Fiber tractography using machine learning

Peter F. Neher a,*, Marc-Alexandre Côté b, Jean-Christophe Houde b, Maxime Descoteaux b, Klaus H. Maier-Hein a

a Medical Image Computing (MIC), German Cancer Research Center (DKFZ), Heidelberg, Germany
b Sherbrooke Connectivity Imaging Laboratory (SCIL), Computer Science Department, Université de Sherbrooke, Sherbrooke, Quebec, Canada

ARTICLE INFO

Keywords:
Fiber tractography
Machine learning
Diffusion-weighted imaging
Connectomics

ABSTRACT

We present a fiber tractography approach based on a random forest classification and voting process, guiding each step of the streamline progression by directly processing raw diffusion-weighted signal intensities. For comparison to the state-of-the-art, i.e. tractography pipelines that rely on mathematical modeling, we performed a quantitative and qualitative evaluation with multiple phantom and in vivo experiments, including a comparison to the 96 submissions of the ISMRM tractography challenge 2015. The results demonstrate the vast potential of machine learning for fiber tractography.

1. Introduction

Fiber tractography on the basis of diffusion-weighted magnetic resonance imaging (DW-MRI) has been a research topic for almost 20 years. A vast spectrum of tractography algorithms has been presented over the last years ranging from local deterministic approaches (Basser, 1998; Chao et al., 2008; Lazar et al., 2003; Mori et al., 1999; Tournier et al., 2012) through probabilistic methods (Behrens et al., 2007; Berman et al., 2008; Descoteaux et al., 2009; Friman et al., 2006; Vorburger et al., 2013; Zhang et al., 2013) to global tractography (Aganj et al., 2011; Daducci et al., 2015; Fillion et al., 2009; Jbabdi et al., 2007; Lemkadem et al., 2014; Mangin et al., 2013; Reisert et al., 2011). To infer information about the complex microstructure of brain tissue and to optimally exploit the acquisition dependent signal characteristics of diffusion-weighted images, tractography algorithms employ mathematical models calculated from the diffusion-weighted signal. Prominent examples include the diffusion tensor (Basser et al., 1994), multi tensor models (Kreher et al., 2005; Malcolm et al., 2010), spherical deconvolution (Alexander, 2005; Jeurissen et al., 2014; Schultz et al., 2010; Tournier et al., 2007), persistent angular structures (Jansons and Alexander, 2003), Q-ball modeling (Aganj et al., 2009; Descoteaux et al., 2007) as well as a large variety of multi-compartment models (Assaf et al., 2008; Assaf and Basser, 2005; Panagiotaki et al., 2012; Sotiropoulos et al., 2012; Zhang et al., 2012). To obtain such a model representation of certain local tissue properties, the corresponding inverse problem has to be solved using the measured signal. Depending on the model, this imposes different constraints on the data quality and acquisition sequence, such as a minimum number of diffusion-weighting gradients and a high signal to noise ratio (SNR). Usually, the more expressive a model is, the more demanding is its calculation and the higher are the requirements for the signal. Choosing the optimal model is not a trivial problem. Oversimplified modeling can, for example, hamper the ability to resolve crossing fiber situations. Modeling approaches make various assumptions about signal and tissue properties that are highly variable across subjects, locations in the brain and acquisition schemes. This issue has been discussed extensively and there is still no solution that is optimal in all situations (Daducci et al., 2014; Farquharson et al., 2013; Jbabdi and Johansen-Berg, 2011; Neher et al., 2015a; Nimsky, 2014). Furthermore, depending on the model and the dataset, a certain number of tractography parameters, such as a termination criterion, e.g. on the basis of a threshold on the fractional anisotropy (FA), have to be adjusted manually, which requires expert knowledge.

In the context of signal modeling, initial studies have successfully shown the potential of machine learning techniques, e.g. for the tasks of image quality transfer and tissue micro-structure analysis (Alexander et al., 2017; Golkov et al., 2016; Nedjati-Gilani et al., 2017; Reisert et al., 2017) and to estimate the number of distinct fiber clusters per voxel (Schultz, 2012). These methods avoid or alleviate some of the issues that come with the usage of diffusion-signal models.

Here, we present the first approach to fiber tractography based on machine learning, which has several advantages: There is no need to explicitly solve the inverse problem to obtain a representation of the
diffusion propagator or the tissue microstructure from the diffusion-weighted signal. Also, classical modeling approaches often struggle with artifacts that are not included in the mathematical model, such as noise and distortions, while a machine learning based approach can, to a certain extent, deal with such signal imperfections by learning them from the training data. While the desired information has of course to be encoded in the signal, there are no general restrictions on the type of image acquisition, e.g. regarding the number of diffusion-weighting gradients or the b-value. This enables straightforward optimization of the method to a specific acquisition scheme. Machine learning also enables seamless integration of additional information into the tractography process, such as other image contrasts or the previous streamline direction. Furthermore, the distinction between white matter and non-white matter tissue is directly learned from the training data. This means that additional white matter mask images or model derived thresholds, such as on the FA or on the magnitude of the peaks determined by the model, are not necessary to constrain the tractography.

Additionally, in classical streamline tractography, the decision about the next direction of the streamline progression is typically based solely on the signal information at the current streamline position. Some approaches have been presented that use neighborhood information to disentangle asymmetric fiber patterns such as curving and fanning fiber structures (Bastiani et al., 2016; Rowe et al., 2013; Savadjiev et al., 2008). The probabilistic nature of our approach directly enables the meaningful integration of the information obtained from multiple signal samples in the local neighborhood, guiding each step of the streamline progression. Similar to the aforementioned approaches, this enables the disentangling of asymmetric fiber configurations. The focus in this work is, in contrast to these previous works, to enable a better-informed decision about whether to terminate or proceed the fiber progression using the neighborhood information.

This work is based on the preliminary results and methods presented at MICCAI 2015 (Neher et al., 2015b). Here, we introduce new types of classification features and training data. The evaluation of our method was extended to the data and results from the ISMRM tractography challenge 2015, including 96 tractography methods for comparison, including an in-depth analysis of the individual components of the presented method. Furthermore, we extensively assessed the capabilities of our method to generalize from in vivo to in vivo, in vivo to in silico and in silico to in vivo data. In our experiments, we show that the method performs as good as or better than the state-of-the-art and generalizes well to unseen datasets.

2. Materials and methods

Standard streamline tractography approaches reconstruct a fiber by iteratively extending the fiber in a direction depending on the current position. The directional information is usually inferred from a signal model at the respective location, such as the diffusion tensor (DT), fiber orientation distribution functions (fODF) or diffusion orientation distribution functions (dODF). The method presented in this work also iteratively extends the current fiber, but in contrast to standard streamline approaches the determination of the next progression direction relies on a different concept:

1. Instead of mathematically modeling the signal, the presented method employs a random forest classifier working on the raw diffusion-weighted image values to obtain information about local tissue properties, such as tissue type (white matter or not white-matter) and fiber direction (Section 2.1). The classifier learns the relationship between the last fiber direction, the local image data and the next fiber direction.

2. To progress or possibly terminate a streamline, not only the image information at the current location but also at several sampling points distributed in the neighborhood are considered. The final decision on the next action is then based on a voting process among the individual classification results at these sampling points (Section 2.2).

2.1. Learning fiber directions using random forest classification

2.1.1. Classification features

We evaluated two types of input features for the random forest classifier: raw signal intensities as well as the voxel-wise coefficients of the corresponding spherical harmonics fit of the diffusion-weighted signal. In both cases, multiple b-values can be handled by concatenating the feature vectors of the individual b-shells. For the first case, the signal is resampled to 100 directions equally distributed over the hemisphere using spherical harmonics, to become independent of the gradient scheme used to acquire the data. In addition to the diffusion-weighted signal features, the normalized previous streamline direction is added to the list of classification features, thus enabling the method to better overcome ambiguous situations by taking its progression history into account. Each signal feature is used twice for training, one time in conjunction with the directional feature and a second time with a zero-vector instead of the direction feature to enable valid classifications at streamline seed points where no previous direction is available. To render the learning and tractography process invariant to image rotations, the fiber directions are rotated using the inverse of the image rotation matrix before being used as training or classification feature.

We furthermore investigated the effect of adding features such as T1 signal values or scalar indices derived from the diffusion-weighted signal, such as the generalized fractional anisotropy (GFA) (Tuch, 2004). To obtain image values at arbitrary positions, the image voxels are interpolated trilinearly.

2.1.2. Reference directions

To train the classifier, reference fiber tracts corresponding to the respective diffusion-weighted image are necessary. The tangent fiber directions of the reference tractogram are discretized, i.e. assigned to the, in terms of angular deviation, nearest of 100 possible target directions $v_i$ ($1 \leq i \leq 100$) distributed over the hemisphere. Each of these directions corresponds to a label that is learned by the classifier.

In our experiments, we explore two variants of obtaining these reference tracts. (1) We use a previously performed standard tractography to obtain the reference tracts. The impact of the tractography algorithm choice on the presented approach was systematically evaluated in our experiments. While this approach introduces a dependency of the training step on the quality of the reference tractogram, we perform further experiments (2) with simulated datasets and the corresponding known ground truth tracts for training. Multiple variations of training and test data are explored to analyze this aspect of our method.

2.1.3. Classifier output

The classifier produces a probability $P(v_i)$ for each of the 100 target directions $v_i$ as well as a non-fiber probability $P_{\text{nonfib}}$.

2.1.4. Classifier implementation

We used the VIGRA random forest implementation (https://uкоothe.github.io/vigra/) included in the medical imaging interaction toolkit MITK (Nolden et al., 2013).

2.2. Streamline progression using neighborhood sampling

The primary goal of using neighborhood information to determine the next progression direction or a fiber termination is to avoid a premature termination of the streamline. The neighborhood sampling and voting process described in this section stabilizes the streamline progression by making it less sensitive to noise and local signal ambiguities. Furthermore, it enables the streamline to deflect off white matter boundaries that run relatively parallel to the streamline and that could otherwise...
lead to a streamline termination (see Fig. 1a). The process consists of several steps that are detail in the following paragraphs.

2.2.1. Neighborhood sampling

We implemented two neighborhood sampling schemes: (1) Sampling on a complete sphere around the current streamline position $p^0$ and (2) sampling on a hemisphere in front of $p^0$. Parameters are the number $N$ of sampling points $p^j$ ($1 \leq j \leq N$) and the sampling radius $r$. In case (1), only $N/2$ sampling points are used, since only one hemisphere needs to be covered. At seed positions, where no previous streamline direction $v_{old}$ is available, scheme (1) is employed. The current streamline position $p^0$ is included in both schemes.

2.2.2. Proposal generation

Classification is performed at each $p^j$ to infer a weighted local direction proposal $\nu^j$. We implemented a deterministic and a probabilistic method to generate the proposals:

(1) The deterministic method simply sums all possible directions weighted with their respective probability $P_j(v_j)$, as provided by the classifier (see Section 2.1), and their deviation from $v_{old}$:

$$\nu = \sum_i w_i v_i,$$

with $w_i = P_j(v_i, v_{old})$. The dot product is a directional prior that promotes straight fibers. An additional hard curvature threshold is employed that, when exceeded, sets $w_i = 0$.

(2) The probabilistic method directly samples the output direction from the discrete distribution provided by the classifier. Again, a hard curvature threshold is employed that, when exceeded, sets the respective probability to zero. This modified discrete distribution is normalized to sum to 1. The sampled direction $\nu^j$ is weighted by its respective probability $P_j(\nu^j)$.

2.2.3. Consider vote for termination

If the non-fiber probability $P_{nonfib}^j$ of sample $j$ exceeds the cumulated weighted probabilities of all possible directions ($P_{nonfib}^j > \sum_i w_i$) or probability of the sampled direction ($P_{nonfib}^j > P(\nu^j)$) respectively, a potential tract boundary is identified and a vote for termination of sample $j$ is considered.

Now, the position vector $d = p^j - p$ is related to the previous direction $v_{old}$ in order to decide whether termination is preferable or should be avoided. A termination is considered more likely if non-fiber regions lie straight ahead (i.e. in the current direction of streamline progression $v_{old}$). If the streamline progresses more or less parallel to the detected fiber bundle margin, a premature termination is rather avoided (see Fig. 1b). To this end an auxiliary sample position $\hat{p}^j$ is evaluated that is determined by a 180° rotation of $d$ around $v_{old}$:

$$\hat{p}^j = p - d + 2\langle v_{old}, d \rangle v_{old}.$$

If $P_{nonfib}$ at the new position exceeds the respective threshold of the deterministic or probabilistic sample ($P_{nonfib}^j > \sum_i w_i$ or $P_{nonfib}^j > P(\nu^j)$ respectively), $\nu^j$ is set to $(0,0,0)$ (vote for termination, see Fig. 1a), otherwise $\nu^j$ is set to $\hat{d} = \hat{p}^j - p$ to deflect the streamline away from the detected fiber margin (see Fig. 1b).

Voting process: After all local proposals are collected, the final decision about the new progression direction or a fiber termination is made. To decide whether a streamline should terminate, we implemented two voting strategies:

(1) If full spherical sampling is used, a streamline terminates if all sampling points vote for termination.

(2) If the sampling points are distributed on the half sphere, a streamline terminates if the majority of all frontal sampling points, meaning sampling points that are located in a 90° cone around $v_{old}$ ahead of $p$, vote for termination.

Strategy (2) was designed as a compromise between the full neighborhood sampling and the classical approach where only the current streamline position is considered. We evaluate the effect of the different sampling- and corresponding termination-strategies in our experiments.

If the streamline is not terminated in the voting process, the subsequent streamline direction $v$ is calculated as the normalized sum of the proposal directions: $v = \frac{\sum j \nu^j}{\|\sum j \nu^j\|}$. Analogous to the rotation of the fiber

![Fig. 1. Illustration of the voting process leading to a termination after the next step (a) or to a streamline deflection (b). The current streamline position is denoted $p^0$. $p^j$ and $\hat{p}^j$ are the sampling points and corresponding alternative sampling points, $\nu^j$ is the direction proposal at sampling point $p^j$ and $v_{old}$ is the direction of the previous streamline progression step.](image-url)
directions that are used as input features for the classifier (see Section 2.1), the output direction $v$ is multiplied with the original image rotation matrix to obtain alignment with the image.

2.3. Experiments

We performed four types of experiments to evaluate our approach.

2.3.1. Experiment 1

To determine the optimal choice of the tractography algorithm used to create the training data for our approach and to obtain an initial evaluation of its performance, we used a simulated replication of the FiberCap phantom (see Fig. 2) (Fillard et al., 2011). The dataset was simulated using the Fiber simulation tool (Neher et al., 2014) with the following parameters: 30 gradient directions, a b-value $1000 \text{ s mm}^{-2}$, $3 \text{ mm}$ isotropic voxels and a signal-to-noise ratio of about 40. The motivation for this choice of parameters was to create a challenging dataset that is also comparable to a clinical acquisition in terms of acquisition parameters. Tractography was performed with 12 combinations of the following openly available tractography and local modeling techniques (see Table 1 for corresponding toolkits):

- Tractography algorithms:
  - Deterministic streamline tractography (DET)
  - Fiber assignment by continuous tracking (FACT)
  - Tensor deflection tractography (TEND)
  - Probabilistic streamline tractography (PROB)
  - Global Gibbs tractography
- Local modeling techniques:
  - Single-Tensor model (DT)
  - Two-Tensor model (DT-2)
  - Constrained spherical deconvolution (CSD)
  - Constant solid angle Q-ball (CSA)

The presented approach was trained individually with each of the 12 benchmark methods. Based on the results of this analysis, the most promising algorithm was chosen to obtain the training tractogram for all further experiments ($A_{\text{train}}$).

To quantify the performance of the methods, the following metrics from the Tractometer evaluation protocol (Côté et al., 2013) were analyzed: the fraction of no connections (NC), valid connections (VC), and were not optimized.

2.3.2. Experiment 2

The in vivo performance of our approach in comparison to the 12 methods described in Experiment 1 was qualitatively evaluated on basis of reconstructions of the corticospinal tract (CST) and by an analysis of the spatial distribution of fiber end points. The dataset was acquired using 81 gradient directions, a b-value $3000 \text{ s mm}^{-2}$ and $2.5 \text{ mm}$ isotropic voxels. All methods were run with their default parameterization, which is the same as in Experiment 1. In contrast to Experiment 1, no manual adjustment of FA and ODF peak thresholds was necessary in vivo. As in Experiment 1, the interpolated raw signal values were used as input features for the classifier. To obtain the training reference for the in vivo dataset, we used the tractography method determined in Experiment 1 ($A_{\text{train}}$) with its default parameterization.

2.3.3. Experiment 3

In this experiment, we assessed the performance of the presented method using the ISMRM tractography challenge 2015 data (Maier-Hein et al., 2016) (www.tractometer.org/ismrm_2015.challenge/). The ground truth fiber bundles mimic the shape and complexity of 25 well known in vivo fiber bundles (see Fig. 3). The diffusion-weighted dataset was simulated with 32 gradient directions, a b-value $1000 \text{ s mm}^{-2}$ and $2 \text{ mm}$ isotropic voxels.

This experiment enabled us to compare our approach to all 96 original submissions of the tractography challenge comprising a large variety of tractography pipelines with different pre-processing, local reconstruction, tractography, and post-processing algorithms.

As preprocessing step, the dataset was denoised and corrected for distortions using MTrixx (dwi2noise & dwipreproc, http://www.mrtrix.org/).

The random forest classifier was trained using all combinations of the following parameters, resulting in a total of 8 trained classifiers:

- Classification features:
  1. Interpolated raw signal values
  2. Spherical harmonics coefficients (order 6)
- Additional features:
  1. No additional features
  2. T1 signal and GFA
- Training tractograms:
  1. Ground truth fibers ($GT_{\text{train}}$) used to simulate the phantom image
  2. Tractogram obtained using the method $A_{\text{train}}$

To reduce the computational load and since the parameters yielded stable results across a broad range, the maximum tree depth was reduced to 25 and the number of sampling points $N$ during tracking was reduced.
to 30. To obtain balanced classes, the number of non-fiber samples was chosen automatically to match the number of fiber samples.

The tractography process is influenced by several components, namely the classifier, the neighborhood sampling and the voting strategy, as well as the direction proposal. To disentangle the effects of the individual components, tractography was performed with all combinations of the 8 trained classifiers and the different options for the sampling strategy and actual tractography:

- **Sampling strategies:**
  1. No neighborhood sampling. Only the prediction of the classifier at the current streamline position is taken into account. This enables to discern the contributions of the classifier vs. the contribution of the neighborhood sampling and voting mechanism.
  2. Hemispherical sampling with majority voting (see 2.2, voting strategy (2)).
  3. Full spherical sampling (see 2.2, voting strategy (1)).

- **Direction proposal type (see Section 2.2):**
  1. Deterministic
  2. Probabilistic

Furthermore, we repeated the same experiments with classical direction proposal strategies instead of the random forest classification:

1. The voxel-wise peak directions calculated using CSD (deterministic tractography only) and (2) the voxel-wise CSD fODFs (deterministic and probabilistic tractography). CSD was performed using MRtrix and tractography using MITK. As termination criterion for these experiments we used a peak magnitude and CSA-Q-ball GFA threshold of 0.1 respectively (Aganj et al., 2009). Comparing the classification based approach to these classical approaches enables a more in-depth analysis of the contribution of the random forest classifier to the results. All experiments were run with 1 and 3 seed points per voxel. The step size and sampling distance of 0.5 \( f \), \( r = 0.25f \) was kept at the same values as in Experiment 1 and 2.

The same evaluation metrics as presented in the original challenge were analyzed for all 16 tractograms using the official challenge evaluation pipeline: valid bundles (VB), invalid bundles (IB), valid connections (VC), bundle overlap (OL) and bundle overreach (OR).

2.3.4. **Experiment 4**

Our final experiments aim at evaluating the generalization capability of the presented method using in vivo and in silico data:

As in vivo data we used five datasets of the Human Connectome Project (HCP) (Van Essen et al., 2013, 2012; Van Essen and Ugurbil, 2012) for training and five other HCP datasets for testing (HCP\textsubscript{train} [Subject IDs: 984472, 979984, 978578, 994273, 987983] and HCP\textsubscript{test} [Subject IDs: 992774, 991267, 983773, 965771, 965367]). The HCP datasets are acquired using 270 gradient directions, three b-values (1000 s mm\(^{-2}\), 2000 s mm\(^{-2}\), 3000 s mm\(^{-2}\)) and 1.25 mm isotropic voxels.

As in silico data for this experiment, we employed the ISMRM tractography challenge phantom already used in Experiment 3. Since the phantom and the HCP datasets were acquired with different imaging sequences, they feature different image contrasts. Therefore, the phantom dataset was normalized to feature the same signal mean and standard deviation across all weighted volumes inside the white matter as the HCP datasets. To ensure comparability with the results obtained on the in silico data, which only features a b-value of \( b = 1000 \text{ s mm}^{-2} \), we also limited the following in vivo experiments to a b-value of 1000 s mm\(^{-2}\) for training and tractography.

Three types of generalization were tested:

1. In vivo → in vivo: In this part of the experiment, we trained our method using HCP\textsubscript{train} and evaluated its performance on the unseen datasets of HCP\textsubscript{test}. Training was performed using multi-tissue CSD (Jeurissen et al., 2014) deterministic tractography as \( A_{\text{train}} \), about 12 million samples per dataset and spherical harmonics coefficients as classification features. Since the HCP datasets feature a much higher resolution compared to the datasets used in the other experiments, tractography on HCP\textsubscript{test} was performed using an increased sampling distance (0.7f) and step size (1.0f). Streamlines were seeded five times in every brain voxel and proposals were generated deterministically using the hemispherical sampling scheme. Evaluation was performed qualitatively by manually extracting the corticospinal tract (CST), cingulum (Cg) and fornix (Fx) from each of the five test results and a successive visual inspection of the tracts.
3. Results

3.1. Experiment 1

The best results on the phantom image were obtained using the CSD DET tractography parameterization (Tournier et al., 2012, 2007) for training our approach ($A_{\text{train}} =$ CSD DET). With this configuration, the presented approach outperformed all benchmark methods in four out of the five metrics (see Table 1). Only 3% of the tracts terminated prematurely. Furthermore, the presented approach yielded the highest percentage of valid connections (93%), the highest bundle overlap (94%) and the lowest local angular error (4%). All 7 valid bundles in the phantom were reconstructed successfully. Also, the percentage of invalid connections (4%) is rather low compared to the majority of benchmark algorithms (rank 4 out of 13). When varying the method that was used for training, the percentage of prematurely ending fibers and valid connections yielded by the presented approach improved on average by 56% and 36% respectively as compared to the benchmark tractograms. The average percentage of invalid connections, however, was increased by 21%.

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>NC</th>
<th>VC</th>
<th>IC</th>
<th>OL</th>
<th>AE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT$^2$</td>
<td>DET$^1$</td>
<td>60%</td>
<td>25%</td>
<td>15%</td>
<td>21%</td>
<td>5'</td>
</tr>
<tr>
<td>DT$^3$</td>
<td>FACT$^1$</td>
<td>62%</td>
<td>23%</td>
<td>14%</td>
<td>24%</td>
<td>6'</td>
</tr>
<tr>
<td>DT$^4$</td>
<td>TEND$^1$</td>
<td>84%</td>
<td>8%</td>
<td>8%</td>
<td>23%</td>
<td>10'</td>
</tr>
<tr>
<td>DT$^5$</td>
<td>PROB$^2$</td>
<td>57%</td>
<td>23%</td>
<td>20%</td>
<td>27%</td>
<td>8'</td>
</tr>
<tr>
<td>DT$^6$</td>
<td>Global$^1$</td>
<td>82%</td>
<td>10%</td>
<td>8%</td>
<td>42%</td>
<td>12'</td>
</tr>
<tr>
<td>CSA$^1$</td>
<td>DET$^2$</td>
<td>24%</td>
<td>67%</td>
<td>9%</td>
<td>70%</td>
<td>8'</td>
</tr>
<tr>
<td>CSA$^2$</td>
<td>PROB$^3$</td>
<td>91%</td>
<td>5%</td>
<td>4%</td>
<td>83%</td>
<td>18'</td>
</tr>
<tr>
<td>CSA$^3$</td>
<td>Global$^1$</td>
<td>81%</td>
<td>14%</td>
<td>5%</td>
<td>74%</td>
<td>13'</td>
</tr>
<tr>
<td>DT$^2+2$</td>
<td>DET$^1$</td>
<td>60%</td>
<td>37%</td>
<td>3%</td>
<td>58%</td>
<td>6'</td>
</tr>
<tr>
<td>CSD$^3$</td>
<td>DET$^1$</td>
<td>21%</td>
<td>78%</td>
<td>1%</td>
<td>86%</td>
<td>4'</td>
</tr>
<tr>
<td>CSD$^2$</td>
<td>PROB$^3$</td>
<td>66%</td>
<td>28%</td>
<td>7%</td>
<td>93%</td>
<td>4'</td>
</tr>
<tr>
<td>CSD$^3$</td>
<td>Global$^1$</td>
<td>81%</td>
<td>17%</td>
<td>2%</td>
<td>72%</td>
<td>12'</td>
</tr>
<tr>
<td>Proposed</td>
<td>3%</td>
<td>93%</td>
<td>4%</td>
<td>94%</td>
<td>4'</td>
<td></td>
</tr>
</tbody>
</table>

3.2. Experiment 2

Based on the results of Experiment 1, the CSD DET tractography was used as training method $A_{\text{train}}$ for the in vivo experiments. In vivo, our approach successfully reconstructed a whole brain tractogram including challenging regions such as the crossing between the corpus callosum, the CST and the superior longitudinal fasciculus. Our method was furthermore able to reconstruct parts of the CST that other approaches often missed (e.g. lateral projections of the CST in Fig. 4b). In comparison to the benchmark algorithms, most of the fibers reconstructed by the presented approach correctly terminated in the cortex (see Fig. 4a).

3.3. Experiment 3

The following paragraphs describe the performance of the presented method with respect to its individual components and elucidate specific aspects of the results. We focus on the performance of the deterministic proposal generation in conjunction with training the classifier using the classical CSD streamline tractogram $A_{\text{train}}$, if not stated otherwise. The results of the probabilistic proposal generation and the training using the ground truth tracts $GT_{\text{train}}$ are analyzed separately in the respective paragraphs. All results were tested for significance using a significance level of $a = 0.05$ (Bonferroni corrected $a = 4.47 \times 10^{-11}$), if not stated otherwise.

General observations and comparison to the original challenge submissions: The presented approach is the only method that was able to reconstruct all 25 valid bundles. Interestingly, it was able to outperform its training tractogram $A_{\text{train}}$ with respect to most metrics (see Fig. 5). Compared to the mean scores over all 96 benchmark submissions, the proposed approach performed better in terms of overlap (+19.9%), valid bundles (+2.9), valid connections (+7.5%) and invalid bundles (-0.7), albeit the latter two improvements were not statistically significant. The overreach on the other hand was increased by 11%. Significance tests were performed with a Mann-Whitney $U$-test.

Table 2 shows how many of the originally submitting teams could be outperformed with respect to the team-mean of the respective score (not tested for significance due to a partially very small number of submissions per team). The scores show that our approach consistently outperformed more than half of the challenge teams. In case of the VB score, all 20 teams could be outperformed.

Classification vs. CSD based tractography: This paragraph presents the results of the proposed classification based approach compared to classical CSD based proposal generation. On average over all sampling schemes, the classification based approach outperformed classical sampling with respect to the valid connections (+16.1%) and the valid bundles (+1.1). Significance tests were performed with a Mann-Whitney $U$-test. Statistically not significant changes were observed for the bundle overlap (+3.6%), the overreach (+3.4%) and the number of invalid bundles (+7). The remaining results of this paragraph were not tested for significance due to the low number of samples in the individual groups.

Table 3 shows the score differences between the two approaches, individually calculated for each sampling scheme to disentangle the effects of the respective component. For all schemes, valid bundles, valid connections and bundle overlap were improved using the classification based approach. On the other hand, the number of invalid bundles and the overreach were increased. While out of the classical methods, the peak based approach performed best in terms of valid connections (58%) and valid bundles (23.5), it was still outperformed by the classification based approach (61.1% and 24.25 respectively).

Neighborhood sampling: Surprisingly, the neighborhood sampling only had a very little effect on the scores. Using hemispherical sampling and voting, the only significant score change was an increased percentage of valid connections (+0.7%). Using full spherical sampling, the valid connections score was decreased by 2.6% while overlap and overreach were increased by 2.3% and 3.5% respectively. Significance tests were performed with a paired $t$-test.

Probabilistic vs. deterministic proposal generation: To compare the results between the probabilistic and deterministic proposal generation, only the pipelines without neighborhood sampling were analyzed to isolate the effect of the probabilistic sampling. Furthermore, using neighborhood sampling in conjunction with probabilistic sampling is somewhat counterproductive, since the sampling process generates a weighted average of the proposals, which counteracts the probabilistic nature of the proposal generation. The deterministic approach yielded a much higher valid connections rate (+14.6%), while the overreach was
lower using the probabilistic approach (−4.4%). Significance tests were performed with a paired t-test. The approaches showed no significant differences in the other scores.

Training data: In this paragraph, we analyzed the effect of the used training data ($A_{\text{train}}$ or $GT_{\text{train}}$). As expected, the presented method performs clearly better when using the $GT_{\text{train}}$ tracts as training data as compared to the training tractogram obtained with $A_{\text{train}}$ (see Fig. 5).

Especially the valid connections (−14.5%), bundle overlap (−15.7%) and bundle overreach (−5.7%) were distinctly improved using $GT_{\text{train}}$. Significance tests were performed with a paired t-test.

Effect of the different diffusion-weighted features: Using spherical harmonics coefficients instead of the raw signal values as classification features resulted in consistently but only slightly higher valid connection rates (−2.3%) and a higher overlap (−3%). Significance tests were
performed with a paired $t$-test.

Effect of additional features: The differences are very small and consistent effects could only be detected when training on the ground truth, where inclusion of T1 and GFA features lead to a slight increase of valid connections (+1.3%), a higher overlap (+0.7%) and a lower overreach (-2.6%). Significance tests were performed with a paired $t$-test.

Effect of the varying number of seed points: Using a larger number of seed points resulted in a higher bundle overlap (+6.6%), while at the same time increasing the overreach (+5.2%) and the number of invalid bundles (+9.5). Significance tests were performed with a paired $t$-test.

### 3.4. Experiment 4

Overall, the presented approach was able to generalize well to unseen datasets. The following paragraphs describe the individual results of the three parts of this experiment.

1. **In vivo → in vivo**: Fig. 6 shows the tracts extracted from the five HCP test tractograms. All tracts were reconstructed successfully, which demonstrates that the classification based approach is capable of generalization to unseen datasets.

2. **In vivo → in silico**: The method trained on five HCP datasets was able to reconstruct 23/25 valid bundles in the ISMRM tractography challenge phantom, while reconstructing 94 invalid bundles. Further scores were a fraction of valid connections of 52%, a bundle overlap of 59% and a bundle overreach of 37%, which is comparable to the results of Experiment 3, where the approach was directly trained on the phantom dataset.

3. **In silico → in vivo**: As in (1), the approach trained on simulated data was able to reconstruct the 3 tracts of interest in all five unseen in vivo test datasets (see Fig. 7).

**Fig. 5.** Scores of the 16 new tractograms obtained using the presented method (color) in comparison to the original challenge submissions (gray). The tractograms in Team 21 (green x) were obtained using $A_{\text{train}}$ to generate the training reference and the tractograms in Team 22 (blue, +) with the ground truth tracts $G_{\text{test}}$ as training reference. The scores of training tractogram $A_{\text{test}}$ itself are shown for comparison (red, star). For reasons of clarity, only the results obtained with the frontal neighborhood sampling scheme are shown.

**Fig. 8.** shows a comparison between the results of classical CSD tractography (MRtrix) and the results of Experiment 4 (3). The classical streamline method had more difficulties to reconstruct the three tracts to the full extent and struggled with prematurely ending fibers.

### 4. Discussion and conclusion

We presented a random-forest classification-based approach to fiber tractography using neighborhood information that guides each step of the streamline progression. The presented approach is the first to utilize machine learning for fiber tractography. The method systematically exploits the diffusion-weighted signal not only locally but also in the neighborhood of the current streamline position.

We thoroughly evaluated the performance of the presented method in comparison to over 100 state-of-the-art tractography pipelines on simulated phantom datasets as well as in vivo.

In the in vivo experiments (Experiment 2), our approach yielded very good results in reconstructing difficult tracts (e.g. the lateral projections of the CST) and a much lower number of fibers ending prematurely inside the brain. As expected, tensor based approaches had difficulties in detecting the lateral projections of the CST. However, even the benchmark method that performed best in the phantom experiments ($A_{\text{train}}$ = CSD DET) was unable to detect these projection fibers. The benchmark methods that showed a relatively high sensitivity in this region (CSA + CSD PROB and DT-2 DET) displayed a very low specificity in the phantom experiments (Experiment 1) as well as with respect to the in vivo end-point distribution. In contrast, the presented algorithm showed a constantly high sensitivity and specificity.

The quantitative analysis on the ISMRM tractography challenge phantom dataset (Experiment 3) yielded various valuable insights into the

<table>
<thead>
<tr>
<th>Score</th>
<th>VB</th>
<th>IB</th>
<th>VC</th>
<th>OL</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No neighborhood sampling</td>
<td>+0.9</td>
<td>+8.8</td>
<td>+21%</td>
<td>+3%</td>
<td>+4%</td>
</tr>
<tr>
<td>Frontal neighborhood sampling</td>
<td>+1.1</td>
<td>+6.6</td>
<td>+15%</td>
<td>+3%</td>
<td>+2%</td>
</tr>
<tr>
<td>Full neighborhood sampling</td>
<td>+1.2</td>
<td>+5.7</td>
<td>+12%</td>
<td>+4%</td>
<td>+4%</td>
</tr>
</tbody>
</table>
performance of our approach and its individual components.

One interesting aspect that catches the eye is the fact that the presented approach outperforms its respective training tractogram $A_{\text{train}}$ in many aspects, even without the use of neighborhood sampling. Why the classifier is able to outperform the training tractogram is hard to pin down exactly, but there are a number of factors that contribute to this behavior: (1) The classifier can directly relate the previous progression direction to the local signal, which is a clear advantage over $A_{\text{train}}$. (2) During training, the classifier sees a lot of data from various situations and can rely on this accumulated knowledge when encountering new situations. While the signal the classifier sees at the moment is the basis of its decision, it did not learn by heart what to do when encountering this exact signal configuration. The method generalizes very well, which has a regularizing effect. This also enables the method to work well on previously unseen data. (3) The neighborhood sampling amplifies these advantages by enabling the resolution of asymmetric fiber configurations.

Fig. 6. Cingulum (left), corticospinal tract (middle) and fornix (right) reconstructed from the five HCP test subjects (see Section 2.3 Experiment 4 (1)). The results were obtained with the presented approach after training on five different HCP subjects.
and a better informed decision about whether to terminate or proceed with the streamline progression. In general, the presented approach yields one of the best performances in terms of valid connections and valid bindles, while keeping a high overlap/overreach ratio.

Our evaluation further showed that the good performance of the presented approach is mainly attributable to the random forest classifier and only to a very small degree to the neighborhood sampling process, particularly in Experiment 3. The reasons for the good performance of the classifier are already discussed in the previous paragraph. Our results show that the neighborhood sampling mainly affects the distribution of fiber endpoints, but does not necessarily increase the number of valid connections. While the effect of the neighborhood sampling was only analyzed systematically in case of the more complex 3D phantom used in Experiment 3, it is likely that the effect is stronger in case of the much simpler 2D phantom (Experiment 1). The sampling basically keeps the fibers alive as long as possible, which is more likely to have a positive

Fig. 7. Cingulum (left), corticospinal tract (middle) and fornix (right) reconstructed from the five HCP test subjects (see Section 2.3, Experiment 4 (3)). The results were obtained with the presented approach after training on the ISMRM tractography challenge phantom.
image contrast and datasets to obtain more comprehensive training data with respect to datasets. One important aspects in this regard is the simulation of further challenges to address. Further work is necessary to quantify and improve the effect in case of a simple phantom with well-defined and easier to recognize bundle endpoint regions. Unsurprisingly, the probabilistic proposal generation yielded much lower valid connection scores. Nevertheless, this approach may be useful in conjunction with higher order integration methods or approaches that exploit and quantify the local uncertainty, e.g. to yield probabilistic connectivity maps.

The capability of the presented approach to generalize to unseen datasets was successfully demonstrated in Experiment 4. We showed that the generalization is possible between different in vivo images acquired with the same MR sequence as well as between in vivo images and simulated datasets (in both directions). The latter experiment, generalizing from simulated datasets to in vivo, is especially important, since it enables us to train on datasets with known ground truth. This eliminates the bias introduced by training on tracts obtained with a conventional tractography method. We could also show that we can yield better results, in terms of tract-completeness, with this approach as compared to the classical approach used to generate the training data, albeit only qualitatively. Nevertheless, the clearly observable differences in the tractograms obtained in Experiment 4 (1) and (3) (compare Figs. 6 and 7) emphasize the impact of the training data on the resulting tractogram, this is an expected phenomenon. Furthermore, the phantom and the in vivo datasets have quite different structural properties, e.g. the phantom contains much more non-fiber regions since it only includes 25 major fiber bundles.

While the presented results are promising, there are still some challenges to address. Further work is necessary to quantify and improve the performance of the presented approach when training on simulated datasets. One important aspects in this regard is the simulation of further datasets to obtain more comprehensive training data with respect to image contrast and fiber structure. Another aspect where methodological improvements are definitely possible is the currently employed naïve approach to generalize between the simulated and in vivo domain, using approaches of unsupervised domain adaptation and transfer learning (Götz et al., 2014; Heimann et al., 2014; Long et al., 2016, 2014; McKeough et al., 2013; Pan and Yang, 2010; Sener et al., 2016). This also includes the possibility to generalize to images acquired with different settings such as b-value and number of shells. Initial experiments already showed a relatively good generalizability between b-values using only a simple image contrast normalization. In the context of multiple b-shells, it would also be interesting to evaluate feature sets other than spherical harmonics coefficients that are more suitable for multiple shells. Another challenge is the still high number of invalid bundles, which is a known issue of current fiber tractography approaches (Maier-Hein et al., 2016). This is a challenge the whole fiber tractography community is facing and that does not have a simple solution. However, it seems promising to incorporate as much additional knowledge in the tractography process as possible, for which machine learning based approaches seem to be well suited. Interesting candidates would be functional MRI data or prior knowledge in form of cortical parcellations. An extension of the presented method to directly include a distinction between different non-white matter tissue types, such as gray matter and corticospinal fluid, seems promising to further improve the decision on where to terminate the fiber progression. This includes addressing further interesting aspects of the fiber termination such as orthogonality to and uniform coverage of the gray-white-matter interface. We are also planning to analyze how the process of including neighborhood information can be improved further, e.g. by applying patch-based classification at the sampling positions in order to include more context information into the individual proposals. Before implementing the voting process presented in this work, we discussed the possibility to jointly process of the feature vectors of all sampling positions using one classifier. We ultimately decided against this for one main reason. With the presented approach we have a better control about how the neighborhood information is incorporated into the decision process. It would be difficult to provide suitable training data that could teach the classifier our intended behavior of avoiding prematurely ending streamlines. Nevertheless, a combination of both approaches, patch-based classification and sample voting, seems promising and will be investigated in the future. In this context we are also testing other variations of the sampling and voting strategy to improve that could improve the results. Another issue that we are currently investigating is the handling of image rotations that are not accounted for in the image rotation matrix. We are currently looking into the possibility of augmenting the training data with random rotations. This would enable the classifier to handle images with arbitrary rotations equally well, but it would also increase the training effort. A more straightforward alternative is the alignment of the respective rotated test image with the training data using a standard rigid registration technique.

The source-code of all methods presented in this work is available...
open-source and integrated into the Medical Imaging Interaction Toolkit (MITK) (Fritzsche et al., 2012; Nolden et al., 2013). The datasets used in Experiment 1 and 2 are available for download at www.nitrc.org/projects/diffusion-data/. The dataset used in Experiment 3 as well as many other resources regarding the ISMRM tractography challenge are available at www.tractometer.org/ismrm_2015_challenge/. The HCP datasets used in Experiment 4 are available at www.humanconnectome.org/data/.

Acknowledgments

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 11 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. This work was supported by the Collaborative Research Center (SFB/TMR 125 Cognition-Guided Surgery) of the German Research Foundation (DFG) grant number INST 35/1120-1, DFG grant MA 6340/10-1 and DFG grant MA 6340/12-1.

References


P.F. Neher et al.


