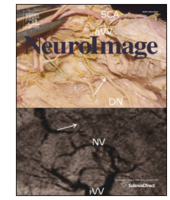
NeuroImage xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg




# A new compression format for fiber tracking datasets

Caroline Presseau, Pierre-Marc Jodoin, Jean-Christophe Houde, Maxime Descoteaux *

*Computer Science Department, Faculty of Science, Université de Sherbrooke, 2500 Boulevard Université, Sherbrooke, QC J1K 2R1, Canada*
*Centre d'Imagerie Moléculaire de Sherbrooke (CIMS), Centre de Recherche CHUS, Canada*

A R T I C L E   I N F O

*Article history:*
Accepted 22 December 2014
Available online xxxx

*Keywords:*
Compression
Diffusion MRI
Tractography
Diffusion Tensor Imaging (DTI)
High Angular Resolution Diffusion Imaging (HARDI)

A B S T R A C T

A single diffusion MRI streamline fiber tracking dataset may contain hundreds of thousands, and often millions of streamlines and can take up to several gigabytes of memory. This amount of data is not only heavy to compute, but also difficult to visualize and hard to store on disk (especially when dealing with a collection of brains). These problems call for a fiber-specific compression format that simplifies its manipulation. As of today, no fiber compression format has yet been adopted and the need for it is now becoming an issue for future connectomics research. In this work, we propose a new compression format, *.zfib*, for streamline tractography datasets reconstructed from diffusion magnetic resonance imaging (dMRI). Tracts contain a large amount of redundant information and are relatively smooth. Hence, they are highly compressible. The proposed method is a processing pipeline containing a linearization, a quantization and an encoding step. Our pipeline is tested and validated under a wide range of DTI and HARDI tractography configurations (step size, streamline number, deterministic and probabilistic tracking) and compression options. Similar to JPEG, the user has one parameter to select: a worst-case maximum tolerance error in millimeter (mm). Overall, we find a compression factor of more than 96% for a maximum error of 0.1 mm without any perceptual change or change of diffusion statistics (mean fractional anisotropy and mean diffusivity) along bundles. This opens new opportunities for connectomics and tractometry applications.





## 1. Introduction

Diffusion magnetic resonance imaging (dMRI) tractography is an increasingly popular research area that helps in understanding structural connectivity between brain regions. The great success of dMRI tractography comes from its capability to accurately describe some of the neural architecture in vivo (Descoteaux and Poupon, 2014). Streamline fiber tracking datasets contain thousands, if not millions of streamlines and each streamline contains hundreds to thousands of 3D points. These streamlines are often called "*tracts*". Here, we prefer to use the term streamline for a set of 3D points that represent virtual anatomical fiber representations (Côté et al., 2013). For example, a dataset generated with the deterministic DTI (or tensorline) algorithm (Weinstein et al., 1999; Lazar et al., 2003) using a 0.2 mm step size and 500,000 streamlines requires roughly 1.3 gigabytes (GB) of space.

As such, some datasets are so large that they cannot be visualized due to the limited amount of RAM (Random Access Memory) available on most computers. Thus, visualization, storage, and handling of such a dataset require heavy processing and a lot of memory. Unfortunately, no fiber compression format has yet been adopted and the need for it is now becoming a glaring issue for future research.

Compression algorithms are categorized into two distinct families, *lossy* and *lossless*. Lossless data compression algorithms allow the original signal (or document) to be recovered from the compressed one with no loss of quality (Nelson and Jean-Loup, 1995; Sayood, 2006; Salomon and Motta, 2010). ZIP, GIF and PNG (Taubman and Marcellin, 2002; Sayood, 2006; Schelkens et al., 2009; Salomon and Motta, 2010) file formats are typical examples of lossless compression algorithms. On the other hand, lossy algorithms perform a compression by removing elements from the original signal. Hence, the exact original data cannot be retrieved from the compressed version (Nelson and Jean-Loup, 1995; Sayood, 2006; Salomon and Motta, 2010). Multimedia file formats such as *mp3*, *JPEG* and *MPEG* are typically associated to lossy compression. As opposed to lossless compression, lossy compression techniques have no limit on the amount by which they can compress a signal. Lossy compression thus involves a trade-off between quality and compression ratio (Nelson and Jean-Loup, 1995; Sayood, 2006; Salomon and Motta, 2010).

What differentiates a lossy compression method from another one is often the end application it was designed for. Although *mp3*, *JPEG* and *MPEG* share a common ground, they nonetheless have their own specifications. Due to the very nature of their signal, compression methods used for audio files (1D), image files (2D) or movie files (2D + time)

* Corresponding author.
*E-mail address:* m.descoteaux@usherbrooke.ca (M. Descoteaux).

http://dx.doi.org/10.1016/j.neuroimage.2014.12.058
1053-8119/© 2015 Elsevier Inc. All rights reserved.

cannot be used interchangeably without significant modification. The situation is the same with streamline tracking datasets which also call for an application-specific compression scheme. Blindly applying a compression method to a streamline dataset could only lead to sub-optimal results.

Streamlines are 3D curves represented as a collection of 3D points. Unlike voice and music, streamlines are both smooth and defined in a 3D space. Unlike images and videos, streamlines are not defined over a regular lattice and, most importantly, are not functions of space. Neuroimaging tracking applications have also their own specifications. One important feature for a streamline compression method is to preserve perfect accuracy at the end points. For many neuroimaging applications, the accuracy of the starting and ending points is more important than any other points along the curve (e.g. *connectomics*). On the other hand, for neurosurgical planning (Chamberland and Descoteaux, 2012; Fortin et al., 2012), the full path is of interest and the error must be controlled. Also, unlike *JPEG* or *MPEG*, whose focus is to keep low visual compression artifacts, streamline datasets are meant for medical applications where qualitative perceptual errors are not a primary factor. Instead, medical users prefer to account for compression errors quantified in millimeters.

The aim of this work is to provide a new lossy compression format called *.zfib* for streamline tracking datasets. We propose a complete, simple and powerful compression scheme validated under a wide range of tractography configurations and compression options. We demonstrate that streamlines are smooth and often represented with a large number of points that can be removed without changing the pictorial view of streamlines on the screen nor changing the average fractional anisotropy (FA) and mean diffusivity (MD) along white matter bundles. Careful experiments are performed on real datasets of different sizes (24 megabytes (MB) to 15 GB), from different tractography algorithms (deterministic HARDI, probabilistic HARDI, and deterministic DTI[1]), different step sizes (0.1 to 1 mm) and different number of streamlines (60,000 to 3,000,000). Overall, with a 0.1 mm maximum error, which is very small considering the voxel size (usually 2 mm isotropic), we can reach a compression ratio of more than 96%.

Our findings open new perspectives for future connectomics and group studies using tractography results with large number of streamlines (Hagmann et al., 2008; Honey et al., 2009) but also for future tractography methods. Datasets of several gigabytes of memory that were before impossible to visualize and hard to store on disk may now be visualized and stored with only few megabytes of memory. For example, the 3,000,000 deterministic DTI and probabilistic HARDI streamline files of size 7.83 GB and 5.92 GB respectively, can be compressed down to 95.6 MB and 94.4 MB respectively with a maximum error of 0.5 mm.

The remainder of this paper is organized as follows. In Section 2, we introduce a generic four-step compression pipeline made of a linearization step, a transformation/approximation step, a quantization step, and an encoding step. These steps are commonly used in well-known compression algorithms such as mp3 and JPEG (Pennebaker and Mitchell, 1993; Sayood, 2006, Chap. 13; Salomon and Motta, 2010, Chap. 7; Nelson and Jean-Loup, 1995, Chap. 11). Since these steps can accommodate different algorithms and include a variety of parameters, Sections 3 and 4 thoroughly validate the impact of each step on compression ratios, speed, and accuracy on a variety of streamline datasets. In Section 5, we discuss the results as well as the pros and cons of every step. We then present the final compression algorithm and provide parameters which produce high compression ratios, low processing time, and high accuracy. We then draw conclusions in Section 6.

[1] DTI: Diffusion Tensor Imaging HARDI: High-Angular-Resolution Diffusion Imaging.

## 2. Methodology

### 2.1. Definitions

A fiber tracking dataset is a set of 3D curves in which each streamline is represented as a series of 3D points. The definitions used in this paper are as follows:

- Streamline point: A point in 3D space, $p_i = (x_i, y_i, z_i) \in \mathbf{R}^3$.
- Streamline: a curve in 3D space containing a finite, ordered and connected sequence of 3D points $f_i = \{p_1, \ldots, p_k\}$.
- Fiber tracking dataset: A set of streamlines $F_n = \{f_1, \ldots, f_n\}$ for a finite $n \in \mathbf{N}$.
- $|.|$: Cardinality operator which returns number of elements in a set or a sequence. For example, $|f|$ stands for the total number of points $p_i$ in the streamline $f$.

### 2.2. Piece-wise linearization of streamlines

Tractography algorithms find structural white matter connections by following the principal directions of diffusion at each voxel (Descoteaux and Poupon, 2014). One important tractography parameter is the step size $\delta$. Step size determines the distance between two consecutive points $p_i$ and $p_{i+1}$. As such, a small $\delta$ leads to a tight set of points $p_i$ for each streamline. This implies that more points among each streamline are likely to be collinear in space. Hence, removing some collinear points would produce a compression without affecting much the accuracy of the data. Decreasing the number of points not only helps in reducing the file size, it also allows for faster rendering and the use of less RAM at runtime. In fact, this step could even be introduced within tractography algorithms themselves.

The goal of this step is to remove as many points as possible within a certain margin of millimetric error ($\epsilon$ mm). This is achieved with a piece-wise linearization procedure over each streamline according to a tolerance error, $\varepsilon$. Thus, depending on how severely the streamlines are linearized, the pictorial view of streamlines remains almost unchanged.

Formally, $\hat{f}$ is a valid linearized version of $f$ (the raw input streamline) according to a tolerance value $\epsilon$ if and only if every point $p_i \in f$ respects the following constraint:

$$\forall p_i \in f \;:\; dist\left(\hat{f}, p_i\right) < \varepsilon \qquad i = 1, \ldots, |f|, \tag{1}$$

where *dist* is the shortest distance between a streamline point $p_i$ and the corresponding fiber $\hat{f}$. In other words, $\hat{f}$ is a valid linearized version of $f$ if one cannot remove a point from $\hat{f}$ without creating an error of more than $\epsilon$ mm.

Fig. 1 illustrates the linearization process for different $\epsilon$ values. The reader shall note that $\hat{f}$ always contains the starting and the ending points $p_1$ and $p_{|f|}$ of $f$. Hence, the accuracy of the end points is preserved, which is crucial not to change the connectivity profile of the tractography dataset.

### 2.3. Streamline transformation and approximation

A signal transformation is a generic term to describe a reversible mathematical operation which projects a signal from a set of basis functions to another set of basis functions. These basis functions define the domain on which the signal is represented. For example, a Fourier transformation converts a signal from a spatial domain to a frequency domain, and vice versa. A transformation function does not compress data per se. It rather concentrates its energy on a smaller number of

Please cite this article as: Presseau, C., et al., A new compression format for fiber tracking datasets, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2014.12.058