation was reached with just approximately $20 - 25$ prototypes. Thus, it claimed that the dissimilarity representation can produce compact feature spaces for this kind of data.

In order to handle the computational burden of clustering a large number of streamline under strong time constraints for real-time interaction when doing segmentation, we proposed a solution based on the dissimilarity representation and the MBKM (mini-batch k-mean) algorithm. Experiments on real dMRI data showed that the time required to cluster the streamlines with the proposed solution was always the lowest and always < 1s during interactive use, thus it met the requirements for a comfortable user experience. Conversely, the time required by the standard k-means algorithm was inadequate. At the first step of the segmentation session the clustering of the whole tractography requires ≈ 20s with the proposed method. This may be an issue with interactive use, but can be solved by pre-computing this clustering once and then by storing the result together with the actual dataset for future use.

When studying tractography data across subjects, it is usually necessary to align, i.e. to register, tractographies together. This registration step is most often performed by applying the transformation resulting from the registration of other volumetric images (T1, FA). In contrast with registration methods that *transform* tractographies, in this work, we tried to find which streamline in one tractography corresponds to which streamline in the other tractography, without any transformation. In other words, we tried to find a *mapping* between the tractographies. We proposed a graph-based solution for the tractography mapping problem and we explained similarities and differences with the related well-known graph matching problem. Specifically, we defined a loss function based on the pairwise streamline distance and reformulate the mapping problem as combinatorial optimization of that loss function. We showed preliminary promising results where we compared the proposed method, implemented with simulated annealing, against other standard registration techniques in the task of segmentation of the corticospinal tract.

Although dissimilarity representation is able to build a fast and accurate vectorial representation for streamline, it limits to only intra-subject while most of neuroscientific analyses of tractography require inter-subject comparisons. In this work, we proposed the algorithmic solution to build the common vectorial representation of streamlines across subject. The core of the proposed solution was to combine two state-of-the-art elements: first using the recently proposed tractography mapping approach to align the prototypes across subjects; then applying the dissimilarity representation based on the aligned prototypes to build the common vectorial representation for streamline. We evaluated our proposed solution in the context of tractography segmentation. Results from CST (Cortical Spinal Tract) segmentation showed that our method can produce a good vectorial representation for streamlines across subjects comparing to the original tractography registration method.

We also provided an implementation of our framework for tract segmentation, call TRACTOME. In this scientific interaction tool we presented a novel, simple way to interactively visualize and segment streamlines from large tractogaphy in 3D space. We solved the problem of interacting with tractographies by creating real-time simplifications in terms of the underlying bundle structures. The process that we proposed works recursively: starting from a small number of clusters of streamlines the user decides which clusters to explore. Exploring a cluster means that the application re-clusters its content at a finer grained level. This goes, as far as we know, beyond any other available medical imaging software. Moreover, this demonstration also integrated many utility functions, such as undo, log, zoom, save the works, load the result, etc. This enables medical practitioners and researchers to meaningfully navigate the entire space of the tractography and perform the segmentation task more easily and accuracy.

6.2 Future works

Here, we will describe our future plans and what the research extension that we want to take after finishing this thesis.

Tractography mapping: currently, as a preliminary strategy to approximate the minimal loss function to get the optimal mapping, we implemented the simulated annealing (SA) [54] meta-heuristic, a reference method for combinatorial optimization. We are aware that this method of combinatorial optimization can be significantly improved by using different optimization method, or even we should propose a specific optimizer for our problem.

Common representation: in this work, we proposed a new method for building a common vectorial representation for streamlines across subjects. However, the evaluation process was only based on the segmentation task. We need to investigate a general method for evaluating the common vectorial representation for streamlines.

Clinical application: up to present, the task of finding the difference between healthy and ALS diseased brains has not completed yet. We believe that the satisfactory result will be published in near future. After that, the general framework for clinical diagnosing based on the differences between two folders of interesting tracks must be presented and applied for other brain diseases.

Software improvement: we plan to investigate further pattern recognition algorithms to better support the expert during tractography segmentation. One example is to use supervised machine learning to automatic identify the candidate of a specific tract from a given a set of labelled-streamlines, instead of working on the whole tractography at the beginning. Working on this direction is able to help the segmentation task to get a high accuracy with less computational cost. Moreover, integrating the libraries used in our software in order to run it in different platform is also another work which has to be done in near future.

Bibliography

[1] *Mapping Tractography Across Subjects*, 2014.

[2] A. L. Alexander, K. M. Hasan, M. Lazar, J. S. Tsuruda, and D. L. Parker. Analysis of partial volume effects in diffusion-tensor MRI. *Magnetic resonance in medicine*, 45(5):770–780, May 2001.

[3] H. A. Almohamad and S. O. Duffuaa. A linear programming approach for the weighted graph matching problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(5):522–525, May 1993.

[4] David Arthur, Bodo Manthey, and Heiko Röglin. k-Means Has Polynomial Smoothed Complexity. In *Proceedings of the 2009 50th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '09, pages 405–414, Washington, DC, USA, 2009. IEEE Computer Society.

[5] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1):89–112, August 2008.

[6] P. J. Basser, J. Mattiello, and D. LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, January 1994.

[7] Pierre-Louis L. Bazin, Chuyang Ye, John A. Bogovic, Navid Shiee, Daniel S. Reich, Jerry L. Prince, and Dzung L. Pham. Direct segmentation of the major white matter tracts in diffusion tensor images. *NeuroImage*, 58(2):458–468, September 2011.

[8] T. E. Behrens, H. Johansen Berg, S. Jbabdi, M. F. Rushworth, and M. W. Woolrich. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34(1):144–155, January 2007.

[9] T. E. J. Behrens, M. W. Woolrich, M. Jenkinson, H. Johansen-Berg, R. G. Nunes, S. Clare, P. M. Matthews, J. M. Brady, and S. M. Smith. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.*, 50(5):1077–1088, November 2003.

[10] W. D. Bidgood and S. C. Horii. Introduction to the ACR-NEMA DICOM standard. *RadioGraphics*, 12(2):345–355, March 1992.

[11] LÃ©on Bottou and Yoshua Bengio. Convergence Properties of the K-Means Algorithms. In *Advances in Neural Information Processing Systems 7*, pages 585–592, 1995.

[12] M. Bozzali, A. Falini, M. Franceschi, M. Cercignani, M. Zuffi, G. Scotti, G. Comi, and M. Filippi. White matter damage in Alzheimer's disease assessed in vivo using diffusion tensor magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(6):742–746, June 2002.

[13] Anders Brun, Hans Knutsson, Hae-Jeong Park, Martha E. Shenton, and Carl-Fredrik Westin. Clustering Fiber Traces Using Normalized Cuts. pages 368–375. 2004.

[14] P. T. Callaghan, C. D. Eccles, and Y. Xia. NMR microscopy of dynamic displacements: k-space and q-space imaging. *Journal of Physics E: Scientific Instruments*, 21(8):820–822, November 2000.

[15] Rich Caruana and Alexandru Mizil. An Empirical Comparison of Supervised Learning Algorithms. In *Proc. of the 23rd Intl. conf. on Machine learning*, ICML '06, pages 161–168, NY, 2006. ACM.

[16] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based Classification: Concepts and Algorithms. *J. Mach. Learn. Res.*, 10:747–776, June 2009.

[17] Jonathan D. Clayden. Imaging connectivity: MRI and the structural networks of the brain. *Functional neurology*, 28(3):197–203, 2013.

[18] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *Int. J. Patt. Recogn. Artif. Intell.*, 18(03):265–298, May 2004.

[19] I. Corouge, S. Gouttard, and G. Gerig. Towards a shape model of white matter fiber bundles using diffusion tensor MRI. In *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*. IEEE, April 2004.

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[21] Mirco Cosottini, Marco Giannelli, Francesca Vannozzi, Ilaria Pesaresi, Selina Piazza, Gina Belmonte, and Gabriele Siciliano. Evaluation of corticospinal tract impairment in the brain of patients with amyotrophic lateral sclerosis by using diffusion tensor imaging acquisition schemes with different numbers of diffusion-

weighting directions. *Journal of computer assisted tomography*, 34(5):746–750, 2010.

[22] R. W. Cox, J. Ashburner, H. Breman, K. Fissell, C. Haselgrove, C. J. Holmes, J. L. Lancaster, D. E. Rex, S. M. Smith, J. B. Woodward, and S. C. Strother. A (sort of) new image data format standard: NIfTI-1. In *Tenth Annual Meeting of the Organization for Human Brain Mapping*, 2004.

[23] C. Demiralp, J. F. Hughes, and D. H. Laidlaw. Coloring 3D Line Fields Using Boy&#x02019;s Real Projective Plane Immersion. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1457–1464, November 2009.

[24] M. Descoteaux, R. Deriche, T. R. Knosche, and A. Anwander. Deterministic and Probabilistic Tractography Based on Complex Fibre Orientation Distributions. *Medical Imaging, IEEE Transactions on*, 28(2):269–286, February 2009.

[25] R. Douglas Fields. White Matter Matters. *Sci Am*, 298(3):54–61, March 2008.

[26] Marie-Pierre P. Dubuisson and Anil K. Jain. A modified Hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision &amp;amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568 vol.1. IEEE, October 1994.

[27] Robert P. W. Duin and Elżbieta Pkalska. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, 33(7):826–832, May 2012.

[28] Stephen G. Eick and Alan F. Karr. Visual Scalability. *Journal of Computational and Graphical Statistics*, 11(1):22–43, March 2002.

[29] E. W. Forgy. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[30] E. Garyfallidis. *Towards an accurate brain tractography*. PhD thesis, University of Cambridge, 2012.

[31] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan van der Walt, Maxime Descoteaux, Ian Nimmo-Smith, and Dipy Contributors. Dipy, a library for the analysis of diffusion MRI data. *Frontiers in Neuroinformatics*, 8(8):1+, February 2014.

[32] Eleftherios Garyfallidis, Matthew Brett, Marta M. Correia, Guy B. Williams, and Ian Nimmo-Smith. QuickBundles, a Method for Tractography Simplification. *Frontiers in neuroscience*, 6(175), 2012.

[33] G. Gerig, S. Gouttard, and I. Corouge. Analysis of brain white matter via fiber tract modeling. In *In: Proc. IEEE Int. Conf. EMBS. 2004*, volume 2, pages 4421–4424. IEEE, 2004.

[34] Zoubin Ghahramani. Unsupervised Learning. In *Advanced Lectures on Machine Learning*, pages 72–112. 2004.

[35] Alvina Goh and RenÃ© Vidal. Algebraic Methods for Direct and Feature Based Registration of Diffusion Tensor Images. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3953 of *Lecture Notes in Computer Science*, pages 514–525. Springer Berlin Heidelberg, 2006.

[36] Steven Gold and Anand Rangarajan. A Graduated Assignment Algorithm for Graph Matching. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 18(4):377–388, April 1996.

[37] Dan Golding, Marc Tittgemeyer, Alfred Anwander, and Tania Douglas. A comparison of methods for the registration of tractographic fibre images. In P. Robinson and A. Nel, editors, *Proceedings of the Twenty-Second Annual Symposium of the Pattern Recognition Association of South Africa*, pages 55–59, 2011.

[38] Casey B. Goodlett, P. Thomas Fletcher, John H. Gilmore, and Guido Gerig. Group analysis of DTI fiber tract statistics with application to neurodevelopment. *NeuroImage*, 45(1 Suppl), March 2009.

[39] P. Guevara, C. Poupon, D. Rivière, Y. Cointepas, M. Descoteaux, B. Thirion, and J-F F. Mangin. Robust clustering of massive tractography datasets. *NeuroImage*, 54(3):1975–1993, February 2011.

[40] N. S. Hageman, D. W. Shattuck, K. Narr, and A. W. Toga. A diffusion tensor imaging tractography algorithm based on Navier-Stokes fluid mechanics. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 798–801. IEEE, April 2006.

[41] P. Hagmann, J. P. Thiran, L. Jonasson, P. Vandergheynst, S. Clarke, P. Maeder, and R. Meuli. DTI mapping of human brain connectivity: statistical fibre tracking and virtual dissection. *NeuroImage*, 19(3):545–554, July 2003.

[42] Patric Hagmann, Lisa Jonasson, Philippe Maeder, Jean-Philippe P. Thiran, Van J. Wedeen, and Reto Meuli. Understanding diffusion

MR imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 26 Suppl 1(suppl 1):S205–S223, October 2006.

[43] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Physics in medicine and biology*, 46(3), March 2001.

[44] Dorit S. Hochbaum and David B. Shmoys. A Best Possible Heuristic for the k-Center Problem. *Mathematics of Operations Research*, 10(2):180–184, May 1985.

[45] R. Hussein, U. Engelmann, A. Schroeter, and H. P. Meinzer. DICOM structured reporting: Part 1. Overview and characteristics. *Radiographics*, 24(3):891–896, 2004.

[46] S. Jbabdi, M. W. Woolrich, J. L. Andersson, and T. E. Behrens. A Bayesian framework for global tractography. *NeuroImage*, 37(1):116–129, August 2007.

[47] M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, June 2001.

[48] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2):825–841, October 2002.

[49] Mark Jenkinson, Christian F. Beckmann, Timothy E. Behrens, Mark W. Woolrich, and Stephen M. Smith. Fsl. *NeuroImage*, 62(2):782–790, August 2012.

[50] L. Jonasson, P. Hagmann, J. Thiran, and V. Wedeen. Fiber tracts of high angular resolution dMRI are easily segmented with spectral clustering. In *ISMRM*, 2005.

[51] Ning Kang, Jun Zhang, E. S. Carlson, and D. Gembris. White matter fiber tractography via anisotropic diffusion simulation in the human brain. *Medical Imaging, IEEE Transactions on*, 24(9):1127–1137, August 2005.

[52] D. A. Keim. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):1–8, January 2002.

[53] DanielA Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual Analytics: Scope and Challenges. In SimeonJ Simoff, MichaelH Böhlen, and Arturas Mazeika, editors, *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, chapter 6, pages 76–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[54] P. J. M. Laarhoven and E. H. L. Aarts, editors. *Simulated Annealing: Theory and Applications*. Kluwer Academic Publishers, Norwell, MA, USA, 1987.

[55] E. W. Lang, A. M. Tomé, I. R. Keck, J. M. Górriz Sáez, and C. G. Puntonet. Brain Connectivity Analysis: A Short Survey. *Intell. Neuroscience*, 2012, January 2012.

[56] Mariana Lazar. Mapping brain anatomical connectivity using white matter tractography. *NMR Biomed.*, 23(7):821–835, August 2010.

[57] Denis Le Bihan and Heidi Johansen-Berg. Diffusion MRI at 25: Exploring brain tissue structure and function. *NeuroImage*, 61(2):324–341, June 2012.

[58] A. Leemans, J. Sijbers, S. De Backer, E. Vandervliet, and P. Parizel. Multiscale white matter fiber tract coregistration: A new feature-based approach to align diffusion tensor data. *Magnetic Resonance in Medicine*, 55(6):1414–1423, 2006.

[59] Alexander Leemans, Jan Sijbers, Steve De Backer, Everhard Vandervliet, and PaulM Parizel. Affine Coregistration of Diffusion Tensor Magnetic Resonance Images Using Mutual Information. In Jacques Blanc-Talon, Wilfried Philips, Dan Popescu, and Paul Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, volume 3708 of *Lecture Notes in Computer Science*, pages 523–530. Springer Berlin Heidelberg, 2005.

[60] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, June 1995.

[61] Lorenzo Livi and Antonello Rizzi. The graph matching problem. *Pattern Analysis & Applications*, 16(3):253–283, August 2013.

[62] Mahnaz Maddah, W. Eric Grimson, Simon K. Warfield, and William M. Wells. A unified framework for clustering and quantitative analysis of white matter fiber tracts. *Medical image analysis*, 12(2):191–202, April 2008.

[63] Mahnaz Maddah, Andrea U. J. Mewes, Steven Haker, Grimson, and Simon K. Warfield. Automated Atlas-Based Clustering of White Matter Fiber Tracts from DTMRI. In James S. Duncan and

Guido Gerig, editors, *MICCAI 2005*, volume 3749 of *Lecture Notes in Computer Science*, chapter 24, pages 188–195. Springer, Berlin, Heidelberg, 2005.

[64] Mahnaz Maddah, Lilla Zöllei, Eric E. Grimson, and William M. Wells. Modeling of Anatomical Information in Clustering of White Matter Fiber Trajectories Using Dirichlet Distribution. *IEEE on Pattern Analysis and Machine Intelligence.*, 2008:1–7, July 2008.

[65] V. R. S. Mani and Dr Rivazhagan. Survey of Medical Image Registration. *Journal of Biomedical Engineering and Technology*, 1(2):8–25, 2013.

[66] Arnaldo Mayer and Hayit Greenspan. *Direct Registration of White Matter Tractographies with Application to Atlas Construction*. 2007.

[67] Peter Mildenberger, Marco Eichelberg, and Eric Martin. Introduction to the DICOM standard. *European radiology*, 12(4):920–927, April 2002.

[68] Bart Moberts, Anna Vilanova, and Jarke J. van Wijk. Evaluation of Fiber Clustering Methods for Diffusion Tensor Imaging. In *IEEE Visualization*, pages 65–72, 2005.

[69] Susumu Mori and Peter C. M. van Zijl. Fiber tracking: principles and strategies a technical review. *NMR Biomed.*, 15(7-8):468–480, November 2002.

[70] M. Mustra, K. Delac, and M. Grgic. Overview of the DICOM standard. In *ELMAR, 2008. 50th International Symposium*, volume 1, pages 39–44. IEEE, 2008.

[71] Radhouène Neji, Ahmed Besbes, Nikos Komodakis, Jean Deux, Mezri Maatouk, Alain Rahmouni, Guillaume Bassez, Gilles Fleury,

and Nikos Paragios. Clustering of the human skeletal muscle fibers using linear programming and angular Hilbertian metrics. *Info. processing in Med. Imaging*, 21:14–25, 2009.

[72] Lauren J. O'Donnell, William M. Wells, Alexandra J. Golby, and Carl-Fredrik F. Westin. Unbiased groupwise registration of white matter tractography. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 15(Pt 3):123–130, 2012.

[73] Lauren J. O'Donnell and Carl-Fredrik F. Westin. Automatic tractography segmentation using a highdimensional white matter atlas. In *IEEE Trans. Med. Imag*, pages 1562–1575, 2007.

[74] E. Olivetti, T. B. Nguyen, E. Garyfallidis, N. Agarwal, and P. Avesani. Fast Clustering for Interactive Tractography Segmentation. In *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*, pages 42–45. IEEE, June 2013.

[75] Emanuele Olivetti and Paolo Avesani. Supervised segmentation of fiber tracts. In *Proceedings of SIMBAD'11*, SIMBAD'11, pages 261–274, Berlin, Heidelberg, 2011. Springer-Verlag.

[76] Emanuele Olivetti, Thien B. Nguyen, and Eleftherios Garyfallidis. The Approximation of the Dissimilarity Projection. In *IEEE International Workshop on Pattern Recognition in NeuroImaging*, volume 0, pages 85–88, Los Alamitos, CA, USA, 2012. IEEE.

[77] Emanuele Olivetti, Sriharsha Veeramachaneni, Susanne Greiner, and Paolo Avesani. Brain connectivity analysis by reduction to pair classification. In *Proceeding of Cognitive Information Processing (CIP), 2010*, pages 275–280, June 2010.

[78] Stephen M. Omohundro. Five Balltree Construction Algorithms, 1989.

[79] Mauricio Orozco-Alzate and César G. Castellanos-Dom′ınguez. Clustering on dissimilarity representations for detecting mislabelled seismic signals at nevado del ruiz volcano. *Earth Sciences Research Journal*, 11(2):135–140, December 2007.

[80] Evren Özarslan and Thomas H. Mareci. Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging. *Magnetic Resonance in Medicine*, 50(5):955–965.

[81] S. Pajevic and C. Pierpaoli. Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: application to white matter fiber tract mapping in the human brain. *Magnetic resonance in medicine*, 42(3):526–540, September 1999.

[82] H. J. Park, M. Kubicki, M. E. Shenton, A. Guimond, R. W. McCarley, S. E. Maier, R. Kikinis, F. A. Jolesz, and C. F. Westin. Spatial normalization of diffusion tensor MRI using multiple channels. *Neuroimage*, 20(4):1995–2009, 2003.

[83] Geoffrey J. M. Parker, Hamied A. Haroon, and Claudia A. M. Wheeler-Kingshott. A framework for a streamline-based probabilistic index of connectivity (PICo) using a structural interpretation of MRI diffusion measurements. *J. Magn. Reson. Imaging*, 18(2):242–254, August 2003.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and

E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[85] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011.

[86] E. Pekalska, R. Duin, and P. Paclik. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, February 2006.

[87] Elzbieta Pekalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence).* World Scientific Publishing Company, December 2005.

[88] Elzbieta Pekalska, Pavel Paclik, and Robert P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *J. Mach. Learn. Res.*, 2:175–211, 2002.

[89] Harald Piringer. *Large Data Scalability in Interactive Visual Analysis*. PhD thesis, Institute of Computer Graphics and Algorithms, University of Technology, Vienna, Favoritenstrasse 9-11/186, A-1040 Vienna, Austria, May 2011.

[90] Anand Rangarajan, Steven Gold, and Eric Mjolsness. A Novel Optimizing Network Architecture with Applications. *Neural Comput.*, 8(5):1041–1060, July 1996.

[91] Christian Ros, Daniel Güllmar, Martin Stenzel, Hans-Joachim Mentzel, and Jürgen R. Reichenbach. Atlas-Guided Cluster Analysis of Large Tractography Datasets. *PLoS ONE*, 8(12):e83847+, December 2013.

[92] J. Ruiz-Alzola, C-F F. Westin, S. K. Warfield, C. Alberola, S. Maier, and R. Kikinis. Nonrigid registration of 3D tensor medical data. *Medical image analysis*, 6(2):143–161, June 2002.

[93] Peter Savadjiev, Jennifer Campbell, G. Pike, and Kaleem Siddiqi. Streamline Flows for White Matter Fibre Pathway Segmentation in Diffusion MRI. pages 135–143. 2008.

[94] D. Sculley. Web-scale K-means Clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1177–1178, New York, NY, USA, 2010. ACM.

[95] SimeonJ Simoff, MichaelH Böhlen, and Arturas Mazeika. Visual Data Mining: An Introduction and Overview. In SimeonJ Simoff, MichaelH Böhlen, and Arturas Mazeika, editors, *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2008.

[96] Stephen M. Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, November 2002.

[97] Stamatios N. Sotiropoulos, Saad Jbabdi, Junqian Xu, Jesper L. Andersson, Steen Moeller, Edward J. Auerbach, Matthew F. Glasser, Moises Hernandez, Guillermo Sapiro, Mark Jenkinson, David A. Feinberg, Essa Yacoub, Christophe Lenglet, David C. Van Essen, Kamil Ugurbil, Timothy E. Behrens, and WU-Minn HCP Consortium. Advances in diffusion MRI acquisition and processing in

the Human Connectome Project. *NeuroImage*, 80:125–143, October 2013.

[98] Frederic Stahl, Bogdan Gabrys, Mohamed M. Gaber, and Monika Berendsen. An overview of interactive visual data mining techniques for knowledge discovery. *WIREs Data Mining Knowl Discov*, 3(4):239–256, July 2013.

[99] David S. Tuch, Timothy Reese, Mette Wiegell, Nikos Makris, John Belliveau, and Van Wedeen. High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magn. Reson. Med.*, 48(4):577–582, October 2002.

[100] D. Turnbull and C. Elkan. Fast recognition of musical genres using RBF networks. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):580–584, April 2005.

[101] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, and E. Yacoub. The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231, October 2012.

[102] Setsu Wakana, Arvind Caprihan, Martina Panzenboeck, James Fallon, Michele Perry, Randy Gollub, Kegang Hua, Jiangyang Zhang, Hangyi Jiang, Prachi Dubey, Ari Blitz, Peter Zijl, and Susumu Mori. Reproducibility of quantitative tractography methods applied to cerebral white matter. *NeuroImage*, 36(3):630–644, July 2007.

[103] Qian Wang, Pew-Thian Yap, Guorong Wu, and Dinggang Shen. Application of neuroanatomical features to tractography clustering. *Hum. Brain Mapp*, 34(9):2089–2102, September 2013.

[104] R. Wang, T. Benner, A. Sorensen, and V. Wedeen. Diffusion Toolkit: A Software Package for Diffusion Imaging Data Processing and Tractography - 03720.pdf, May 2007.

[105] Xiaogang Wang, W. Eric Grimson, and Carl-Fredrik F. Westin. Tractography segmentation using a hierarchical Dirichlet processes mixture model. *NeuroImage*, 54(1):290–302, January 2011.

[106] Yi Wang, Aditya Gupta, Zhexing Liu, Hui Zhang, Maria L. Escolar, John H. Gilmore, Sylvain Gouttard, Pierre Fillard, Eric Maltbie, Guido Gerig, and Martin Styner. DTI registration in atlas based fiber analysis of infantile Krabbe disease. *NeuroImage*, 55(4):1577–1586, April 2011.

[107] D. Wassermann, L. Bloy, E. Kanterakis, R. Verma, and R. Deriche. Unsupervised white matter fiber clustering and tract probability map generation: applications of a Gaussian process framework for white matter fibers. *NeuroImage*, 51(1):228–241, May 2010.

[108] Brandon Whitcher, Volker J. Schmid, and Andrew Thorton. Working with the DICOM and NIfTI Data Standards in R. *Journal of Statistical Software*, 44(6):1–29, October 2011.

[109] M. Zaslavskiy, F. Bach, and J. P. Vert. A Path Following Algorithm for the Graph Matching Problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2227–2242, December 2009.

[110] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. Many-to-Many Graph Matching: A Continuous Relaxation Approach.

In JoséLuis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, chapter 33, pages 515–530. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[111] Song Zhang, S. Correia, and D. H. Laidlaw. Identifying White-Matter Fiber Bundles in DTI Data Using an Automated Proximity-Based Fiber-Clustering Method. *IEEE Transactions on Visualization and Computer Graphics*, 14(5):1044–1053, September 2008.

[112] Feng Zhou and F. De la Torre. Factorized graph matching. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 127–134. IEEE, June 2012.

[113] Ulas Ziyan, Mert R. Sabuncu, W. Eric Grimson, and Carl-Fredrik F. Westin. Consistency Clustering: A Robust Algorithm for Group-wise Registration, Segmentation and Automatic Atlas Construction in Diffusion MRI. *International journal of computer vision*, 85(3):279–290, 2009.

[114] Lilla Zöllei, Erik Learned-Miller, Eric Grimson, and William Wells. Efficient Population Registration of 3D Data. In Yanxi Liu, Tianzi Jiang, and Changshui Zhang, editors, *Computer Vision for Biomedical Image Applications*, volume 3765 of *Lecture Notes in Computer Science*, pages 291–301. Springer Berlin Heidelberg, 2005.

[115] O. Zvitia, A. Mayer, and H. Greenspan. Adaptive mean-shift registration of white matter tractographies. In *the 5th IEEE Intl. Symposium on Biomedical Imaging, 2008. ISBI 2008.*, pages 692–695, May 2008.