COMET – A framework for the large scale Cluster analysis Of Major Equivalent Tracts

C. Ros[1], D. Güllmar[1], and J. R. Reichenbach[1]

[1]Medical Physics Group, Department of Diagnostic and Interventional Radiology I, Jena University Hospital, Jena, Thuringia, Germany

Introduction - Fiber tractography is an exciting tool to study the inherent complexity of white matter structures in the brain. Due to the steady technological progress, both deterministic and probabilistic whole brain tractography (WBT) approaches became feasible. WBT reconstructs a multitude of fiber tracts. In typical data sets the number of fiber tracts can easily exceed several thousand up to more than a million of fiber tracts. Due to this wealth of information, WBT has the potential to recover unknown paths that may remain undetected with pure ROI-based tractography methods. To analyze neural connectivity or structural changes in underlying white matter fiber bundles, quantitative tractography-based analysis is applicable [1-4], as long as the reconstructed fiber tracts are bundled together in a way that they represent the microscopic axonal pathways [5] correctly. Instead of manual processing (e.g., ROI drawing strategies [3]), which is prone to errors and highly time-consuming, machine learning methods such as cluster analysis (CA) are auspicious techniques for bundling of fiber tracts. Even though huge progress has been made, fully automated, fast and unsupervised CA of axonal pathways with high quality still poses a problem. Especially, if large data sets are employed, CA methods are not able to handle the huge amount of raw data and therefore limit the application severely and may even hinder quantitative tract-based analysis. To cope with such large data sets, we propose COMET – a new framework that is capable to perform autonomous, high quality Cluster analysis Of Major Equivalent Tracts (COMET), even in large data sets. The framework facilitates exploratory data analysis of fiber tracts and enables the exploration of axonal pathways and connections, as well as quantitative analysis in large groups of subjects.

COMET – Cluster analysis Of Major Equivalent Tracts

The COMET framework focuses on various aspects of cluster analysis and takes several considerations for the processing of large data sets into account. To asses the similarity of fiber tracts, COMET contains a variety of cluster analysis methods and facilitates the use of a multitude of proximity measures (e.g. Hausdorff distance, minimum distance, combined distance measures, etc.) that can either be used independently or in conjunction with other proximity measures to perform clustering in higher dimensions [6, 7]. To perform efficient as well as effective clustering of large tractography data sets, COMET introduces CATSER for Cluster Analysis Through Smart Extracted Representatives. CATSER employs basically agglomerative hierarchical clustering based on the CURE technique [8]. First, Chernoff Bounds [8] are employed to obtain a sample of tracts, before Local Outlier Factors (LOFs) [9] are extracted for the sample. LOFs are an estimate on how isolated tracts are with respect to their surrounding neighbourhood and are used in CATSER to adjust the distances between tracts/clusters as well as to minimize the effect of outlying fiber tracts. In the next step, CA is performed for the sample by starting with a disjoint set of clusters, where every tract is a separate cluster. In the following CA iterations, most similar clusters are merged successively. Similarity between two clusters is defined as the similarity between the most similar representative tracts of each cluster. Due to unique LOF properties, the influence of outlying fiber tracts is effectively reduced and robust representative tracts are extracted. Finally, prototype clusters are extracted and remaining tracts that are not part of the sample are assigned to the nearest prototype cluster. To facilitate fast processing of large data sets, COMET was implemented in C++ for 64-bit Linux. It heavily relies on symmetric-multi-processing architectures with multiple CPU cores. By employing parallel processing, computation time is reduced due to the acceleration of computationally demanding processing stages.

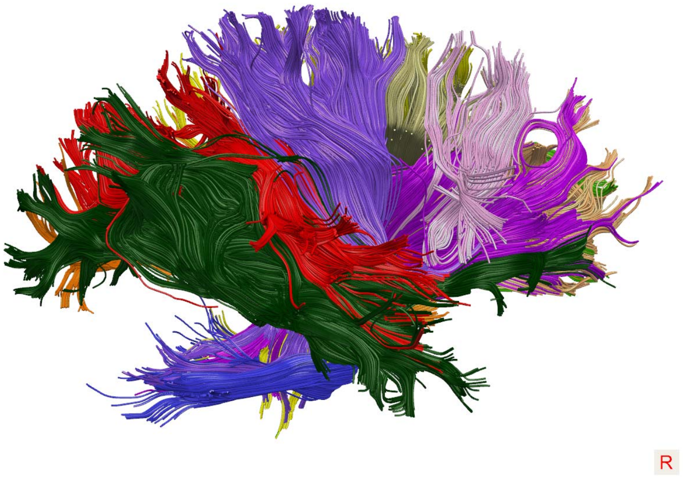

Fig. 1 Clustering of fiber tracts with the COMET framework. More than 900 000 tracts were processed with the CATSER cluster analysis method by employing three proximity measures. Processing time was ~45 minutes.

Materials and Methods - DTI data sets of a healthy volunteer were acquired on a clinical 3 T whole body MR-Scanner (Magnetom Tim Trio, Siemens Healthcare, Erlangen, Germany), using a conventional twice refocused Echo Planar Imaging (EPI) sequence [10]. A 12 channel phased array matrix head coil was employed and the following parameters were used: T_E=113 ms, T_R=7900 ms, α = 90°, iPAT=2, matrix of 96×96, 55 slices with a thickness of 2.5 mm, resulting in a voxel size of 2.5×2.5×2.5 mm³. Five b_0 images without diffusion weighting as well as 70 diffusion weighted images sampled with different gradient directions at b=1000 s/mm² were acquired. In-plane interpolation was performed on the MR-Scanner, resulting in a nominal voxel size of 1.25×1.25×2.5 mm³. The Diffusion Toolkit [11] was used to perform whole brain fiber tractography and to reconstruct 1.8 million fibers. Tracts with lengths less than 30 mm were subsequently removed from the data set. Cluster analysis was finally performed for the remaining >900 000 fiber tracts with the COMET framework by using CATSER as well as three different proximity measures (orientation similarity, hemispheric affiliation of tracts [6, 7] and the Hausdorff distance).

Results - In Fig. 1 the 20 largest clusters are shown with different colors. By using a multiprocessing system with 32 cores and sampling size of 10 000 tracts, the computation time was around ~45 minutes for the whole data set. The outlier elimination in CATSER successfully identified and removed spurious tracts. The resulting clusters are clearly separated, easy to discriminate and not corrupted by spurious tracts.

Discussion & Conclusion - By exploiting the capabilities of modern multiprocessor system and using new clustering techniques as well as a variety of similarity measures, our toolkit is able to cluster large data sets on the order of minutes. To further speed up cluster analysis, alternative architectures for massive parallel data processing e.g., Graphics Processing Units (GPUs), can be employed [12]. The toolkit offers great flexibility due to various clustering methods and multiple proximity measures. New proximity measures can easily be implemented, while maintaining the ability to parallelize the computations. The framework can either be used to study single subjects (e.g., tumor patients) as well as large groups of subjects. Due to fast processing, data exploration in multi-subject imaging studies becomes feasible and may reveal useful information about axonal pathways and connectivity. By using cluster analysis, brain connectivity quantification can be employed to obtain plenty of information and to study neuropsychiatric disorders, such as schizophrenia or obsessive compulsive disorder.

Acknowledgements - This study was supported by the German Federal Ministry of Education and Research (BMBF), project number: 01GW0740. We like to thank Siemens Healthcare, Erlangen, Germany for their support.

References

[1] Berman et al, 2005, Neuroimage 27, 862–871 [2] Wakana et al, 2007, Neuroimage 36, 630–644 [3] Partridge et al, 2005, J Magn Reson Imaging 22, 467–474 [4] Voineskos et al, 2009, Neuroimage 45, 370–376 [5] Gray, 1973, Anatomy of the human body, 29th edition [6] Güllmar et al, 2008, Proc Intl Soc Mag Reson Med, 16 [7] Ros et al, 2010, Proc Intl Soc Mag Reson Med, 18 [8] Guha et al, 2001, Inf Syst 26, 35–58 [9] Breunig et al, 2000, SIGMOD Conf, ACM 93–104 [10] Heid, 2000, Proc Intl Soc Mag Reson Med, 8 [11] Wang et al, 2007, Proc Intl Soc Mag Reson Med 15, #3720 [12] Ros et al, 2011, Proc Intl Soc Mag Reson Med 19