Quality Assurance using Outlier Detection on an Automatic Segmentation Method for the Cerebellar Peduncles

Ke Li[*a], Chuyang Ye[b], Zhen Yang[a], Aaron Carass[a], Sarah H. Ying[c], and Jerry L. Prince[a]

[a]Dept. Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218
[b]Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190
[c]The Johns Hopkins School of Medicine, Baltimore, MD 21205

ABSTRACT

Cerebellar peduncles (CPs) are white matter tracts connecting the cerebellum to other brain regions. Automatic segmentation methods of the CPs have been proposed for studying their structure and function. Usually the performance of these methods is evaluated by comparing segmentation results with manual delineations (ground truth). However, when a segmentation method is run on new data (for which no ground truth exists) it is highly desirable to efficiently detect and assess algorithm failures so that these cases can be excluded from scientific analysis. In this work, two outlier detection methods aimed to assess the performance of an automatic CP segmentation algorithm are presented. The first one is a univariate non-parametric method using a box-whisker plot. We first categorize automatic segmentation results of a dataset of diffusion tensor imaging (DTI) scans from 48 subjects as either a success or a failure. We then design three groups of features from the image data of nine categorized failures for failure detection. Results show that most of these features can efficiently detect the true failures. The second method—supervised classification—was employed on a larger DTI dataset of 249 manually categorized subjects. Four classifiers—linear discriminant analysis (LDA), logistic regression (LR), support vector machine (SVM), and random forest classification (RFC)—were trained using the designed features and evaluated using a leave-one-out cross validation. Results show that the LR performs worst among the four classifiers and the other three perform comparably, which demonstrates the feasibility of automatically detecting segmentation failures using classification methods.

Keywords: quality assurance, segmentation, cerebellar peduncles, outlier detection, box-whisker plot, classification

1. INTRODUCTION

Cerebellar peduncles (CPs) are major white matter tracts that connect the cerebellum and other brain parts, including the cerebral cortex and the spinal cord[1]. They consist of the superior cerebellar peduncles (SCPs), the middle cerebellar peduncle (MCP), and the inferior cerebellar peduncles (ICPs). Automatic segmentation methods for CPs are necessary for studying their structures and functions objectively and efficiently. Fortunately, diffusion tensor imaging (DTI)[2] has made this achievable. However, while algorithms for automatically segmenting the cerebellar peduncles based on DTI have been proposed[3-8], only the method of Ye et al.[7] correctly segments the decussation of the SCPs (dSCP), the region where the SCPs cross. This method consists of a random forest classifier (RFC) and a multi-object geometric deformable model (MGDM). The random forest classifier uses features extracted from the DTI scans to provide an initial segmentation of the peduncles. MGDM is then used to refine the random forest classification, leading to smoother and more accurate segmentations. Results show that this method is able to resolve the dSCPs and accurately segments the other cerebellar peduncles as well.

Usually performance evaluation of automatic segmentation methods is conducted by comparing the segmentation results with manual delineations (ground truth). However, while this approach characterizes the performance in an average sense, when the method is run on new data (for which no ground truth exists) it is highly desirable to be able to assess

[*] kli26@jhu.edu; Image Analysis and Communications Laboratory, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at http://SPIE.org/manuscripts

Return to the Manage Active Submissions page at http://spie.org/submissions/tasks.aspx and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

algorithm failures so that these cases can be excluded from analysis or rerun with different parameters. Considering the large size of some datasets and the heavy workload of visual inspection, finding a way to automatically and accurately detect algorithm failures of these automatic segmentation methods is highly desirable.

In this work, we aim to better understand the performance of an automatic segmentation algorithm of CPs using two outlier detection methods. The first one is a simple univariate non-parametric method using box-whisker plots. To extract features for outlier detection, we first categorized the segmentation results of a dataset (48 subjects) as either a success or a failure. We then extracted three groups of features from image data of nine categorized failures in this dataset. Outliers were detected from these features using box-whisker plots and the features' performance was compared. The other method is the supervised classification. We first categorized the segmentation results of a larger dataset (249 subjects) as a success or a failure. We then trained four classifiers—linear discriminant analysis (LDA), logistic regression (LR), support vector machine (SVM), and random forest classification[9] (RFC)—using the extracted features on two training sets and evaluated their performance using a leave-one-out cross validation. We describe the two methods in the next section with a focus on the feature extraction from categorized failures.

2. METHODS

Automatic segmentation labels of two DTI datasets of whole heads were used in this work. The first dataset consists of 48 subjects: 18 controls and 30 patients with neurological diagnoses affecting the cerebellum. A larger dataset contains 249 subjects: 49 controls and 154 patients with different kinds of ataxia. Diffusion weighted images (DWIs) of the two datasets were acquired using a multi-slice, single-shot EPI sequence on a 3T MR scanner (Intera, Philips Medical Systems, Netherlands) on which we run the CP segmentation pipeline of Ye et al[7].

To automatically detect segmentation failures of this CP segmentation method, we need to study these failures first. A segmentation failure is defined as one that looks very different from the normal segmentations or does not have all the six labels. We manually categorized the segmentation results of the 48 subjects in the first dataset as either a success or a failure by a Principle Eigenvector (PEV) edge map and a linear Westin index[10] (computed from the diffusion tensor). The reason for using a linear Westin index is because it can show the tracts with a relatively high contrast. The crossing tracts have lower linear Westin index values while the noncrossing tracts have higher ones. Due to the imperfect quality of DTI scans and algorithm itself, nine failures were found among the 48 subjects. One reference and the nine failures are shown in Figure 1.

We then designed three groups of features of the image data from the nine failures. The first group of features is object oriented and characterizes the failures on the peduncle level. Volumes and surface areas of the six CPs are two features in this category, notated as $\boldsymbol{v} = [v_{lSCP}, v_{rSCP}, v_{dSCP}, v_{MCP}, v_{lICP}, v_{rICP}]$ and $\boldsymbol{s} = [s_{lSCP}, s_{rSCP}, s_{dSCP}, s_{MCP}, s_{lICP}, s_{rICP}]$. The second group of features is data quality oriented. We chose diffusion tensor related features including the means and standard deviations of the fractional anisotropy (FA), the mean diffusivity (MD), the linear Westin index C_l, the planer Westin index C_p, and the spherical Westin index C_s of the whole brain, notated as $\boldsymbol{FA} = [u_{FA}, std_{FA}]$, $\boldsymbol{MD} = [u_{MD}, std_{MD}]$, and $\boldsymbol{C} = \left[u_{C_l}, std_{C_l}, u_{C_p}, std_{C_p}, u_{C_s}, std_{C_s}\right]$. We found that dim or abnormal linear Westin indices were highly correlated with the failures. The third group of features is brain mask related. We found that abnormal brain masks can make the linear Westin index incomplete, which can cut out some structures of the CPs and lead to a segmentation failure. To detect these failures, volumes of the left cerebrum, right cerebrum, and whole brain mask and the symmetry of a brain mask were used, notated as $\boldsymbol{BM} = [v_{lBM}, v_{rBM}, v_{BM}, sym_{BM}]$, where $sym_{BM} = \left|\frac{v_{lBM}}{v_{rBM}} - 1\right|$. In summary, the final feature vector $\boldsymbol{f}$ to be used in the two outlier detection methods is a 26-dimensional vector composed of the volumes and surface areas of the six CPs, the means and standard deviations of the FA, the MD, and the three Westin indices, and the brain mask features, i.e., $\boldsymbol{f} = (\boldsymbol{v}, \boldsymbol{s}, \boldsymbol{FA}, \boldsymbol{MD}, \boldsymbol{C}, \boldsymbol{BM})$.

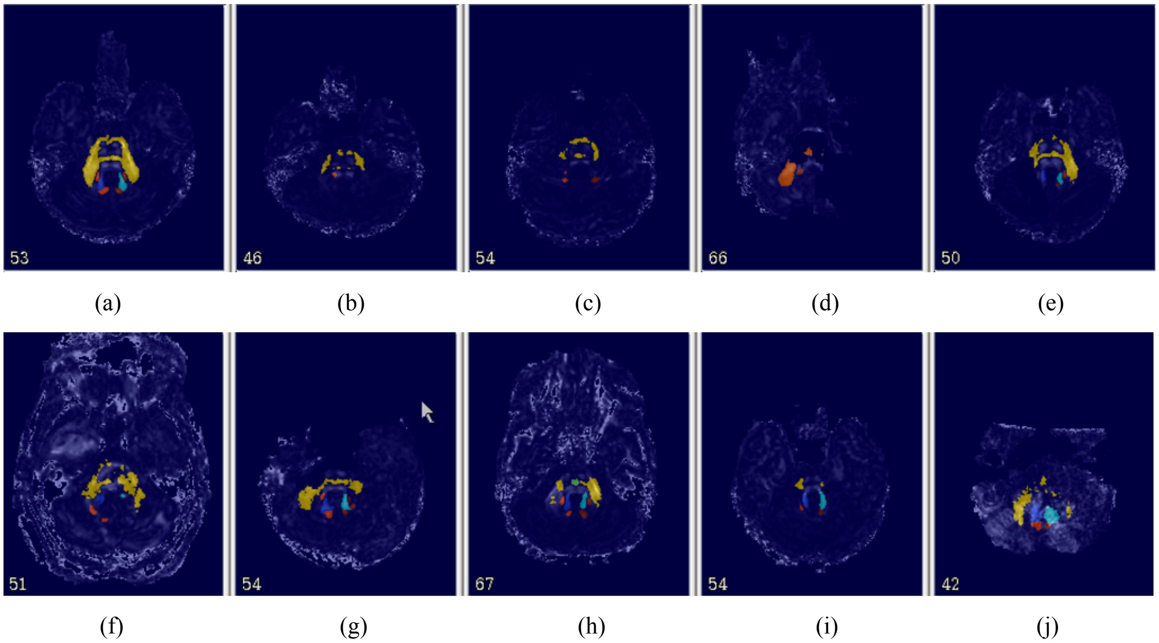

Figure 1. One reference and nine segmentation failures in the first data set with 48 subjects: (a) is a successful segmentation of a subject as a reference. (b)–(j) are nine segmentation failures from nine subjects.

With well-defined features above, we can apply them to outlier detection tasks. We take the definition of an outlier from Grubbs[11]—“An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.” Outliers in our numerical data of features were detected using a box-whisker plot. The bottom and top of the box are the first and third quartiles of the data. Between them is the Interquartile Range (IQR), namely 50% of the data. If a data point is 1.5×IQR or more above the third quartile, or 1.5×IQR or more below the first quartile, it is detected as an outlier. The notch in the boxplots displays a confidence interval around the median. If two boxes’ notches do not overlap, there is strong evidence (95% confidence) that their median differ. Outliers in the first dataset were detected using the box-whisker plots.

The other outlier detection method is the supervised classification. Four classifiers—LDA, LR, SVM, and RFC—were used on a larger dataset containing 249 subjects. To train these classifiers, we manually categorized the segmentation results of this dataset as either a success or a failure and found a total of 12 failures. Two training sets were used. One is very unbalanced which consists of the 12 failures and the 237 successes. The other contains the 12 failures and 24 randomly selected successes. Each classifier’s performance was evaluated using a leave-one-out cross validation.

3. RESULTS

Boxplots with outliers detected by some selected features in the first data set containing 48 subjects are shown in Figures 2–5. Volumes of the six CPs of the 48 subjects with diagnoses are shown in Figure 2. The 10 manual delineations are connected with the corresponding 10 automatic segmentations from the algorithm of Ye et al. by dashed lines. Means and standard deviations of the FA, the MD, and the three Westin indices of the whole brain are shown in Figures 3 and 4 without ground truth to compare (since we only have manual delineated segmentation labels, not other parameters). Brain mask features including the volumes of the right, left, and whole brain mask and the symmetry of the brain masks of the 48 subjects are shown in Figure 5.

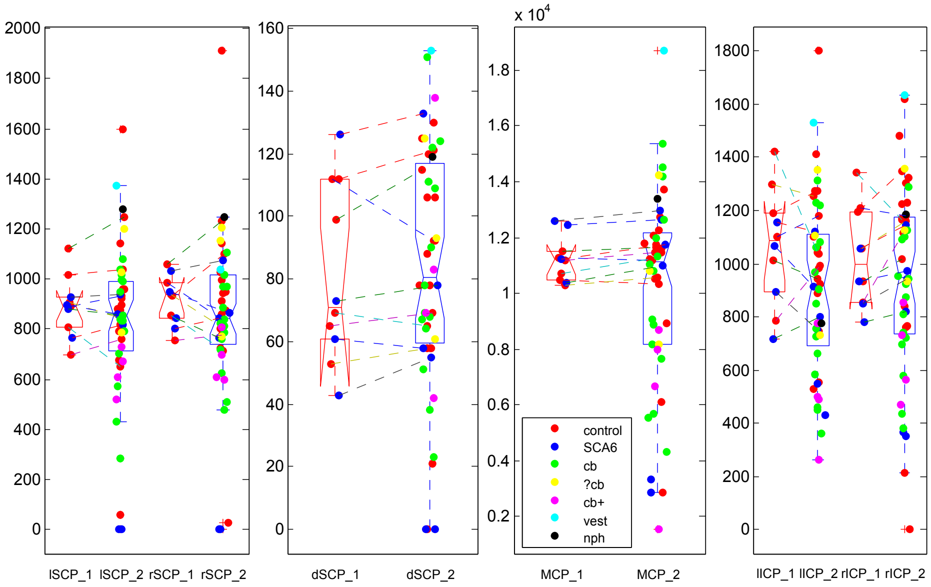

Figure 2. Volumes of six CPs of manual delineations of 10 subjects in the first dataset (red boxes) and automatic segmentations of the 48 subjects (blue boxes), respectively.

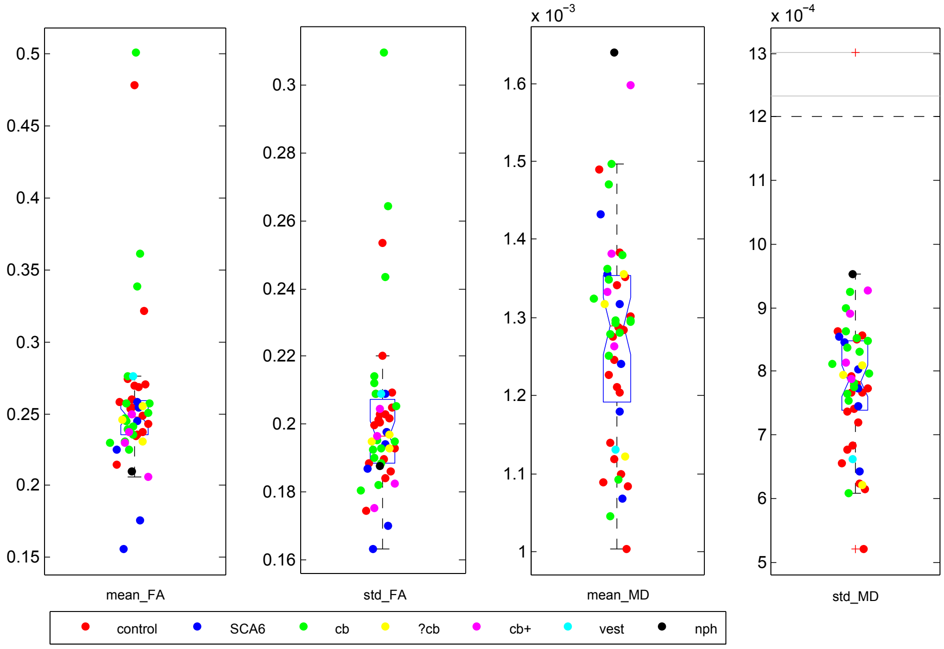

Figure 3. Means and standard deviations of FA and MD of the whole brains of the 48 subjects.

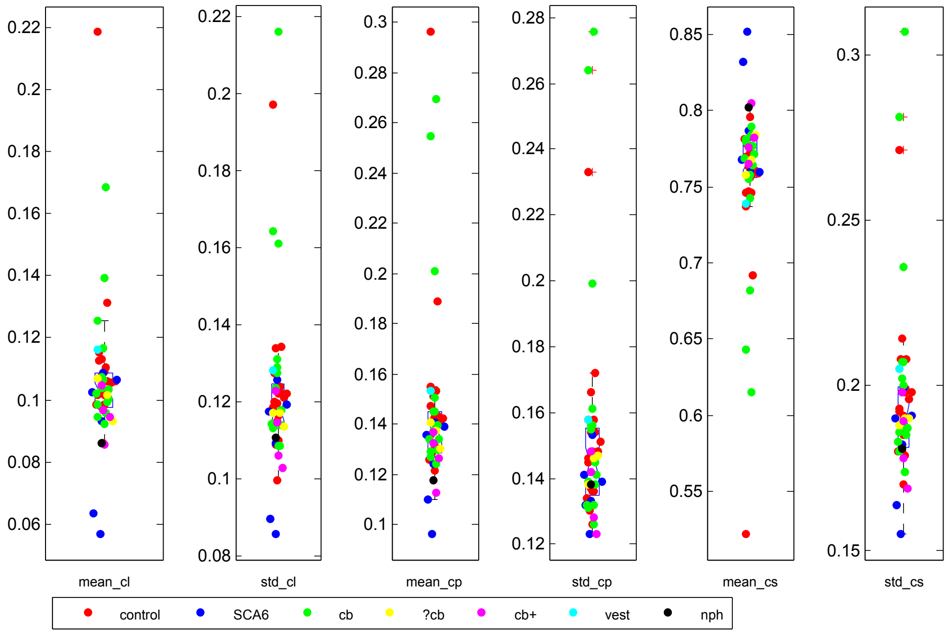

Figure 4. Means and standard deviations of the three Westin indices of the whole brains of the 48 subjects.

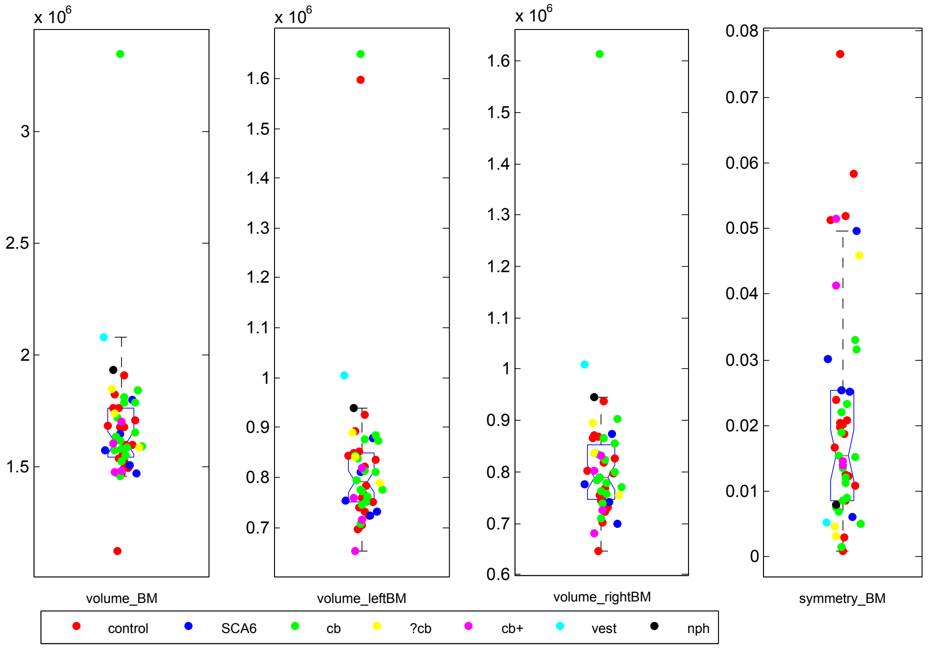

Figure 5. Brain mask features: volumes of the left, right, and whole brain masks (the left three boxplots) and the symmetry of the brain masks (the rightest boxplot) of the 48 subjects.

We evaluated the performance of our selected features in the task of finding segmentation failures. The detected outliers by these features and the categorized segmentation failures (ground truth) were compared. For each feature, we computed the true positive and false positive rates, as shown in Table 1. To note, “volume” and “surface area” in Table 1 are the volumes and surface areas of the six CPs. As long as one CP is detected as an outlier by its volume or surface area, this segmentation result with this CP is detected as an outlier (since we assume a successful segmentation should include all the six CPs).

Table 1. Performance of three groups of features on detecting outliers in the 48 subjects (nine failures).

Features	surface area	u_{FA}	u_{C_s}	volume	u_{C_l}	std_{C_l}	u_{C_p}	std_{FA}
# TP	7	7	7	6	6	6	6	4
# FP	2	0	0	2	0	0	0	0
Features	std_{C_p}	std_{C_s}	std_{MD}	v_{BM}	v_{lBM}	v_{rBM}	sym_{BM}	u_{MD}
# TP	4	4	2	2	2	2	2	0
# FP	0	0	0	0	1	0	3	2

The four classifiers were evaluated using a leave-one-out cross-validation. The misclassification rate (MCR), the numbers of true positive (TP) and false positive (FP), and the TP and FP rates of the four classifiers on the two training sets are shown in Table 2. In the first training set, the RFC performs best and the LR performs worst. The performances of the LDA and the linear SVM are comparable. While in the second training set, the LDA, the LR, and the linear SVM perform comparably and the RFC performs worse than the other three.

Table 2. Performance comparison of the four classifiers (LDA, LR, SVM, and RFC) on two training sets.

	Set 1:12 failures+237 successes					Set 2: 12 failures + 24 successes				
	MCR	# TP	# FP	TP rate	FP rate	MCR	# TP	# FP	TP rate	FP rate
LDA	0.028	8	3	0.667	0.013	0.056	11	1	0.917	0.042
LR	0.044	7	6	0.583	0.025	0.056	11	1	0.917	0.042
SVM	0.028	7	2	0.583	0.008	0.056	10	0	0.833	0
RFC*	0.021(0.002)	9	2	0.75(0)	0.009(0.002)	0.075(0.013)	10	1	0.863(0.041)	0.044(0.009)

* 100 trees, mtry = 4; 20 runs, data in the parenthesis are standard deviations.

4. DISCUSSION

As for the boxplot results of peduncles' volumes, ideally the dashed lines connecting the volumes and surface areas of the six CPs of the 10 manual delineations and the corresponding automatic segmentations should be parallel. However, since the pipelines used for processing the manual delineations and the automatic segmentations are different, volume differences can be expected. The positions of the notches in the paired results show that their medians are statistically the same. Also outliers in these boxplots cover several kinds of diagnoses rather than a specific one. This indicates that the segmentation algorithm can perform well on different diseases and is not biased to a certain one.

The true positive and false positive rates of each feature in Table 1 show that the object oriented features (volumes and surface areas of each peduncle) and some of the data quality related features (mean FA, the mean spherical Westin index, and the mean and standard deviation of the linear Westin index) generally perform better than the brain mask feature. Among the data quality features, mean MD fails to detect any true failure. Considering the small size of the failures, this can be reasonable and we cannot conclude further about the mean MD.

5. CONCLUSION

We present two outlier detection methods for assessing the performance of an automatic CPs segmentation algorithm. The method based on box-whisker plots can detect segmentations failures effectively using the two object features and most of the data quality features. As for the classification method, considering the four classifiers' performance on both training sets, we can conclude that the LR performs worst among the four classifiers and the other three perform comparably. In summary, this general classification approach that we have pursued here can be applied to other medical image segmentation algorithms. While our application is very specific (to the segmentation of the cerebellar peduncles) there are numerous automatic segmentation algorithms used on medical imaging data in neuroscience and in many other fields of study which would benefit from automatic quality assurance. Our approach suggests an overall methodology that could be adapted and used in many other applications.

REFERENCES

[1] Sivaswamy, Lalitha, et al. "A diffusion tensor imaging study of the cerebellar pathways in children with autism spectrum disorder." Journal of child neurology (2010).

[2] Le Bihan, Denis, et al. "Diffusion tensor imaging: concepts and applications."Journal of magnetic resonance imaging 13.4 (2001): 534-546.

[3] Bazin, Pierre-Louis, et al. "Direct segmentation of the major white matter tracts in diffusion tensor images." NeuroImage 58.2 (2011): 458-468.

[4] Ye, Chuyang, et al. "Labeling of the cerebellar peduncles using a supervised Gaussian classifier with volumetric tract segmentation." SPIE Medical Imaging. International Society for Optics and Photonics, 2012.

[5] Ye, Chuyang, et al. "Segmentation of the complete superior cerebellar peduncles using a multi-object geometric deformable model." In Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on (pp. 49–52).

[6] Hao, Xiang, et al. "Improved segmentation of white matter tracts with adaptive Riemannian metrics." Medical image analysis 18.1 (2014): 161-175.

[7] Ye, Chuyang, et al. "Segmentation of the Cerebellar Peduncles Using a Random Forest Classifier and a Multi-object Geometric Deformable Model: Application to Spinocerebellar Ataxia Type 6." Neuroinformatics (2015): 13:367–381.

[8] Zhang, Song, Stephen Correia, and David H. Laidlaw. "Identifying white-matter fiber bundles in DTI data using an automated proximity-based fiber-clustering method." Visualization and Computer Graphics, IEEE Transactions on 14.5 (2008): 1044-1053.

[9] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[10] Westin, Carl-Fredrik, et al. "Geometrical diffusion measures for MRI from tensor basis analysis." Proceedings of ISMRM. Vol. 97. 1997.

[11] Grubbs, Frank E. "Procedures for detecting outlying observations in samples." Technometrics 11.1 (1969): 1-21.