MICCAI 2012
Workshop on Multi-Atlas Labeling

Edited by Bennett A. Landman and Simon K. Warfield

Alexander Hammers, Alireza Akhondi-Asl, Andrew J. Asman, Annemie Ribbens, Bennett A. Landman, Blake Lucas, Brian Avants, Christian Ledig, Christos Davatzikos, Da Ma, Daniel Rueckert, Dirk Vandermeulen, Frederik Maes, Guray Erus, Holly Holmes, Hongzhi Wang, Jiahui Wang, Jimit Doshi, Joe Kornegay, Jose Manjon, Joseph Hajnal, Katherine Gray, Louis Collins, M. Jorge Cardoso, Marc Modat, Mark Lythgoe, Martin Styner, Mehran Armand, Michael Miller, Paul Aljabar, Paul Suetens, Paul Yushkevich, Pierrick Coupe, Robin Wolz, Rolf Heckemann, Russell Taylor, Sébastien Ourselin, Shiva Keihaninejad, Simon Eskildsen, Simon K. Warfield, Susumu Mori, Vladimir Fonov, Xiaoying Tang, Yael Shiloh-Malawsky, Yangming Ou, Yoshito Otake, Zheng Fan

Co-Chairs

Bennett Landman, Vanderbilt
Simon Warfield, Harvard

Workshop Committee

Paul Aljabar, Imperial College London
Dzung Pham, Henry M. Jackson Foundation
Hakmook Kang, Vanderbilt University
Arno Klein, Columbia University
Torsten Rohlfing, Stanford University
Dinggang Shen UNC Chapel Hill
T. Robin Langerak, Utrecht
Paul Thompson, UCLA
Paul Yushkevich, UPENN

Grand Challenge Description

1 Rules

1.1 Contest Data

On approximately April 1, 2012, the following data will be made available after registration (see below).

- ***Training Atlas Pairs***: **15** datasets for distinct human data consisting of a de-faced T1-weighted structural MRI dataset and as associated manually labeled volume with one label per voxel. Each volume (MRI and label) will be stored in a separate 3D NiFTI file. These files will be properly interpreted by the MIPAV software package (freely available). (from the "reliability" OASIS data set).
- ***Testing Target MRI***: **20** T1-weighted structural MRI datasets of 15 distinct subjects..

1.2 Evaluation Procedures

After submission of a preliminary manuscript, contestants will be given an access code to upload labeled target MRI datasets through this website. Within 24 hours of upload, the contact author will be e-mailed a PDF of quantitative performance results. The primary metric for the grand challenge contest is the mean Dice similarity coefficient across all brain labels and all subjects in the "testing target" cohort. **The list of anatomical labels is here.**

Other metrics and visualizations will be included. Specifically, informal comparative criteria will be produced for:

- tissue segmentation (GM, WM and CSF)
- basic structure segmentation (subcortical structures, cerebellum, brain stem and unparcellated cortex)

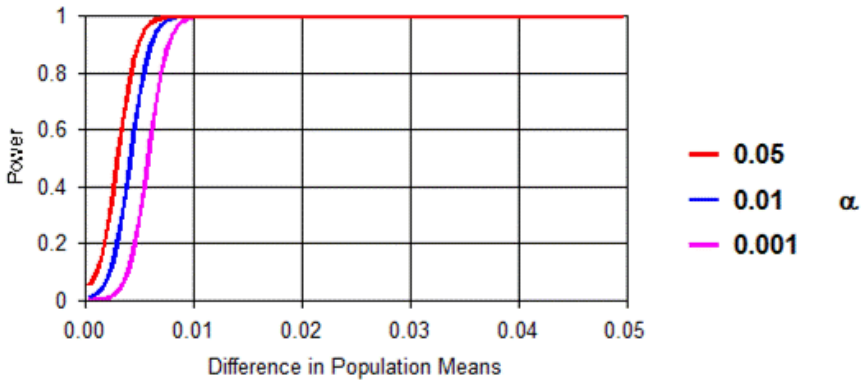
Resubmissions of different label results are permitted within the evaluation window. We understand that technical glitches, format errors, etc. may lead to misleading error metrics. The organizers encourage (and will assist with) multiple submissions to address such errors. The last submission of label results under an access code will be considered the authors final submission.

We *strongly* advocate against parameter tuning based on the hidden testing labels during the evaluation window. The number of submissions will be recorded on the summary PDF and must be reported in the final publication. Submission of multiple similar algorithms for evaluation is **encouraged**. Please use a different preliminary manuscript for each algorithm variant. Text *may* be shared among manuscript submissions.

1.3 Statistical Power of This Challenge

To assess the statistical power of the 2012 Grand Challenge on Multi-Atlas Labeling to differentiate between two labeling algorithms, we consider a two-sided, paired t-test scenario. In a pilot study of 15 labeled whole-brain MRI datasets, we performed all-pairs registration with VABRA (as implemented in the Java Image Science Toolkit – JIST). For each image, we randomly partitioned the remaining 14 registered datasets into two groups of seven. Then, we applied log-odds locally weighted vote (as described in Sabuncu, et al, TMI 2010) to independently fuse each set of labeled atlases. The mean Dice Similarity Coefficient (DSC) over all labels was 0.885±0.0051 for the 15 brains. The standard deviation of the difference between the first and second set of seven atlases was 0.0054 DSC.

The challenge will be evaluated on a study of 15 subjects. The hypothetical t-test would be between two methods on the same subjects; hence would form paired observations. The above prior data indicate that the difference in the response of matched pairs is approximately normally distributed with standard deviation 0.0054 DSC. We will be able to detect a true difference in the mean response of matched pairs of -0.005 or 0.005 DSC with probability (power) 0.9. The alpha-rate (Type I error probability) associated with this test of the null hypothesis that this response difference is zero is 0.05. The plot below shows power as a function of effect size (mean DSC between methods) and alpha-rate. The x-axis is in units of DSC difference.



1.4 Validation (Hidden) Data

The hidden true labels will be revealed September 1, 2012 and made available to all authors who submitted a response to the workshop challenge. Additionally, the full source code of the program used to generate the evaluation PDF's will be made public.

1.5 Terms of Use for Data Provided During the Contest

The data will be released under the Creative Commons Attribution-NonCommercial (CC BY-NC) with no end date. Users should credit the MRI scans as originating from the OASIS project and the labeled data as "provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/) under academic subscription". These references should be included in all workshop and final publications.

1.6 Availability of Data after Contest Completion

After the completion of the contest, testing/training data will be available from this website after users agree to the license terms.

After the completion of the contest, please contact Neuromorphometrics (http://neuromorphometrics.com/) for additional details relevant to this dataset.

1.7 Eligibility

No individual who has access to the labeled testing datasets is permitted to submit a response to this grand challenge. Specifically, employees, affiliates or contractors for Neuromorphometrics are not permitted to participate in any contest entry.

It is assumed that the majority of entrants will seek to use multi-atlas method to respond to this challenge. However, any method of labeling the testing datasets is permitted as long as the approach is described in a reproducible manner.

1.8 Software Sharing

Contestants are encouraged (but not required) to make available the software, tools, and source code for the methods used in response to this challenge. However, software and code sharing are **not** required for participation.

1.9 Contest Results

After the workshop (after October 5, 2012), a summary of all methods will be presented here. We will publish a citable book (with ISBN) with a compilation of all results and a summary of rankings. This book will be available at cost (or no cost, if possible) – neither editors nor their institutions will receive any compensation for this publication. On this website, we will make available copies of all papers, the label masks submitted for each method, the PDF result files for each method, and (if the authors desire) additional resources (software/code/etc.) that a reader may find relevant to each paper.

2 Challenge Demographics

2.1 Training Data

	Number	Age	
Total F	10	19	min
Total M	5	23	average
Total	15	34	max

2.2 Training Details

Case #	train/test	Training	Subject	Gender	Age
1000	train		OAS1_0061	F	20
1001	train		OAS1_0080	F	25
1002	train		OAS1_0092	M	22
1006	train		OAS1_0145	M	34
1007	train		OAS1_0150	F	20
1008	train		OAS1_0156	F	20
1009	train		OAS1_0191	F	21
1010	train		OAS1_0202	F	23
1011	train		OAS1_0230	F	19
1012	train		OAS1_0236	F	20
1013	train		OAS1_0239	F	29
1014	train		OAS1_0249	F	28
1015	train		OAS1_0285	M	20
1036	train		OAS1_0353	M	22
1017	train		OAS1_0368	M	22

2.3 Testing Data

	Number	Age	
Total (unique) F	10	18	min
Total (unique) M	5	45.7	average
Total (unique)	16	90	max

2.4 Testing Details

Case #	train/test	Testing	Subject	Gender	Age	Notes
1003	test		OAS1_0101	M	29	1st Scan
1004	test		OAS1_0111	M	23	1st Scan

1005	test	OAS1_0117	M	25	1st Scan
1018	test	OAS1_0379	F	20	1st Scan
1019	test	OAS1_0395	F	26	1st Scan
1023	test	OAS1_0101	M	29	2nd Scan
1024	test	OAS1_0111	M	23	2nd Scan
1025	test	OAS1_0117	M	25	2nd Scan
1038	test	OAS1_0379	F	20	2nd Scan
1039	test	OAS1_0395	F	26	2nd Scan
1101	test	OAS1_0091	F	18	
1104	test	OAS1_0417	F	30	
1107	test	OAS1_0040	F	38	
1110	test	OAS1_0282	F	45	
1113	test	OAS1_0331	F	54	
1116	test	OAS1_0456	M	61	
1119	test	OAS1_0300	M	68	
1122	test	OAS1_0220	F	75	
1125	test	OAS1_0113	F	83	
1128	test	OAS1_0083	F	90	

Summary of Challenge Results Based on All Testing Data

Overall Rank †	Repro. Rank‡	Team Name	Mean DSC Overall	Mean DSC Cortical	Mean DSC Non-Cortical
1	1	PICSL_BC	0.7654	0.7388	0.8377
2	2	NonLocalSTAPLE	0.7581	0.7318	0.8296
3	3	MALP_EM	0.7576	0.7328	0.8252
4	4	PICSL_Joint	0.7499	0.7216	0.8271
5	6	MAPER	0.7413	0.7144	0.8144
6	7	STEPS	0.7372	0.7107	0.8095
7	5	SpatialSTAPLE	0.7372	0.7093	0.8130
8	9	CIS_JHU	0.7357	0.7131	0.7971
9	8	CRL_Weighted_STAPLE_ANTS+Baloo	0.7344	0.7122	0.7950
10	10	CRL_Weighted_STAPLE_ANTS	0.7308	0.7066	0.7966
11	11	CRL_STAPLE_ANTS+Baloo	0.7290	0.7064	0.7919
12	12	CRL_STAPLE_ANTS	0.7280	0.7033	0.7951
13	15	CRL_Probabilistic_STAPLE_ANTS+Baloo	0.7251	0.7009	0.7911
14	14	CRL_MV_ANTS+Baloo	0.7247	0.6966	0.8012
15	16	CRL_MV_ANTS	0.7243	0.6951	0.8035
16	13	DISPATCH	0.7243	0.6965	0.8000
17	18	CRL_Probabilistic_STAPLE_ANTS	0.7223	0.6972	0.7907
18	22	SBIA_SimRank+NormMS+WtROI	0.7212	0.6940	0.7953
19	19	SBIA_BrainROIMaps_MV_IntCorr	0.7193	0.6933	0.7904
20	23	SBIA_BrainROIMaps_JaccDet_IntCorr	0.7186	0.6913	0.7927
21	20	BIC-IPL-HR	0.7173	0.6888	0.7948
22	21	SBIA_SimMSVoting	0.7172	0.6898	0.7918
23	17	UNC-NIRAL	0.7171	0.6869	0.7992
24	24	SBIA_SimRank+NormMS	0.7162	0.6884	0.7919
25	25	BIC-IPL	0.7107	0.6829	0.7864

† Overall Rank is computed based on the relative "mean DSC" over all labels over all subjects (this table).
‡ Reproducibility Rank is computed based on the relative "mean DSC" over all labels for only subjects in the reproducibility cohort (next table).

Summary of Challenge Results Based on Reproducibility Data

Overall Rank †	Repro. Rank‡	Team Name	Mean DSC Overall	Mean DSC Cortical	Mean DSC Non-Cortical
1	**1**	PICSL_BC	0.782	0.7528	0.8614
2	**2**	NonLocalSTAPLE	0.7764	0.7473	0.8554
3	**3**	MALP_EM	0.7708	0.7416	0.8504
4	**4**	PICSL_Joint	0.7663	0.7361	0.8482
7	**5**	SpatialSTAPLE	0.7576	0.7278	0.8388
5	**6**	MAPER	0.7518	0.7207	0.8364
6	**7**	STEPS	0.7500	0.7202	0.8311
9	**8**	CRL_Weighted_STAPLE_ANTS+Baloo	0.7470	0.7225	0.8137
8	**9**	CIS_JHU	0.7440	0.7178	0.8151
10	**10**	CRL_Weighted_STAPLE_ANTS	0.7424	0.7160	0.8142
11	**11**	CRL_STAPLE_ANTS+Baloo	0.7412	0.7158	0.8103
12	**12**	CRL_STAPLE_ANTS	0.7393	0.7125	0.8121
16	**13**	DISPATCH	0.7388	0.7091	0.8199
14	**14**	CRL_MV_ANTS+Baloo	0.7378	0.7063	0.8234
13	**15**	CRL_Probabilistic_STAPLE_ANTS+Baloo	0.7372	0.7108	0.8092
15	**16**	CRL_MV_ANTS	0.7360	0.7038	0.8236
23	**17**	UNC-NIRAL	0.7350	0.7030	0.8220
17	**18**	CRL_Probabilistic_STAPLE_ANTS	0.7334	0.7062	0.8075
19	**19**	SBIA_BrainROIMaps_MV_IntCorr	0.7313	0.7007	0.8145
21	**20**	BIC-IPL-HR	0.7299	0.6965	0.8209
22	**21**	SBIA_SimMSVoting	0.7283	0.6977	0.8116
18	**22**	SBIA_SimRank+NormMS+WtROI	0.7282	0.6957	0.8164
20	**23**	SBIA_BrainROIMaps_JaccDet_IntCorr	0.7265	0.6932	0.8172
24	**24**	SBIA_SimRank+NormMS	0.7236	0.6909	0.8125
25	**25**	BIC-IPL	0.7225	0.6900	0.8112

† Overall Rank is computed based on the relative "mean DSC" over all labels over all subjects (last table).
‡ Reproducibility Rank is computed based on the relative "mean DSC" over all labels for only subjects in the reproducibility cohort (this table).

BIC-IPL

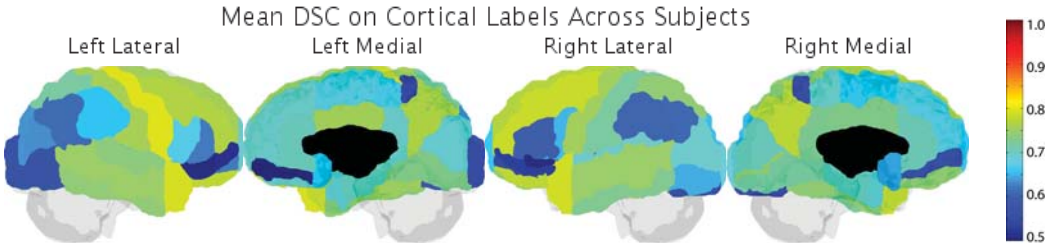
Attempt Number: 1

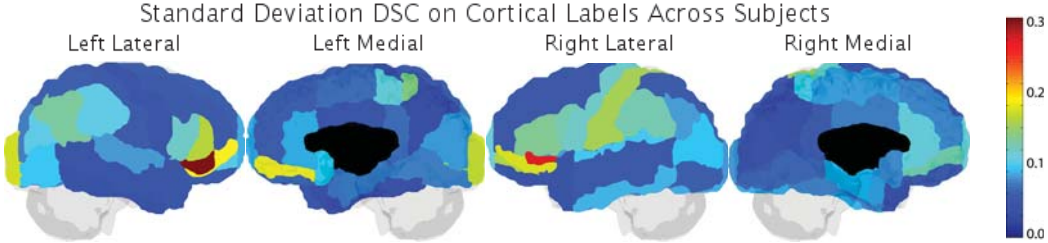
Date: 31-Jul-2012

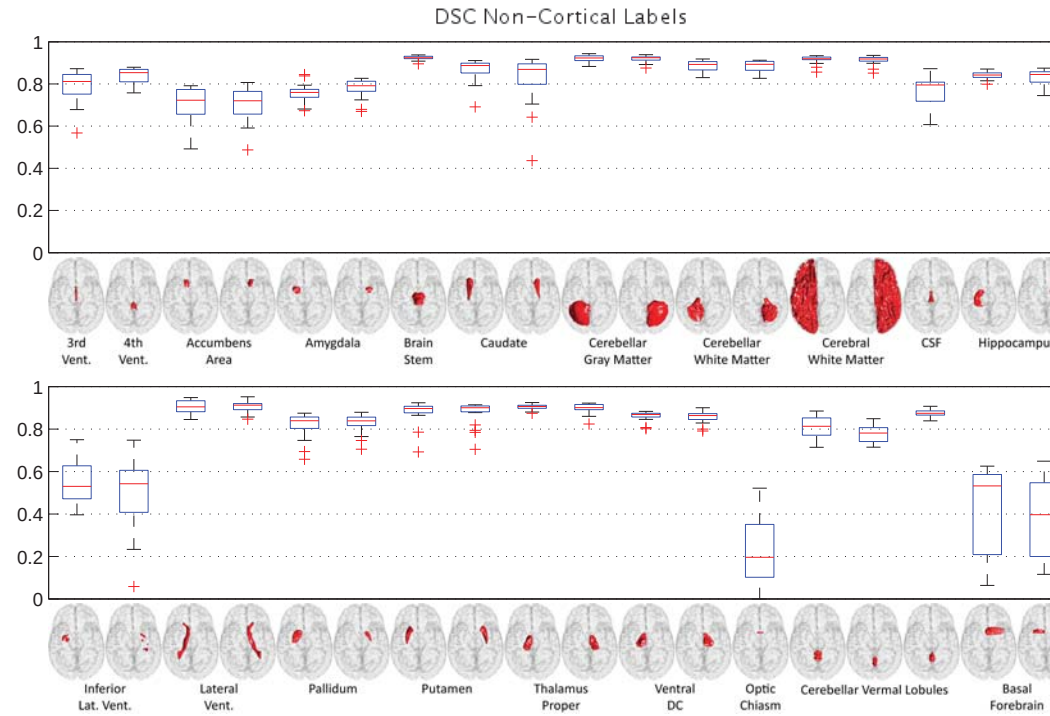
Mean DSC Overall: 0.7107 +/- 0.0217
Rep: Mean DSC Overall: 0.7225 +/- 0.0120

Mean DSC Cortical: 0.6829 +/- 0.0210
Rep: Mean DSC Cortical: 0.6900 +/- 0.0147

Mean DSC Non-Cortical: 0.7864 +/- 0.0303
Rep: Mean DSC Non-Cortical: 0.8112 +/- 0.0069







BIC-IPL-HR

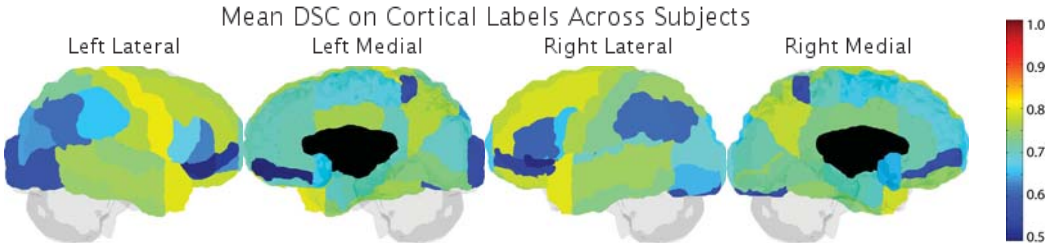
Attempt Number: 1

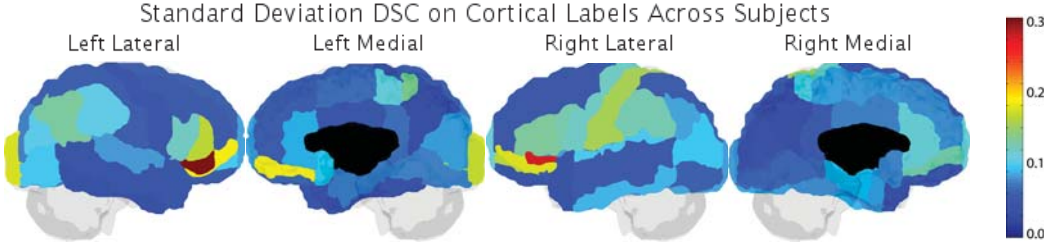
Date: 31-Jul-2012

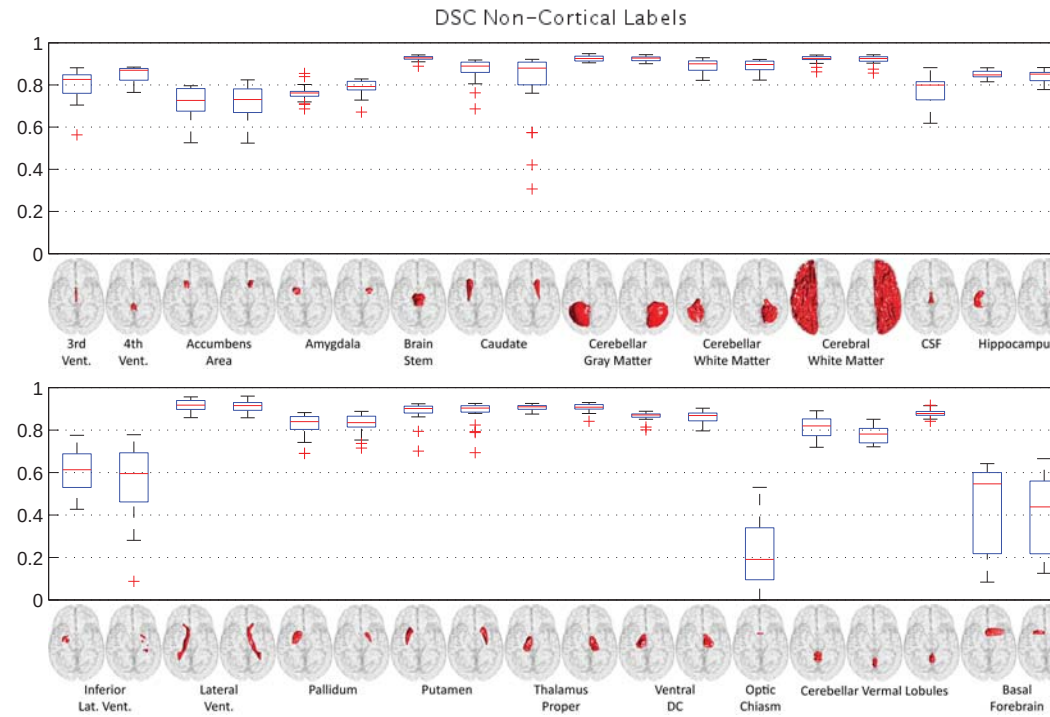
Mean DSC Overall: 0.7173 +/- 0.0227
Rep: Mean DSC Overall: 0.7299 +/- 0.0128

Mean DSC Cortical: 0.6888 +/- 0.0218
Rep: Mean DSC Cortical: 0.6965 +/- 0.0155

Mean DSC Non-Cortical: 0.7948 +/- 0.0320
Rep: Mean DSC Non-Cortical: 0.8209 +/- 0.0087







CIS_JHU

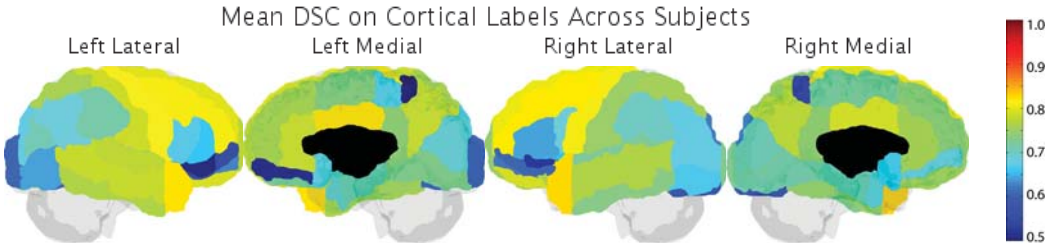
Attempt Number: 1

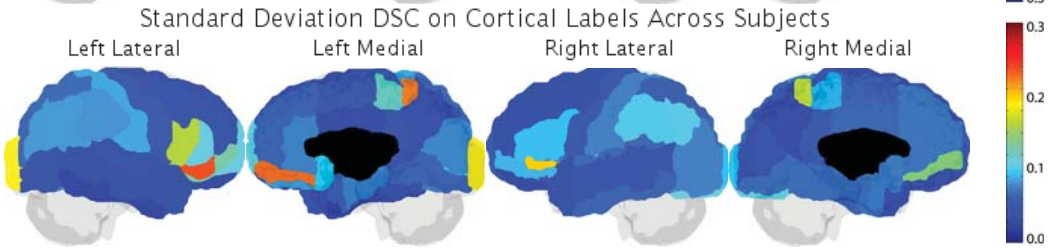
Date: 26-Jul-2012

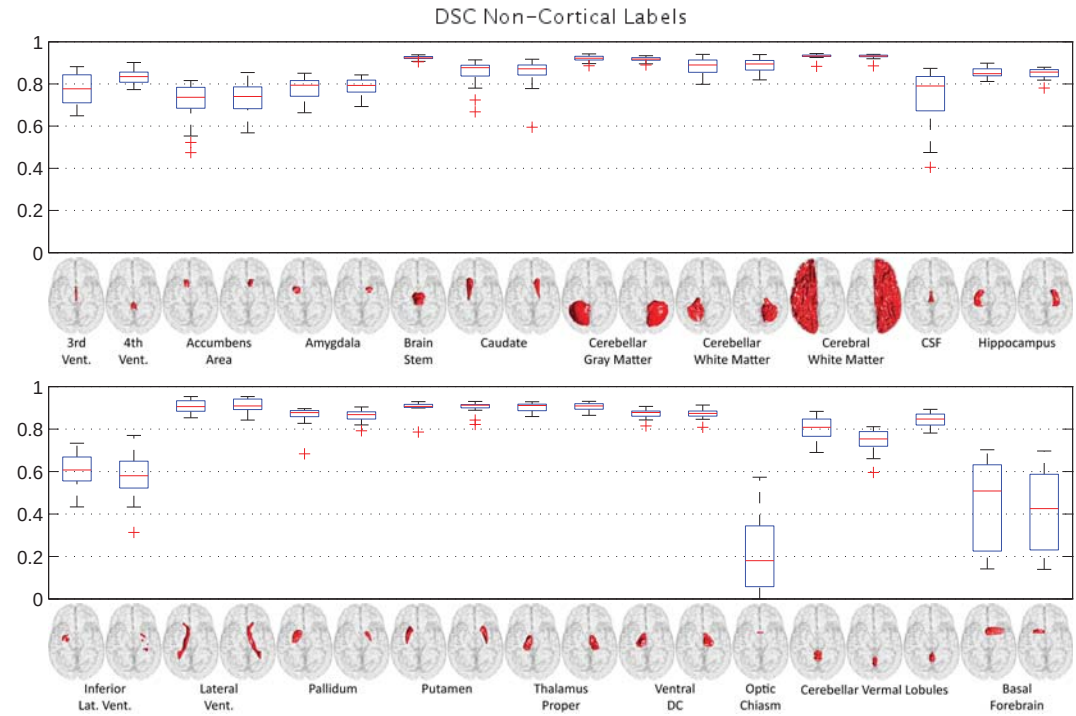
Mean DSC Overall: 0.7357 +/- 0.0183
Rep: Mean DSC Overall: 0.7440 +/- 0.0042

Mean DSC Cortical: 0.7131 +/- 0.0193
Rep: Mean DSC Cortical: 0.7178 +/- 0.0051

Mean DSC Non-Cortical: 0.7971 +/- 0.0230
Rep: Mean DSC Non-Cortical: 0.8151 +/- 0.0076







CRL_MV_ANTS

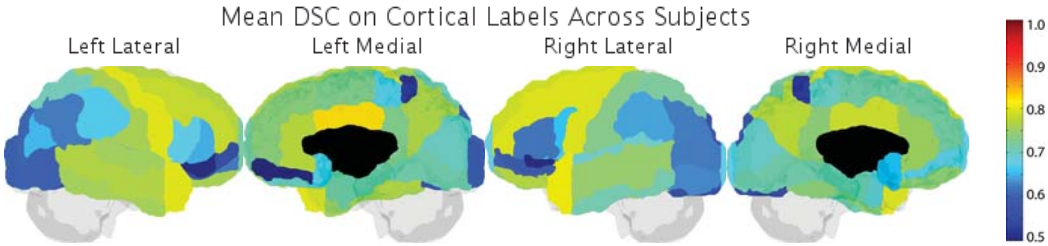
Attempt Number: 1

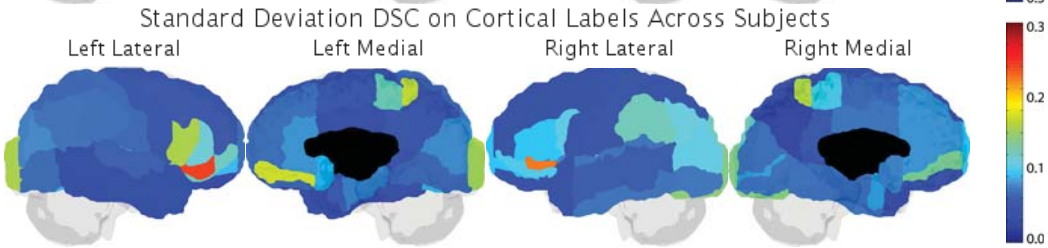
Date: 26-Jul-2012

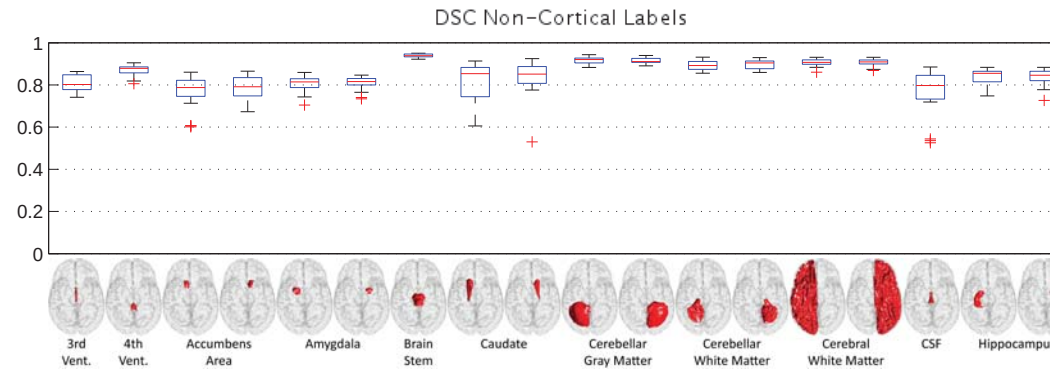
Mean DSC Overall: 0.7243 +/- 0.0252
Rep: Mean DSC Overall: 0.7360 +/- 0.0100

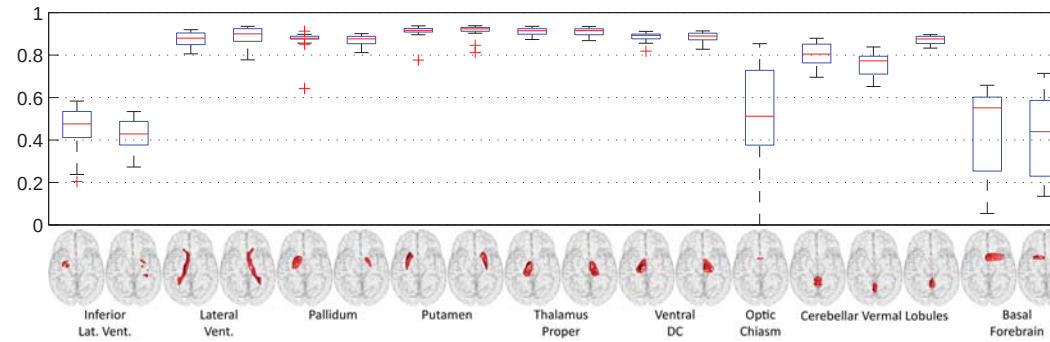
Mean DSC Cortical: 0.6951 +/- 0.0267
Rep: Mean DSC Cortical: 0.7038 +/- 0.0098

Mean DSC Non-Cortical: 0.8035 +/- 0.0285
Rep: Mean DSC Non-Cortical: 0.8236 +/- 0.0174









CRL_MV_ANTS+Baloo

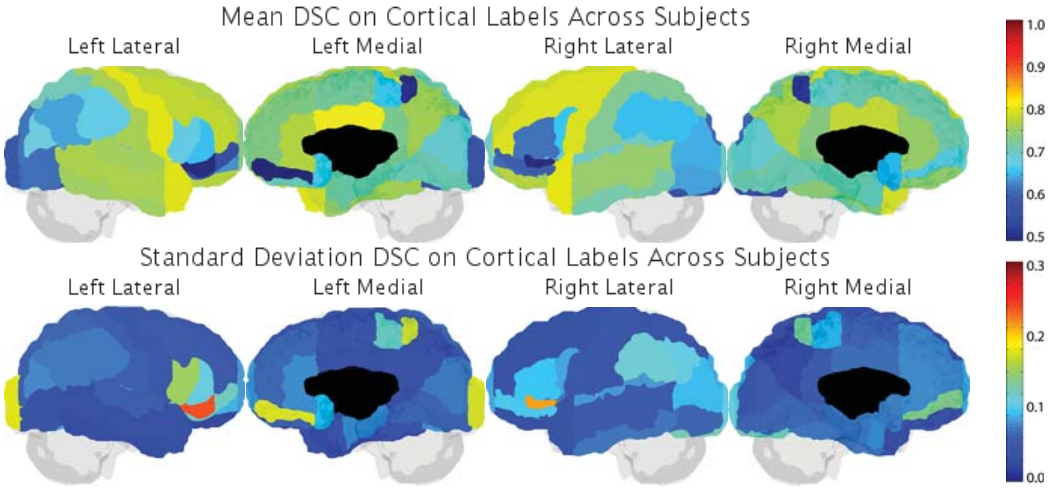
Attempt Number: 1

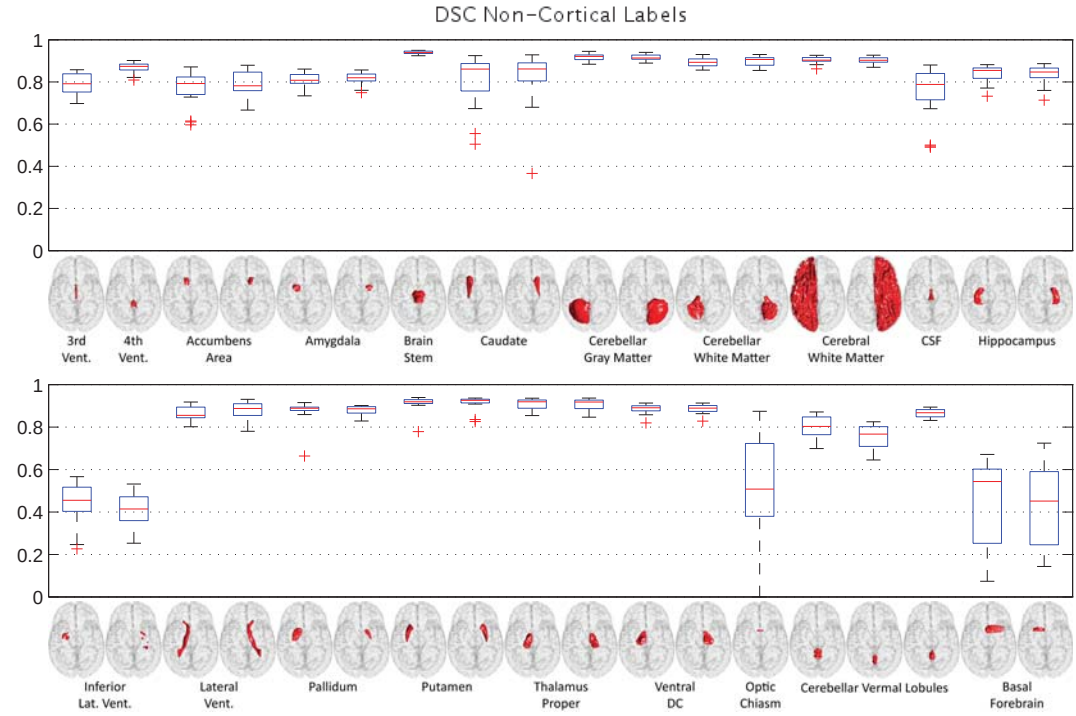
Date: 26-Jul-2012

Mean DSC Overall: 0.7247 +/- 0.0249
Rep: Mean DSC Overall: 0.7378 +/- 0.0094

Mean DSC Cortical: 0.6966 +/- 0.0248
Rep: Mean DSC Cortical: 0.7063 +/- 0.0073

Mean DSC Non-Cortical: 0.8012 +/- 0.0316
Rep: Mean DSC Non-Cortical: 0.8234 +/- 0.0183





CRL_Probabilistic_STAPLE_ANTS

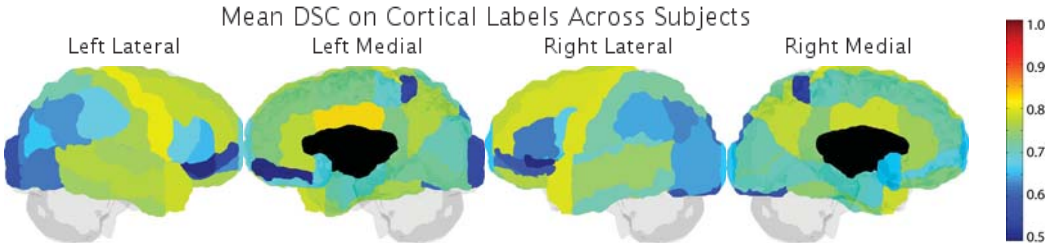
Attempt Number: 1

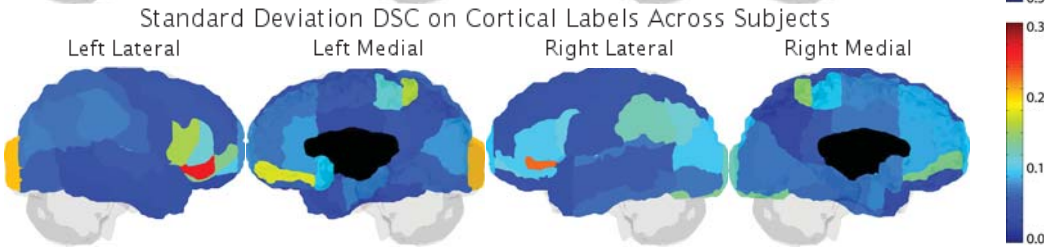
Date: 26-Jul-2012

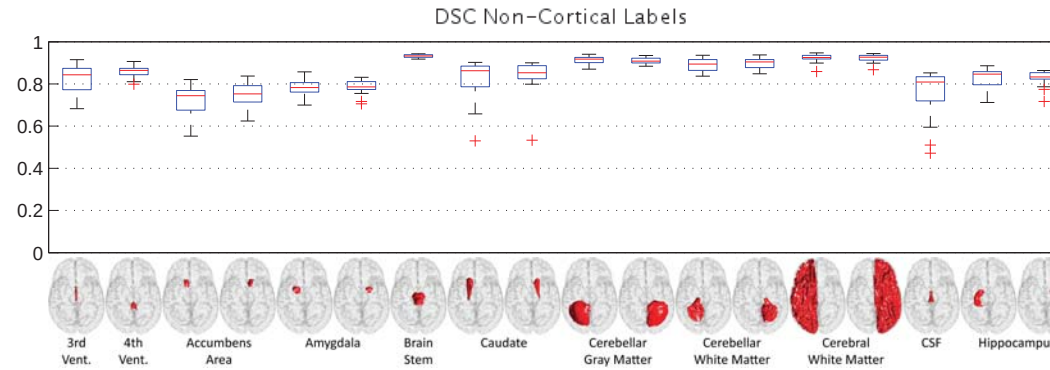
Mean DSC Overall: 0.7223 +/- 0.0249
Rep: Mean DSC Overall: 0.7334 +/- 0.0087

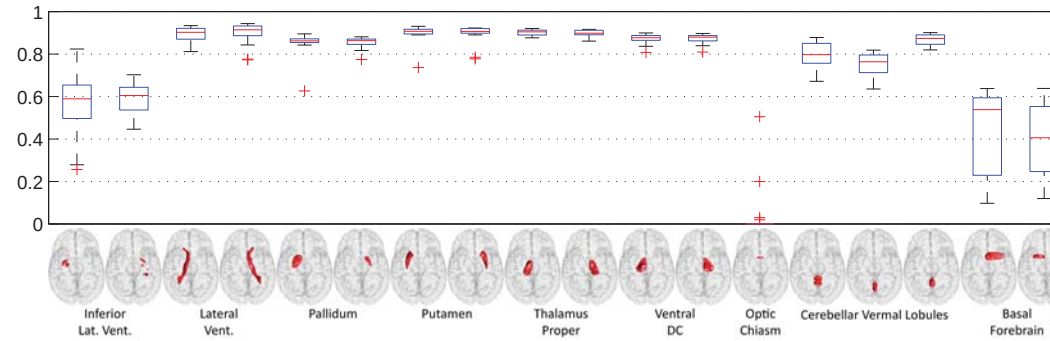
Mean DSC Cortical: 0.6972 +/- 0.0279
Rep: Mean DSC Cortical: 0.7062 +/- 0.0084

Mean DSC Non-Cortical: 0.7907 +/- 0.0234
Rep: Mean DSC Non-Cortical: 0.8075 +/- 0.0161









CRL_Probabilistic_STAPLE_ANTS+Baloo

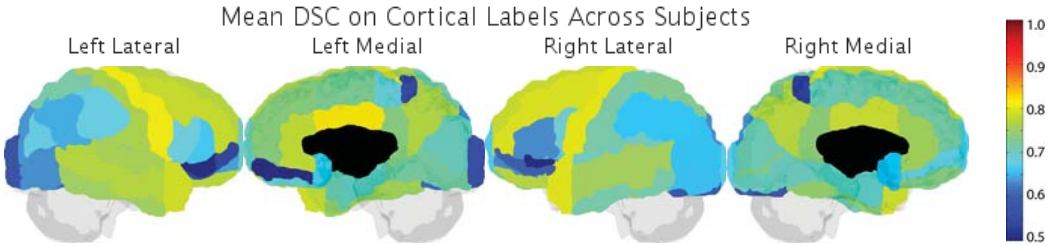
Attempt Number: 1

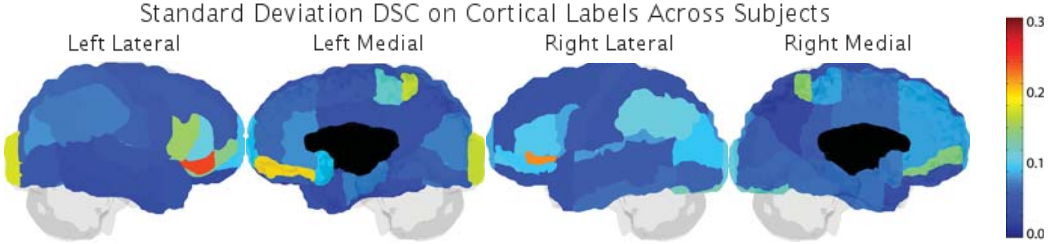
Date: 26-Jul-2012

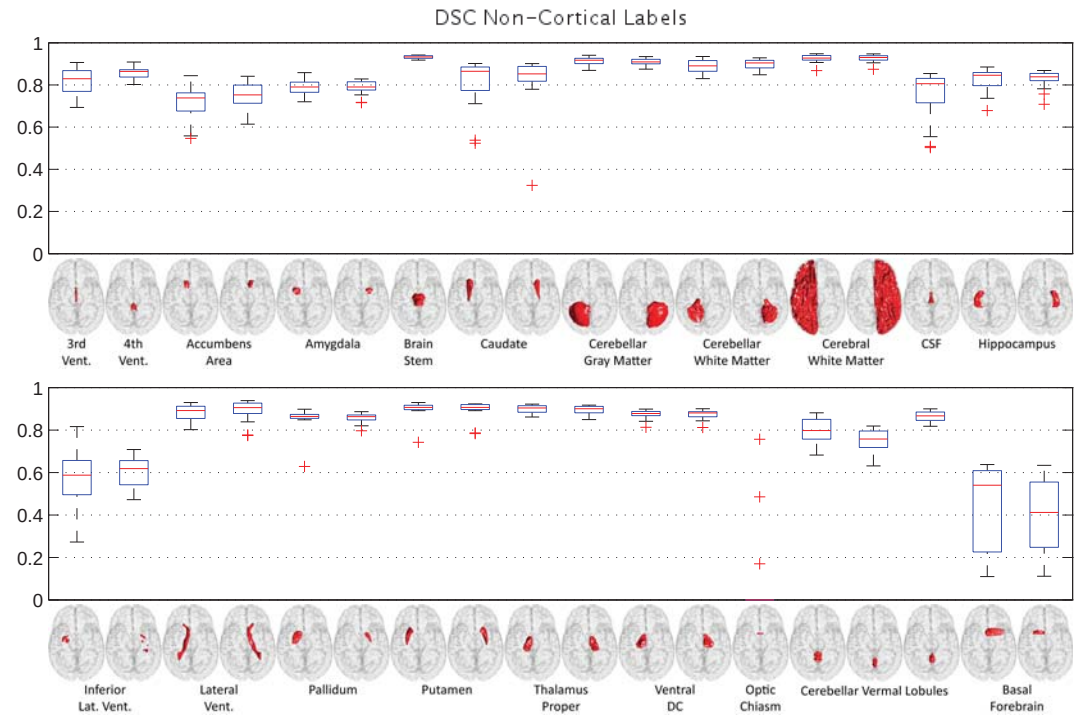
Mean DSC Overall: 0.7251 +/- 0.0244
Rep: Mean DSC Overall: 0.7372 +/- 0.0103

Mean DSC Cortical: 0.7009 +/- 0.0260
Rep: Mean DSC Cortical: 0.7108 +/- 0.0074

Mean DSC Non-Cortical: 0.7911 +/- 0.0260
Rep: Mean DSC Non-Cortical: 0.8092 +/- 0.0196







CRL_STAPLE_ANTS

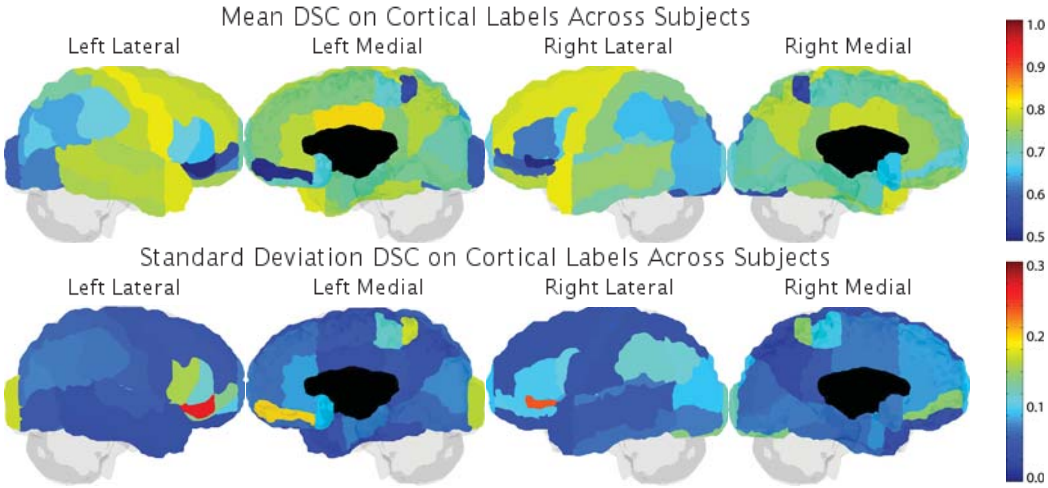
Attempt Number: 1

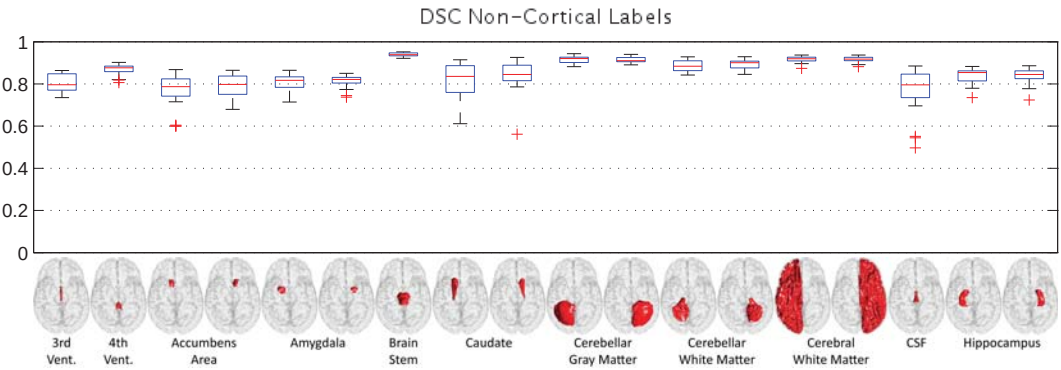
Date: 26-Jul-2012

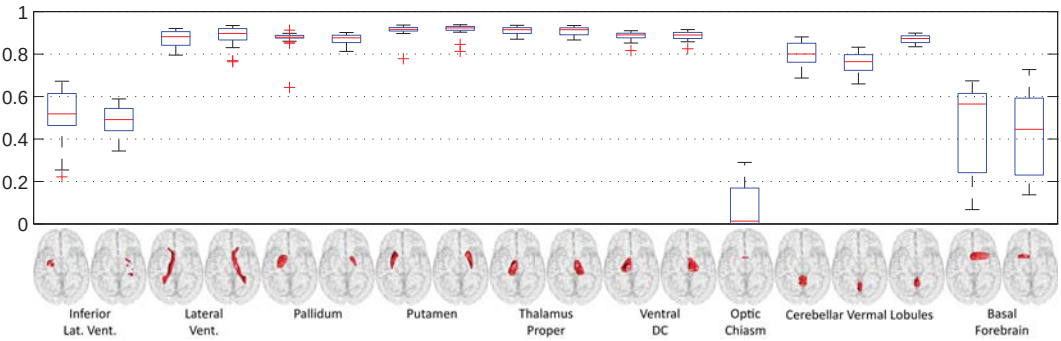
Mean DSC Overall: 0.7280 +/- 0.0236
Rep: Mean DSC Overall: 0.7393 +/- 0.0086

Mean DSC Cortical: 0.7033 +/- 0.0254
Rep: Mean DSC Cortical: 0.7125 +/- 0.0075

Mean DSC Non-Cortical: 0.7951 +/- 0.0245
Rep: Mean DSC Non-Cortical: 0.8121 +/- 0.0167







CRL_STAPLE_ANTS+Baloo

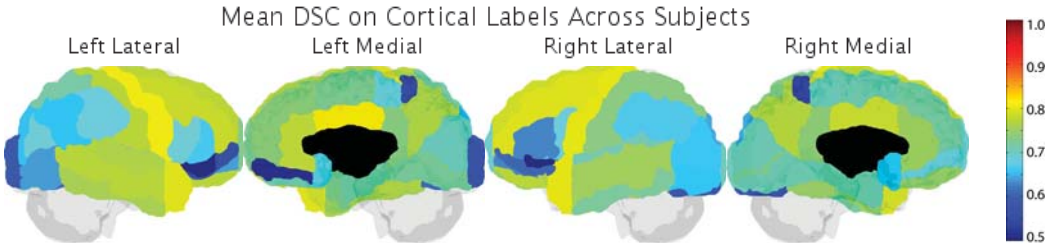
Attempt Number: 1

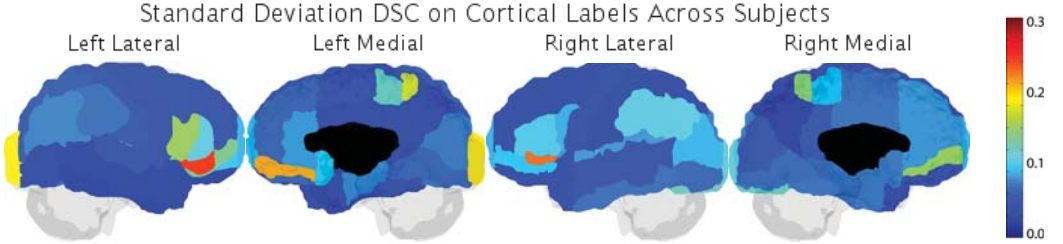
Date: 26-Jul-2012

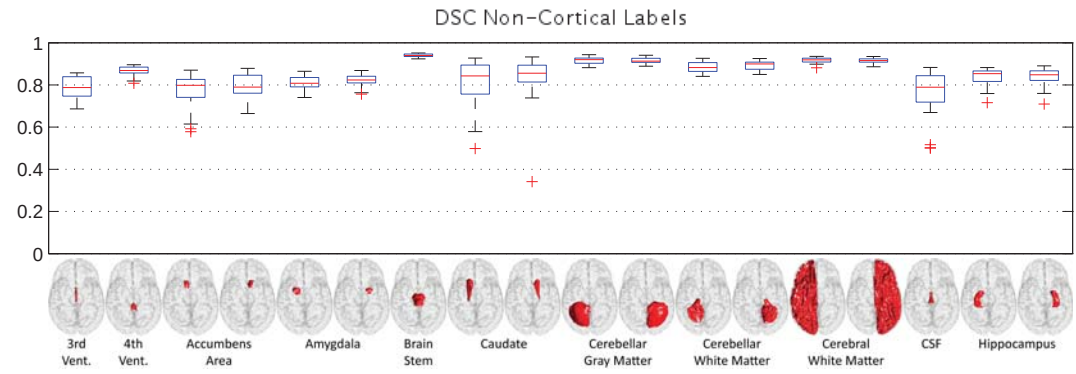
Mean DSC Overall: 0.7294 +/- 0.0235
Rep: Mean DSC Overall: 0.7412 +/- 0.0092

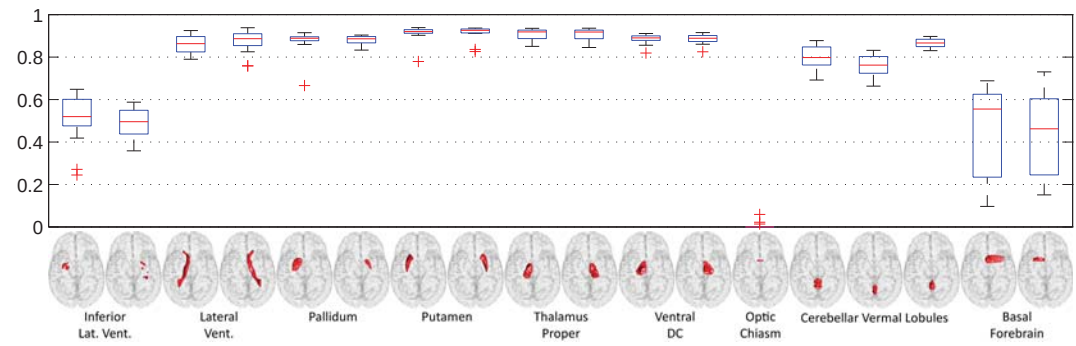
Mean DSC Cortical: 0.7064 +/- 0.0239
Rep: Mean DSC Cortical: 0.7158 +/- 0.0067

Mean DSC Non-Cortical: 0.7919 +/- 0.0267
Rep: Mean DSC Non-Cortical: 0.8103 +/- 0.0179









CRL_Weighted_STAPLE_ANTS

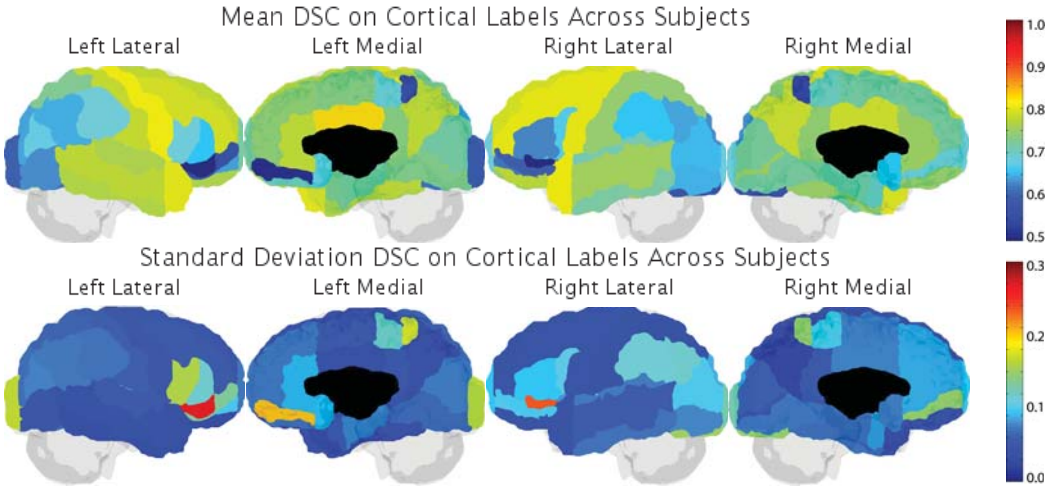
Attempt Number: 1

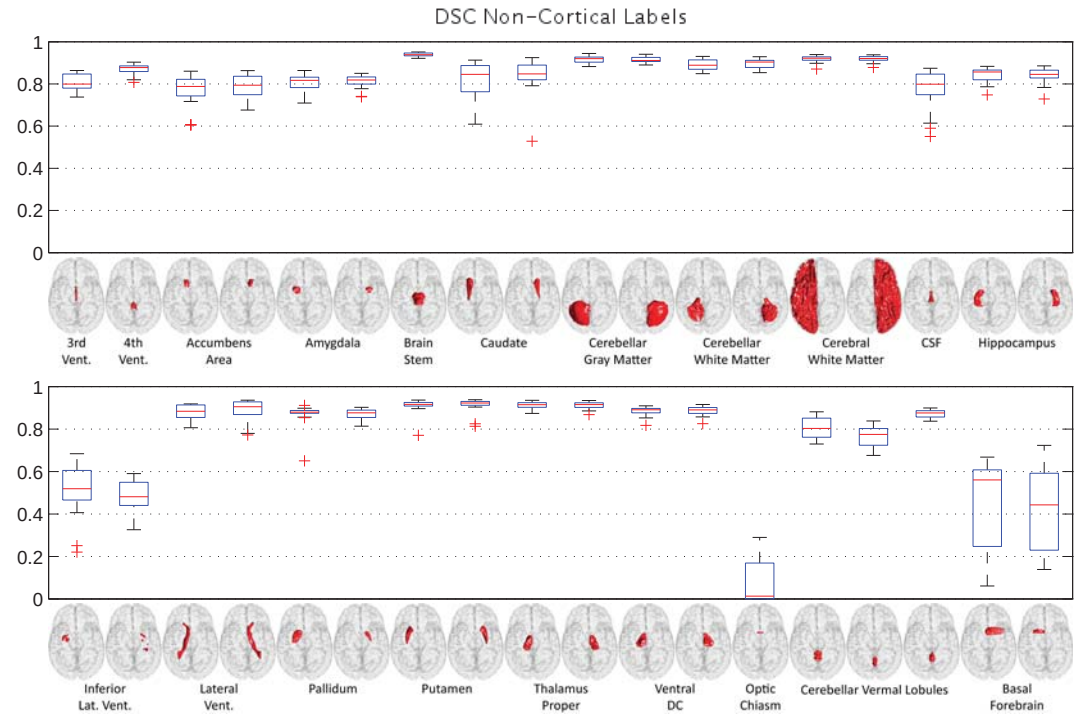
Date: 26-Jul-2012

Mean DSC Overall: 0.7308 +/- 0.0239
Rep: Mean DSC Overall: 0.7424 +/- 0.0085

Mean DSC Cortical: 0.7066 +/- 0.0257
Rep: Mean DSC Cortical: 0.7160 +/- 0.0077

Mean DSC Non-Cortical: 0.7966 +/- 0.0244
Rep: Mean DSC Non-Cortical: 0.8142 +/- 0.0151





CRL_Weighted_STAPLE_ANTS+Baloo

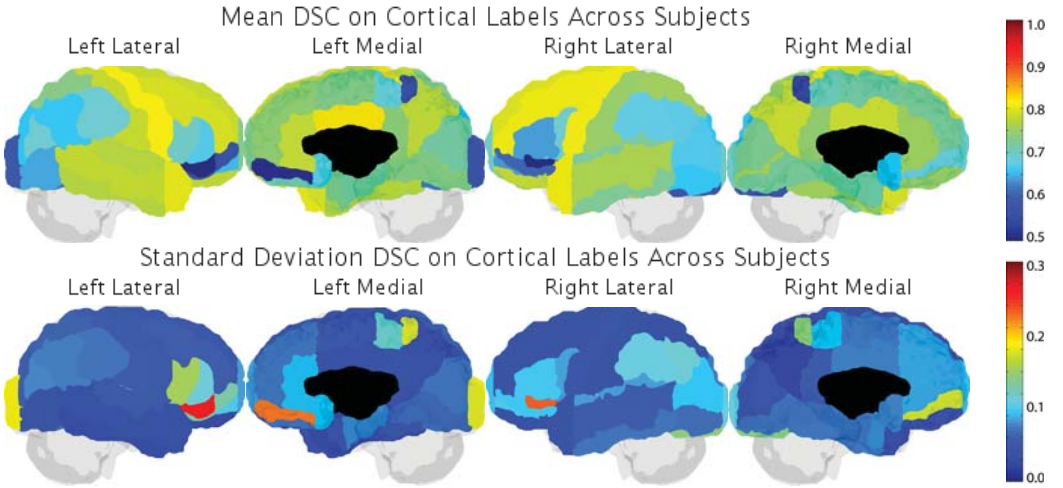
Attempt Number: 1

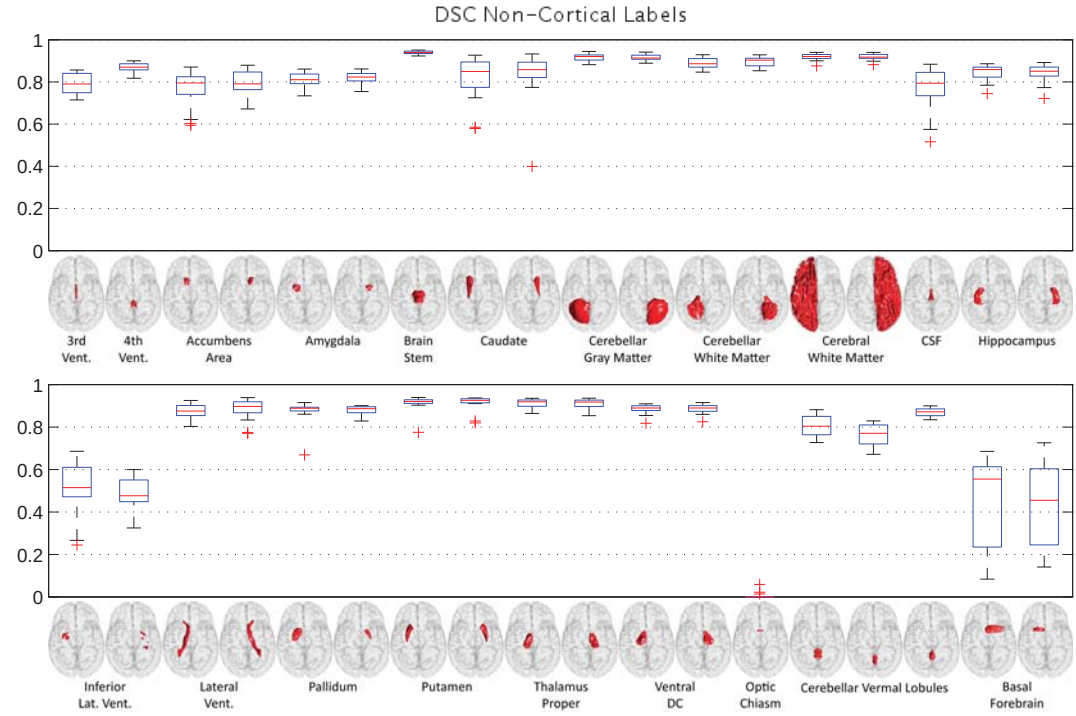
Date: 26-Jul-2012

Mean DSC Overall: 0.7344 +/- 0.0238
Rep: Mean DSC Overall: 0.7470 +/- 0.0091

Mean DSC Cortical: 0.7122 +/- 0.0245
Rep: Mean DSC Cortical: 0.7225 +/- 0.0076

Mean DSC Non-Cortical: 0.7950 +/- 0.0258
Rep: Mean DSC Non-Cortical: 0.8137 +/- 0.0157





DISPATCH

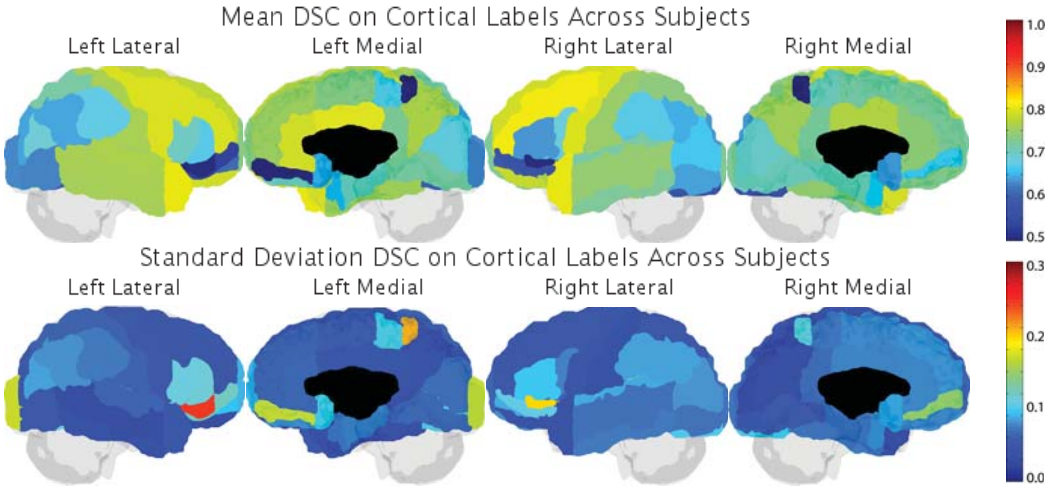
Attempt Number: 1

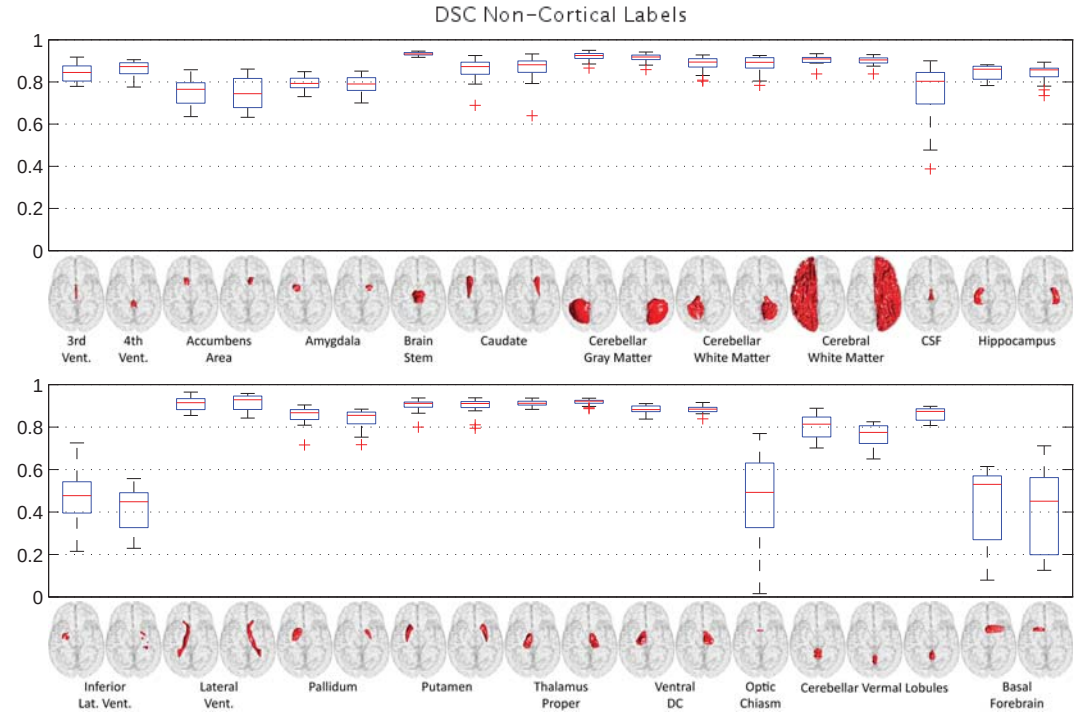
Date: 26-Jul-2012

Mean DSC Overall: 0.7243 +/- 0.0252
Rep: Mean DSC Overall: 0.7388 +/- 0.0097

Mean DSC Cortical: 0.6965 +/- 0.0264
Rep: Mean DSC Cortical: 0.7091 +/- 0.0097

Mean DSC Non-Cortical: 0.8000 +/- 0.0261
Rep: Mean DSC Non-Cortical: 0.8199 +/- 0.0098





MALP_EM

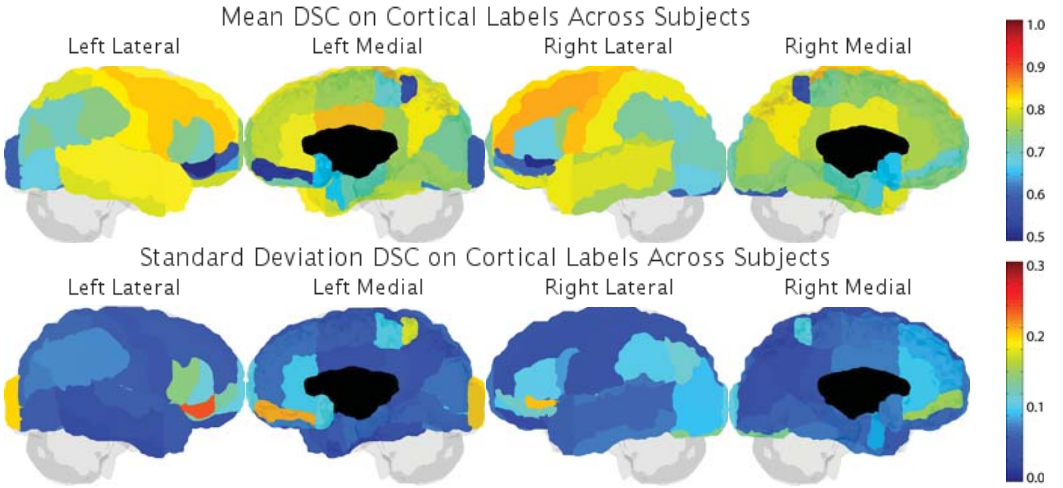
Attempt Number: 3

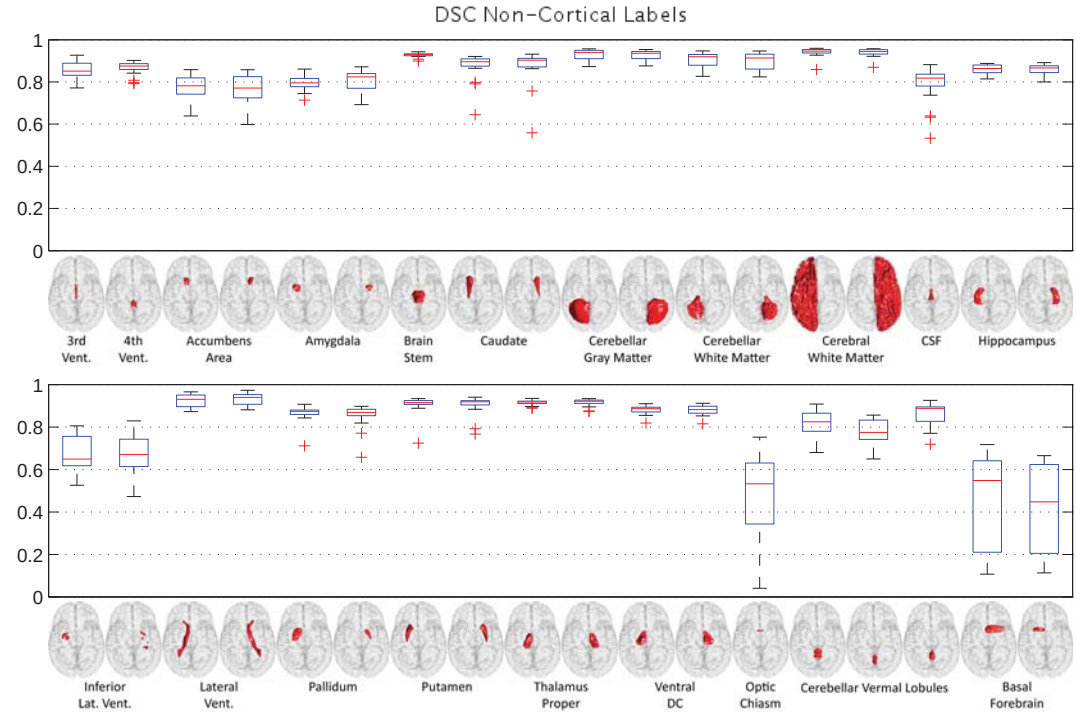
Date: 26-Jul-2012

Mean DSC Overall: 0.7576 +/- 0.0246
Rep: Mean DSC Overall: 0.7708 +/- 0.0058

Mean DSC Cortical: 0.7328 +/- 0.0251
Rep: Mean DSC Cortical: 0.7416 +/- 0.0072

Mean DSC Non-Cortical: 0.8252 +/- 0.0302
Rep: Mean DSC Non-Cortical: 0.8504 +/- 0.0031





maper

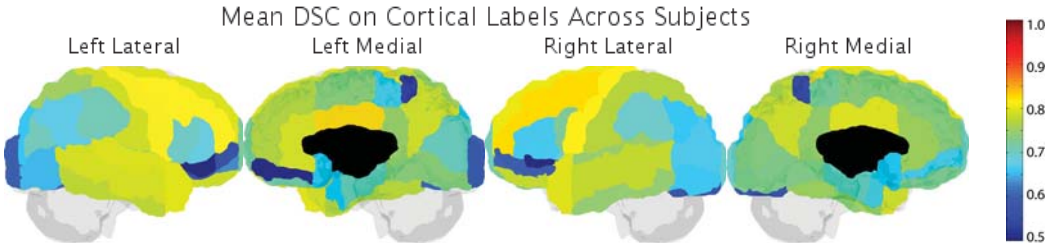
Attempt Number: 1

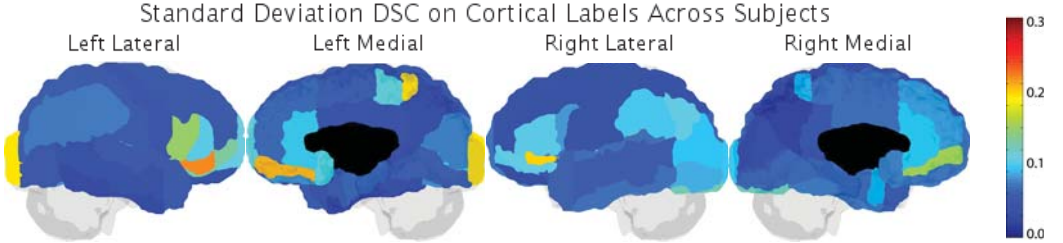
Date: 26-Jul-2012

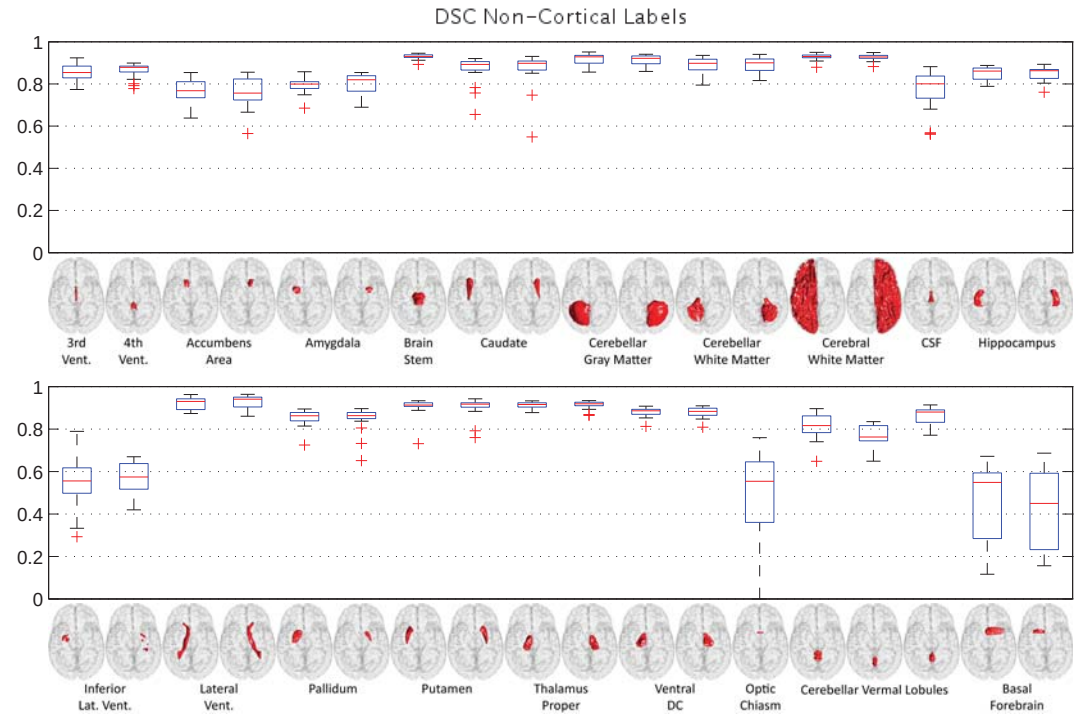
Mean DSC Overall: 0.7413 +/- 0.0228
Rep: Mean DSC Overall: 0.7518 +/- 0.0057

Mean DSC Cortical: 0.7144 +/- 0.0230
Rep: Mean DSC Cortical: 0.7207 +/- 0.0066

Mean DSC Non-Cortical: 0.8144 +/- 0.0293
Rep: Mean DSC Non-Cortical: 0.8364 +/- 0.0084







NonLocalSTAPLE

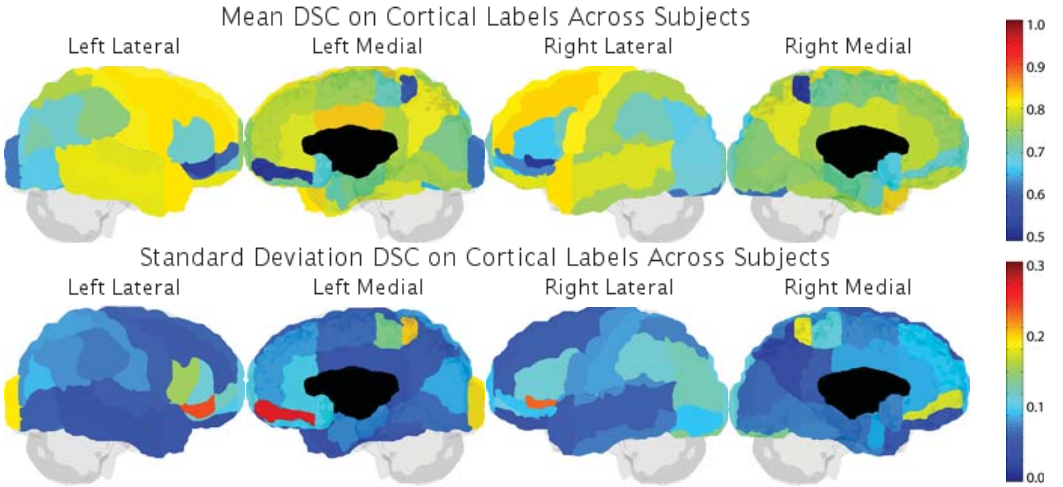
Attempt Number: 2

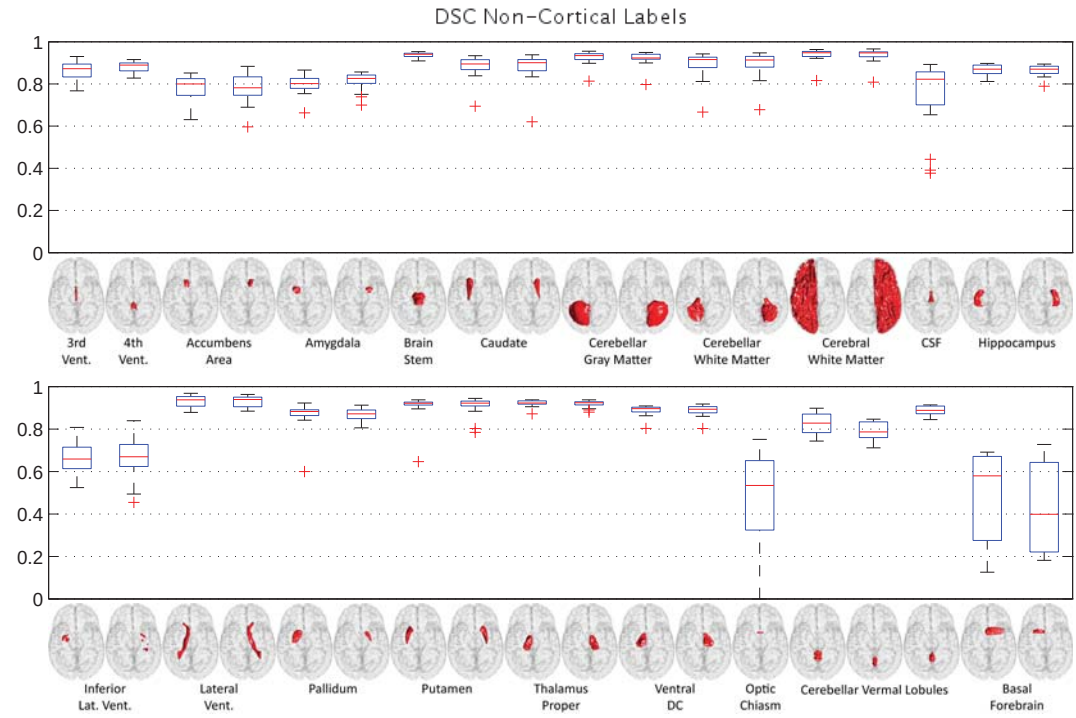
Date: 26-Jul-2012

Mean DSC Overall: 0.7581 +/- 0.0327
Rep: Mean DSC Overall: 0.7764 +/- 0.0064

Mean DSC Cortical: 0.7318 +/- 0.0345
Rep: Mean DSC Cortical: 0.7473 +/- 0.0068

Mean DSC Non-Cortical: 0.8296 +/- 0.0325
Rep: Mean DSC Non-Cortical: 0.8554 +/- 0.0067





PICSL_BC

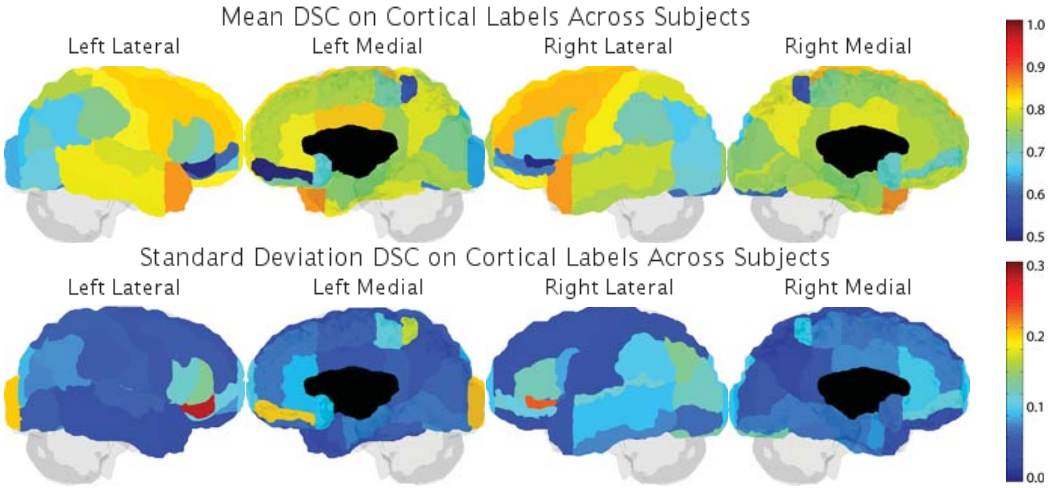
Attempt Number: 3

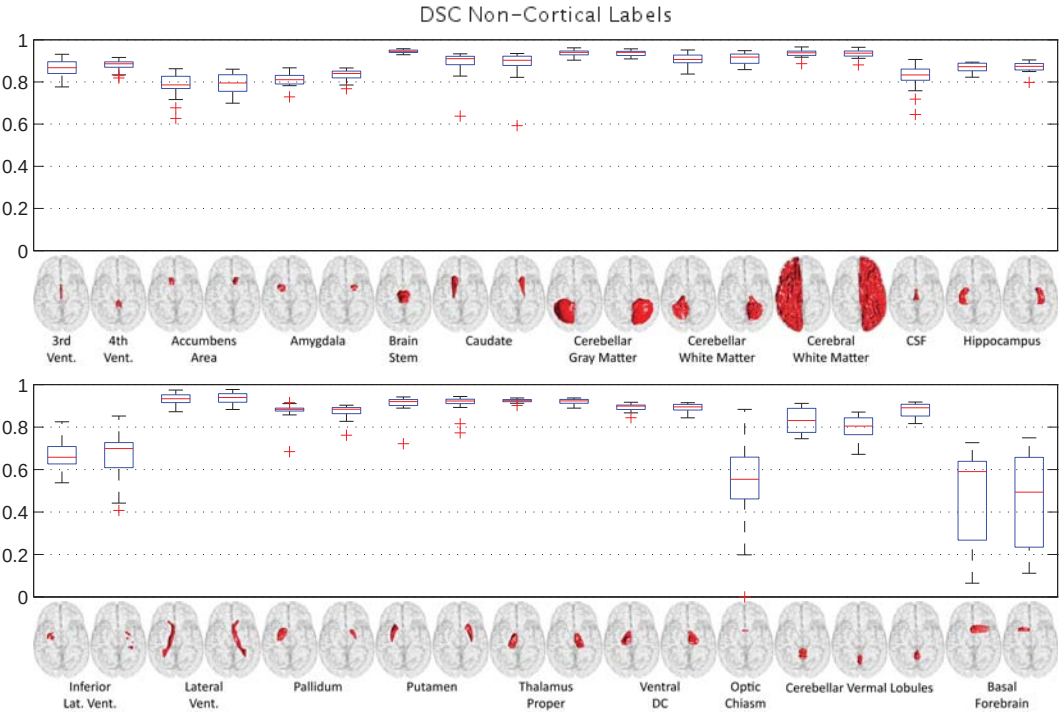
Date: 26-Jul-2012

Mean DSC Overall: 0.7654 +/- 0.0263
Rep: Mean DSC Overall: 0.7820 +/- 0.0098

Mean DSC Cortical: 0.7388 +/- 0.0278
Rep: Mean DSC Cortical: 0.7528 +/- 0.0113

Mean DSC Non-Cortical: 0.8377 +/- 0.0277
Rep: Mean DSC Non-Cortical: 0.8614 +/- 0.0076





PICSL_Joint

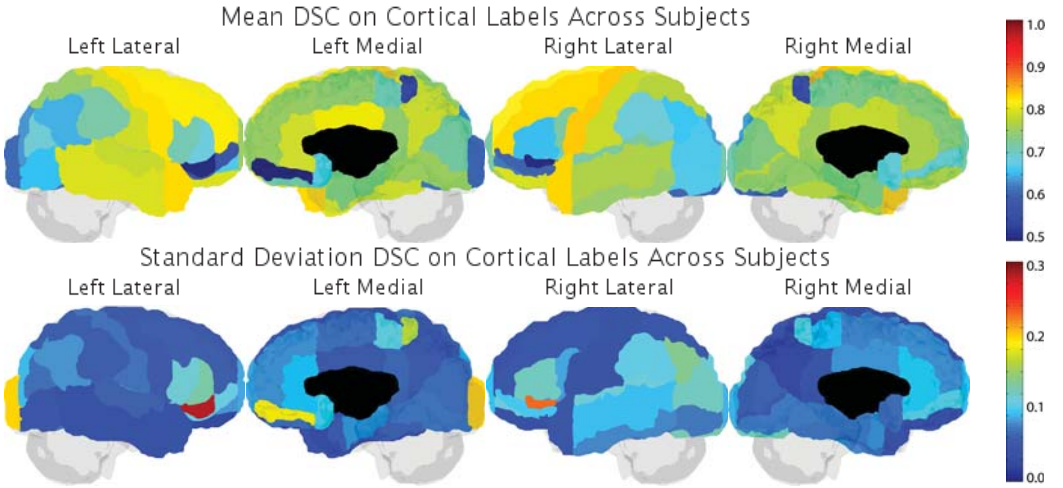
Attempt Number: 1

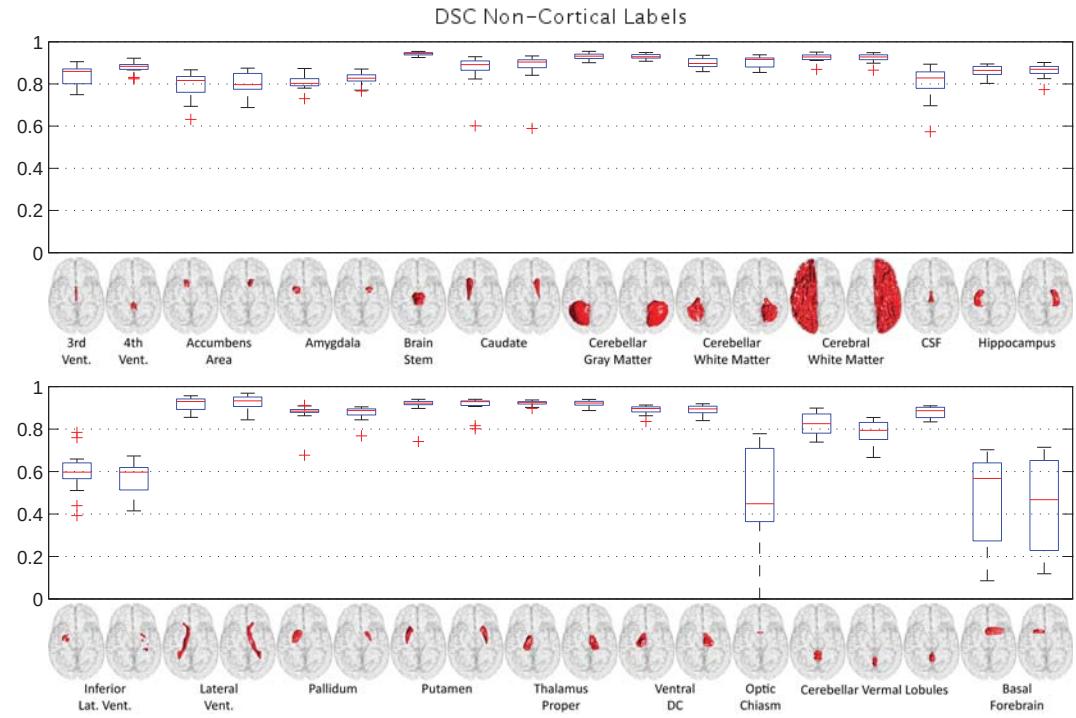
Date: 26-Jul-2012

Mean DSC Overall: 0.7499 +/- 0.0281
Rep: Mean DSC Overall: 0.7663 +/- 0.0119

Mean DSC Cortical: 0.7216 +/- 0.0311
Rep: Mean DSC Cortical: 0.7361 +/- 0.0141

Mean DSC Non-Cortical: 0.8271 +/- 0.0256
Rep: Mean DSC Non-Cortical: 0.8482 +/- 0.0080





SBIA_BrainROIMaps_JaccDet_IntCorr

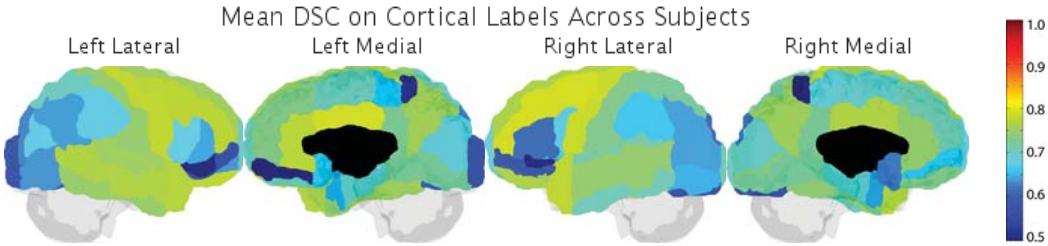
Attempt Number: 1

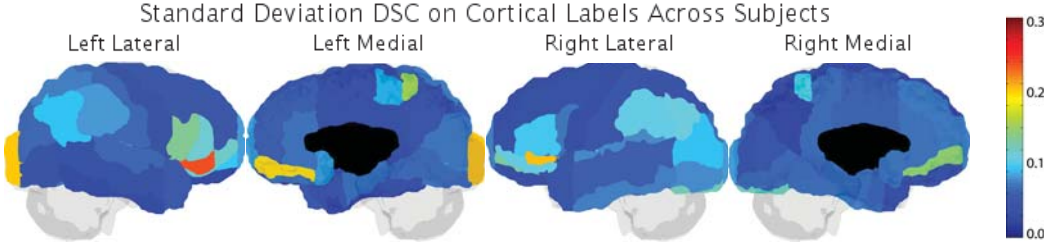
Date: 26-Jul-2012

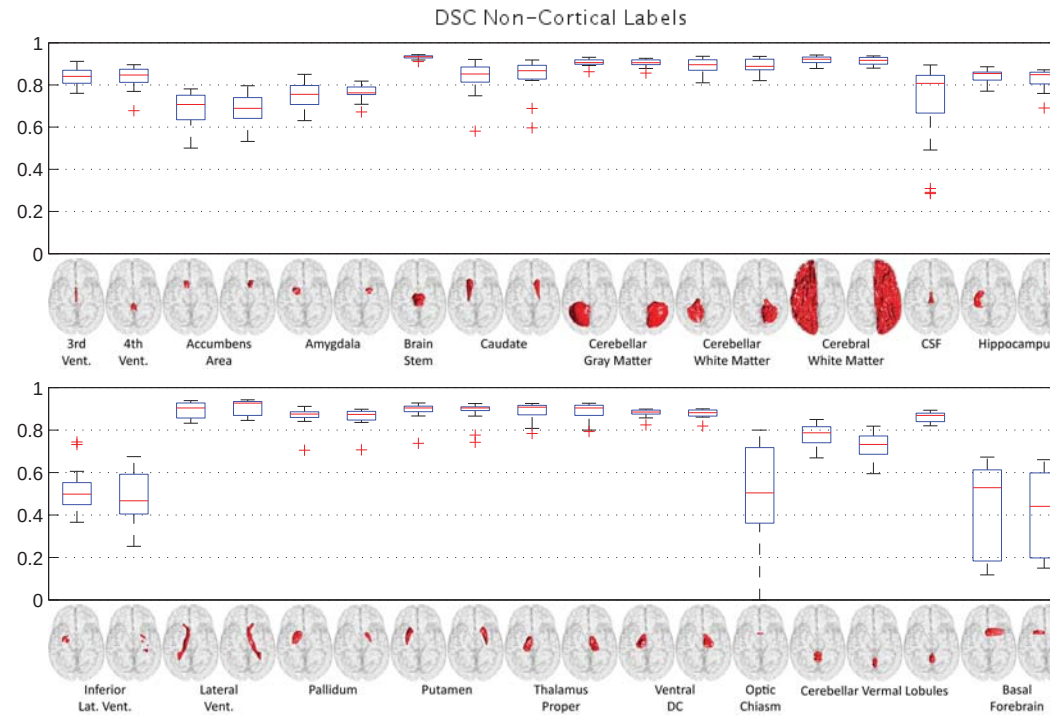
Mean DSC Overall: 0.7186 +/- 0.0238
Rep: Mean DSC Overall: 0.7265 +/- 0.0157

Mean DSC Cortical: 0.6913 +/- 0.0244
Rep: Mean DSC Cortical: 0.6932 +/- 0.0137

Mean DSC Non-Cortical: 0.7927 +/- 0.0334
Rep: Mean DSC Non-Cortical: 0.8172 +/- 0.0222







SBIA_BrainROIMaps_MV_IntCorr

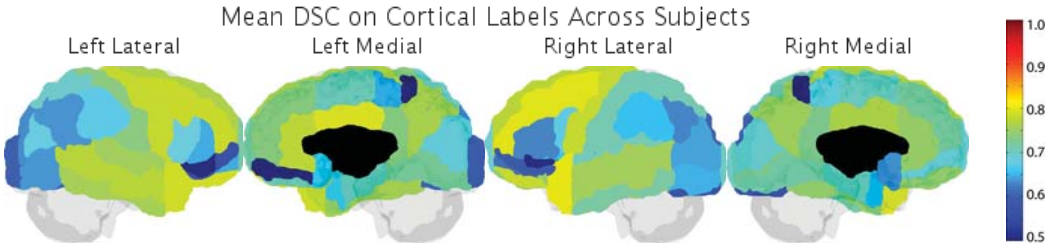
Attempt Number: 1

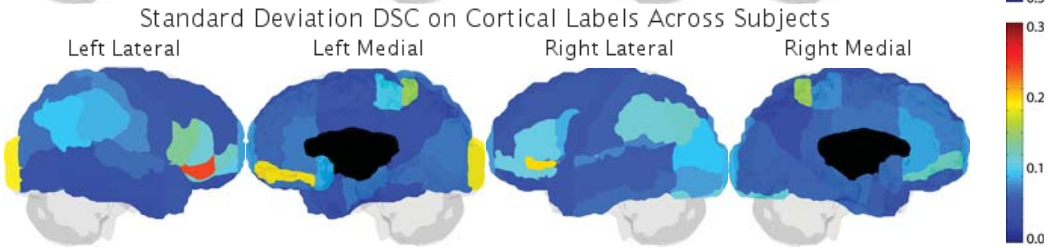
Date: 26-Jul-2012

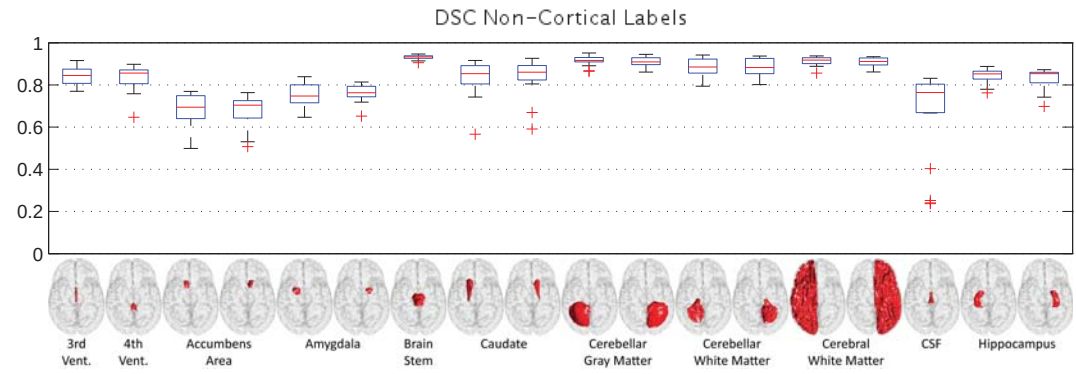
Mean DSC Overall: 0.7193 +/- 0.0284
Rep: Mean DSC Overall: 0.7313 +/- 0.0165

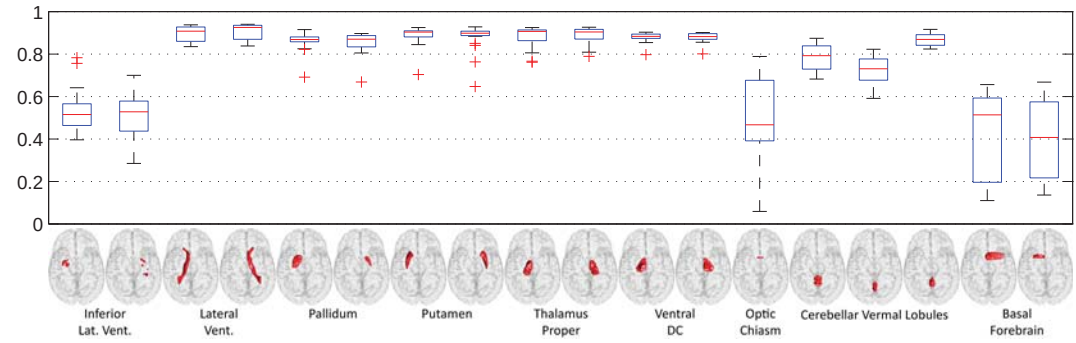
Mean DSC Cortical: 0.6933 +/- 0.0288
Rep: Mean DSC Cortical: 0.7007 +/- 0.0139

Mean DSC Non-Cortical: 0.7904 +/- 0.0347
Rep: Mean DSC Non-Cortical: 0.8145 +/- 0.0244









SBIA_SimMSVoting

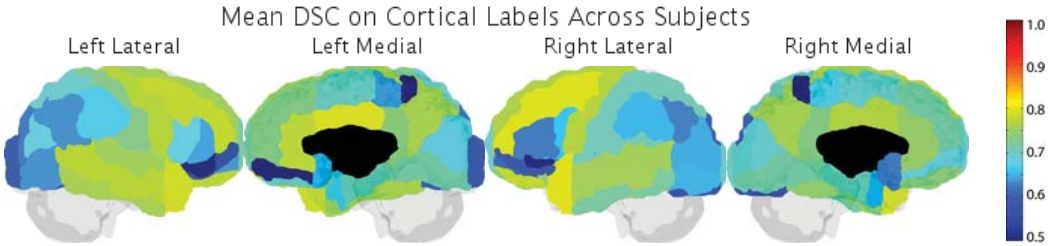
Attempt Number: 1

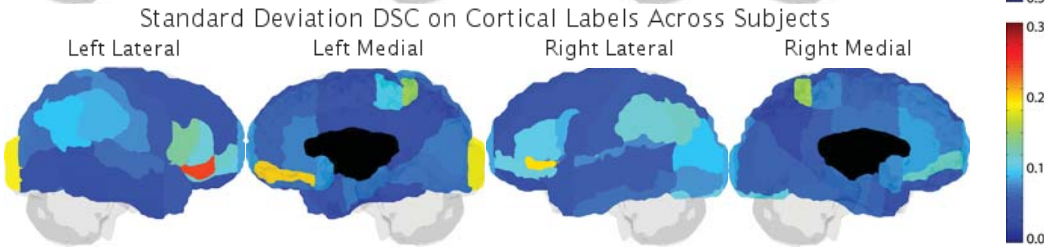
Date: 26-Jul-2012

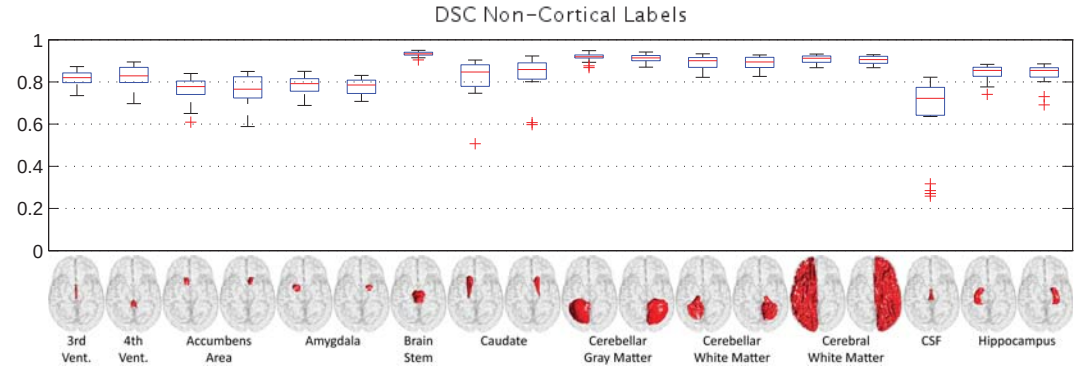
Mean DSC Overall: 0.7172 +/- 0.0286
Rep: Mean DSC Overall: 0.7283 +/- 0.0170

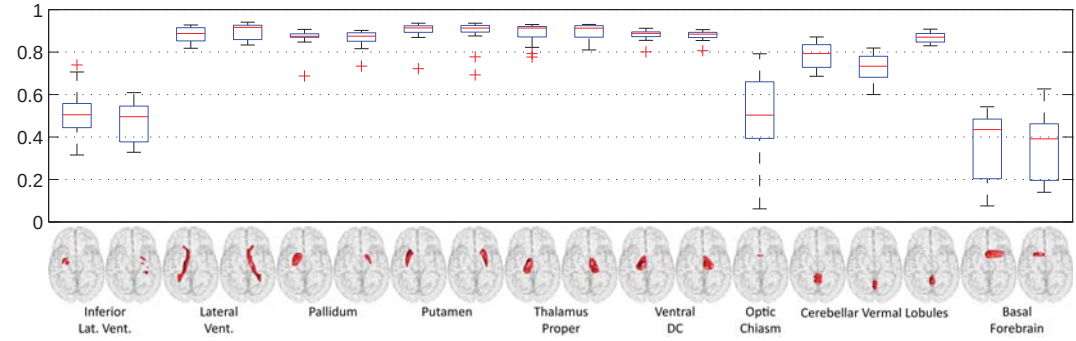
Mean DSC Cortical: 0.6898 +/- 0.0291
Rep: Mean DSC Cortical: 0.6977 +/- 0.0144

Mean DSC Non-Cortical: 0.7918 +/- 0.0326
Rep: Mean DSC Non-Cortical: 0.8116 +/- 0.0244









SBIA_SimRank+NormMS

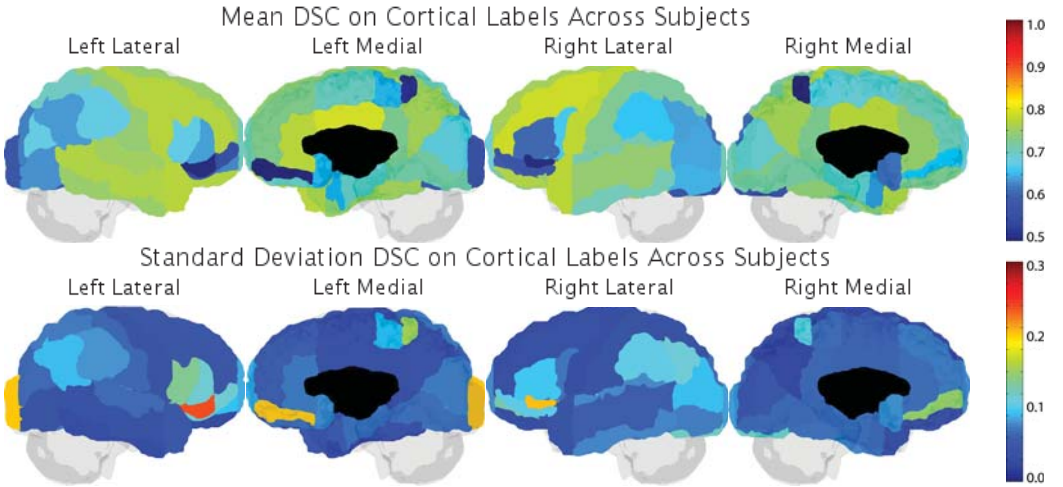
Attempt Number: 1

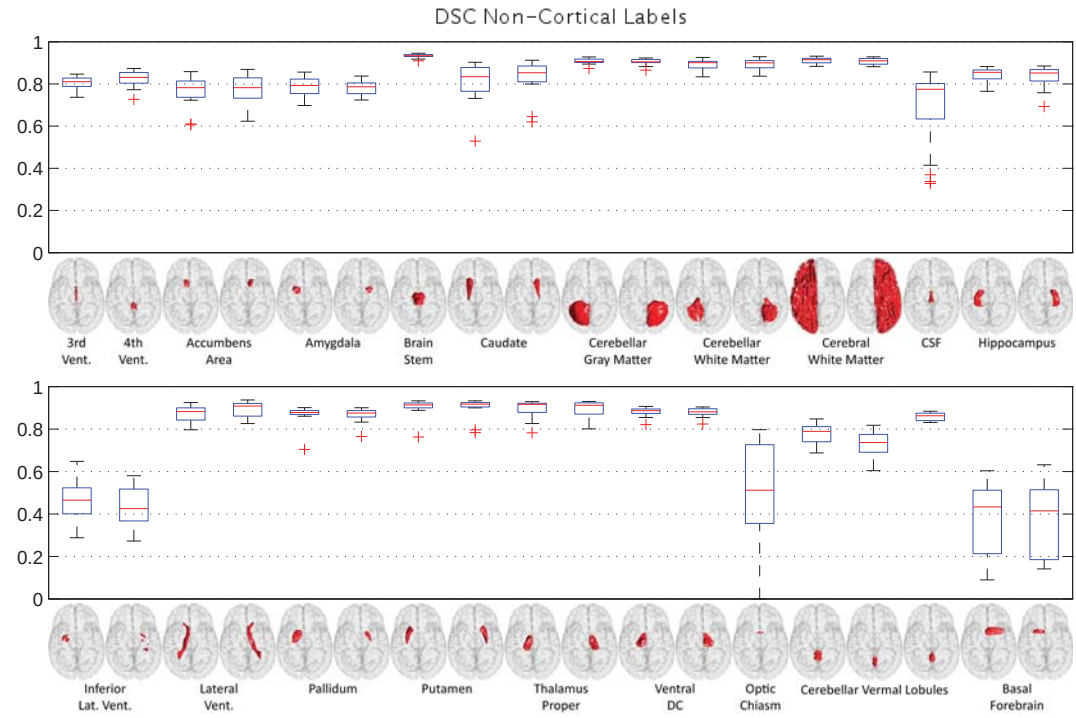
Date: 26-Jul-2012

Mean DSC Overall: 0.7162 +/- 0.0235
Rep: Mean DSC Overall: 0.7236 +/- 0.0152

Mean DSC Cortical: 0.6884 +/- 0.0239
Rep: Mean DSC Cortical: 0.6909 +/- 0.0131

Mean DSC Non-Cortical: 0.7919 +/- 0.0308
Rep: Mean DSC Non-Cortical: 0.8125 +/- 0.0212





SBIA_SimRank+NormMS+WtROI

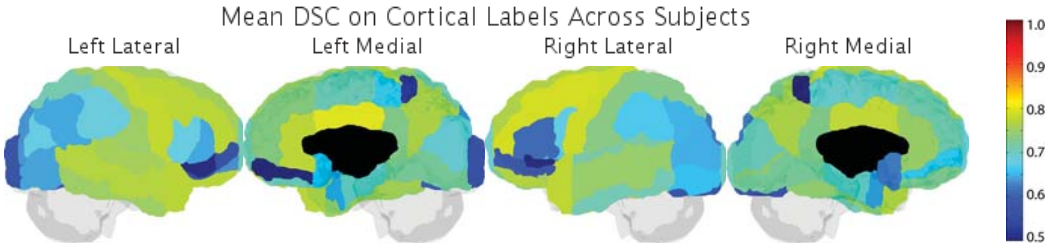
Attempt Number: 1

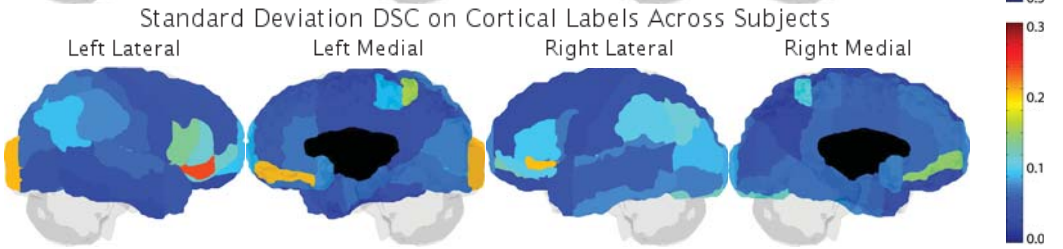
Date: 26-Jul-2012

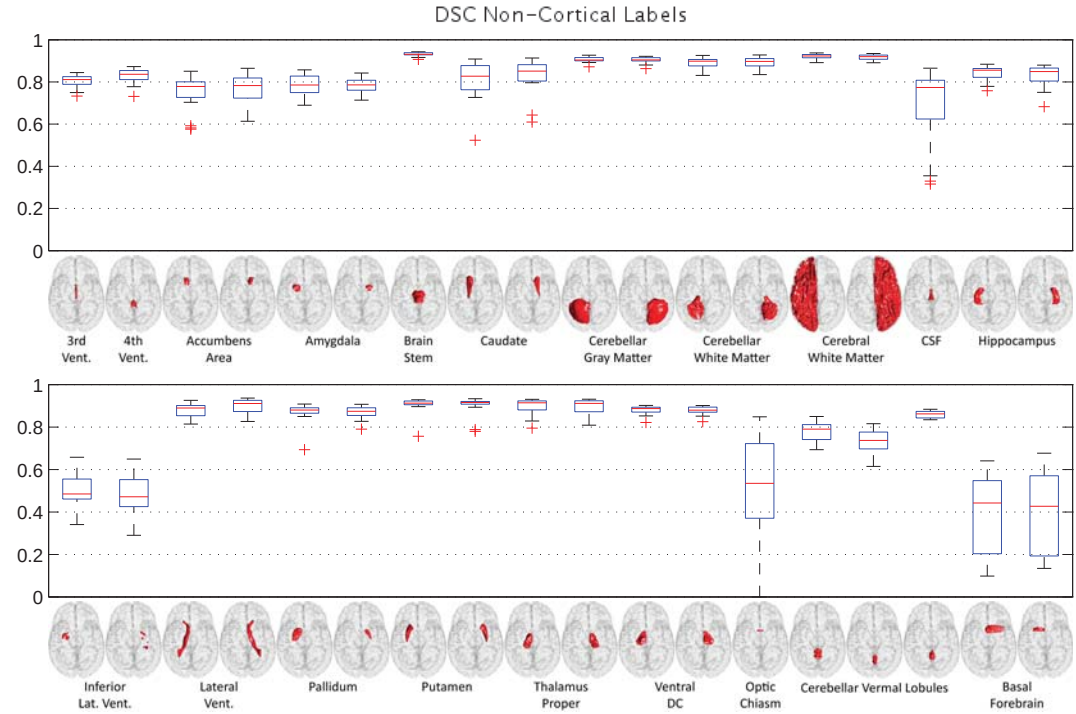
Mean DSC Overall: 0.7212 +/- 0.0225
Rep: Mean DSC Overall: 0.7282 +/- 0.0135

Mean DSC Cortical: 0.6940 +/- 0.0226
Rep: Mean DSC Cortical: 0.6957 +/- 0.0112

Mean DSC Non-Cortical: 0.7953 +/- 0.0308
Rep: Mean DSC Non-Cortical: 0.8164 +/- 0.0202







SpatialSTAPLE

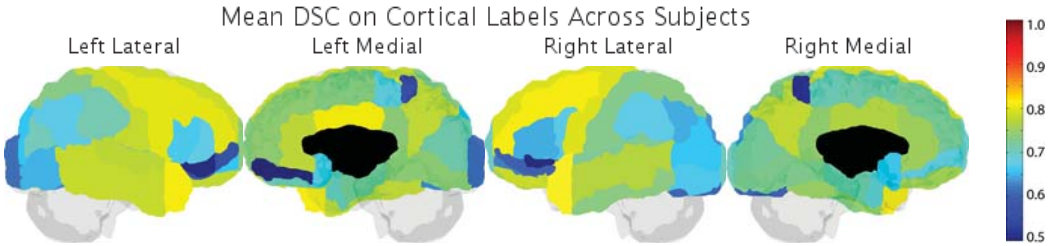
Attempt Number: 1

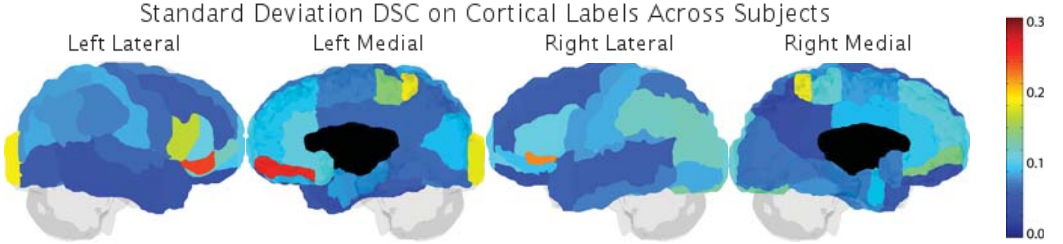
Date: 26-Jul-2012

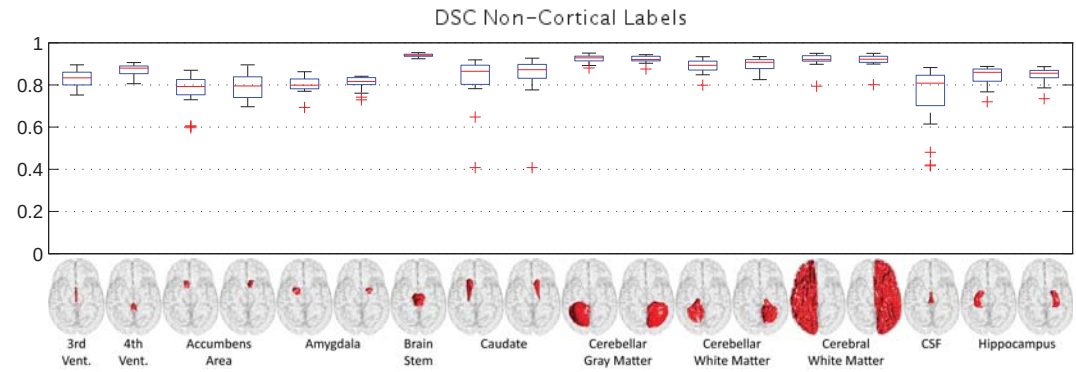
Mean DSC Overall: 0.7372 +/- 0.0377
Rep: Mean DSC Overall: 0.7576 +/- 0.0092

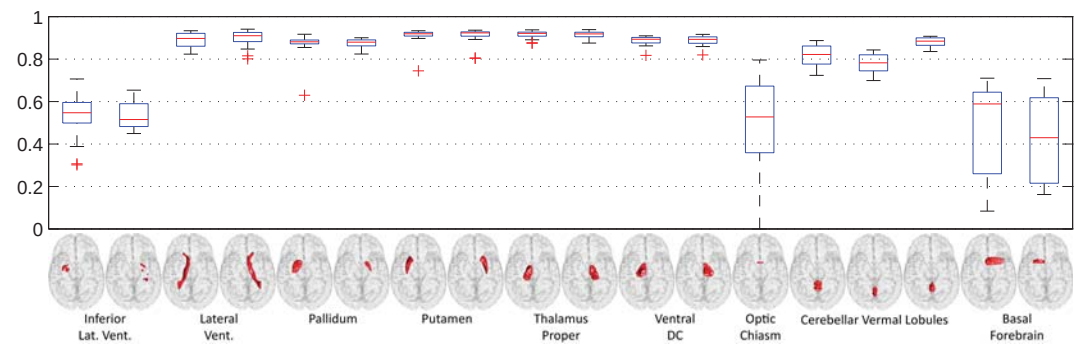
Mean DSC Cortical: 0.7093 +/- 0.0410
Rep: Mean DSC Cortical: 0.7278 +/- 0.0084

Mean DSC Non-Cortical: 0.8130 +/- 0.0341
Rep: Mean DSC Non-Cortical: 0.8388 +/- 0.0157









STEPS

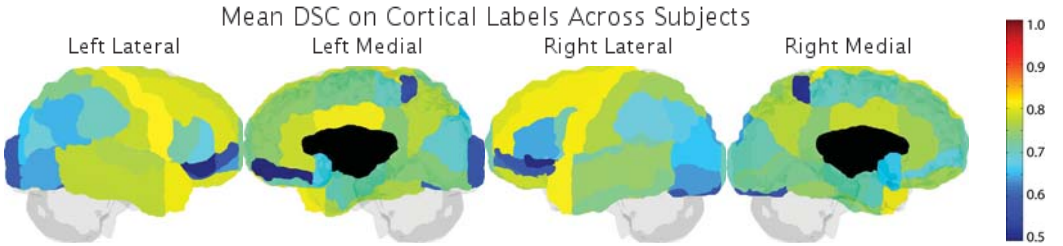
Attempt Number: 2

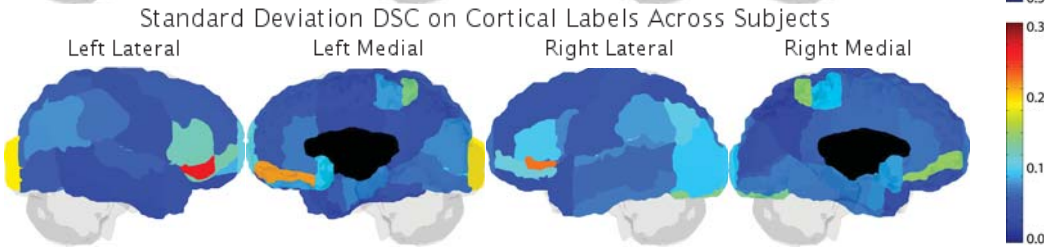
Date: 26-Jul-2012

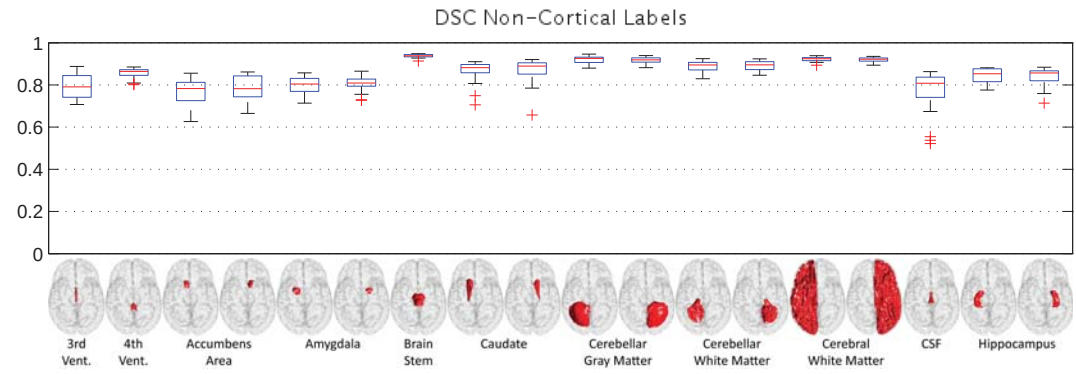
Mean DSC Overall: 0.7372 +/- 0.0217
Rep: Mean DSC Overall: 0.7500 +/- 0.0062

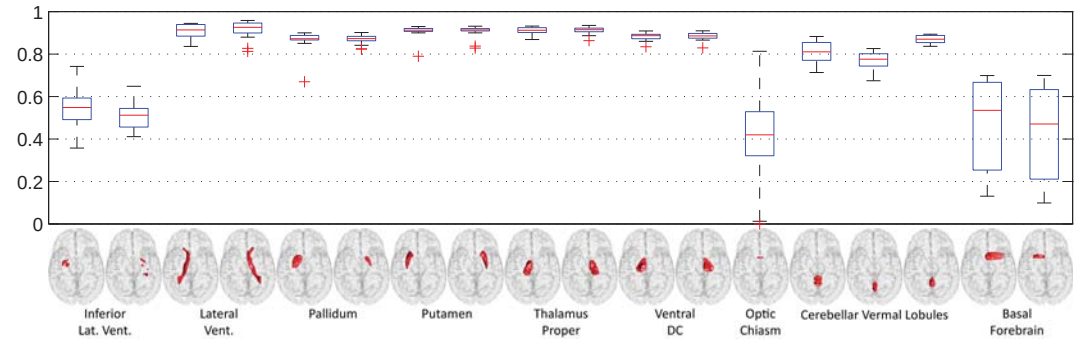
Mean DSC Cortical: 0.7107 +/- 0.0213
Rep: Mean DSC Cortical: 0.7202 +/- 0.0050

Mean DSC Non-Cortical: 0.8095 +/- 0.0270
Rep: Mean DSC Non-Cortical: 0.8311 +/- 0.0101









UNC_NIRAL

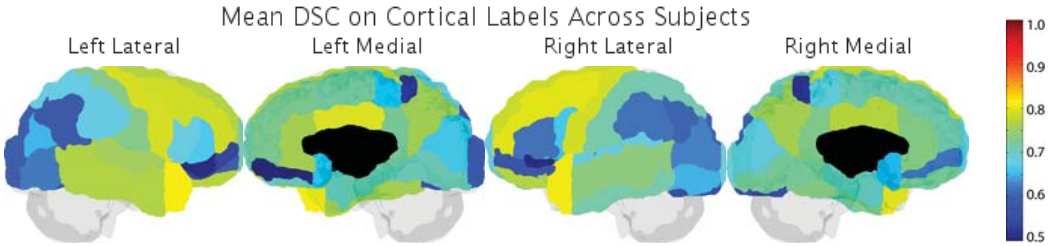
Attempt Number: 1

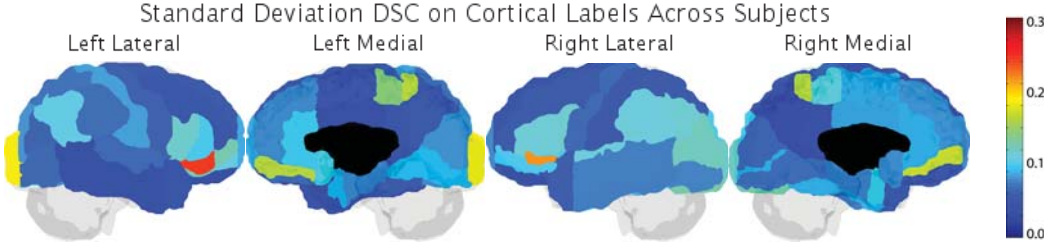
Date: 03-Aug-2012

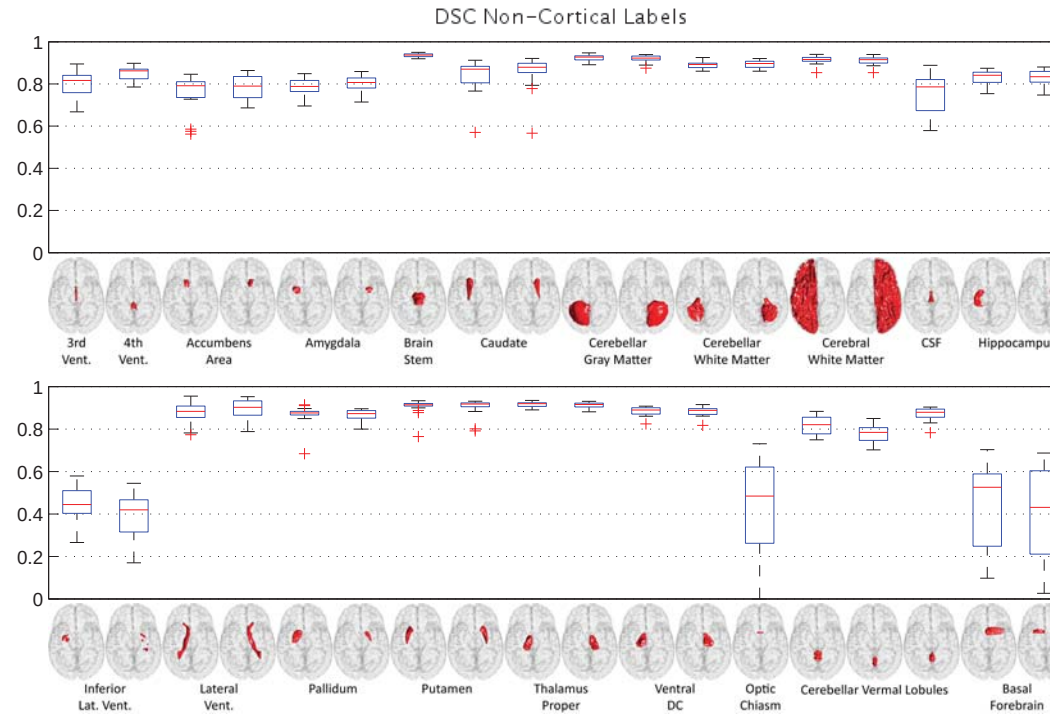
Mean DSC Overall: 0.7171 +/- 0.0320
Rep: Mean DSC Overall: 0.7350 +/- 0.0142

Mean DSC Cortical: 0.6869 +/- 0.0343
Rep: Mean DSC Cortical: 0.7030 +/- 0.0163

Mean DSC Non-Cortical: 0.7992 +/- 0.0292
Rep: Mean DSC Non-Cortical: 0.8220 +/- 0.0100







BIC-IPL

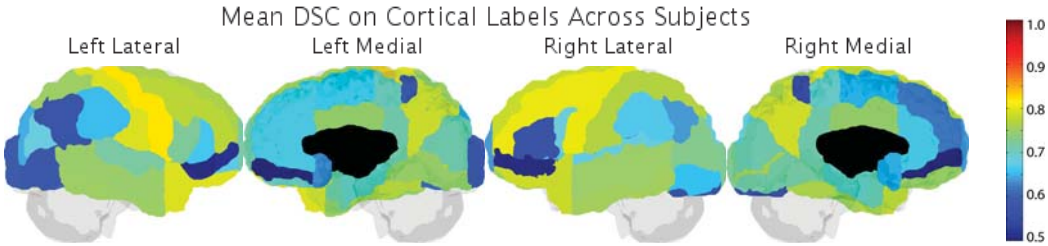
Attempt Number: 1

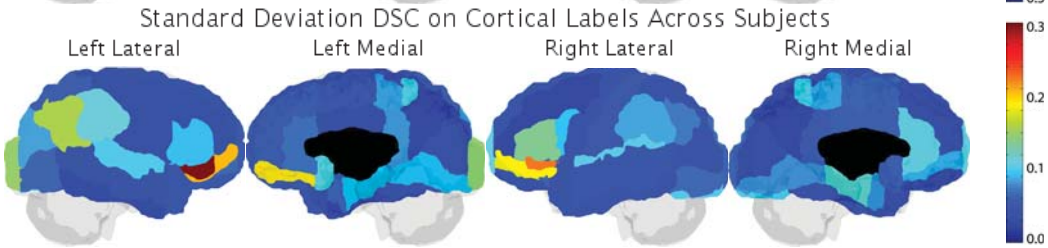
Date: 31-Jul-2012

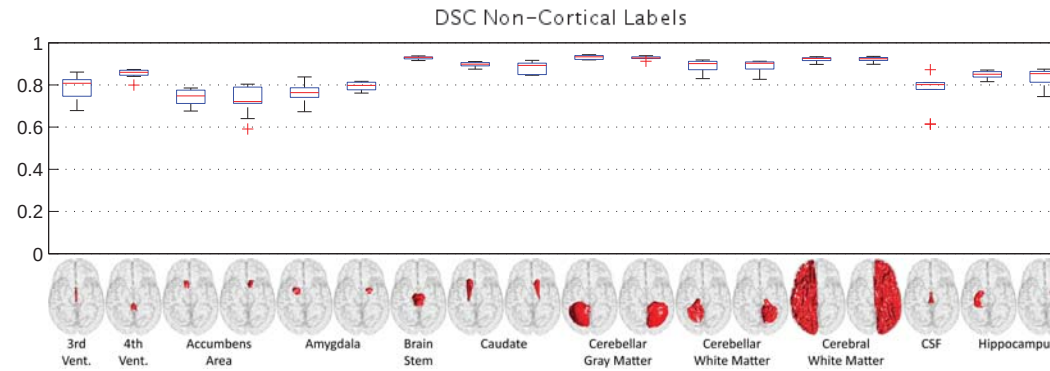
Mean DSC Overall: 0.7225 +/- 0.0128
Rep: Mean DSC Overall: 0.7225 +/- 0.0120

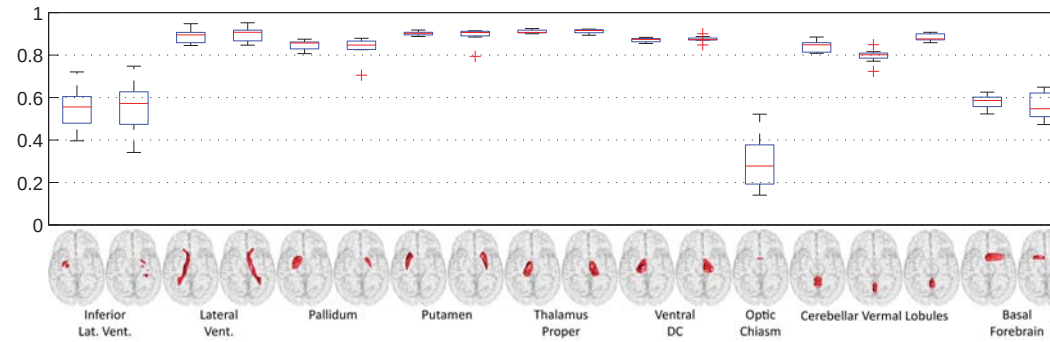
Mean DSC Cortical: 0.6900 +/- 0.0155
Rep: Mean DSC Cortical: 0.6900 +/- 0.0147

Mean DSC Non-Cortical: 0.8112 +/- 0.0089
Rep: Mean DSC Non-Cortical: 0.8112 +/- 0.0069









BIC-IPL-HR

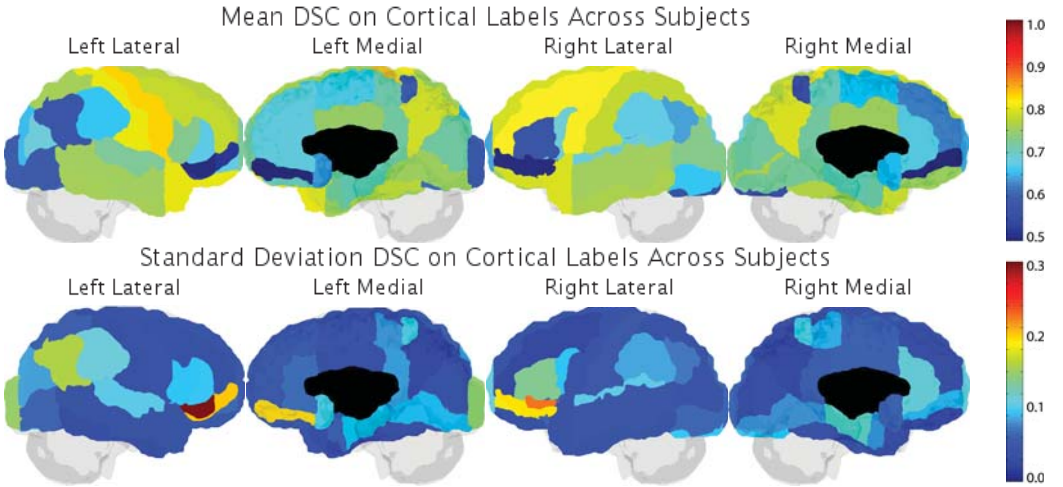
Attempt Number: 1

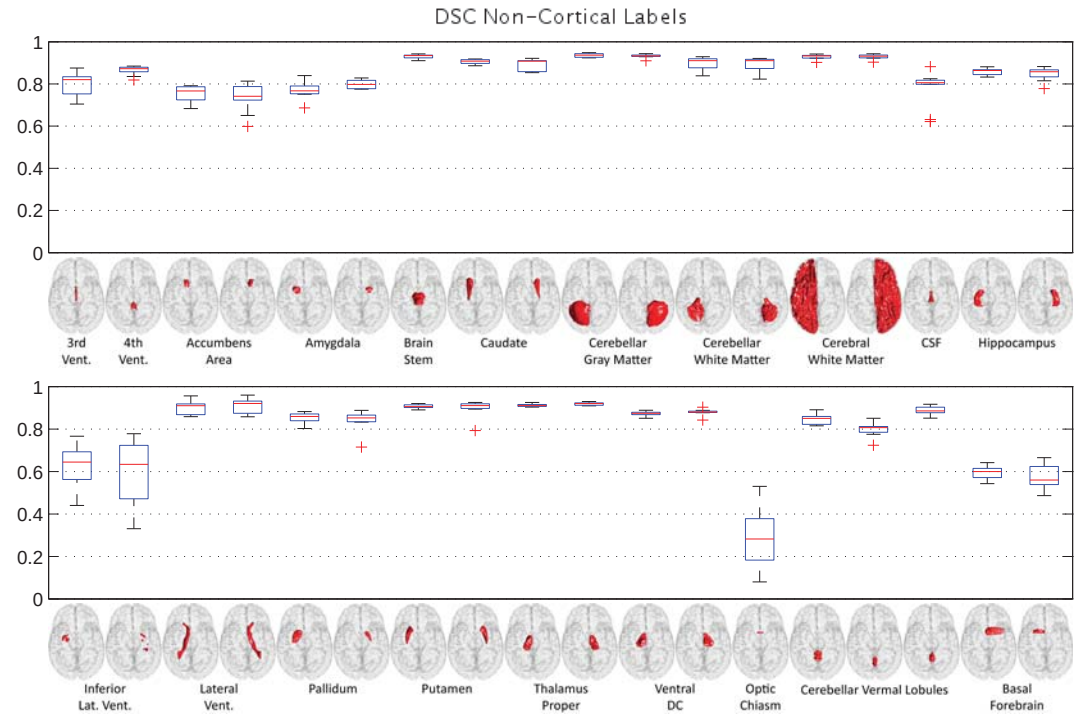
Date: 31-Jul-2012

Mean DSC Overall: 0.7299 +/- 0.0137
Rep: Mean DSC Overall: 0.7299 +/- 0.0128

Mean DSC Cortical: 0.6965 +/- 0.0164
Rep: Mean DSC Cortical: 0.6965 +/- 0.0155

Mean DSC Non-Cortical: 0.8209 +/- 0.0101
Rep: Mean DSC Non-Cortical: 0.8209 +/- 0.0087





CIS_JHU

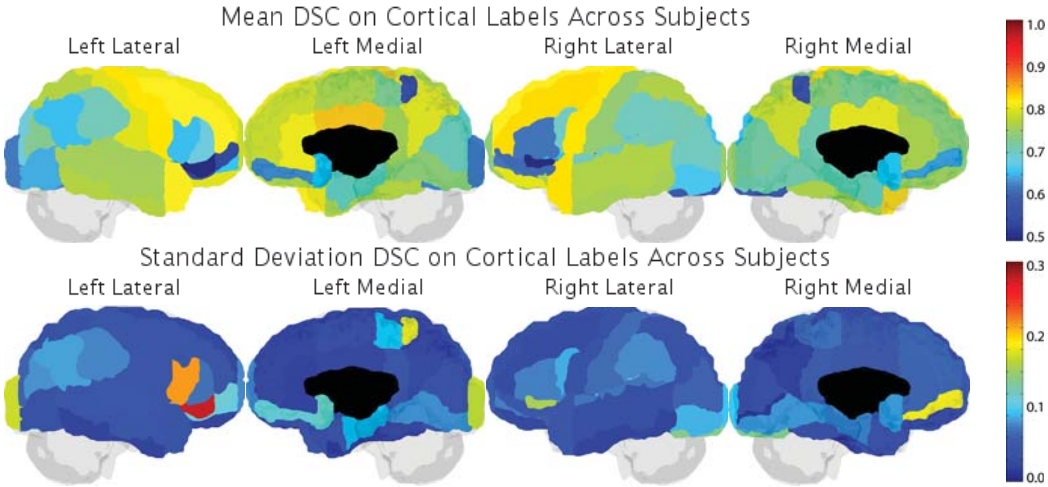
Attempt Number: 1

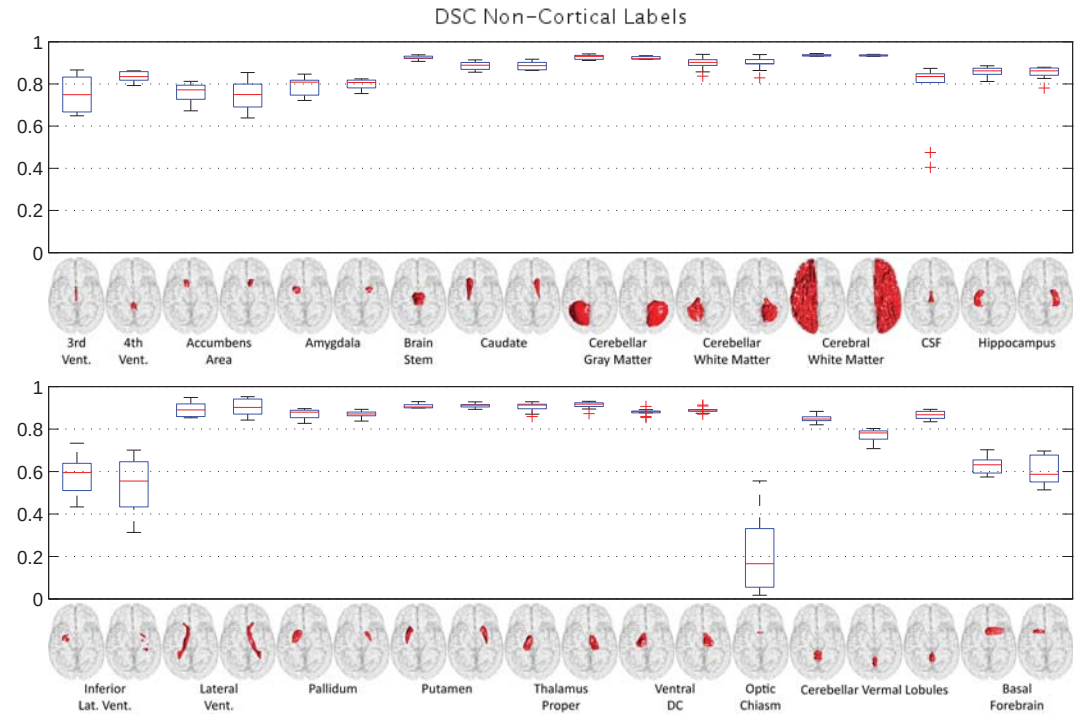
Date: 06-Jul-2012

Mean DSC Overall: 0.7440 +/- 0.0059
Rep: Mean DSC Overall: 0.7440 +/- 0.0042

Mean DSC Cortical: 0.7178 +/- 0.0064
Rep: Mean DSC Cortical: 0.7178 +/- 0.0051

Mean DSC Non-Cortical: 0.8151 +/- 0.0108
Rep: Mean DSC Non-Cortical: 0.8151 +/- 0.0076





CRL_MV_ANTS

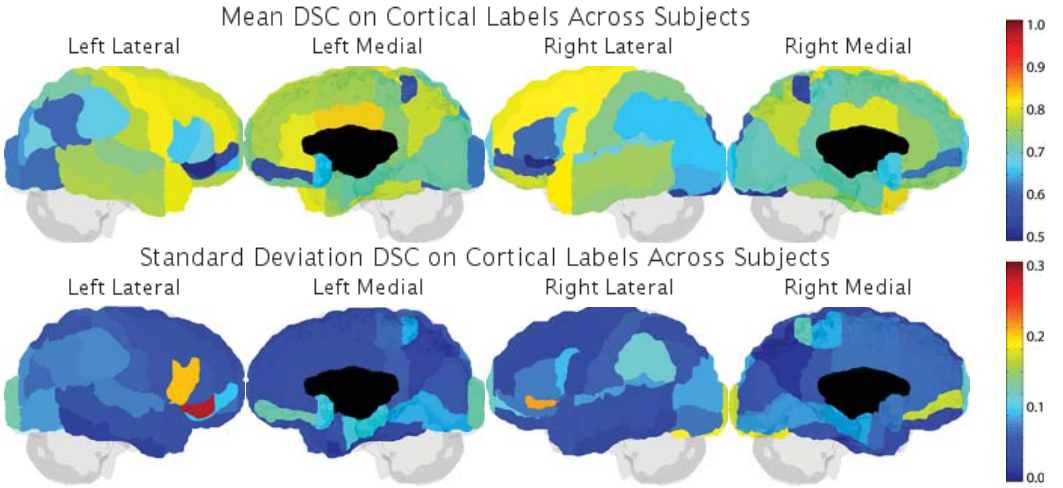
Attempt Number: 1

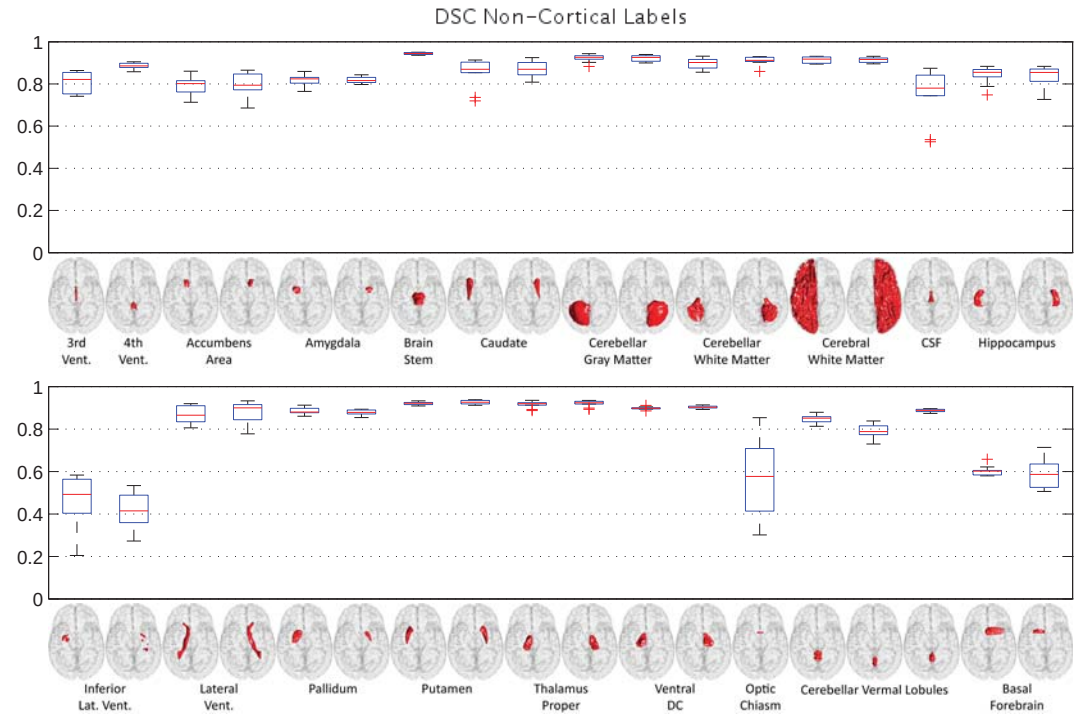
Date: 09-Jul-2012

Mean DSC Overall: 0.7360 +/- 0.0102
Rep: Mean DSC Overall: 0.7360 +/- 0.0100

Mean DSC Cortical: 0.7038 +/- 0.0105
Rep: Mean DSC Cortical: 0.7038 +/- 0.0098

Mean DSC Non-Cortical: 0.8236 +/- 0.0174
Rep: Mean DSC Non-Cortical: 0.8236 +/- 0.0174





CRL_MV_ANTS+Baloo

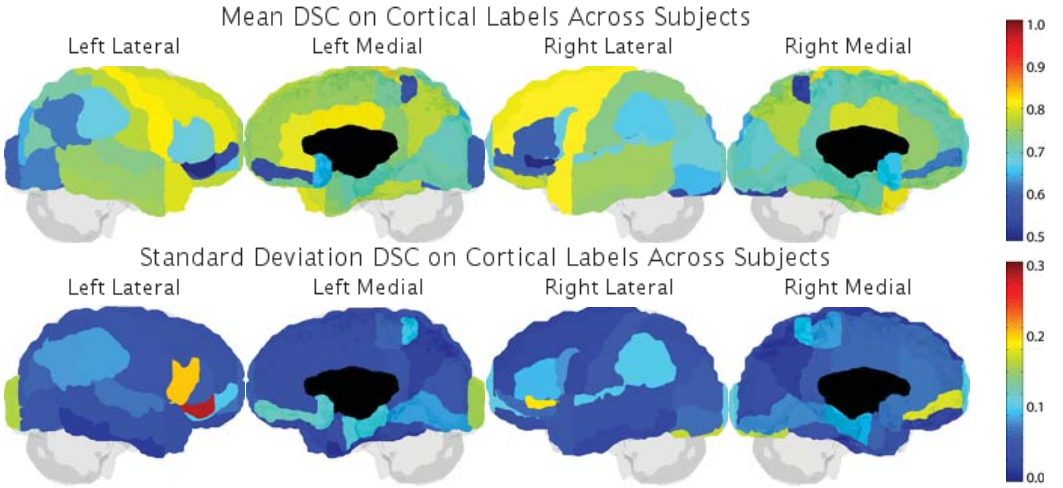
Attempt Number: 1

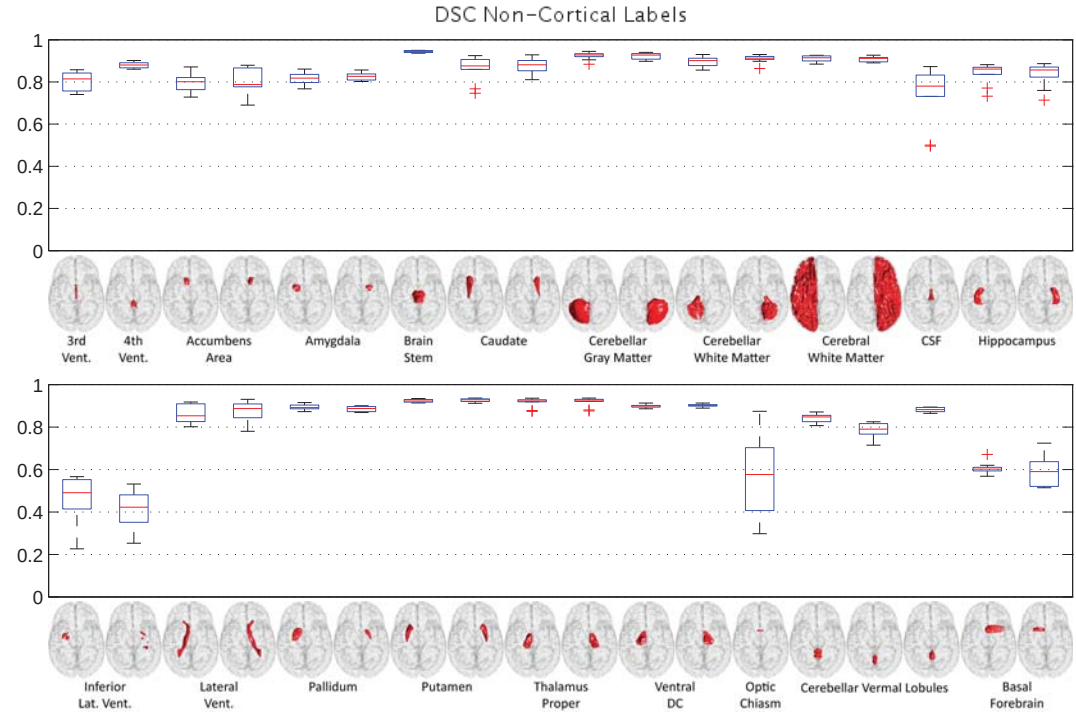
Date: 09-Jul-2012

Mean DSC Overall: 0.7378 +/- 0.0098
Rep: Mean DSC Overall: 0.7378 +/- 0.0094

Mean DSC Cortical: 0.7063 +/- 0.0086
Rep: Mean DSC Cortical: 0.7063 +/- 0.0073

Mean DSC Non-Cortical: 0.8234 +/- 0.0180
Rep: Mean DSC Non-Cortical: 0.8234 +/- 0.0183





CRL_Probabilistic_STAPLE_ANTS

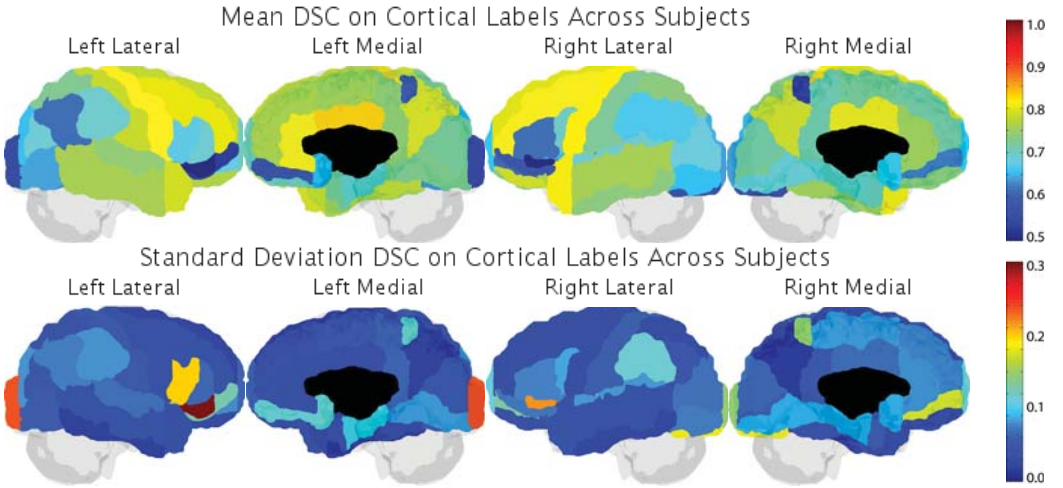
Attempt Number: 1

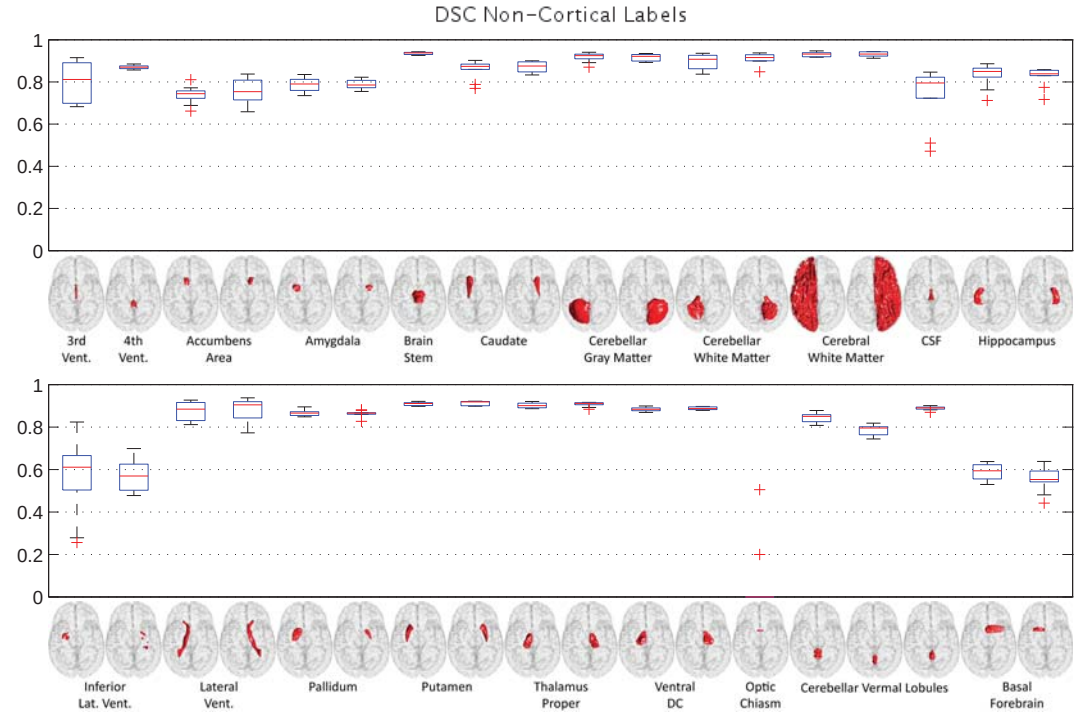
Date: 09-Jul-2012

Mean DSC Overall: 0.7334 +/- 0.0090
Rep: Mean DSC Overall: 0.7334 +/- 0.0087

Mean DSC Cortical: 0.7062 +/- 0.0096
Rep: Mean DSC Cortical: 0.7062 +/- 0.0084

Mean DSC Non-Cortical: 0.8075 +/- 0.0166
Rep: Mean DSC Non-Cortical: 0.8075 +/- 0.0161





CRL_Probabilistic_STAPLE_ANTS+Baloo

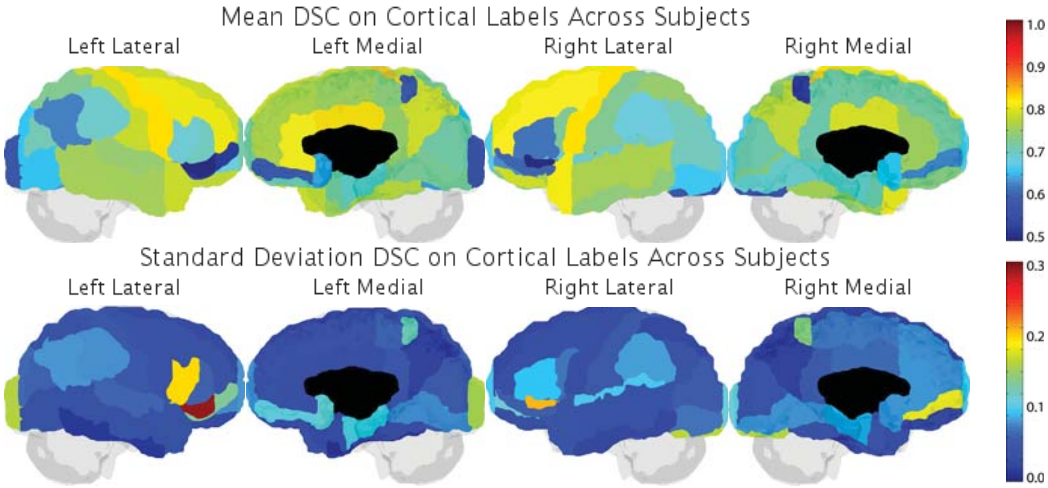
Attempt Number: 1

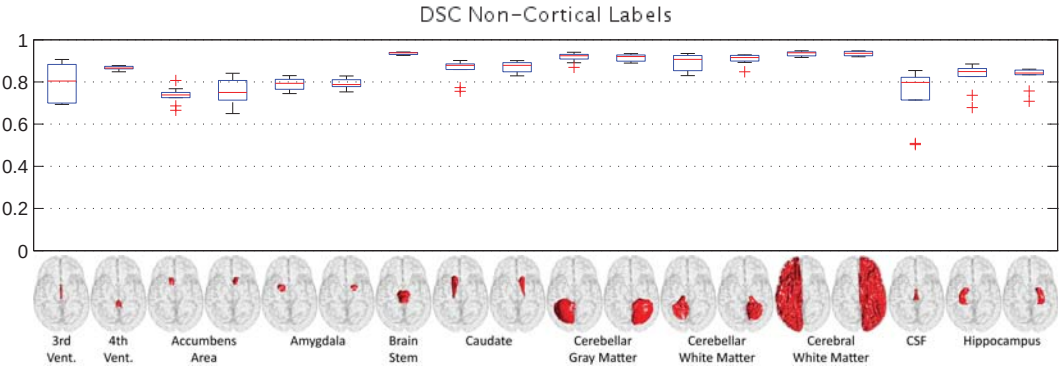
Date: 09-Jul-2012

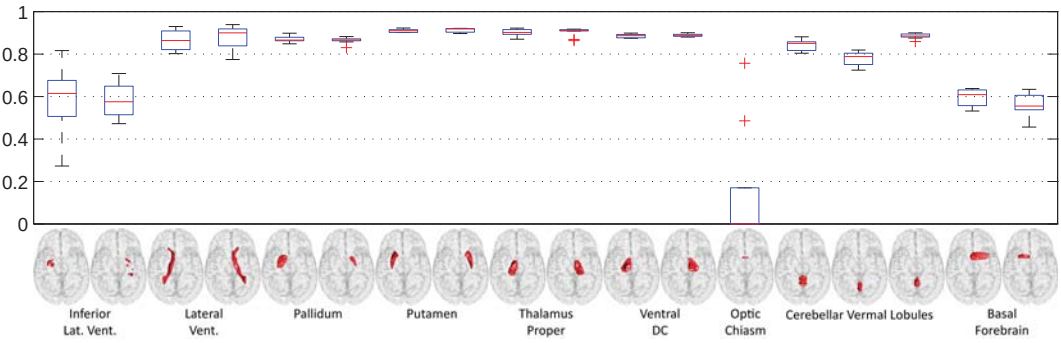
Mean DSC Overall: 0.7372 +/- 0.0106
Rep: Mean DSC Overall: 0.7372 +/- 0.0103

Mean DSC Cortical: 0.7108 +/- 0.0093
Rep: Mean DSC Cortical: 0.7108 +/- 0.0074

Mean DSC Non-Cortical: 0.8092 +/- 0.0203
Rep: Mean DSC Non-Cortical: 0.8092 +/- 0.0196







CRL_STAPLE_ANTS

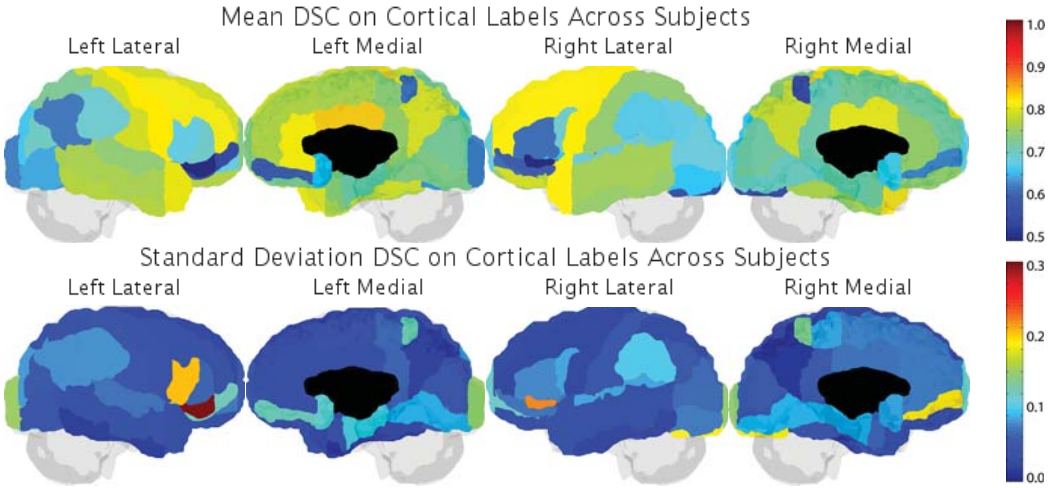
Attempt Number: 1

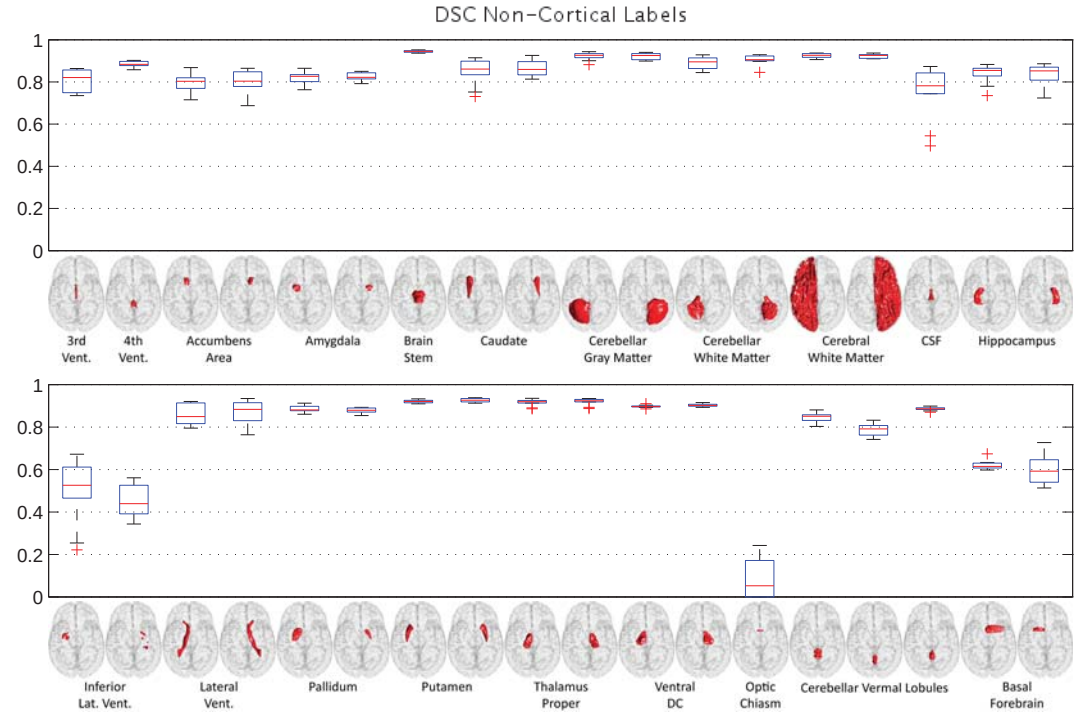
Date: 09-Jul-2012

Mean DSC Overall: 0.7393 +/- 0.0093
Rep: Mean DSC Overall: 0.7393 +/- 0.0086

Mean DSC Cortical: 0.7125 +/- 0.0095
Rep: Mean DSC Cortical: 0.7125 +/- 0.0075

Mean DSC Non-Cortical: 0.8121 +/- 0.0165
Rep: Mean DSC Non-Cortical: 0.8121 +/- 0.0167





CRL_STAPLE_ANTS+Baloo

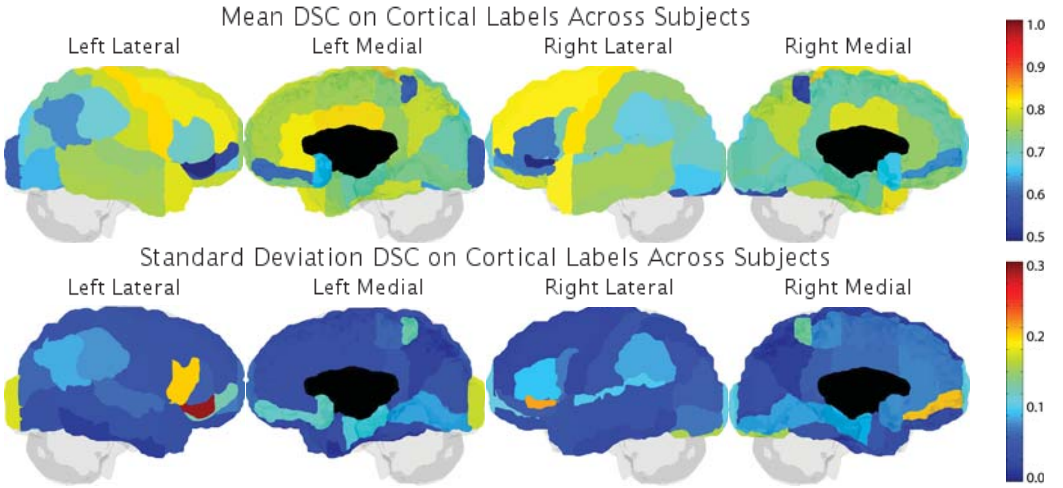
Attempt Number: 1

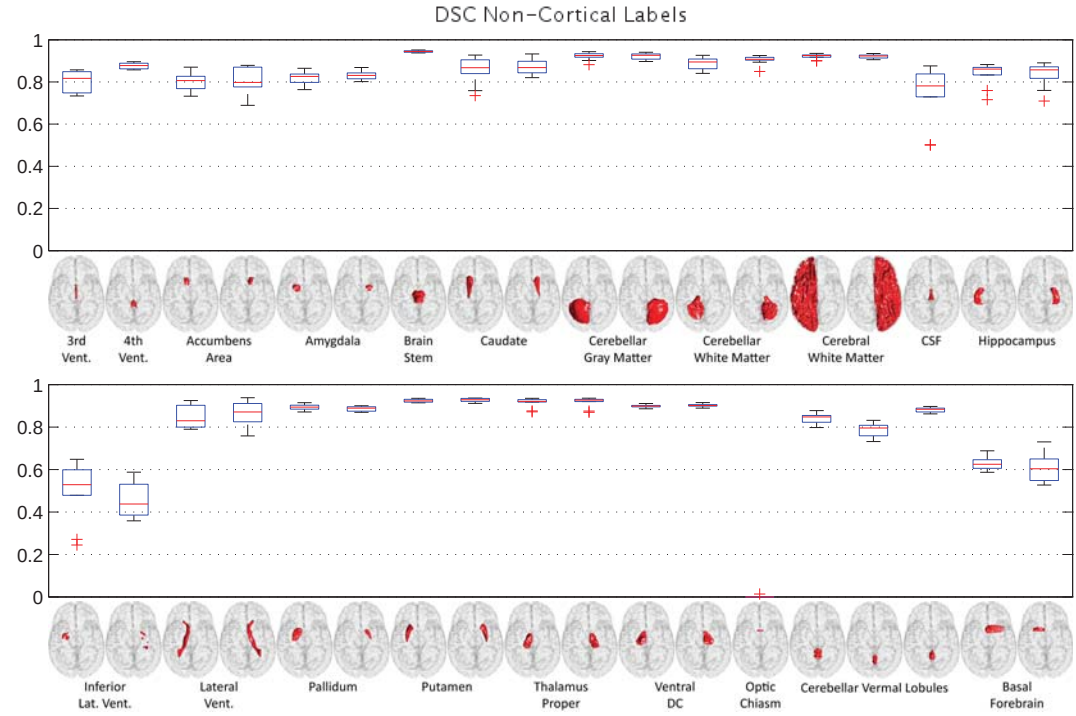
Date: 09-Jul-2012

Mean DSC Overall: 0.7412 +/- 0.0099
Rep: Mean DSC Overall: 0.7412 +/- 0.0092

Mean DSC Cortical: 0.7158 +/- 0.0090
Rep: Mean DSC Cortical: 0.7158 +/- 0.0067

Mean DSC Non-Cortical: 0.8103 +/- 0.0172
Rep: Mean DSC Non-Cortical: 0.8103 +/- 0.0179





CRL_Weighted_STAPLE_ANTS

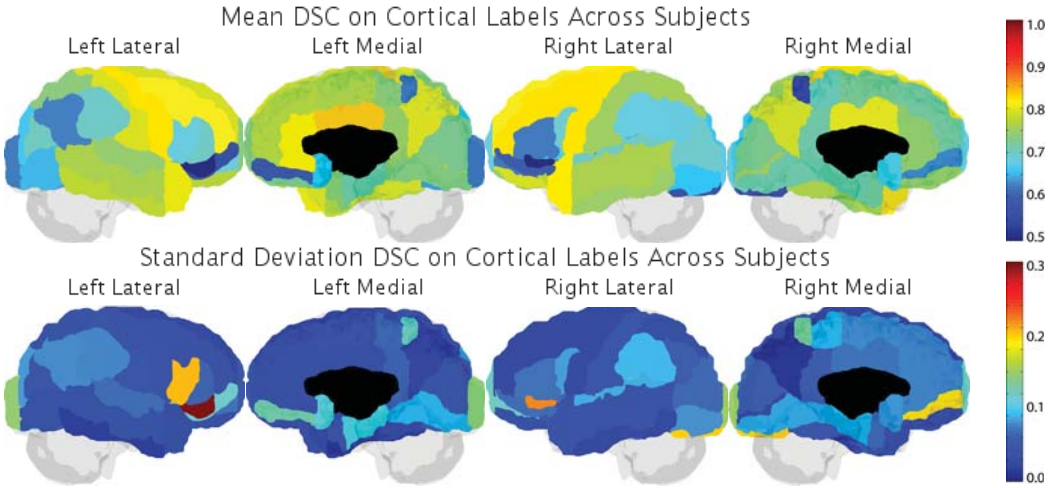
Attempt Number: 1

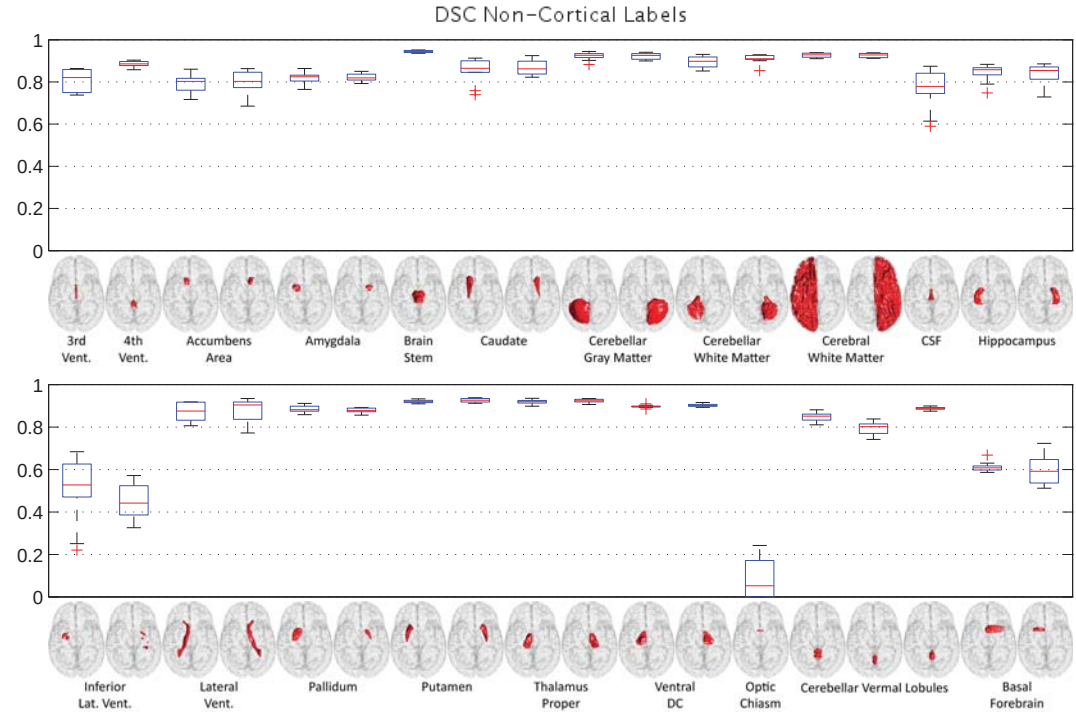
Date: 09-Jul-2012

Mean DSC Overall: 0.7424 +/- 0.0091
Rep: Mean DSC Overall: 0.7424 +/- 0.0085

Mean DSC Cortical: 0.7160 +/- 0.0092
Rep: Mean DSC Cortical: 0.7160 +/- 0.0077

Mean DSC Non-Cortical: 0.8142 +/- 0.0151
Rep: Mean DSC Non-Cortical: 0.8142 +/- 0.0151





CRL_Weighted_STAPLE_ANTS+Baloo

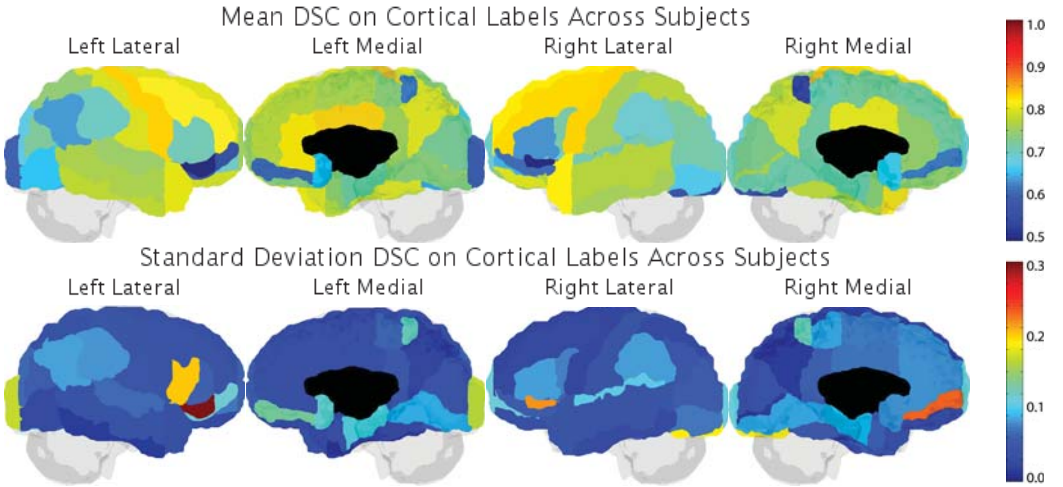
Attempt Number: 1

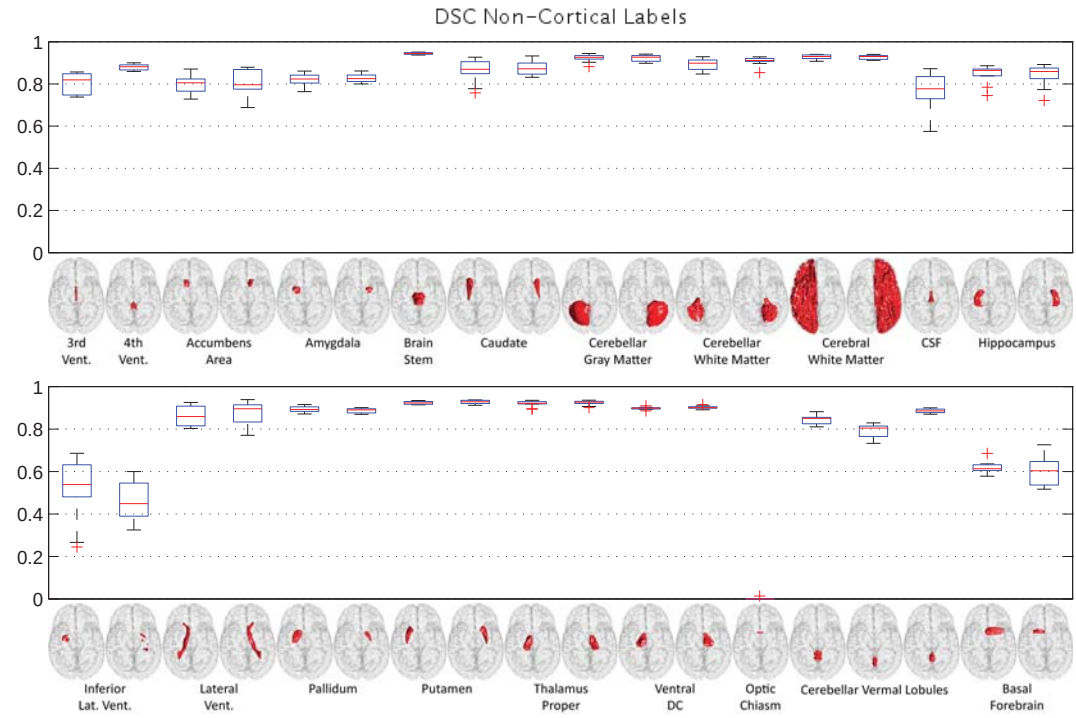
Date: 09-Jul-2012

Mean DSC Overall: 0.7470 +/- 0.0096
Rep: Mean DSC Overall: 0.7470 +/- 0.0091

Mean DSC Cortical: 0.7225 +/- 0.0092
Rep: Mean DSC Cortical: 0.7225 +/- 0.0076

Mean DSC Non-Cortical: 0.8137 +/- 0.0152
Rep: Mean DSC Non-Cortical: 0.8137 +/- 0.0157





DISPATCH

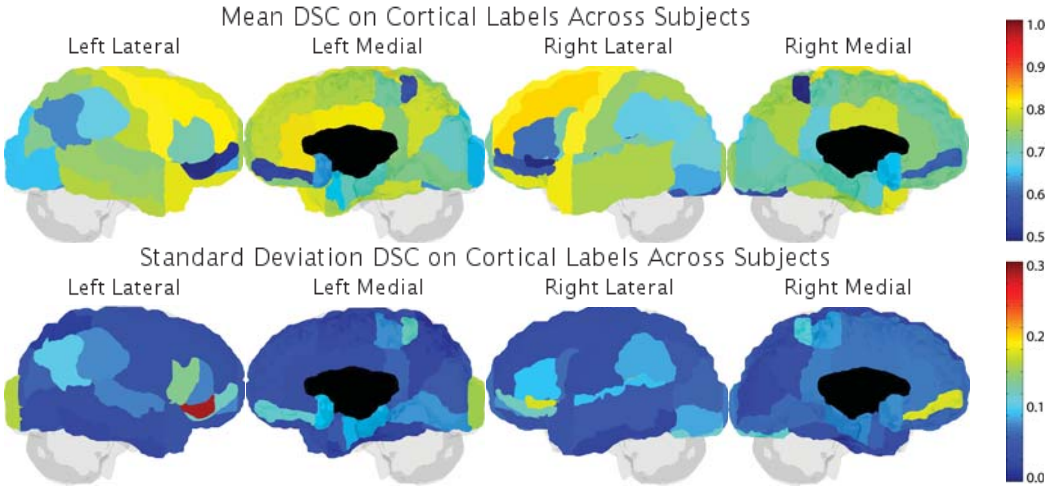
Attempt Number: 1

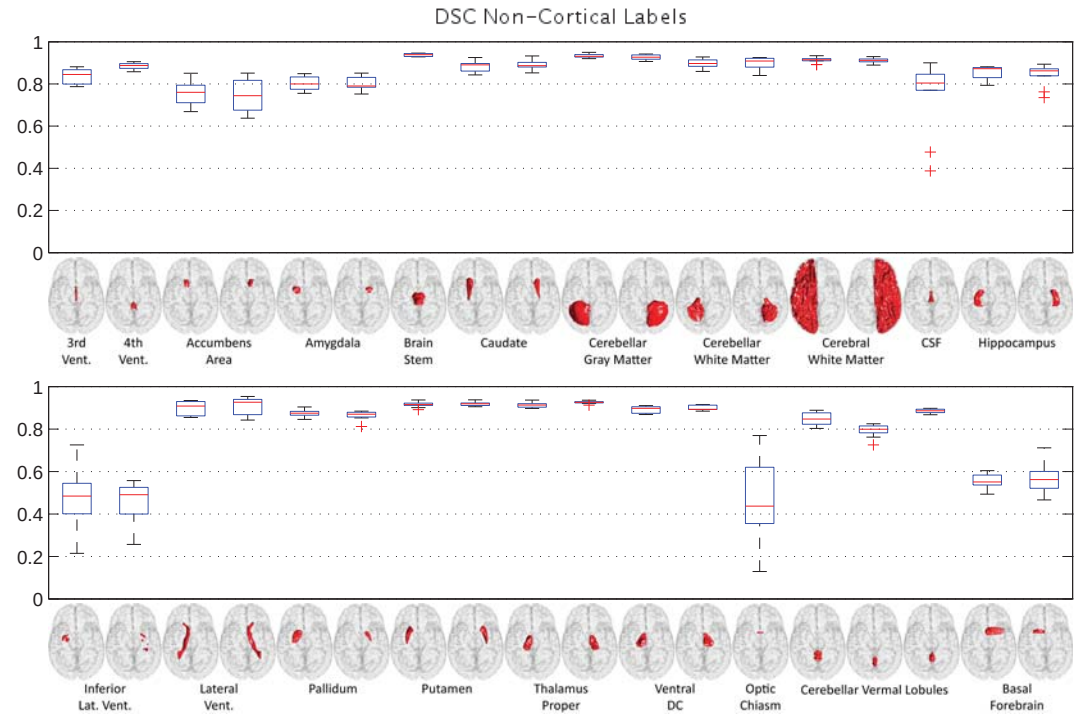
Date: 06-Jul-2012

Mean DSC Overall: 0.7388 +/- 0.0095
Rep: Mean DSC Overall: 0.7388 +/- 0.0097

Mean DSC Cortical: 0.7091 +/- 0.0099
Rep: Mean DSC Cortical: 0.7091 +/- 0.0097

Mean DSC Non-Cortical: 0.8199 +/- 0.0113
Rep: Mean DSC Non-Cortical: 0.8199 +/- 0.0098





MALP_EM

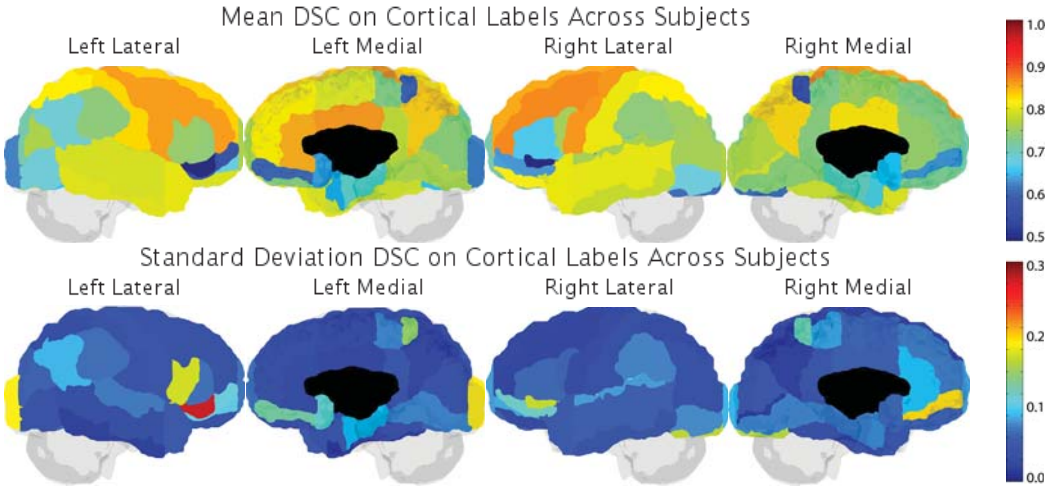
Attempt Number: 3

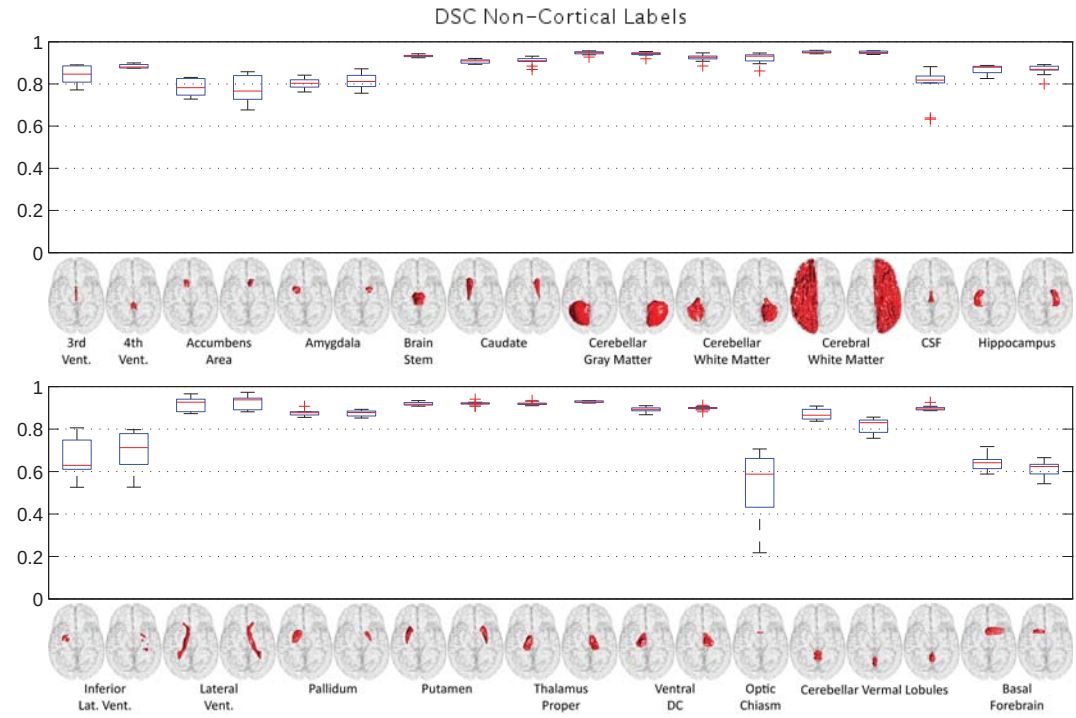
Date: 05-Jul-2012

Mean DSC Overall: 0.7708 +/- 0.0074
Rep: Mean DSC Overall: 0.7708 +/- 0.0058

Mean DSC Cortical: 0.7416 +/- 0.0101
Rep: Mean DSC Cortical: 0.7416 +/- 0.0072

Mean DSC Non-Cortical: 0.8504 +/- 0.0049
Rep: Mean DSC Non-Cortical: 0.8504 +/- 0.0031





maper

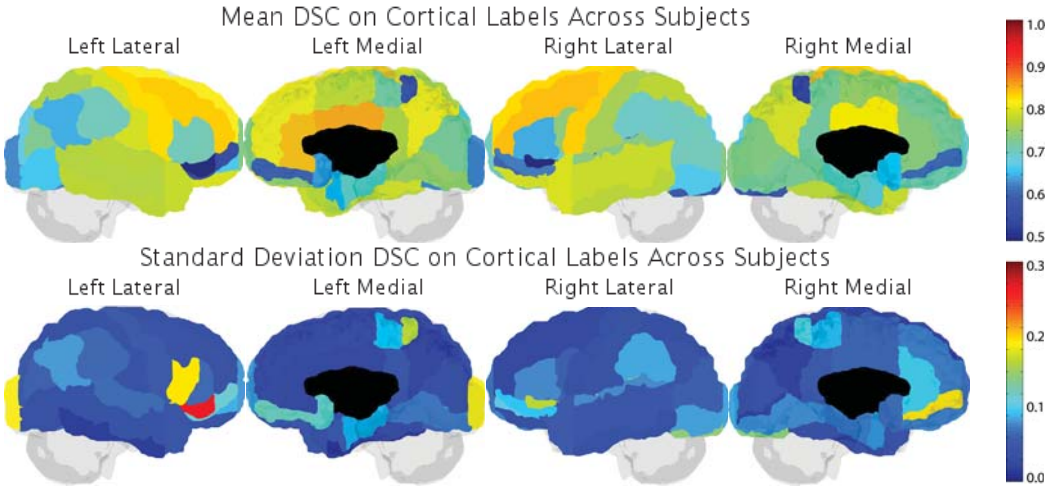
Attempt Number: 1

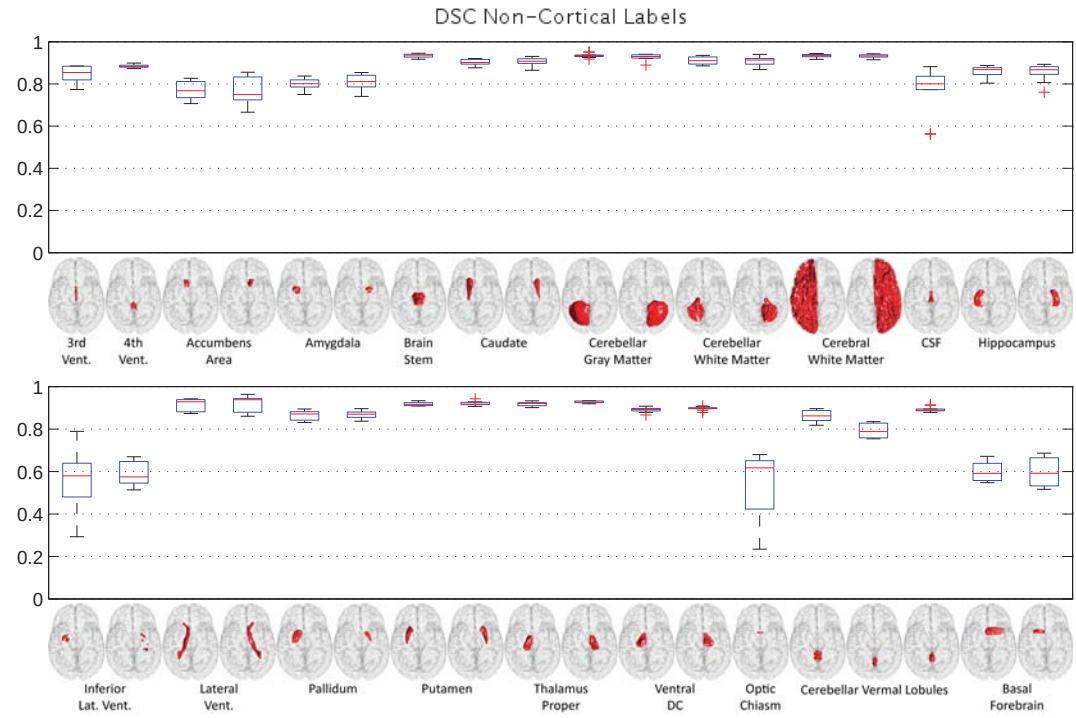
Date: 03-Jul-2012

Mean DSC Overall: 0.7518 +/- 0.0077
Rep: Mean DSC Overall: 0.7518 +/- 0.0057

Mean DSC Cortical: 0.7207 +/- 0.0109
Rep: Mean DSC Cortical: 0.7207 +/- 0.0066

Mean DSC Non-Cortical: 0.8364 +/- 0.0096
Rep: Mean DSC Non-Cortical: 0.8364 +/- 0.0084





NonLocalSTAPLE

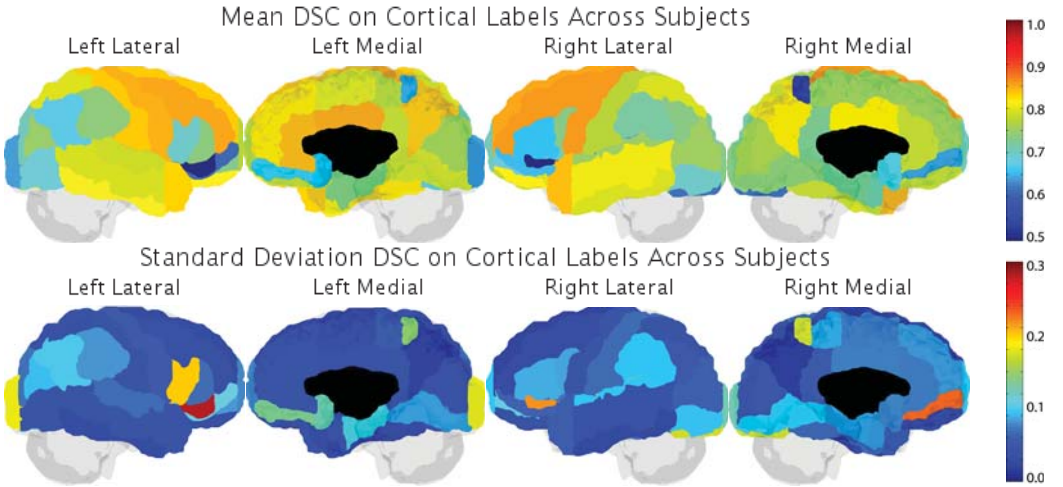
Attempt Number: 2

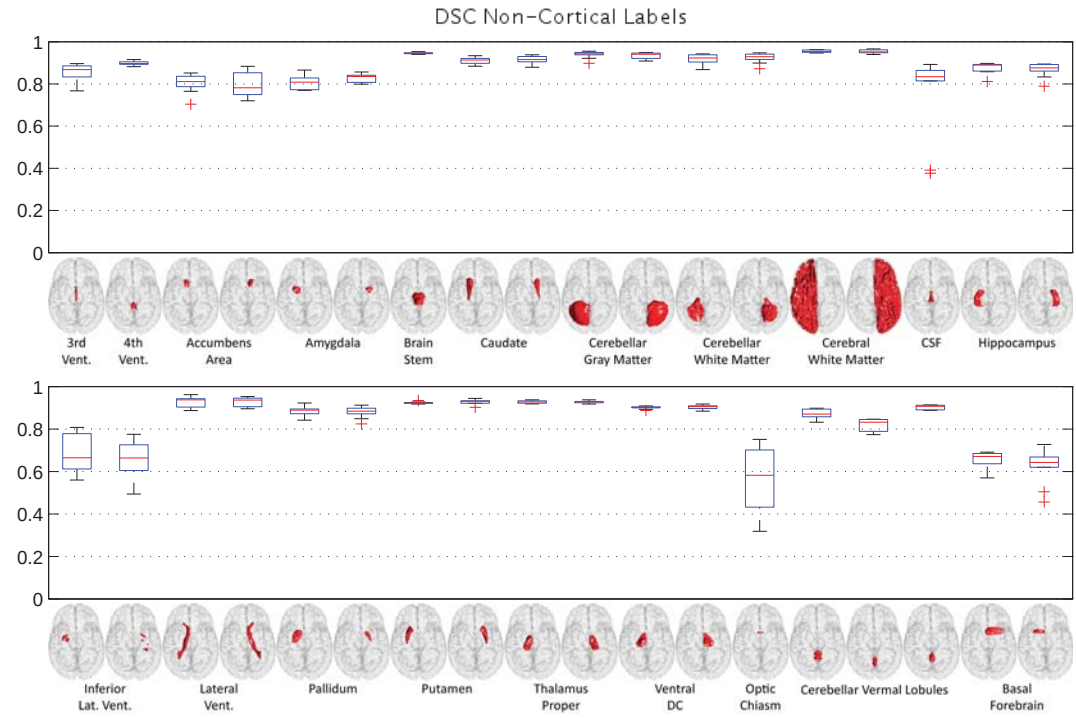
Date: 04-Jul-2012

Mean DSC Overall: 0.7764 +/- 0.0069
Rep: Mean DSC Overall: 0.7764 +/- 0.0064

Mean DSC Cortical: 0.7473 +/- 0.0077
Rep: Mean DSC Cortical: 0.7473 +/- 0.0068

Mean DSC Non-Cortical: 0.8554 +/- 0.0087
Rep: Mean DSC Non-Cortical: 0.8554 +/- 0.0067





PICSL_BC

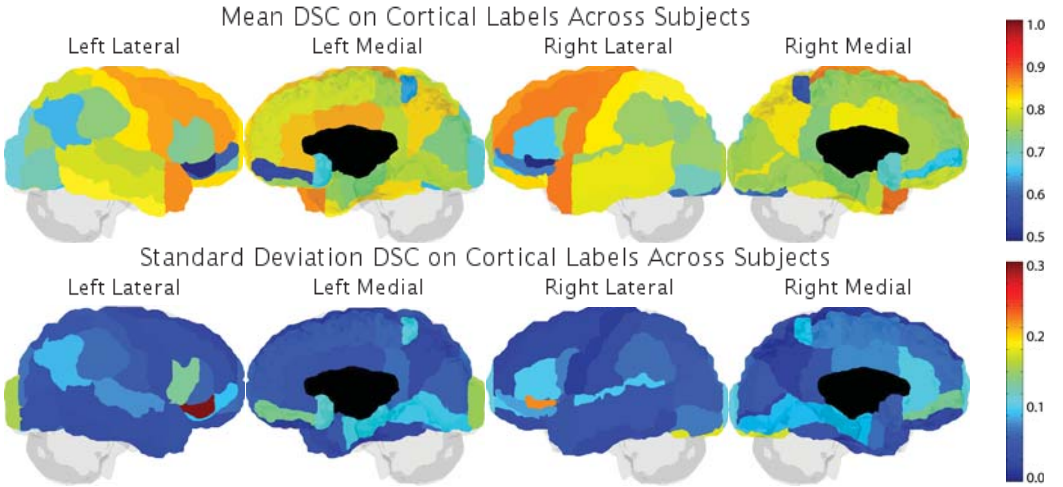
Attempt Number: 3

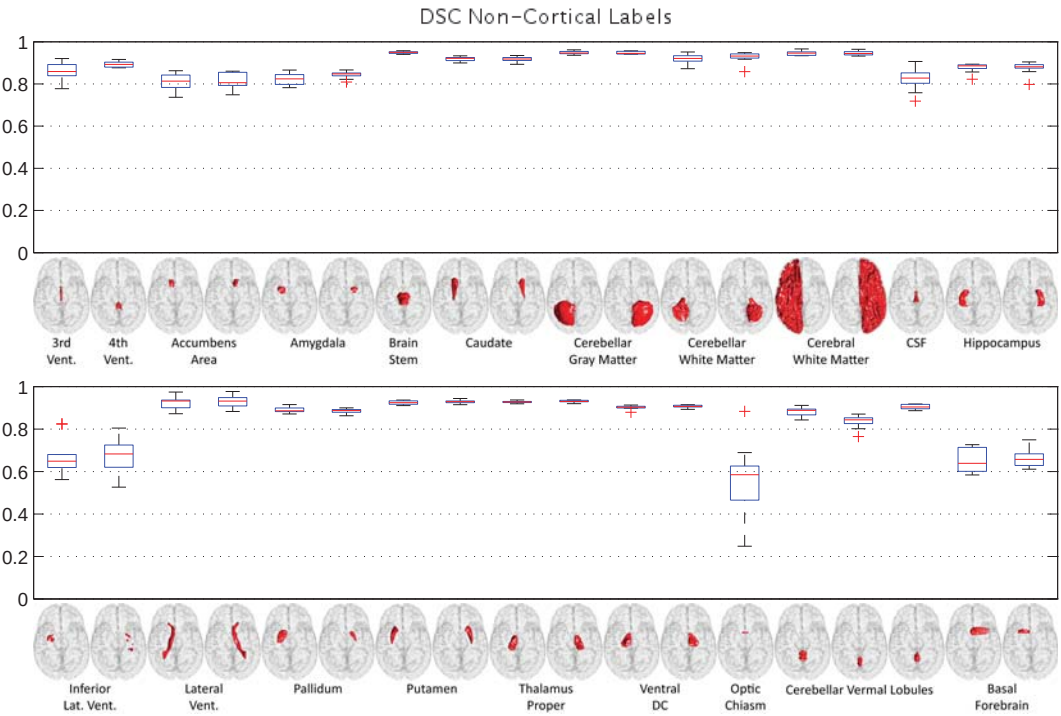
Date: 05-Jul-2012

Mean DSC Overall: 0.7820 +/- 0.0104
Rep: Mean DSC Overall: 0.7820 +/- 0.0098

Mean DSC Cortical: 0.7528 +/- 0.0122
Rep: Mean DSC Cortical: 0.7528 +/- 0.0113

Mean DSC Non-Cortical: 0.8614 +/- 0.0093
Rep: Mean DSC Non-Cortical: 0.8614 +/- 0.0076





PICSL_Joint

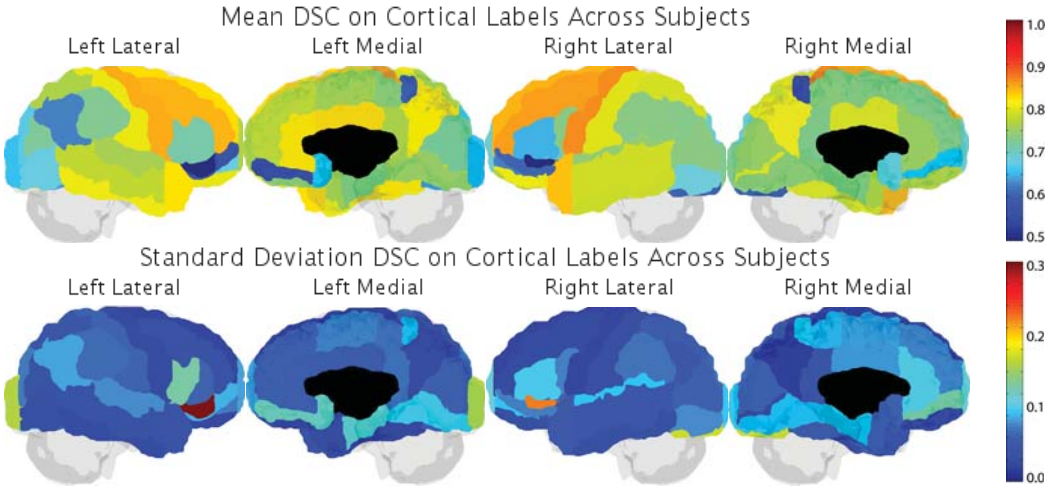
Attempt Number: 1

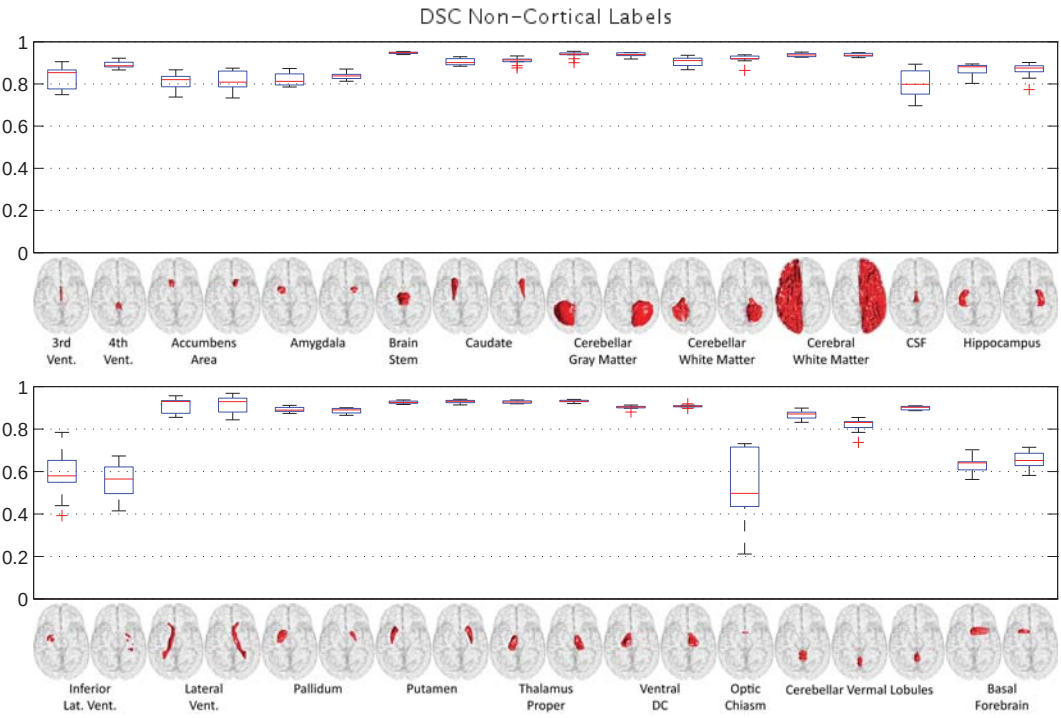
Date: 03-Jul-2012

Mean DSC Overall: 0.7663 +/- 0.0128
Rep: Mean DSC Overall: 0.7663 +/- 0.0119

Mean DSC Cortical: 0.7361 +/- 0.0150
Rep: Mean DSC Cortical: 0.7361 +/- 0.0141

Mean DSC Non-Cortical: 0.8482 +/- 0.0099
Rep: Mean DSC Non-Cortical: 0.8482 +/- 0.0080





SBIA_BrainROIMaps_JaccDet_IntCorr

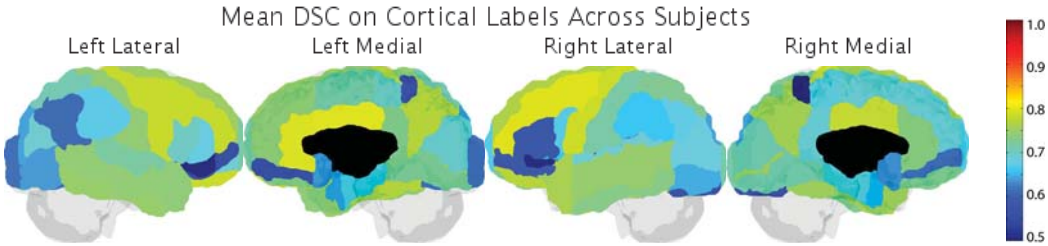
Attempt Number: 1

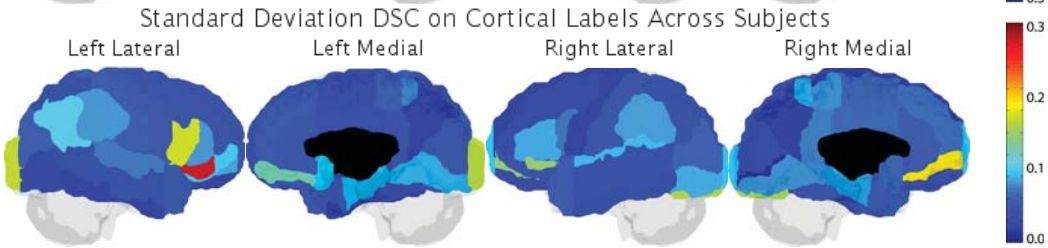
Date: 09-Jul-2012

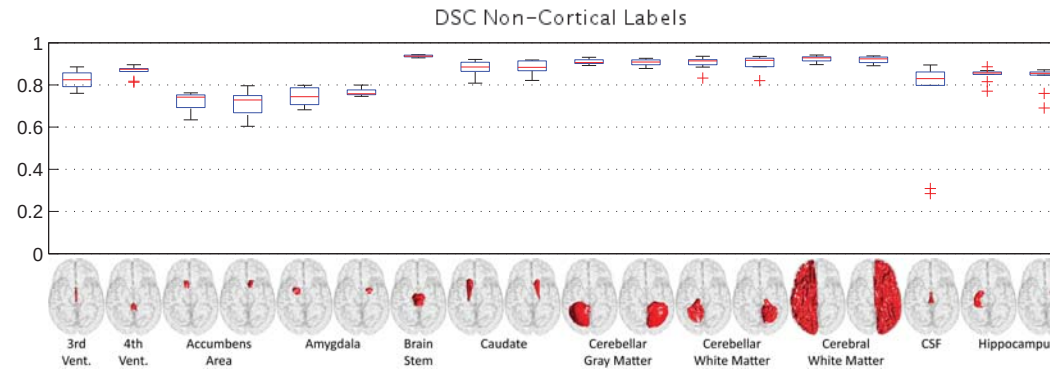
Mean DSC Overall: 0.7265 +/- 0.0154
Rep: Mean DSC Overall: 0.7265 +/- 0.0157

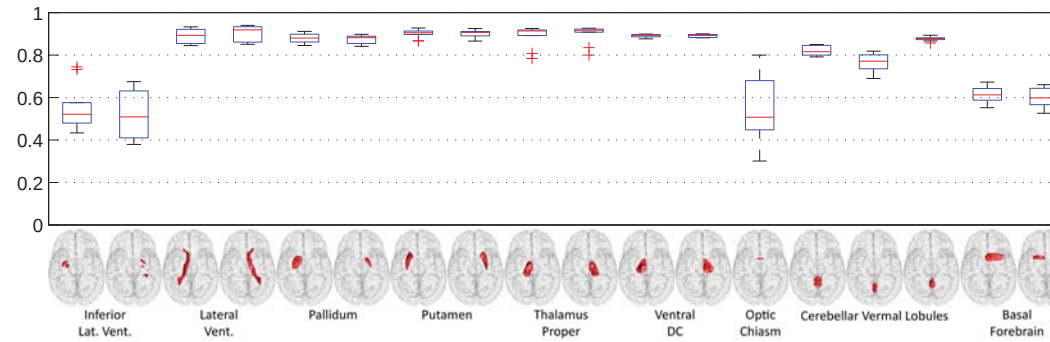
Mean DSC Cortical: 0.6932 +/- 0.0143
Rep: Mean DSC Cortical: 0.6932 +/- 0.0137

Mean DSC Non-Cortical: 0.8172 +/- 0.0214
Rep: Mean DSC Non-Cortical: 0.8172 +/- 0.0222









SBIA_BrainROIMaps_MV_IntCorr

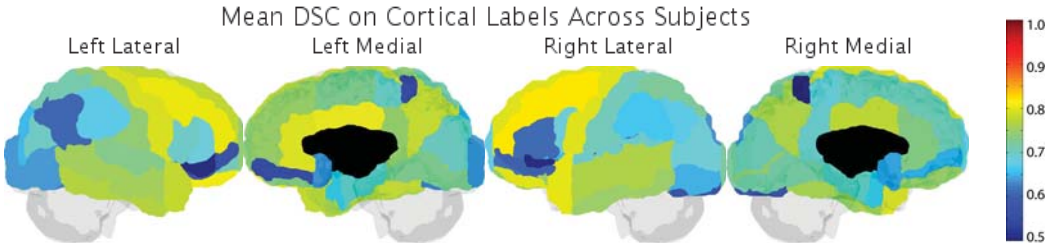
Attempt Number: 1

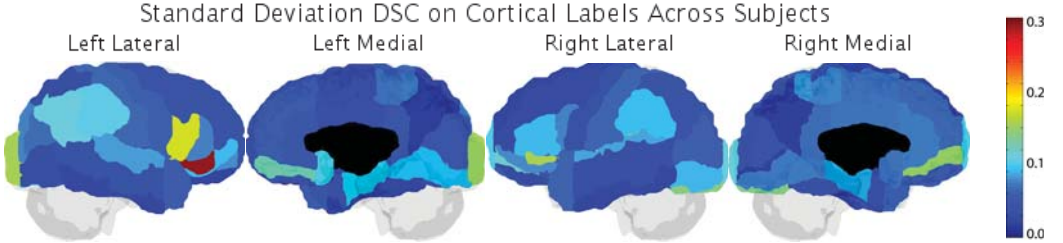
Date: 09-Jul-2012

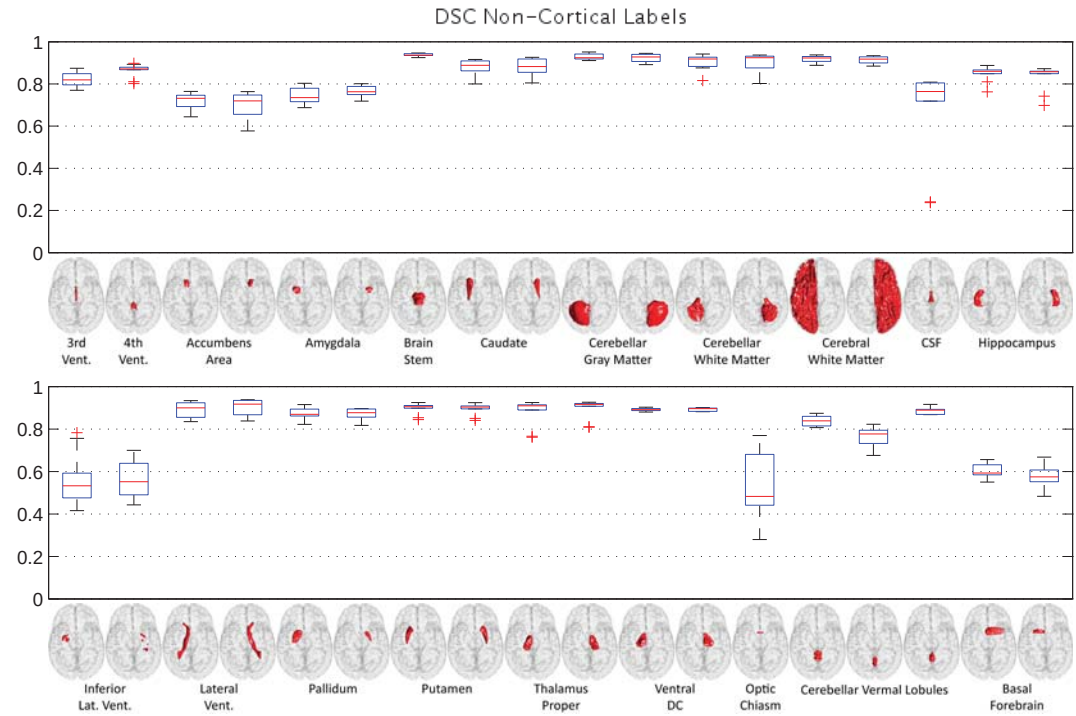
Mean DSC Overall: 0.7313 +/- 0.0161
Rep: Mean DSC Overall: 0.7313 +/- 0.0165

Mean DSC Cortical: 0.7007 +/- 0.0140
Rep: Mean DSC Cortical: 0.7007 +/- 0.0139

Mean DSC Non-Cortical: 0.8145 +/- 0.0233
Rep: Mean DSC Non-Cortical: 0.8145 +/- 0.0244







SBIA_SimMSVoting

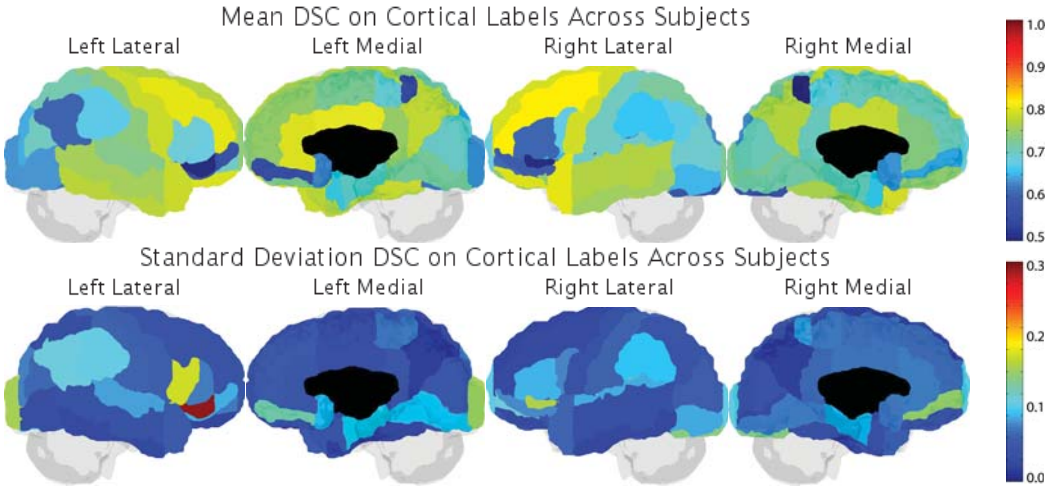
Attempt Number: 1

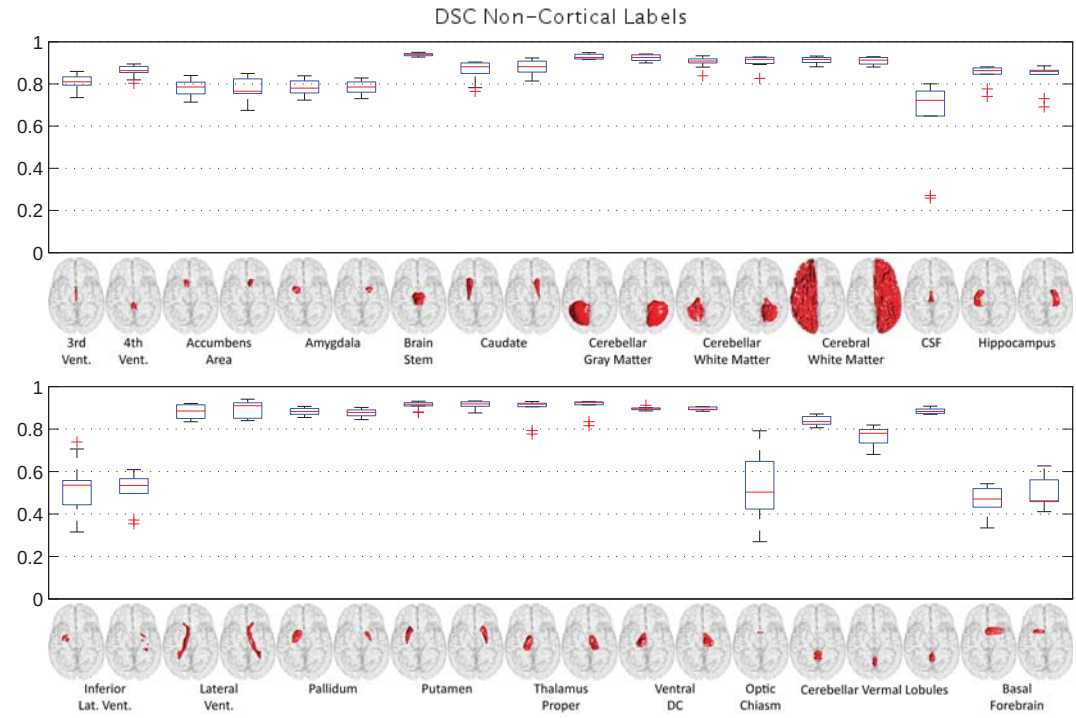
Date: 06-Jul-2012

Mean DSC Overall: 0.7283 +/- 0.0164
Rep: Mean DSC Overall: 0.7283 +/- 0.0170

Mean DSC Cortical: 0.6977 +/- 0.0143
Rep: Mean DSC Cortical: 0.6977 +/- 0.0144

Mean DSC Non-Cortical: 0.8116 +/- 0.0237
Rep: Mean DSC Non-Cortical: 0.8116 +/- 0.0244





SBIA_SimRank+NormMS

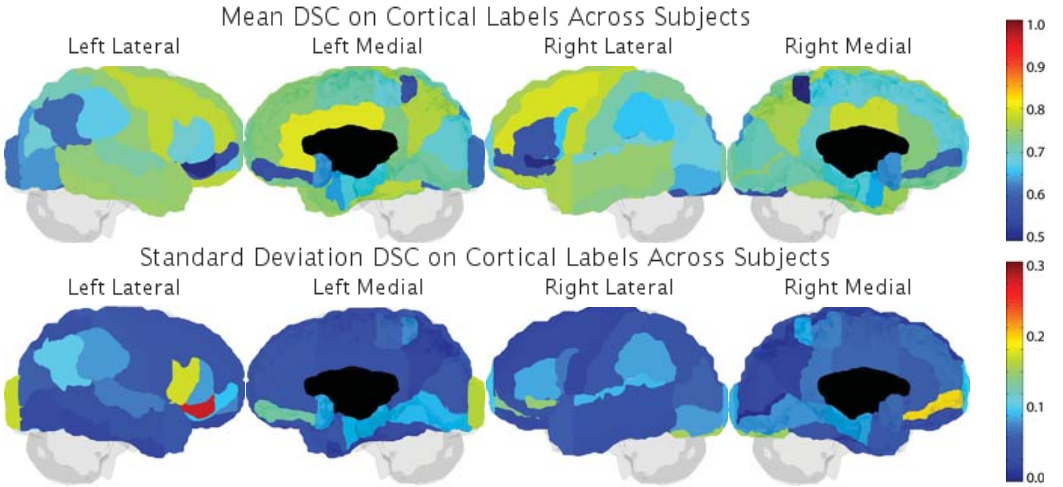
Attempt Number: 2

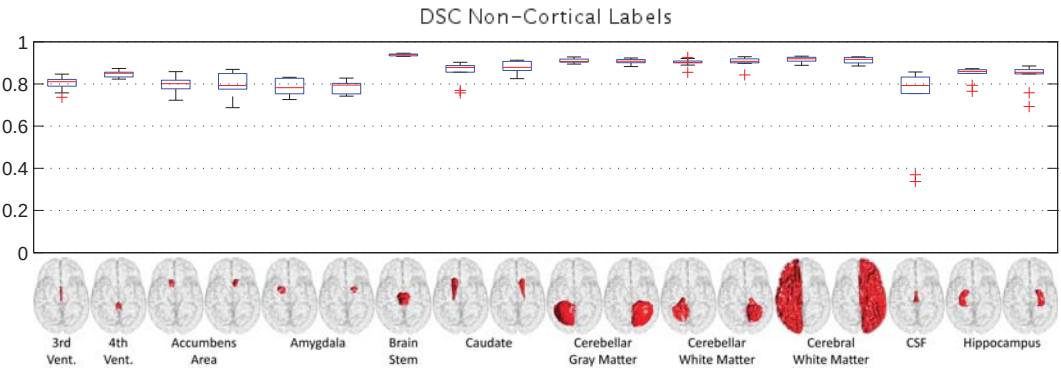
Date: 06-Jul-2012

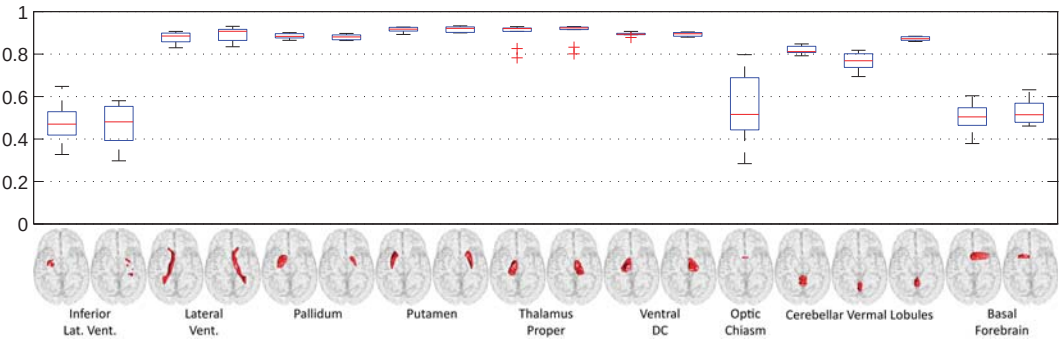
Mean DSC Overall: 0.7236 +/- 0.0150
Rep: Mean DSC Overall: 0.7236 +/- 0.0152

Mean DSC Cortical: 0.6909 +/- 0.0137
Rep: Mean DSC Cortical: 0.6909 +/- 0.0131

Mean DSC Non-Cortical: 0.8125 +/- 0.0208
Rep: Mean DSC Non-Cortical: 0.8125 +/- 0.0212







SBIA_SimRank+NormMS+WtROI

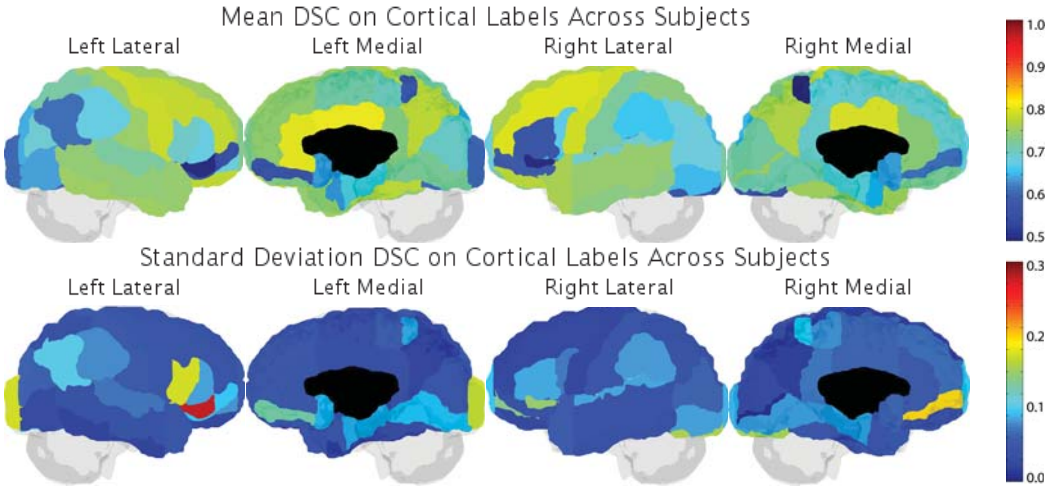
Attempt Number: 1

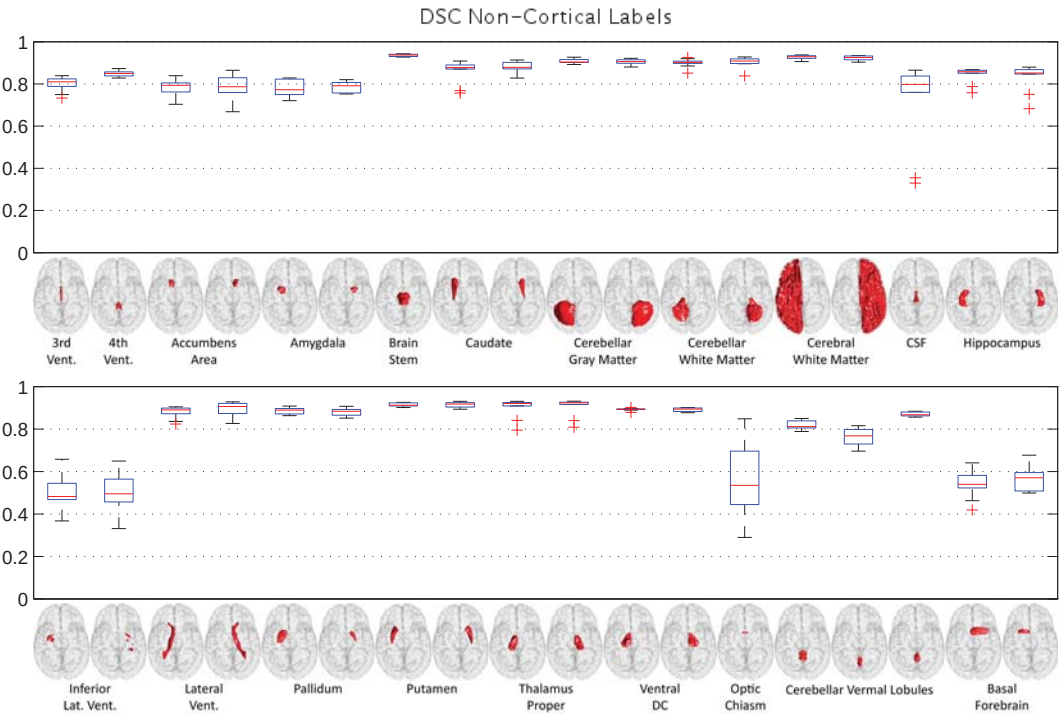
Date: 05-Jul-2012

Mean DSC Overall: 0.7282 +/- 0.0138
Rep: Mean DSC Overall: 0.7282 +/- 0.0135

Mean DSC Cortical: 0.6957 +/- 0.0125
Rep: Mean DSC Cortical: 0.6957 +/- 0.0112

Mean DSC Non-Cortical: 0.8164 +/- 0.0200
Rep: Mean DSC Non-Cortical: 0.8164 +/- 0.0202





SpatialSTAPLE

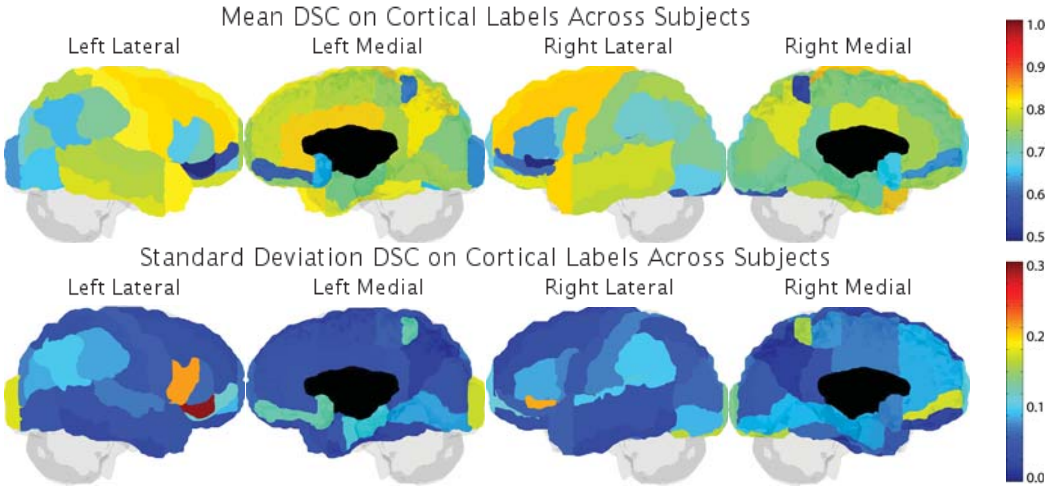
Attempt Number: 1

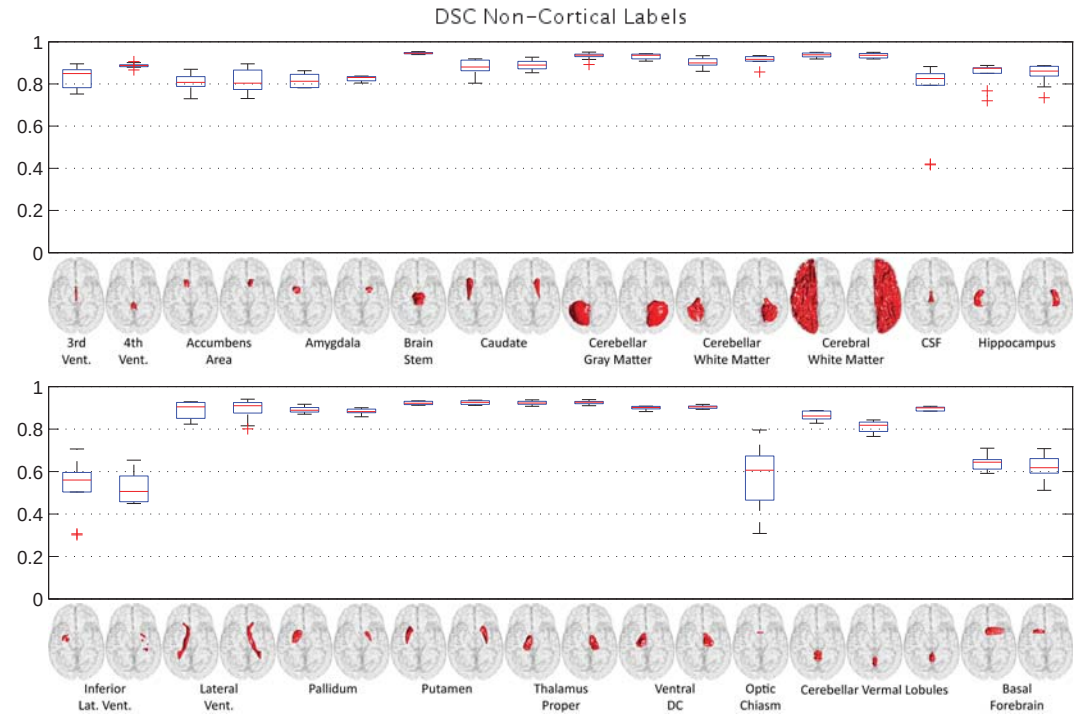
Date: 03-Jul-2012

Mean DSC Overall: 0.7576 +/- 0.0095
Rep: Mean DSC Overall: 0.7576 +/- 0.0092

Mean DSC Cortical: 0.7278 +/- 0.0093
Rep: Mean DSC Cortical: 0.7278 +/- 0.0084

Mean DSC Non-Cortical: 0.8388 +/- 0.0158
Rep: Mean DSC Non-Cortical: 0.8388 +/- 0.0157





STEPS

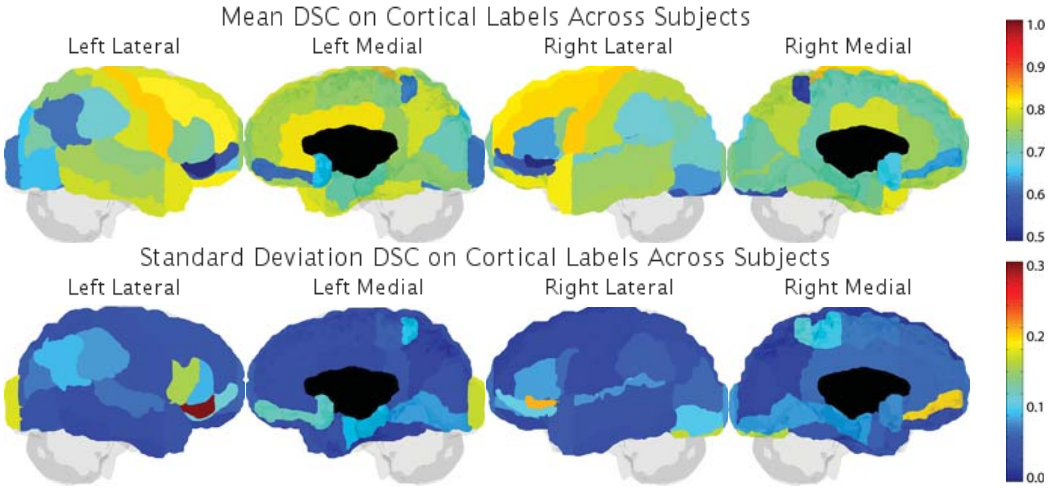
Attempt Number: 2

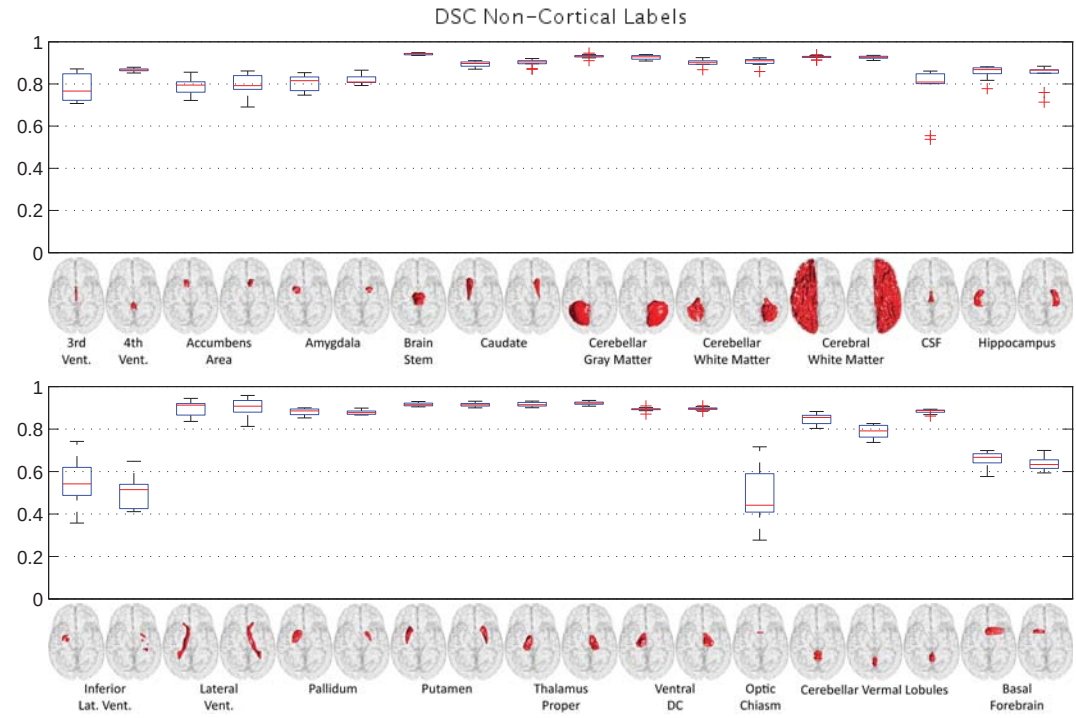
Date: 06-Jul-2012

Mean DSC Overall: 0.7500 +/- 0.0063
Rep: Mean DSC Overall: 0.7500 +/- 0.0062

Mean DSC Cortical: 0.7202 +/- 0.0056
Rep: Mean DSC Cortical: 0.7202 +/- 0.0050

Mean DSC Non-Cortical: 0.8311 +/- 0.0107
Rep: Mean DSC Non-Cortical: 0.8311 +/- 0.0101





UNC_NIRAL

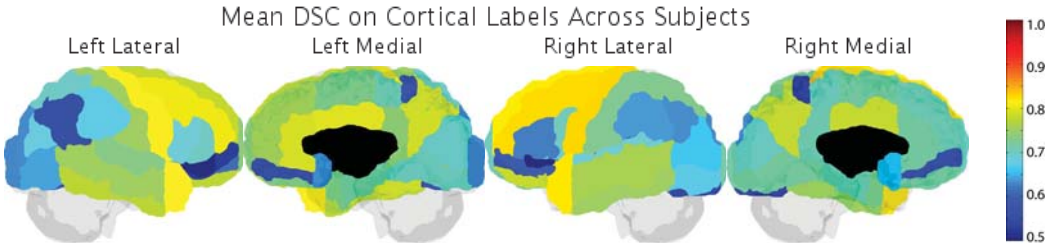
Attempt Number: 1

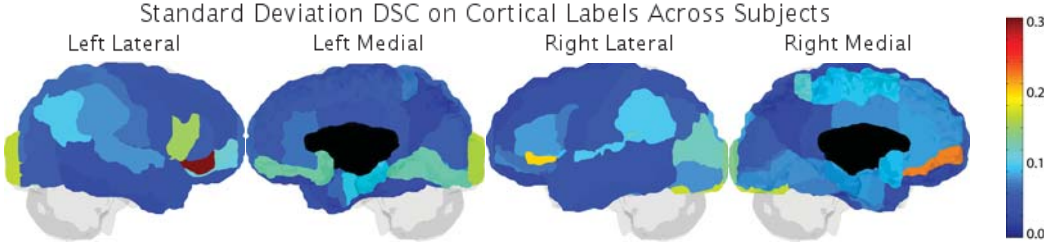
Date: 13-Jul-2012

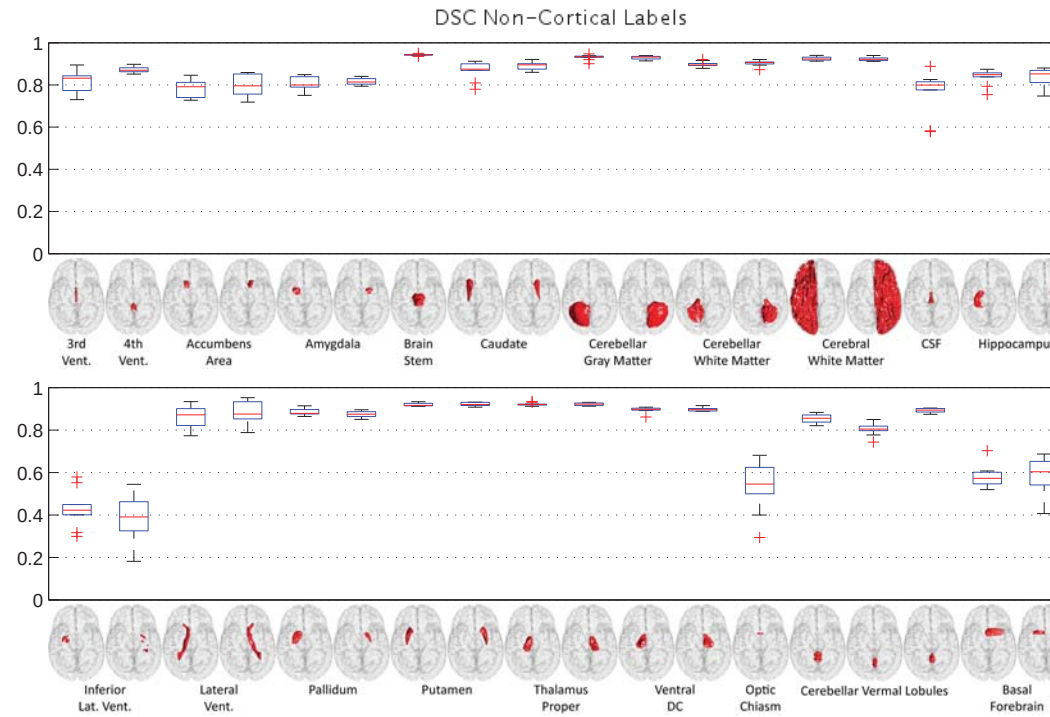
Mean DSC Overall: 0.7350 +/- 0.0151
Rep: Mean DSC Overall: 0.7350 +/- 0.0142

Mean DSC Cortical: 0.7030 +/- 0.0178
Rep: Mean DSC Cortical: 0.7030 +/- 0.0163

Mean DSC Non-Cortical: 0.8220 +/- 0.0103
Rep: Mean DSC Non-Cortical: 0.8220 +/- 0.0100







Multi-atlas labeling with population-specific template and non-local patch-based label fusion

Vladimir S. FONOV[1], Pierrick Coupé[2], Simon F. Eskildsen[3], Jose Vicente Manjon Herrera[4], D. Louis Collins[1]

[1]McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada. University, 3801 University Street, Montreal, Canada H3A 2B4
[2]LaBRI UMR CNRS 5800, 33405 Bordeaux, FRANCE
[3]Center of Functionally Integrative Neuroscience, Aarhus University, Aarhus, Denmark
[4]Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain

Abstract. We propose a new method combining a population-specific non-linear template atlas approach with non-local patch-based structure segmentation for whole brain segmentation into individual structures. This way, we benefit from the efficient intensity-driven segmentation of the non-local means framework and from the global shape constraints imposed by the nonlinear template matching.

Keywords: Non-linear registration, average anatomical template, non-local patch segmentation

1 Introduction

Label fusion segmentation methods have been recently become very popular for solving the automatic structure segmentation problem. Several strategies have been proposed to propagate expert manual segmentations of multiple templates onto a new subject for structure segmentation [1], [2], [3]. In this study, we propose to combine our recently published non-local patch-based segmentation method [4] with population-specific nonlinear template construction [5].

2 Methods

Available anatomical scans from 15 “training” datasets were used to create a left-right symmetric non-linear average anatomical template (see Fig. 1), using the technique described in [5]. Using a left-right symmetric template increases the training library twofold for the patch-based segmentation. The resulting non-linear transformations were applied to the manual segmentations, warping them into a common space forming an anatomical library with twice the number of samples of as in the “training dataset”. Thus, the training library used for the non-local patch-based seg-

adfa, p. 1, 2011.

mentation algorithm [4] consists of 30 pairs of nonlinearly warped T1w images with their corresponding warped segmentation samples.

2.1 Segmentation method

The procedure segments a new image using the following steps:

1. Image pre-processing includes non-uniformity correction [6], linear intensity normalization using histogram matching between the image and the average template, and affine registration to the template [7].
2. Non-linear registration of the subject's scan to the template [8], using a hierarchical framework, with parameters as described in [5].
3. The scan under study is warped into the template space and the patch-based segmentation algorithm is applied using the anatomical training library (see Fig 1 D,E)
4. The patch-based segmentation is warped back into native scan space using the inverse of the non-linear transformation estimated in step 2.

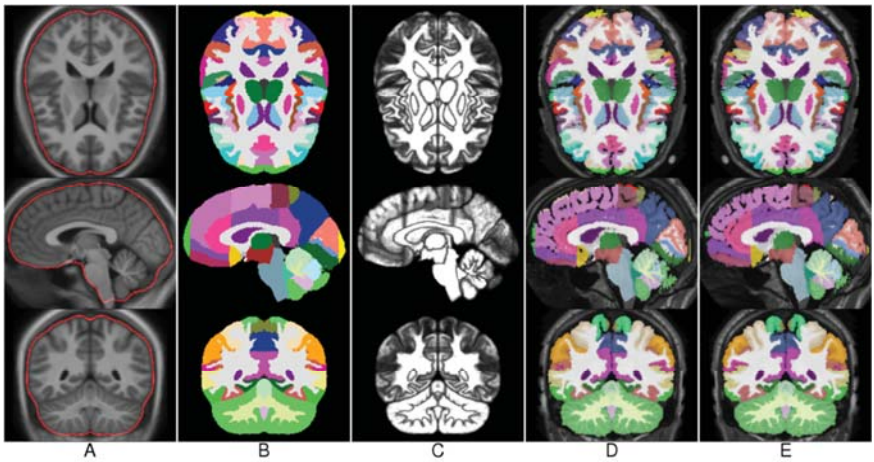


Fig. 1. Average anatomical template, constructed from the training dataset. A: average T1w template; B: majority overlap of anatomical labels for average template; C: pixel-wise generalized overlap, D: example of a one training template in the anatomical library (subject 1000, non-linearly warped into the template space); E: the same template flipped to increase the training library (left-right flipped subject 1000, non-linearly warped into a common space).

2.2 Parameter optimization and validation

The parameters of the non-local patch-based segmentation, i.e., patch size and search area, were chosen using a leave-one-out (LOO) experimental design. For each of the scans from the training dataset, the rest of the available segmentations were used in the algorithm above and the result was compared with the gold standard segmentation provided, using the generalized overlap metric [9].

3 Results

The leave-one-out experiments showed that a patch size of 3x3x3 voxels, and a search area of 7x7x7 voxels provided the best median generalized overlap metric, see see Fig. 2 and Fig. 3. These parameters were then used to segment the "training" dataset.

Average spatial distribution in the errors in LOA experiment is shown on Fig. 4, to produce it each voxel where results of automatic segmentation method disagreed with the ground truth was assigned value of 1, and to 0 otherwise. Resulting binary maps were non-linearly warped into a common space of the template using linear interpolation, and averaged, producing density map of the errors in LOA experiment.

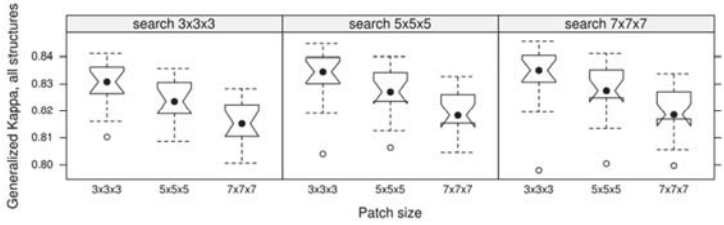


Fig. 2. Results of the leave-one-out experiment, varying parameters of the non-local patch segmentation for all structures.

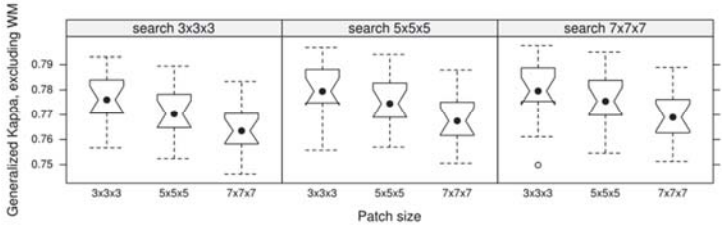


Fig. 3. Results of the leave-one-out for all structures except white matter (excluding IDs 40,41,44,45).

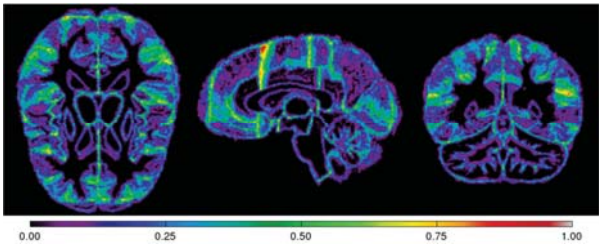


Fig. 4. Spatial distribution of the errors in LOA experiment, 0.00 corresponds to agreement in all cases, 1.00 to a disagreement in all cases.

4 Conclusions

We have created a whole-brain segmentation method which produces promising results in a LOO experiment. It is worth noting that in our LOO experiment the results for the whole brain segmentation may be biased towards structures with relatively large volumes (White Matter of cerebrum and cerebellum), excluding them from analysis reduces generalized overlap ratio (see Fig. 2,3).

The spatial distribution of errors (Fig 4) indicates that majority of segmentation errors occurs on the boundary of two structures, thus overall structures with smaller boundary-to-volume ration are expected to have lower overlap ratio. Furthermore, looking at the Figs. 4, 1C and 1B one can note that the most severe disagreement between automatic segmentation and the ground truth happen on the edge of structures which are defined by a straight line between different elements of the anatomy, not following any visible landmarks.

5 References

1. Aljabar, P., et al., *Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy.* NeuroImage, 2009. **46**(3): p. 726-738.
2. Heckemann, R.A., et al., *Automatic anatomical brain MRI segmentation combining label propagation and decision fusion.* NeuroImage, 2006. **33**(1): p. 115-126.
3. Collins, D.L. and J.C. Pruessner, *Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion.* NeuroImage, 2010. **52**(4): p. 1355-1366.
4. Coupe, P., et al., *Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation.* NeuroImage, 2011. **54**(2): p. 940-954.
5. Fonov, V., et al., *Unbiased average age-appropriate atlases for pediatric studies.* NeuroImage, 2011. **54**(1): p. 313-327.
6. Sled, J.G., A.P. Zijdenbos, and A.C. Evans, *A nonparametric method for automatic correction of intensity nonuniformity in MRI data.* Medical Imaging, IEEE Transactions on, 1998. **17**(1): p. 87-97.
7. Collins, D.L., et al., *Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space.* Journal of Computer Assisted Tomography, 1994. **18**(2): p. 192-205.
8. Collins, D.L., et al., *Automatic 3-D model-based neuroanatomical segmentation.* Human Brain Mapping, 1995. **3**(3): p. 190-208.
9. Crum, W.R., O. Camara, and D.L.G. Hill, *Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis.* Medical Imaging, IEEE Transactions on, 2006. **25**(11): p. 1451-1461.

Segmentation via the Random Multi-atlas Orbit Model in Computational Anatomy

Xiaoying Tang[1], Susumu Mori[2], and Michael I. Miller[1]

[1]Center for Imaging Science, Johns Hopkins University
[2]Department of Radiology, Johns Hopkins University

Abstract. In this paper, we examine a multi-atlas random orbit model in which imagery is modeled as conditional Gaussian random fields, conditioned on both the random atlas which generates it and the random diffeomorphism associated with the atlas. The model is examined for segmenting T1 imagery in which an iterative algorithm is employed for simultaneously estimating the unknown atlas-diffeomorphism pair and generating the maximum-a-posteriori (MAP) estimator of the subject labels. Since the goal is to generate the MAP estimator of the segmentation labels, the iterative algorithm is a derivative of the EM algorithm thereby removing the conditioning on the unknown atlas labels and the diffeomorphism. The segmenting accuracy of our method is evaluated for whole brain segmentations of 136 structures in the fifteen training datasets provided by the organizer of the workshop using a leave-one-out test. Our results are shown to be appealing.

Keywords: Multi-atlas, LDDMM, Random orbit model

1 Introduction

Segmenting cortical and subcortical structures of the human brain is important in clinical neuroimaging studies. The segmentation problem is usually handled in the setting of Bayesian maximum a posteriori (MAP). There are typically two approaches, in both of which appearance models (usually Gaussian appearance models) are defined. The first approach models various features of the voxels such as the intensity value as Gaussian distributions and then perform MAP estimation combined with other techniques such as markov random fields [1]. The second approach tries to incorporate shape priors into the intensity models with a weighting matrix estimated from a training set [2].

Our method utilizes large deformation diffeomorphic metric mapping (LDDMM) [3] [4]. A single atlas is likely to cause local misclassifications of the target, especially when the shape difference between the atlas and the subject is large. Using multiple atlases is capable of avoiding the local misclassifications from various atlases [5]. Our method is based on multi-atlas LDDMM image mapping. Given a set of pre-labeled T1-weighted atlas images, we model the to-be-segmented target image as a conditional Gaussian random field, conditioned on, in this multi-atlas setting, both the unknown atlas and the corresponding

unknown diffeomorphism between the atlas and the target. The atlas selection is iteratively optimized using the expectation-maximization (EM) algorithm, which gives rise to the maximum a posteriori estimation problem via a mixture of atlases.

2 Method

2.1 Principles of LDDMM-image

Given an atlas T1-weighted image I_0 and a target T1-weighted image I_1, where I_0 and I_1 are functions defined on the image domain $\Omega \subseteq R^3$, the algorithm LDDMM-image [4] computes a diffeomorphic transformation $\varphi : \Omega \to \Omega$ as the end-point of the flow of an energy-minimizing velocity vector field $v_t : \Omega \to R^3, t \in [0,1]$. The velocity vector field is specified by the ordinary differential equation (ODE) $\dot{\phi}_t = v_t(\phi_t)$, which starts with $\phi_0 = Id$, where Id is the identity transformation such that $Id(x) = x, \forall x \in \Omega$. The diffeomorphic deformation φ is thus calculated as: $\varphi = \phi_1 = \int_0^1 v_t(\phi_t)\, dt$ with $\phi_0 = Id$. The optimal deformation is estimated by solving the variational problem:

$$\hat{v} = \underset{v:\dot{\phi}_t = v_t(\phi_t)}{\arg\min} \left(\int_0^1 \|Lv_t\|_{L^2}^2 \, dt + \frac{1}{\sigma^2} \left\| I_0 \circ \phi_1^{-1} - I_1 \right\|_{L^2}^2 \right) \tag{1}$$

To ensure that the solution lies in the space of diffeomorphisms, a sufficient amount of smoothness is achieved by defining the operator L as: $L = (-\alpha \nabla^p + \gamma)\, \mathrm{I}_{3\times 3}$, where $p \geq 1.5$ in 3-dimensions, γ is usually fixed to be 1, α affects the degree of smoothness of the deformation, and ∇ is the gradient operator.

2.2 Probabilistic Model

Let A be a set of T1-weighted atlas images, paired with its manual labels $A = (I, W)$, where I denotes the gray-scaled T1 image and W denotes the manual segmentations of I. Given a to-be-segmented subject with image intensity I_i of the subject at voxel x_i modeled as conditional Gaussian random field, conditioned on the unknown atlas and the corresponding unknown diffeomorphism (A_i, φ_{A_i}). The algorithm for segmentation involves iterative atlas selection and diffeomorphism construction which is a variant of the expectation-maximization (EM) method. The iteration procedure is briefly described as follows:

1. Initialize: for each voxel i of the target image, for each atlas $a \in A$, set the diffeomorphism to identity $\hat{\varphi}_a = Id$ and set initial weights to uniform conditional probability as:

$$\alpha^{old}(a) = \frac{1}{|A|}, \tag{2}$$

where $|A|$ denotes the total number of atlases.

2. For each voxel i, in terms of each atlas a, calculate:

$$\log p(I_i, W_i|a, \hat{\varphi}_a) = \log p(I_i|W_i, a, \hat{\varphi}_a) + \log p\left(W_i|a, \hat{\varphi}_a\right), \tag{3}$$

where

$$p(I_i|W_i, a, \hat{\varphi}_a) = \frac{1}{\sqrt{2\pi}\sigma(W^a \circ \hat{\varphi}_a^{-1})} e^{-\frac{\left|I_i - \mu(W^a \circ \hat{\varphi}_a^{-1})\right|^2}{2\sigma(W^a \circ \hat{\varphi}_a^{-1})^2}}, \tag{4}$$

with

$$\mu(W^a)_i = \frac{\sum\limits_{j \in structure} I_j^{(a)}}{\sum\limits_{j \in structure} 1}, (\sigma(W^a)_i)^2 = \frac{\sum\limits_{j \in structure} \left(I_j^{(a)} - \mu(W^a)_i\right)^2}{\sum\limits_{j \in structure} 1}, \tag{5}$$

where i indexes different structures. The quantity $p\left(W_i|a, \hat{\varphi}_a\right)$ is calculated by performing trilinear interpolation when transferring the manual labels $W^{(a)}$ of the atlases under the action of diffeomorphism $\hat{\varphi}_a(\cdot)$ – composition with $\hat{\varphi}_a^{-1}$.

3. Update the label classification of each voxel in the target via:

$$W_i^{new} = \arg\max_{W_i} \sum_j \sum_a \alpha_j^{new}(a) log p(I_j, W_j|a, \hat{\varphi}_a), \tag{6}$$

where i indexes voxels.

4. Update segmentation $W^{old} \leftarrow W^{new}$, and compute optimum diffeomorphism for each $A_i = a$ via:

$$\hat{\varphi}_a = \arg\max_{\varphi} p(a, \varphi|W^{old}, I) \tag{7}$$

$$= \arg\max_{\varphi} \log p(W^{old}|a, \varphi, I) + \log \pi(a, \varphi) \tag{8}$$

where $\pi(a, \varphi)$ is the prior probability of the atlas a and its diffeomorphism to the subject. We use the metric distance in LDDMM [3] to estimate this prior probability.

5. Update $\alpha_i^{old}(a) \leftarrow \frac{p(a, \hat{\varphi}_a|W^{old}, I_i)}{\sum\limits_a p(a, \hat{\varphi}_a|W^{old}, I_i)}$ for each $A_i = a$, go to 2.

3 Results

We evaluate the accuracy of our algorithm for segmenting the whole brain based on the 15 training datasets provided by the organizer of the workshop using a leave-one-out test. Based on the manual labelings of the training datasets, we segment the whole brain into 136 cortical and subcortical regions. The automated

results have been compared with those of manually labeling the same datasets. Due to the limitation of space, we only list the Kappa overlaps of 41 regions including all the subcortical structures, ventricular structures, and some of the cortical and white matter regions. According to the results shown in Fig. 1, Multi-atlas LDDMM is capable of achieving Kappa overlaps higher than 0.87 for a majority of subcortical structures and 0.8 for cortical regions.

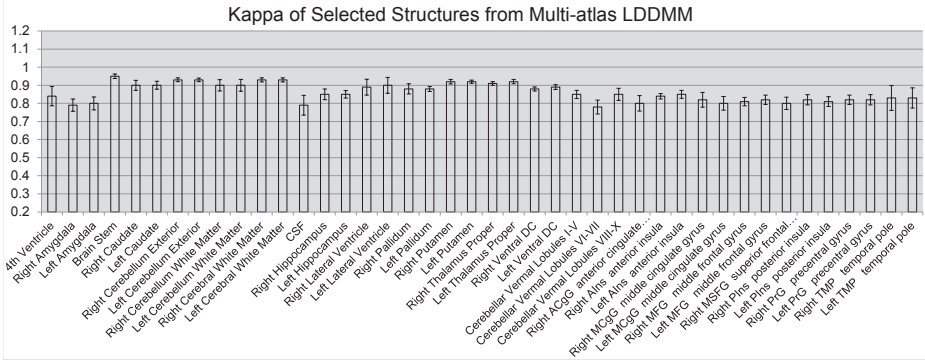


Fig. 1. The averaged Kappa Overlaps and the standard deviations of 15 subjects for 41 different brain structures obtained from Multi-atlas LDDMM

4 Acknowledgments

The work presented here was supported by grants: NIH R01 EB000975, NIH P41 EB015909 and NIH R01 EB008171.

References

1. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.: Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. Neuron 33, 341–355 (2002)
2. Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M.: A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 907–922 (2011)
3. Miller, M.I., Trouv, A., Younes, L.: On the metrics and EulerCLagrange equations of computational anatomy. Annu. Rev. Biomed. Eng. 4, 375–405 (2002)
4. Beg, M.F., Miller, M.I., Trouv, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. 61, 139–157 (2005)
5. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126 (2006)

Evaluation of Some STAPLE Based Fusion Algorithms

Alireza Akhondi-Asl, Simon K. Warfield

Computational Radiology Laboratory, Department of Radiology,Children's Hospital, 300 Longwood Avenue, Boston, MA, 02115, USA

Abstract. In this paper, we have evaluated some newly developed STAPLE based fusion algorithms which utilize intensity information of the target image and the templates in the fusion process. Since fusion algorithms are sensitive to the registration and skull stripping approaches, to have a fair comparison, we have also reported the segmentation results of the classic STAPLE and majority voting. In addition, we have utilized two different registration algorithms and have shown that for all of the fusion methods the segmentation results are superior when both of the registration algorithms are utilized. Finally, based on the preliminary results the weighted STAPLE algorithm is superior to the other three fusion algorithms.

1 Introduction

Fusion algorithms have been widely used in a variety of medical image segmentation problems, in particular, brain segmentation. The key purpose of such algorithms is to fuse set of templates and their corresponding segmentations to have an accurate and robust estimation of the segmentation of the target image. The simplest way to fuse the templates, known as the majority voting, is to count the number of votes for each label and assign the label with the highest number of votes to the voxel. However, it is known that the templates have different performances which considering these differences may lead to a more accurate estimation of the segmentation of the target image.

Intensity based locally weighted voting methods and STAPLE algorithm with its extensions and variations are two categories of the fusion methods which have been introduced for this purpose. In the former approach, intensity similarity of the target image and the templates are used to estimate the local weight of the templates. Mean square error and normalized cross correlation based functions are mainly used as the similarity metrics in this type of approach, where the intensity similarities are considered as the weights of the decisions at each voxel. However, these algorithms can be sensitive to the intensity normalization and more importantly, cannot compensate some of the intrinsic weaknesses of the majority voting approach. In addition, there is no well-defined relation between the weights and the performance of the segmentations [1–4].

In the latter method, the performance of the raters and the hidden ground truth are estimated iteratively using an Expectation-Maximization (EM) algorithm

[5]. The method was first introduced for the estimation of the performance of the raters. Since then, many variations of the algorithm have been introduced [6–8]. In this challenge we evaluate some novel STAPLE based fusion frameworks to use intensity information of the target image and the templates to improve the accuracy of the estimated segmentation of the target image. Since fusion algorithms are sensitive to the registration and skull stripping process, we have also generated segmentation results of the classic STAPLE and majority voting using the same registration and skull stripping approach. In the next section, we briefly describe our methods.

2 Methods

2.1 Probabilistic STAPLE

In this approach, to get higher accuracy, we propose modifying each one of the input segmentations to correct errors due to uncaptureable inter-individual anatomical dissimilarities between the target image and the templates and also the errors due to intrinsic registration inaccuracies. To this end, intensity and label map images of each one of the aligned templates are used to train a classifier. For this challenge, we use a local Gaussian mixture model (GMM) as the classifier. In the next step, each one of the trained classifiers is used to segment the target image. The output of each one of the classifiers is a probabilistic segmentation of the target image. To use these probabilistic segmentations, we use a novel extension of the STAPLE algorithm, which uses an EM framework to simultaneously estimate the performance of the segmentations and the hidden ground truth from a collection of probabilistic segmentations.
In general, the new fusion algorithm can be used for the fusion and evaluation of statistical classifiers, manual segmentations specified as confidence levels, or any set of probabilistic segmentations of a target image.

2.2 Weighted STAPLE

In this approach, similar to any other fusion algorithm, we assume that the each rater makes a decision at each voxel. However, in our new model, we assume that a confidence level is associated with each decision at each voxel for each rater. This confidence can be any number between zero and one where confidence level of one indicates the highest certainty in the decision. Moreover, we assume that all of the raters use the same standard to describe their confidence level to avoid any bias due to dissimilarities in the definition of the confidence. Similar to the STAPLE algorithm, the fusion algorithm uses an EM framework to simultaneously estimate the performance of the segmentations and the hidden ground truth from a collection of segmentations with associated confidence levels. In this challenge, to generate the confidence levels, we use normalized cross correlation to find the local similarity between the templates and the target image. There are two main constraints in our approach: Confidence levels should be between $[0 - 1]$ and higher similarities should indicate higher confidence level.

2.3 Skull Stripping

Since we are interested in the segmentation of the cortical grey matter structures, one of the important steps in the automatic brain parcellation is the skull stripping. In this challenge, we have used a novel approach for the brain extraction. For the skull stripping we use 15 training datasets and register them to each one of the test images using the registration approach described in [9]. We have used the manual segmentations of the training data to generate the brain mask for each one of 15 training datasets. Finally, we have utilized a local version of the approach described in 2.2 to generate the brain mask for each one of the testing datasets. It should be mentioned that we have used the same parameters for all of the cases and the generated masks are directly used for the skull stripping of the images.

2.4 Registration

After skull stripping of a test image, templates should be aligned to the target image using an appropriate registration algorithm. We have examined different non-rigid registration algorithms and based on these experiments, we have used ANTS [10] and the method described in [9] for the alignment of the templates to the target image. The output transformations were applied to the manual segmentation of the templates. In general, for each one of the fusion algorithms, we have used two sets of input segmentations. The first one is based on the ANTS registration algorithm [10] and the second one is using both ANTS and the method described in [9]. This means that in the second approach, we have used 30 registered templates as the input segmentation of the fusion algorithms while for the first approach there are only 15 segmentations. Due to some practical limitations, we have used binary version of algorithms described above and combined the output of the binary fusions to get the multi-category segmentation.

3 Results

We have evaluated four fusion algorithms using two different registration approaches. Based on the generated results, it can be seen that for all of the fusion algorithms, the results are superior when both registration algorithms have been utilized. In other words, by using more registration algorithms, fusion errors due to the registration inaccuracies are decreased. This suggests that using more registration algorithms can improve the fusion performance. In addition, it can be seen that weighted STAPLE generated the best results compared to the other fusion approaches.

4 Conclusions

We have introduced two new STAPLE based fusion algorithms and have compared them to the majority voting and classic STAPLE algorithm. In addition,

we have examined the effect of increasing the number of registration algorithms on the fusion process and have shown that this strategy may improve the accuracy of the estimated segmentation. It should be mentioned that we have used binary fusions and have combined the results to generate the multi-category segmentations. In the future, we will utilize the approach described in [11] to further improve the segmentation accuracies. In addition, both weighted STAPLE and probabilistic STAPLE have parameters that should be optimized. In the future works, we will optimize these parameters for specific application to have more accurate segmentation results.

References

1. Lötjonen, J., Wolz, R., Koikkalainen, J., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D.: Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage **49**(3) (2010) 2352–2365
2. van Rikxoort, E., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M., Pluim, J., van Ginneken, B.: Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. Medical Image Analysis **14**(1) (2010) 39–49
3. Sabuncu, M., Yeo, B., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. Medical Imaging, IEEE Transactions on **29**(10) (2010) 1714–1729
4. Artaechevarria, X., Munoz-Barrutia, A.: Combination strategies in multi-atlas image segmentation: Application to brain MR data. Medical Imaging, IEEE Transactions on **28**(8) (2009) 1266–1277
5. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. Medical Imaging, IEEE Transactions on **23**(7) (2004) 903–921
6. Langerak, T., van der Heide, U., Kotte, A., Viergever, M., van Vulpen, M., Pluim, J.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). Medical Imaging, IEEE Transactions on **29**(12) (2010) 2000–2008
7. Asman, A., Landman, B.: Robust statistical label fusion through consensus level, labeler accuracy and truth estimation (collate). Medical Imaging, IEEE Transactions on (99) (2011) 1779–1794
8. Landman, B., Asman, A., Scoggins, A., Bogovic, J., Xing, F., Prince, J.: Robust statistical fusion of image labels. Medical Imaging, IEEE Transactions on **31**(2) (2012) 512–522
9. Suarez, R., Commowick, O., Prabhu, S., Warfield, S.: Automated delineation of white matter fiber tracts with a multiple region-of-interest approach. NeuroImage (2011)
10. Avants, B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J.: The optimal template effect in hippocampus studies of diseased populations. NeuroImage **49**(3) (2010) 2457–2466
11. Commowick, O., Akhondi-Asl, A., Warfield, S.: Estimating a reference standard segmentation with spatially varying performance parameters: Local map staple. (2012)

Label propagation using group agreement – DISPATCH

Rolf A. Heckemann[1,2], Christian Ledig[3], Paul Aljabar[3,4], Katherine R. Gray[3], Daniel Rueckert[3], Joseph V. Hajnal[4], and Alexander Hammers[1,2]

[1] The Neurodis Foundation, Lyon, France, soundray@fondation-neurodis.org,
[2] Centre for Neuroscience (Hammersmith), Imperial College London, UK
[3] Department of Computing, Imperial College London, UK,
[4] Division of Imaging Sciences and Biomedical Engineering, Kings College London, UK

Abstract. To address the challenge of applying expert anatomical knowledge captured in brain atlases to unseen brain images, we propose "DISPATCH". **D**ISPATCH **i**s **s**egmentation by **p**ropagating **at**lases with **c**oerced **h**armony; a multi-level, multi-resolution label propagation approach that exploits per-level groupwise agreement.
The method relies on pairwise label propagation at a given resolution level. The resulting segmentations are combined using selective vote-rule decision fusion. The consolidated label set serves as a common target for a label-based registration using label consistency as the similarity metric. The resulting transformations are used as the starting point for iterating the registration-fusion-registration sequence at the next resolution level. We participate in the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling" with this novel approach.

1 Introduction

Magnetic resonance (MR) scanning of the human brain with state-of-the-art equipment generates large quantities of data. These are typically represented as 3D grey scale images that map the spatial signal distribution. For many applications, visual section-by-section review of these images is still the preferred method of processing such data. It does not, however, scale to large numbers of images. To extract information from large sets of images or multi-centre image repositories, such as ADNI, AIBL, IXI, OASIS, etc., efficient automatic methods are required.

Automatic anatomical segmentation provides an important avenue towards dimensionality reduction, feature extraction, and identification of imaging biomarkers. To segment a brain image of a given study subject or patient anatomically, most approaches rely on atlases generated by experts through manual segmentation of equivalent images. The optimal strategy for transferring this expert knowledge from the atlases to the new image is a matter of scientific debate. However, label propagation with decision fusion has consistently been among the best performing methods.

The approach presented here, "DISPATCH" (**D**ISPATCH **i**s **s**egmentation by **p**ropagating **at**lases with **c**oerced **h**armony) is a multilevel labelling procedure relying on forcing agreement of all atlases at each refinement step. It was developed from a brain extraction method called "pincram" (**p**yramidal **in**tra**cra**nial **m**asking). On the Segmentation Validation Engine (http://sve.loni.ucla.edu/), pincram is currently ranked second after the gold-standard labels for the site's test set (accessed 5 July 2012).

2 Method

2.1 Material

Data provided in the course of the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling" were used. A total of 35 images was supplied, originating from the OASIS project (http://www.oasis-brains.org/). Training data consisted of 15 T1-weighted images, with spatially corresponding, expertly generated maps identifying 138 labels, 113 of which had been declared relevant for the challenge[1]. Testing data consisted of 20 T1-weighted images, with labels that remained hidden from the participants.

2.2 Iterative labelling procedure

For a given target, labels were generated at progressive levels of refinement, termed affine, coarse, medium and fine, according to the detail level of the image registration step. At each level, the following calculations were carried out:

1. Label sets were generated from all 15 atlases using a standard label propagation approach based on image registration (cf. Section 2.3).
2. The 15 label sets were consolidated in target space using vote-rule decision fusion, generating fused set F1.
3. The 15 individual label sets were ranked in descending order of label agreement (Jaccard index) with F1.
4. These ranked values were used to determine an acceptability threshold $\theta = L_1 - 2(L_1 - L_7)$. L_7 was a coarse approximation of the median.
5. Individual segmentations passing the threshold were consolidated, generating fused set F2 and a mask M indicating non-unanimously labelled target voxels. Both F2 and M are in target space.
6. M was applied to the target image, creating an image with original grey scale values in voxels that were non-unanimously labelled and zero beyond. This ensured that only regions where labels had not been unanimously assigned were considered for the similarity calculation during the subsequent iteration.
7. The 15 atlas label sets (in their original space) were registered to F2, maximizing label consistency (cf. Section 2.3). The resulting 15 transformations from each individual atlas space to the target space were retained.

[1] Label sets were provided by Neuromorphometrics, Inc. (http://neuromorphometrics.com/) under academic subscription.

The masked target image and the retained transformations were used as starting points to iterate the procedure at the next level of refinement.

2.3 Image registration

Each pair of images (T1 or label sets) was affine-registered (first iteration) or nonrigidly registered (subsequent iterations). Nonrigid registration consisted in applying displacements to the atlas image via a lattice of control points, blended using B-spline basis functions, maximizing a similarity metric (see below) [1]. The stopping condition for the optimization was either no further improvement in similarity or the reaching of a maximum number of iterations.

The similarity metric for registering pairs of T1-weighted images was normalized mutual information [2]. To ensure the effectiveness of the data reduction step (6.), only values greater than zero were considered in the calculation.

The similarity metric for registering pairs of label sets was label consistency. It is defined as the fraction of voxels with agreed classifications, i.e. if n_{ij} represents the number of voxels given label i by one image and label j by the other, then label consistency is measured as

$$\frac{\sum_i n_{ii}}{\sum_{i,j} n_{ij}}. \tag{1}$$

All registration steps were carried out using the Image Registration Toolkit (IRTK, www.doc.ic.ac.uk/~dr/software/).

3 Discussion

The DISPATCH procedure combines proven techniques in a novel fashion. Previous multi-atlas label propagation methods produced multiple segmentations from individual atlases independently and only combined them in the final step of the procedure (e.g. MAPER, [3]). DISPATCH generates instead a consensus labeling at each of several levels of refinement, using information from the T1-weighted images. Starting estimates for the subsequent level are then created by registering all atlas label sets to the consensus set. These starting estimates constrain and inform the subsequent level of T1-pair registrations by driving them towards the consensus. In addition, efficiency is achieved through data reduction: at subsequent refinement levels, only those parts of the target image are considered which are likely to contain label boundaries. Such a data reduction step has previously been described for BEaST, an accurate brain extraction method that similarly constrains the algorithm's boundary search to regions which, based on information taken from preceding iterations, are likely to contain that boundary [4].

DISPATCH performs best on full head images. Brain extraction, a prerequisite step for many other approaches, is unnecessary.

Another related approach that has previously been described is "SIMPLE" [5]. It also relies on estimating the performance of an individual atlas by

estimating agreement of its propagated label set with a fused label set. SIMPLE discards atlases that fail to pass an agreement threshold, whereas DISPATCH rescues such atlases by registering their labels to the consensus set. The performance estimation in SIMPLE is carried out after a detailed registration has been performed, whereas in DISPATCH, a performance estimation step is inserted at each detail level, integrating the fusion step into each successive refinement procedure.

The evaluation results indicate that the DISPATCH approach works in principle. While the accuracy is somewhat inferior to, for example, MAPER, we are nevertheless encouraged by the findings. Since the development of DISPATCH is still in its infancy, no sophisticated refinement has been attempted yet. We expect that incorporating tissue class information, employing more elaborate label fusion techniques, and choosing more appropriate atlas selection strategies will yield substantial improvements. Also, due to the large number of registrations involved in DISPATCH, the parameter space is large, and further improvements are likely to emerge from parameter optimizations.

References

1. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging **18**(8) (August 1999) 712–721
2. Studholme, C., Hill, D.L.G., Hawkes, D.J.: An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition **32**(1) (January 1999) 71–86
3. Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A.: Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. Neuroimage **51**(1) (May 2010) 221–227
4. Eskildsen, S.F., Coupé, P., Fonov, V., Manjón, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Østergaard, L.R.R., Collins, D.L., Alzheimer's Disease Neuroimaging Initiative: BEaST: brain extraction based on nonlocal segmentation technique. Neuroimage **59**(3) (February 2012) 2362–2373
5. Langerak, T.R., van der Heide, U.A., Kotte, A.N., Viergever, M.A., van Vulpen, M., Pluim, J.P.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Transactions on Medical Imaging **29**(12) (December 2010) 2000–2008

Segmentation of MRI brain scans using MALP-EM

Christian Ledig[1⋆], Rolf A. Heckemann[2,3], Paul Aljabar[1,4], Robin Wolz[1], Joseph V. Hajnal[4], Alexander Hammers[2,3], Daniel Rueckert[1]

[1] Department of Computing, Imperial College London, London, UK
[2] The Neurodis Foundation, Lyon, France
[3] Centre for Neuroscience (Hammersmith), Imperial College London, UK
[4] Imaging Sciences and Biomedical Engineering, Kings College London, London, UK

Abstract. We employ a modification of our previously published method based on multi-atlas label propagation (MALP) and intensity-based refinement through expectation-maximization (EM) to segment magnetic resonance (MR) brain scans of the OASIS database. We had gold-standard segmentations available for 15 subjects of the same database, which we used as atlases in a multi-atlas propagation setup. After propagating the available atlases using transformations obtained with the robust MAPER approach, we use a locally weighted fusion strategy to merge the 15 atlas label sets into a consensus probabilistic segmentation of the unseen image. We use these probabilistic labels as priors in a subsequent EM refinement step, where we improve the segmentations based on the intensity distribution of the images. On top of the common EM refinement we apply a statistical correction based on the intensity characteristics of each individual region. The intensity profile of certain regions and their individual neighborhoods are not suited for an intensity based EM refinement nor a statistical correction. Therefore, we only refine regions for which intensity based refinement is beneficial and obtain a final segmentation by merging the labels obtained through MALP, MALP-EM and the statistical corrected MALP-EM regions. For evaluation, we segment MR brain scans of 20 subjects of the OASIS database.

1 Introduction

The segmentation of brain images into anatomical regions in magnetic resonance (MR) scans is an important task in neuroimaging. It yields regional volumetric information and labeling of different brain structures which can support clinical decision making. Even though manual annotations by a trained specialist are accurate, they are not scalable, time consuming and thus expensive. A fully automated method that calculates brain segmentations without user interaction is thus highly desirable and the basis for the segmentation of large data sets,

⋆ This work is partially funded under the 7th Framework Programme by the European Commision (http://cordis.europa.eu/ist/).

such as the data from the Alzheimer's Disease Neuroimaging Inititative (ADNI, adni.loni.ucla.edu) [1] or OASIS [2].

In this work we employ a recent segmentation method [3] which combines the advantages of both intensity based methods, e.g. [4], and approaches based on multi-atlas label propagation (MALP), e.g. [5], to segment MR brain images of 20 healthy adult subjects of the OASIS database [2]. Our approach refines subject specific spatial priors obtained through MALP and label fusion [6] in a probabilistic intensity model solved via expectation-maximization (EM) [7]. We furthermore refine certain regions based on statistical intensity characteristics.

2 Method

2.1 Material

We used the dataset provided through the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling". The training dataset consists of 15 T_1-weighted images with corresponding labels created by experts[1]. We segmented a testing dataset consisting of 20 otherwise identical T1-weighted images with hidden labels into 138 regions. The performance of our approach was evaluated using an automatic online evaluation interface provided through the Challenge.

2.2 Multi-Atlas Label Propagation with EM refinement (MALP-EM)

We use multi-atlas label propagation to derive a subject-specific probabilistic brain atlas for an unseen T_1 weighted MR scan $\mathbf{I}$ that is to be segmented. We incorporate these probabilistic labels into our EM framework as spatial anatomical priors. We index the n voxels of $\mathbf{I}$ by $i = 1, \ldots, n$, so that for intensities $y_i \in \mathbb{R}$ an image can be defined as $\mathbf{I} = \{y_1, y_2, \ldots, y_n\}$. The probabilistic priors are created by transforming M manually generated atlases to the coordinate space of the unseen image. We calculate the M transformations for the label propagation with a non-rigid registration method based on free-form deformations (FFD) [8, 9], which follows a preceding rigid and affine alignment. In particular we employed MAPER [10], which incorporates tissue probability maps into the registration. The probabilistic atlas is then created with a locally weighted multi-atlas fusion strategy [6], by employing a Gaussian weighted sum of squared differences on rescaled, intensity-normalized images. We followed the approach of van Leemput et al. [7] and estimated the hidden segmentation by means of the observed intensities $\mathbf{y}$. Assuming that the observed log-transformed intensities of voxels belonging to a certain class k are normally distributed with mean μ_k and standard deviation σ_k, yields the model parameters $\mathbf{\Phi} = \{(\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_K, \sigma_K)\}$. We applied regularization of the resulting segmentation using the approach of global and stationary Markov Random Fields (MRF) described in [11].

[1] provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/) under academic subscription.

2.3 Statistical correction of MALP-EM

During our experiments we observed that the EM algorithm tends to produce segmentations with a too low intensity variance within a region (intra-class variance) compared to the gold-standard segmentations. We therefore calculated an expected normalized intra-class variance for each region ($\sigma^2_{\mathrm{Gold},k}$) by averaging the normalized standard deviations $\frac{\sigma_k}{\mu k}$ of each class over the training subjects. We furthermore calculated the averaged (average over all training subjects segmented with a leave-one-out strategy) normalized standard deviation within each region produced by the EM algorithm ($\sigma_{\mathrm{EM},k}$) . By calculating $\Delta_k = (\sigma_{\mathrm{Gold},k} - \sigma_{\mathrm{EM},k})^2$ we estimated by which value the intra-class variance of a certain class should be increased in average to better match the gold-standard characteristics. In a subsequent refinement step we then corrected the intra-class statistics of each class by adding voxels with posterior probability greater than 10%, in decreasing order regarding the label probability, to the region unless the intra-class variance increased by Δ_k. Overlaps of most cortical regions with the gold-standard could be improved using statistical correction.

2.4 Fusion of MALP and MALP-EM

Our experiments revealed that some regions are ill-suited for intensity based refinement, due to either their intensity properties, or to those of their neighborhood. For example, no improvements using EM were obtained for the structures thalamus and putamen which can be explained with the wide overlap of their intensity profile with the profile of white matter. This is also shown in [3]. For these structures it is preferable to rely on the segmentation obtained through MALP alone. By segmenting all available training datasets with a leave-one-out strategy, we determined the subset of regions for which the standard EM refinement or the statistically corrected version is beneficial. We then created a final segmentation by combining the refined labels for this subset with the labels from the MALP approach for the remaining regions. In case of overlapping regions, we labeled a voxel according to the EM-refined label.

2.5 Parameters

To identify neighbouring tissue classes for the implementation of the MRF, we counted the labels of adjacent voxels in the gold-standard segmentations. After thresholding we obtained a 139×139 adjacency matrix G that describes the MRF, with entry (i, i) eqauls 0 and entry (i, j) defined as 1.0 if structures i and j share a boundary and 1.5 if structures i and j are distant. For a voxel size of 1x1x1mm we set for the locally weighted fusion the parameter σ to 2.5. Parameters were optimized using a leave-one-out strategy on the training datasets.

3 Results

The presented approach was evaluated using 20 datasets of the OASIS database with hidden labels. The results were automatically calculated through the Grand

Challenge on Multi-Atlas Labeling. We observe that the MALP approach performs very well on most of the 36 subcortical regions (average Dice similarity coefficient greater than 85%). Since the cohort consists of healthy adults with little intersubject variability, it is not surprising that registration based approaches perform well on this dataset. The EM- and statistical-based refinement is thus particularly relevant in cortical regions where, due to the high structural variability within the brain, registration based approaches are less accurate. Also the high intensity contrast at the cortical boundary between white and grey matter tissue is predestined for intensity based EM refinement. We obtain an average Dice coefficient of 73.28% for cortical and 82.52% for subcortical regions on the testing dataset. This yields an overall average label overlap of 75.76%.

References

1. S. G. Mueller, M. W. Weiner, L. J. Thal, et al., "The Alzheimer's Disease Neuroimaging Initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, pp. 869–877, 2005.
2. D. S. Marcus, T. H. Wang, J. Parker, et al., "Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.
3. C. Ledig, R. Wolz, P. Aljabar, et al., "Multi-class brain segmentation using atlas propagation and EM-based refinement," *Proc. of ISBI 2012*, pp. 896–899, 2012.
4. J. MP. Lötjönen, R. Wolz, J. R. Koikkalainen, et al., "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.
5. R. A. Heckemann, S. Keihaninejad, P. Aljabar, et al., "Automatic morphometry in Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 56, no. 4, pp. 2024–2037, 2011.
6. X. Artaechevarria, A. Munoz Barrutia, and C. Ortiz de Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE TMI*, vol. 28, no. 8, pp. 1266–1277, 2009.
7. K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE TMI*, vol. 18, no. 10, pp. 897–908, 1999.
8. D. Rueckert, L. I. Sonoda, C. Hayes, et al., "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE TMI*, vol. 18, no. 8, pp. 712–721, 1999.
9. M. Modat, G. R. Ridgway, Z. A. Taylor, et al., "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.
10. R. A. Heckemann, S. Keihaninejad, P. Aljabar, et al., "Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation," *NeuroImage*, vol. 51, no. 1, pp. 221–227, 2010.
11. M. J. Cardoso, M. J. Clarkson, G. R. Ridgway, et al., "Load: A locally adaptive cortical segmentation algorithm," *NeuroImage*, vol. 56, no. 3, pp. 1386–1397, 2011.

Multi-atlas propagation with enhanced registration – MAPER

Rolf A. Heckemann[1,2], Shiva Keihaninejad[3], Christian Ledig[4], Paul Aljabar[4,5], Daniel Rueckert[4], Joseph V. Hajnal[5], and Alexander Hammers[1,2]

[1] The Neurodis Foundation, Lyon, France, soundray@fondation-neurodis.org,
[2] Centre for Neuroscience (Hammersmith), Imperial College London, UK
[3] Dementia Research Centre, UCL Institute of Neurology, London, UK
[4] Department of Computing, Imperial College London, UK,
[5] Division of Imaging Sciences and Biomedical Engineering, Kings College London, UK ⋆

Abstract. To address the challenge of applying expert anatomical knowledge captured in brain atlases to unseen brain images, we previously proposed "MAPER" (multi-atlas propagation with enhanced registration).
The approach is based on a pairwise image registration procedure that incorporates tissue class information to obtain a robust anatomical correspondence estimate, even when the target brain is distinctly differently configured from the atlases. Multiple segmentations obtained by propagating individual atlas label sets are combined using a simple procedure (vote-rule decision fusion).
We participate in the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling" with a procedure that remains unchanged in principle from our previous publications. Only at the detail level was the method adapted to the particularities of the challenge.

1 Introduction

The advent of large, publicly available repositories of images of the human brain (ADNI, AIBL, Predict-HD, IXI, OASIS etc.) has changed the playing field for image analysis. Whereas smaller-scale projects could rely on visual review of images by a trained expert, this traditional approach does not scale well to the requirements of data analysis in large multi-centre studies. To extract the information required to answer a defined research question, automatic anatomical segmentation methods are among the most promising and widely applicable avenues.

An established approach for achieving automatic segmentation is to exploit expert knowledge contained in manual segmentations pertaining to magnetic resonance (MR) images. A variety of algorithms have been proposed. Multi-atlas label propagation, followed by a consolidation (fusion) step has repeatedly been shown to be accurate and robust.

⋆ RAH gratefully acknowledges funding from the Dunhill Medical Trust.

We apply here a label propagation method where expert labels are warped into the target space using a geometric transformation determined through pairwise nonrigid image registration. To increase the robustness of the image registration step against large discrepancies that can arise from, e.g., atrophy, initial global and coarse transformations are calculated from pairs of tissue probability maps, rather than native T1 signal maps. Multiple segmentations resulting from processing multiple atlases with a single target set are consolidated in the space of the target using vote-rule decision fusion [1]. We previously described performance characteristics of the underlying multi-atlas method [2] and the tissue-probability based enhancement ("MAPER") [3]. MAPER-generated segmentations of the baseline and screening images acquired by ADNI are publicly available [4].

2 Method

2.1 Material

We downloaded the data for the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling", consisting of 35 images in total, originating from the OASIS project (http://oasis-brains.org). T1-weighted images of 15 subjects had been labelled as training data and supplied with corresponding label sets, which had been generated by manual delineation of 138 regions[1]. Testing data consisted of 20 T1-weighted images of 16 subjects. Label sets for the testing data were hidden from the contestants.

2.2 Image registration

Probabilistic classification of intracranial voxels into tissue classes (grey matter, white matter, and cerebrospinal fluid) was performed on the atlas and target images. The partial volume estimates from the tissue classification were combined into a multispectral image volume, with each channel of the image representing a partial volume estimate for one of the three tissue classes. The atlas and target images were then aligned using affine and coarse nonrigid (20 mm control point spacing, CPS) registration. As a departure from our previous implementations, we did not use the summed cross-correlation as the similarity measure to maximize. Instead, we minimized Kulback-Leibler divergence across all channels of the multi-spectral image volume.

The resulting transformation was then used as a starting point for a more detailed registration (10, 5, and 2.5 mm CPS), where normalized mutual information (NMI) between the signal intensities of a T1 image pair is maximized. Displacements were applied to the atlas image via a lattice of control points and blended using B-spline basis functions [5]. At each resolution level, the output transformation of the previous stage was used as the starting point.

[1] Label sets were provided by Neuromorphometrics, Inc. (http://neuromorphometrics.com/) under academic subscription.

2.3 Label fusion

Each pairing of an atlas with a target set yields a label set that uniquely assigns an anatomical label to each target voxel. To consolidate these multiple label sets, the per-voxel modal value of all label assignments was chosen as the final unique assignment (vote-rule decision fusion [1]). In the case of multiple modes, the final label was chosen at random from the tied label values.

2.4 Parameter modifications

To generate tissue probability maps, we subsampled the input images to a resolution of $2 \times 2 \times 2$ mm before applying FSL FAST. This led to an acceleration of the global and coarse registration steps without loss of accuracy.

2.5 Software toolkits

Tissue probability maps were obtained using FAST from the FSL suite [6] and combined using "fslmerge".

The tools used for affine ("reg_aladin") and nonrigid ("reg_f3d") registration were obtained from the Nifty Reg toolkit, an efficient implementation of B-spline warping [7].

Vote rule decision fusion was applied using "combineLabels" from IRTK (www.doc.ic.ac.uk/~dr/software/).

3 Discussion

Its characteristics predestine the MAPER method for certain application scenarios. For example, using normalized mutual information as a similarity metric in the high-dimensional registration steps entails robustness against acquisition differences. MAPER is thus particularly suitable if atlas and target (training and testing) images have been acquired differently, ie. on different scanners, at different centres, or using different sequences. Using tissue probability maps for coarsely aligning atlas and target images relaxes the usually strict requirement that the atlas set be anatomically representative of the target set. Consequently, MAPER performs better than other approaches when target images with severe atrophy are to be segmented with atlases of young, healthy subjects [3]. Neither of these strengths is relevant in the Grand Challenge. Nevertheless, we participate with this method for two reasons. First, the enhancements have been developed with the stated objective of avoiding sacrifices of accuracy in "easy" application scenarios, so we expect its performance on the Grand Challenge data to be reasonable. Second, MAPER output can serve further development, both as a foundation and as a lower-bounds benchmark: for testing sophisticated segmentation combination strategies, it delivers individual segmentations, plus the result from vote-rule fusion to indicate the level of accuracy that any newly developed method should be able to beat. A separate entry to the Grand Challenge, provided by co-author CL, will use MAPER in this way.

References

1. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans Pattern Analysis and Machine Intelligence **20**(3) (March 1998) 226–239
2. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage **33**(1) (October 2006) 115–126
3. Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A.: Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. Neuroimage **51**(1) (May 2010) 221–227
4. Heckemann, R.A., Keihaninejad, S., Aljabar, P., Gray, K.R., Nielsen, C., Rueckert, D., Hajnal, J.V., Hammers, A.: Automatic morphometry in Alzheimer's disease and mild cognitive impairment. Neuroimage **56**(4) (June 2011) 2024–2037
5. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging **18**(8) (August 1999) 712–721
6. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging **20**(1) (January 2001) 45–57
7. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine **98**(3) (June 2010) 278–284

Multi-Atlas Segmentation using Non-Local STAPLE

Andrew J. Asman[1], Bennett A. Landman[1]

[1] Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235
{andrew.j.asman, bennett.landman}@vanderbilt.edu

Abstract. Multi-atlas segmentation provides a general purpose, fully automated class of techniques for transferring spatial information from an existing dataset ("atlases") to a previously unseen context ("target") through image registration. The method used to combine information after registration ("label fusion") has a substantial impact on the overall accuracy and robustness. We demonstrate the use of a recently proposed label fusion algorithm, Non-Local STAPLE, for use in a general framework for multi-atlas segmentation. Non-Local STAPLE reformulates the traditional STAPLE framework from a non-local means perspective. As a result, Non-Local STAPLE attempts to learn which label a given rater (registered atlas) would have observed given perfect correspondence between the target and the atlas. In the end, we demonstrate a general multi-atlas segmentation framework that results in accurate and robust estimates for whole-brain multi-atlas multi-label segmentation.

Keywords: Simultaneous Truth And Performance Level Estimation (STAPLE), Non-Local STAPLE, Multi-Atlas Segmentation

1 Introduction

The *de facto* standard baseline for large-scale, consistent, and robust segmentation is to perform a multi-atlas segmentation in which a collection of canonical atlases (with labels) are used to segment a target-of-interest [1, 2]. Herein, we focus on the problem of label fusion (i.e., resolving voxelwise label conflicts between the various registered atlases).

In general, there are two families of approaches for performing label fusion: (1) voting fusion (e.g., [3-5]) and (2) statistical fusion (i.e., approaches based upon Simultaneous Truth and Performance Level Estimation – STAPLE [6]). In this manuscript we use Non-Local STAPLE (NLS) [7]. NLS is a recently proposed label fusion algorithm that has been shown to be highly robust, particularly when faced with highly variable anatomy and low quality registration. NLS reformulates the STAPLE framework from a non-local means perspective in order to seamlessly integrate exogenous intensity information into the estimation process to provide a theoretically consistent model of multi-atlas observation error. In words, NLS provides a mechanism for learning which label a given atlas *would have observed* given perfect correspondence between the target and the atlas.

In this manuscript, we (1) outline our approach for performing whole-brain multi-atlas segmentation, (2) present the results of a leave-one-out cross-validation experiment on provided training data, and (3) provide brief concluding remarks.

2 Approach

All studies were run on a 64 bit quad-core 3.07GHz computer with 13GB of RAM running Ubuntu 11.04.

2.1 Data

The provided data consists of a collection 35 (15 training and 20 testing) atlases that are part of the Open Access Series of Imaging Studies (OASIS) [8] dataset. Each atlas was manually labeled by an expert anatomist (courtesy of Neuromorphometrics, Inc. Boston, MA). In total there were approximately 140 structures labeled on each atlas.

2.2 Registration

In a comparison between various deformable registration algorithms [9], the SyN registration algorithm ([10], *http://www.picsl.upenn.edu/ANTS/*) was consistently shown to be a top performer for multi-atlas segmentation. Here, we use the same parameters specified in [9] to perform all pairwise registrations between the targets and the atlases. Syn required approximately 2 hours of runtime per registration.

2.3 Intensity Normalization

As NLS uses the intensity differences between the atlases and the target in order to infer non-local correspondence, intensity normalization between the targets and the atlases is an essential component of the NLS fusion process. Here, we normalize the intensities in a two-step process after the registration is performed. First, both the target and the registered atlas intensities are normalized to the 25^{th} and 75^{th} percentiles within the brain region. Second, a 2^{nd} order polynomial is fit to each atlas by finding a least squares solution for the polynomial coefficients that map the mean of each label on the target (via an initial majority vote) to the corresponding labels on the atlases.

2.4 Non-Local STAPLE

For a full derivation and description of all of the parameters of the NLS fusion algorithm see [7]. Here, NLS was initialized with performance parameters equal to 0.95 along the diagonal and randomly setting the off-diagonal elements to fulfill the required constraints. For all presented results, the voxelwise label prior, $f(T_i = s)$, was initialized using the probabilities from a "weak" log-odds majority vote (i.e., decay coefficient set to 0.5) [4], the search neighborhood, $\mathcal{N}_s(i)$, was initialized to a $13 \times 13 \times 13$ voxel window centered at the target voxel of interest, and the patch neighborhood, $\mathcal{N}_p(\cdot)$, was initialized to a $3 \times 3 \times 3$ voxel window. The values of the standard deviation parameters, σ_i and σ_d, were set to 0.1 and 3, respectively. Consensus voxels (voxels where $\max_s f(T_i = s) > 0.95$) were ignored during the estimation process. Lastly, convergence of the algorithm was detected when the average change in the trace of the performance level parameters fell below 10^{-4}. For all presented results NLS required approximately 3 hours of runtime per target volume to converge.

3 Results on Training Set

The results of a leave-one-out cross-validation experiment on the 15 provided training atlases can be seen in Figure 1. The mean Dice Similarity Coefficient (DSC) across the training subjects on all considered labels is 0.7825 ± 0.0098, on cortical labels is 0.7555 ± 0.0135 and on non-cortical labels is 0.8559 ± 0.0120.

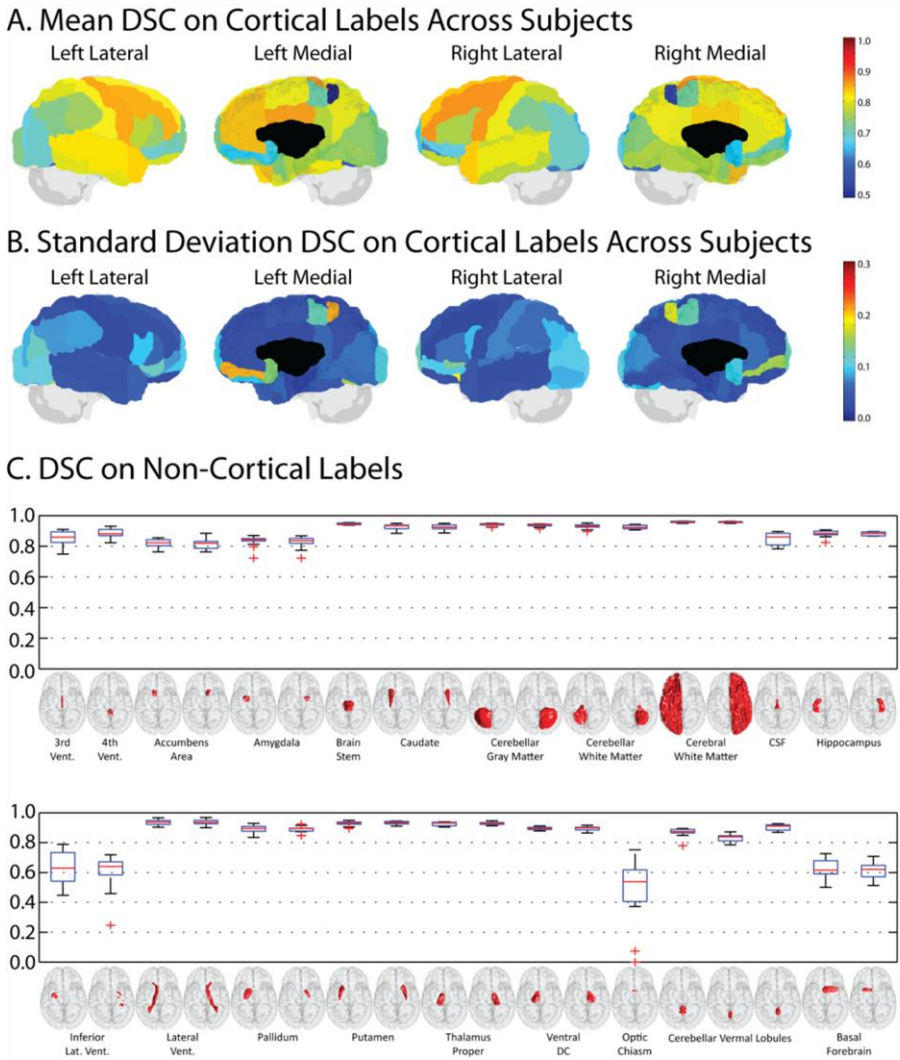


Fig. 1. Results on a leave-one-out cross-validation experiment on the provided training data. The top indicates the mean and standard deviation of the DSC on the cortical labels across the training subjects. The bottom indicates the accuracy on the non-cortical labels.

4 Discussion

Non-Local STAPLE represents a promising statistical fusion algorithm that creates a cohesive theoretical model specifically targeting registered atlas observation behavior. Here, we have presented highly promising results for whole brain segmentation using a fully general multi-atlas segmentation framework.

Nevertheless, several opportunities for future advancement remain. Recently, a myriad of advancements to the STAPLE framework have been suggested (e.g., [11, 12]). Incorporation of these advancements into the NLS fusion model presents fascinating areas of continuing research. Lastly, integration of Markov Random Fields [4, 6] and global/local atlas pre-selection into NLS could provide valuable benefits in terms of segmentation accuracy (e.g., in the presence of highly variable anatomy).

Acknowledgments: This work was supported in part by NIH/NINDS 1R01EB006136, 1R01EB006193, 1R03EB012461, and 1R21NS064534.

References

1. Heckemann, R.A., *et al.*: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115-126 (2006)
2. Rohlfing, T., *et al.*: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. Medical Imaging, IEEE Transactions on 23, 983-994 (2004)
3. Artaechevarria, X., *et al.*: Combination strategies in multi-atlas image segmentation: Application to brain MR data. Medical Imaging, IEEE Transactions on 28, 1266-1277 (2009)
4. Sabuncu, M.R., *et al.*: A generative model for image segmentation based on label fusion. Medical Imaging, IEEE Transactions on 29, 1714-1729 (2010)
5. Coupé, P., *et al.*: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54, 940-954 (2011)
6. Warfield, S.K., *et al.*: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. Medical Imaging, IEEE Transactions on 23, 903-921 (2004)
7. Asman, A.J., Landman, B.A.: Non-Local STAPLE: An Intensity-Driven Multi-Atlas Rater Model. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012, Nice, France (2012)
8. Marcus, D.S., *et al.*: Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience 19, 1498-1507 (2007)
9. Klein, A., *et al.*: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786-802 (2009)
10. Avants, B., *et al.*: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26-41 (2008)
11. Asman, A.J., Landman, B.A.: Formulating Spatially Varying Performance in the Statistical Fusion Framework. IEEE transactions on medical imaging 31, 1326 - 1336 (2012)
12. Asman, A., Landman, B.: Robust Statistical Label Fusion through Consensus Level, Labeler Accuracy and Truth Estimation (COLLATE). Medical Imaging, IEEE Transactions on 30, 1779-1794 (2011)

Grand Challenge on Multi-Atlas Segmentation: A Combined Joint Label Fusion and Corrective Learning Approach

Hongzhi Wang, Brian Avants, and Paul A. Yushkevich

Department of Radiology, University of Pennsylvania

Abstract. We describe technical and implementation details of our algorithm that was used to participate the MICCAI grand challenge on multi-atlas segmentation in 2012.

1 Methods

Our segmentation system contains three sequential components: 1) image-based deformable registration between each atlas image and each testing image, from which the manual labels of the atlas image is propagated to the testing image, 2) label fusion that integrates the labels propagated from different atlases for the same testing image, 3) corrective learning that correct the systematic errors produced by the label fusion technique with respect to the manual segmentation. We describe each component in detail below.

1.1 Image-based deformable registration

The image-based registration between each pair of images were performed in two steps: a global rigid registration and a deformable registration. Global registration was performed using the FSL FLIRT tool [3] with six degrees of freedom and using the default parameters (normalized mutual information similarity metric; search range from -5 to 5 in x, y and z). Based on the global rigid registration, deformable registration was performed using the greedy diffeomorphic Symmetric Normalization (SyN) algorithm implemented by ANTS [1]. SyN registrations used the cross-correlation metric with a $3 \times 3 \times 3$ window; 3 resolution levels with maximum 80 iterations at the coarse and middle levels and 60 iterations at the finest level; step size 0.25; Gaussian regularization with standard deviation of 3 pixels. After registration, reference segmentations from each of the atlases were warped into the target image space. The computation time for registering each pair of images is about 20 hours on a 2G HZ CPU. Overall, 300 registrations between each pair of 15 atlas images and 20 testing images and 210 registrations between each pair of the atlas images were computed.

1.2 Joint Label Fusion

We applied image similarity based local weighted voting for combining the candidate segmentations produced by different atlases for the same target image. The voting weights were computed using the joint label fusion technique [5]. We briefly summarize this technique in this section.

Let T_F be a target image to be segmented and $A^1 = (A^1_F, A^1_S), ..., A^n = (A^n_F, A^n_S)$ be n atlases. A^i_F and A^i_S denote the i_{th} warped atlas image and the corresponding warped manual segmentation of this atlas, obtained by performing deformable image registration to the target image. The segmentation error produced by one atlas is:

$$T_{S,l}(x) = A^i_{S,l}(x) + \delta^i(x) \tag{1}$$

where $T_{S,l}(x), A^i_{S,l}(x) \in \{0, 1\}$ are the observed votes for label l produced by the target image and the i_{th} warped atlas, respectively. Hence, $\delta^i(x) \in \{-1, 0, 1\}$ is the observed label difference. The probability that different atlases produce the same label error at location x is captured by a dependency matrix M_x, with $M_x(i, j) = p(\delta^i(x)\delta^j(x) = 1 \mid T_F, A^i_F, A^j_F)$ measuring the error correlation between i_{th} and j_{th} atlases, which is estimated by intensity differences as:

$$M_x(i, j) \sim \left[\sum_{y \in \mathcal{N}(x)} |A^i_F(y) - T_F(y)||A^j_F(y) - T_F(y)| \right]^{\beta} \tag{2}$$

where $\mathcal{N}(x)$ is a neighborhood centered at x. In our experiment, we normalize the intensity vector obtained from each local image intensity patch, such that the normalized vector has zero mean and a constant norm.

The expected label difference between the combined solution and the target segmentation is:

$$E_{\delta^1(x),...,\delta^n(x)} \left[(T_{S,l}(x) - \sum_{i=1}^{n} w_x(i) A^i_{S,l}(x))^2 \mid F_T, F_1, ..., F_n \right] \approx w^t_x M_x w_x \tag{3}$$

where t stands for transpose and $w_x(i)$ is the voting weight for A^i at x. To minimize the expected label difference, the voting weights are:

$$\mathbf{w}_x = \frac{M_x^{-1} 1_n}{1^t_n M_x^{-1} 1_n} \tag{4}$$

where $1_n = [1; 1; ...; 1]$ is a vector of size n. To avoid inverting an ill-conditioned matrix M_x, we adding an identity matrix weighted by a small positive number $\alpha = 0.1$ to M_x. With the conditioning matrix, we minimize the following objective function instead:

$$\mathbf{w}^t_x \left(M_x + \alpha I\right) \mathbf{w}_x = \mathbf{w}^t_x M_x \mathbf{w}_x + \alpha \|\mathbf{w}_x\|_2 \quad \text{subject to} \quad \sum_{i=1}^{n} \mathbf{w}_x(i) = 1 \tag{5}$$

Hence, adding a small conditioning identity matrix can be interpreted as enforcing a regularization term that prefers more similar voting weights assigned to different atlases.

The local search technique described in [5] were applied to remedy the registration errors. To enhance the spatial consistencies of voting weights for nearby voxels, we apply mean filter smoothing with the smoothing window $\mathcal{N}$, the same neighborhood used for local appearance patches, to spatially smooth the voting weights for each atlas.

Our method has three free parameters: r, the radius of the local appearance window $\mathcal{N}$ used in similarity-based M_x estimation; r_s, the radius of the local searching window $\mathcal{N}'$ used in remedying registration errors; and β, the parameter used to transfer image similarities in the pairwise joint label difference term (2). The parameters are optimized by exhaustive search among a range of values in each parameter ($r \in \{1, 2, 3\}$; $r_s \in \{0, 1, 2, 3, 4\}$; $\beta \in \{0.5, 0.75, ..., 3\}$) using the atlases in a leave-one-out cross-validation strategy. We measure the average overlap between the automatic segmentation of each atlas obtained via the remaining atlases and the reference segmentation of that atlas, and find the optimal parameters that maximize this average overlap. The selected parameter for segmenting the testing images are $(r, r_s, \beta) = (2, 4, 1.75)$. With the selected parameters, our label fusion algorithm finishes processing one brain image in about three hours. The local search component is the most time consuming part, without local search, our algorithm processes one brain image within 15 minutes.

1.3 Error correction by corrective learning

The joint label fusion technique may produce systematic segmentation errors with respect to the manual segmentation (see [6] for one type of spatial bias produced by weighted voting). To reduce such bias, we apply the corrective learning technique described in [4]. This method applies learning as an error correction tool to improve the segmentation produced by a host segmentation method. In our experiment, the implementation by the segadapter software (available at http://www.nitrc.org/projects/segadapter/) was applied.

To apply this approach, a region of interest (ROI) was defined for each label by dilating the set of voxels assigned to the label by joint label fusion by one voxel. One AdaBoost classifier was trained to identify the voxels assigned to the target label by manual segmentation within the label's ROI. The features used in [4], including spatial, appearance and contextual, joint spatial-appearance and joint spatial-contextual features, were applied to train the classifiers, where the contextual features were extracted from the initial segmentation produced by the host method. All features were extracted within a patch of size $5 \times 5 \times 5$. 500 iterations were used to train each classifier.

The classifiers were trained using the atlases in a leave-one-out fashion. Each atlas was segmented by the remaining atlases using the joint label fusion approach with the selected parameters. All atlases were used to learn the classifiers. Overall, we trained 138 classifiers for all the labels. The training time for each

label depends on the volume of the target label, which varies from a few minutes to several hours. In total, 330 CPU hours were used for training all classifiers.

To apply these trained classifiers to correct segmentation errors for a testing image, we apply each classifier to evaluate the confidence of assigning the corresponding label to each voxel within its ROI. If a voxel belongs to the ROI of multiple labels, the label whose classifier gives the maximal response at the voxel is chosen for the voxel.

1.4 Additional implementation details

It is advised that label 42, 43, 126 and 127 should be ignored. Voxels assigned to these four labels were treated as background voxels, i.e. with label 0.

2 Results

15 atlases and 20 testing images were used in this study. The MRI scans are obtained from the OASIS project and the manual segmentations are provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/).

The segmentation accuracy is evaluated in term of Dice coefficient [2]. Results are summarized for all labels, cortical labels and non-cortical labels, respectively. By applying joint label fusion alone, we produced the following accuracy: 0.766 ± 0.013 for all labels, 0.736 ± 0.015 for cortical labels, and 0.848 ± 0.010 for non-cortical labels. Applying corrective learning improved the segmentation accuracy to 0.782 ± 0.010, 0.753 ± 0.012, and 0.861 ± 0.009, respectively.

References

1. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12(1), 26–41 (2008)
2. Dice, L.: Measure of the amount of ecological association between species. Ecology 26, 297–302 (1945)
3. Smith, S., Jenkinson, M., Woolrich, M., Beckmann, C., aand H. JohansenBerg, T.B., Bannister, P., Luca, M., Drobnjak, I., Flitney, D., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., Stefano, N., Brady, J., Matthews, P.: Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23(Suppl 1), S208S219 (2004)
4. Wang, H., Das, S., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.: A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. NeuroImage 55(3), 968–985 (2011)
5. Wang, H., Suh, J.W., Das, S., Pluta, J., Craige, C., Yushkevich, P.: Multi-atlas segmentation with joint label fusion. IEEE Trans. on Pattern Analysis and Machine Intelligence (2012)
6. Wang, H., Yushkevich, P.A.: Spatial bias in multi-atlas based segmentation. In: CVPR (2012)

Attribute Similarity and Mutual-Saliency Weighting for Registration and Label Fusion

Yangming Ou, Jimit Doshi, Guray Erus, and Christos Davatzikos

Section of Biomedical Image Analysis (SBIA)
Department of Radiology, University of Pennsylvania

Abstract. Multi-atlas segmentation relies on two major components: image registration to propagate segmentation labels, and label fusion to combine multiple labels into one at each voxel. In this paper, we propose to drive both components by an attribute-based similarity metric and a mutual-saliency-based reliability metric. The fundamental idea is to improve registration and label fusion by looking for corresponding voxels that are similar (as measured by their Gabor attributes), and more importantly, reliably similar (as measured by the mutual-saliency of their matching) between atlas and target images. We apply this pipeline to segment 140 structures in brain MRI of 15 training subjects and 20 testing subjects in MICCAI Challenge 2012.

1 Introduction

Multi-atlas segmentation has gained increasing interest in recent years [1–5]. One premise in this approach is that it allows using *a priori* knowledge, as encoded in atlas segmentations, to infer segmentation in target image via atlas-to-target image registration. Another premise is that it allows different atlases to correct each other's errors in a process often known as label fusion. The fused segmentation has shown remarkable improvement over single-atlas-based segmentation in various brain, cardiac and prostate structures.

Despite exciting research in recent years, both image registration and label fusion are not without challenges. In registration, a fundamental question is how to find reliable correspondences across images. This is especially important when there exist considerable structural difference between atlas and target images.

In label fusion, recent studies have obtained improved accuracy by assigning higher weights to atlases that are more similar to target at local voxel level [2, 3]. But a fundamental question is how to properly measure similarity between atlas and target at voxel level. Researchers have used correlation or intensity difference to imply voxel similarities [2, 3]. Ideally there can be a more robust similarity measure incorporating richer geometric context of each voxel. In addition, we hypothesize that a proper reliability measure (i.e., whether an atlas voxel and a target voxel are reliably matched) will further improve label fusion accuracy too. Here matching between two voxels are said reliable if they are similar to each other and meanwhile not similar to anything else in the neighborhood [6]. This is a higher level of confidence in the matching and label inheritance.

In this paper, we propose to improve both registration and label fusion by attribute-based similarity and matching reliability metrics. The idea is the following. When registering atlas to target, we rely more on those regions, compared to other regions, that can establish more reliable matching. When fusing labels, we assign higher confidence/weight to those atlases, compared to other atlases, that are more reliably similar to the target at each voxel. All experiments are done using 15 training brain MR images and 20 testing MR images in the MICCAI 2012 Multi-Atlas Segmentation Challenge.

2 Methods

In this section, we first introduce attribute-based similarity metric and mutual-saliency-based matching-reliability metric (Sec. 2.1). Then we describe their use in guiding registration (Sec. 2.2) and guiding label fusion (Sec. 2.3).

2.1 Definition of Attribute Similarity and Mutual-Saliency

These two concepts were proposed in our recent paper [6]. For the completeness of this paper, below is a brief description. First of all, we represent each voxel $\mathbf{x}$ by geometric context around this voxel, in a d-dimensional multi-scale and multi-orientation Gabor attribute vector $A(\mathbf{x})$. This attribute representation has rendered each voxel more distinctive than intensity information alone [6]. Then, we shall say that two voxels $\mathbf{x}$ and $\mathbf{y}$ in two images are similar, if we observe small difference in their attribute representations, i.e., $\text{sim}(\mathbf{x},\mathbf{y}) = \frac{1}{1+\frac{1}{d}\|A(\mathbf{x})-A(\mathbf{y})\|^2}$.

A pair of voxels $\mathbf{x}$, $\mathbf{y}$ in two images is said mutually-salient, if they are similar to each other and meanwhile less similar to any other voxels in the neighborhood. As shown in Fig. 1, similarity map between $\mathbf{x}$ and all voxels in the vicinity of $\mathbf{y}$ exhibit a delta-shape distribution peaking at $\mathbf{y}$. What this means is that the matching between those two voxels are reliable, because no other voxel in the neighborhood of $\mathbf{y}$ can replace it with same high similarity to $\mathbf{x}$.

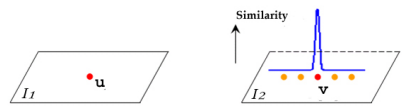


Fig. 1. concept of mutual-saliency to measure matching reliability.

Mathematically, mutual-saliency, $\text{ms}(\mathbf{x},\mathbf{y})$, is approximated by dividing the mean similarity between voxel $\mathbf{x}$ and all voxels in the core neighborhood (CN) of $\mathbf{y}$, by the mean similarity between voxel $\mathbf{x}$ and all voxels in the peripheral neighborhood (PN) of $\mathbf{y}$, where CN and PN are defined in accordance to the scale where Gabor attributes are extracted (see [6] for more details), i.e., $\text{ms}(\mathbf{x},\mathbf{y}) = \frac{\frac{1}{|CN(\mathbf{y})|}\sum_{\mathbf{w}\in CN(\mathbf{y})} sim(\mathbf{x},\mathbf{w})}{\frac{1}{|PN(\mathbf{y})|}\sum_{\mathbf{w}\in PN(\mathbf{y})} \text{sim}(\mathbf{x},\mathbf{w})}$.

Fig. 2 shows a typical set of similarity and mutual-saliency maps from an atlas-to-target registration. Matching in cortical regions observes lower similarity than matching in deep brain structures (as shown in similarity map), and lower reliability (as shown in mutual-saliency map). Contrary is the matching in ventricle and peri-ventricle white matter regions. So we would have more confidence in following the warped segmentation labels in latter regions.

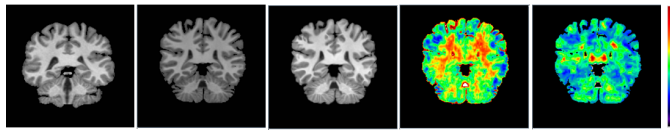


Fig. 2. A typical set of similarity map and mutual-saliency map resulted from registration from an atlas to the target image.

2.2 Registration

The above defined similarity and mutual-saliency are used to modulate registration, as implemented in the DRAMMS software [6]. Specifically, DRAMMS seeks a non-rigid transformation T, based on free form deformation (FFD) model [7], that minimizes the mutual-saliency-weighted attribute differences over target image domain $\Omega \subset \mathrm{R}^3$,

$$\arg\max_{T} \; Energy(T) = \int_{\mathbf{u}\in\Omega} \mathrm{ms}(T^{-1}(\mathbf{u}), \mathbf{u}) \cdot \tfrac{1}{d}\|A(T^{-1}(\mathbf{u})) - A(\mathbf{u})\|^2 \mathrm{d}\mathbf{u} \quad (1)$$

In essence, voxels are matched by their geometric context other than intensity. And, the whole registration is mainly driven by regions/voxels that can reliably match across images.

2.3 Label Fusion

DRAMMS registration maps segmentation regions from different atlases into the same target image. Now we need label fusion to combine those multiple labels at each voxel. Assuming that N atlases, indexed by n, have been each registered to the same target image via a deformation T_n. A voxel $\mathbf{u}$ in the target image space Ω will tentatively have N segmentation labels propagated from all those N atlases, denoted as $\{\mathrm{label}(T_n^{-1}(\mathbf{u}))\}_{n=1}^{N}$. To fuse them into a single segmentation label, we use a similarity and mutual-saliency weighted voting strategy. Specifically, we first calculate the probability of this voxel having each of all L segmentation labels $\{1, 2, \ldots, L\}$, i.e., $\forall l \in 1, 2, ..., L$

$$\Pr(\mathrm{label}(\mathbf{u}) = l) = \frac{\sum_n \mathrm{sim}(T_n^{-1}(\mathbf{u}), \mathbf{u}) \cdot \mathrm{ms}(T_n^{-1}(\mathbf{u}), \mathbf{u}) \cdot \mathbf{1}(\mathrm{label}(T_n^{-1}(\mathbf{u})) = l)}{\sum_n \mathrm{sim}(T_n^{-1}(\mathbf{u}), \mathbf{u}) \cdot \mathrm{ms}(T_n^{-1}(\mathbf{u}), \mathbf{u})} \quad (2)$$

Then, we assign the most likely label $l^\star$ to this voxel $\mathbf{u}$, i.e., $\mathrm{label}(\mathbf{u}) = l^\star$, such that $l^\star = \arg\max_l \Pr(\mathrm{label}(\mathbf{u}) = l)$. In extreme cases, if $\mathrm{sim}(.,.) \equiv 1$ and $\mathrm{ms}(.,.) \equiv 1$, we end up with the classic majority voting, as all atlases are equally trusted at each voxel.

3 Results

Fig. 3 shows leave-one-out results in training dataset (15 subjects from OASIS dataset). We compared the proposed $(sim \times ms)$-weighted voting mechanism with classic majority voting for label fusion. We have several observations:

1) $(sim \times ms)$-weighted voting always improves majority voting.

2) Brain ROI segmentation accuracy is sensitive to initial skull-stripping. A perfect skull-stripping (using ground-truth) helps improve segmenting brain structures. When ground-truth skull-stripping is not used, segmentation accuracy is similar between no skull-stripping and automatic skull-stripping (multi-atlas segmentation with DRAMMS registration and majority voting).

3) Accuracy is also a bit sensitive to how the Dice scores in all 140 brain ROIs are averaged. The ROI-volume-weighted average Dice score (in left figure) is less sensitive to skull-stripping accuracy than the direct average Dice score.

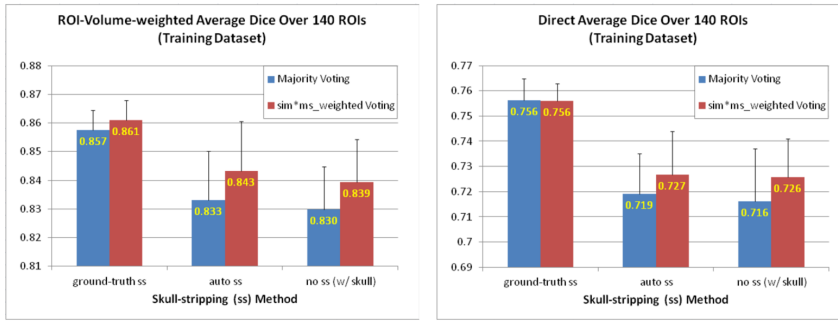


Fig. 3. Leave-one-out resuts in training dataset (15 subjects).

References

1. Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A., Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage. 15;33(1):115-26, (2006).
2. Artaechevarria X, Munoz-Barrutia A, Ortiz-de-Solorzano C., Combination strategies in multi-atlas image segmentation: application to brain MR data, IEEE Trans Med Imaging. 28(8):1266-77, (2009).
3. Sabuncu MR, Yeo BT, Van Leemput K, Fischl B, Golland P., A generative model for image segmentation based on label fusion. IEEE TMI 29(10):1714-29, (2010).
4. Warfield SK, Zou KH, Wells WM, Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation, IEEE Trans Med Imag 23(7), 903-21 (2004).
5. Asman AJ, and Landman BA, Robust Statistical Label Fusion Through Consensus Level, Labeler Accuracy, and Truth Estimation (COLLATE), IEEE Trans Med Imag, 30(10), 1779-1794 (2011).
6. Ou Y, Sotiras A, Paragios N, Davatzikos C, DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. MedIA 15(4):622-39, (2011).
7. Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging. 18(8):712-21. (1999).

Multi-label similarity and truth estimation for propagated segmentations (STEPS) validation

M. Jorge Cardoso[1], Marc Modat[1], Sebastien Ourselin[1,2]

[1] Centre for Medical Image Computing (CMIC), University College London, UK,
[2] Dementia Research Centre (DRC), University College London, UK

Abstract. Quantitative analysis in medical imaging often relies on the segmentation of anatomical images. Several multi-atlas based segmentation propagation methods have recently been published due to the accurate structural segmentations produced by propagating and combining manual delineations from multiple templates in a database. In this paper, we validate a previously published multi-label fusion algorithm named STEPS. This algorithm uses a local ranking strategy for template selection based on the locally normalised cross correlation (LNCC) and an extension to the classical STAPLE algorithm by Warfield et al. [1]. Results show good segmentation accuracy for key cortical and sub-cortical brain areas. Comparison with other fusion strategies will be performed during the MICCAI 2012 Multi-Atlas Label Fusion challenge.

1 Introduction

Recent advances in medical imaging research, mainly improvements in both speed and accuracy of image registration strategies, have enabled a new breed of segmentation strategies based on the propagation of a manually labeled template to a new unseen image.

Several multi-atlas based segmentation propagation methods have recently been published due to the accurate structural segmentations produced by propagating and combining manual delineations from multiple templates in a database. In order to fuse the propagated labels, several techniques have explored either maximum likelihood probabilistic frameworks (STAPLE, STAPLER, COLLATE) or weighted classifier fusion (Majority Voting, Locally Weighted Voting, SIMPLE), with the weights estimated using either local or global similarity metrics. In this work, we validate the recently proposed multi-label local ranking strategy for template selection based on the locally normalised cross correlation (LNCC) and an extension to the classical STAPLE algorithm by Warfield et al. [1] on the *Neuromorphometrics, inc.* dataset. This technique also reduces the bias towards large structures, thus improving the segmentation results. Parameter optimisation is done using a leave-one-out cross-validation methodology on the training set and the optimal parameters are then used to segment the test cases. The performance of STEPS will be compared to other fusion strategies during the MICCAI 2012 Multi-Atlas Label Fusion challenge.

2 Methods

2.1 The STEPS algorithm

In this paper we use the STEPS (Similarity and Truth Estimation for Propagated Segmentations) [2] algorithm. This method uses a probabilistic formulation of the label fusion problem where the likelihood of the complete data (D,T) is maximised given the set of row normalised confusion matrices λ_j, with $\hat{\lambda} = \arg\ \max_\lambda \log(f(D,T|\lambda_j))$. Here, D is a matrix containing a label at each spatial position and for each candidate segmentation, obtained either by manual segmentation or an automatic algorithm, T is an indicator vector representing the hidden true label describing several objects under analysis, and λ_j represents the global degree of agreement or disagreement between the candidate segmentation j and the consensus.

In a segmentation propagation and label fusion setting, not all the propagated labels provide beneficial information due to registration errors. Thus, by curating and selecting the best local labels to use, one can increase the performance of the fusion algorithm. In this work, we only fuse the locally best ranked templates according to the locally normalised cross correlation between the registered template image and the target image. In order to introduce this local ranking information in the above described framework, we introduce a new binary indicator variable L_{ik}, equal to 1 if the image k is in the top X ranked images at position i and equal to 0 otherwise. Here, X controls the number of images to use locally according to the LNCC. Eq. **??** can then be solved by iteratively calculating

$$W_{ia} = \frac{f(T_i = a)\prod_{j:L_{ij}=1}\lambda_j(a, D_{ij})}{\sum_c f(T_i = c)\prod_{j:L_{ij}=1}\lambda_j(c, D_{ij})} \qquad \lambda_j(a,b) = \frac{\sum_{i:D_{ij}=b \cap L_{ij}=1} W_{ia}}{\sum_{i:L_{ij}=1} W_{ia}}$$

until convergence. Here, $f(T_i = a)$ is a Markov Random Field (MRF) prior

$$f(T_i = a) = \frac{\pi_a\, e^{-\beta_i U(a)}}{\sum_{b=1}^{c} \pi_b\, e^{-\beta_i U(b)}} \qquad U(a) = \sum_{b=1}^{R} H_{ab}\left(\sum_{l\in\mathcal{N}_i} W_{lb}\right)$$

and H is a matrix with element H_{ab} containing the transition energy between the class a and the class b, with $\mathcal{N}_i$ being the 6 closest neighbours of voxel i.

Finally, in order to remove the bias introduced in the parameter optimisation due to different sizes (as described in [2]), we assume that if all the classifiers agree on a label at a certain spatial position i, then the voxel is marked as solved and is not taken into account from the estimation of λ_j. Overall, the method can be described as a combination of the LNCC ranking, the MRF and two STAPLE modifications regarding both the introduction of the local indicator function L_{ij} and the removal of consensus voxels from the parameter estimation.

2.2 Software Availability

All the software used for this work is publicly available on the NiftySeg website at `http://niftyseg.sf.com`.

3 Experiments

In this section, we first present the data used for validation. The model parameters are then optimised on a training set using a leave-one-out cross-validation. Finally, these optimised parameters are used to segment a separate test set.

3.1 Data

This study used a set of 31 subjects from the OASIS database (`http://www.oasis-brains.org/`), separated into two sets: 15 training subjects and 16 test subjects. Scanning parameters are described in **??**. All the subjects had had the brain manually parcelated into 143 different labels according to the protocol available at `www.braincolor.org`.

3.2 Image Registration

Each one of the subjects on the training dataset was first affinely registered (12 DOFs) using a block matching approach [3] and then non-rigidly aligned using a fast free-form registration algorithm [4] to all the 31 training and test subjects. Registration parameters are: 5 voxel control-point spacing, 0.1% weight for bending energy and a 0.1% weight for the logarithm of the Jacobian determinant.

3.3 Parameter Optimization

In order to optimise the parameters of the STEPS algorithm, we performed a leave-one-out cross-validation on all the subjects of the training set. For each one of the 15 training subjects, the manual segmentations of the remaining 14 subjects were propagated using the previously estimated transformations and resampled using nearest-neighbour interpolation in order to maintain their binary nature. The mean of all the Dice scores between each structure of the estimated parcelation and the corresponding structure of the manual parcelation was calculated for different values of gaussian kernel size and number of labels used. The registered templates were locally ranked by setting $L_{ik} = 1$ if the registered template k was in the top X ranked images at position i according to the LNCC metric and to 0 otherwise. The value of X was varied between 3 and 14. As the LNCC metric is dependent of the standard deviation (STD) of the Gaussian kernel, for each value of X, the value of the Gaussian STD was varied between 1 and 3 with an increment of 0.25, in order to find the optimal gaussian kernel size. Optimising the STD of the gaussian is equivalent to finding the optimal scale to compare the anatomical features. The registration parameters were not optimised within the same scheme due to computational complexity. The optimal parameters were found to be $X = 6$ and Gaussian STD= 1.75.

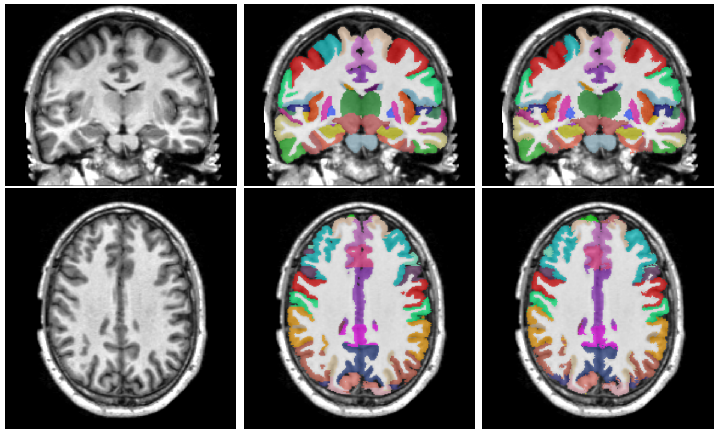


Fig. 1. An example showing a coronal and a axial slice of the anatomical image (left), the automated, the manual segmentation (centre) and the STEPS segmentation (right).

3.4 Results

The proposed framework was then applied to the remaining 16 test subjects. The average Dice score for all the structures was 0.7500 ± 0.0063, with an average Dice score of 0.7202 ± 0.0056 and 0.8311 ± 0.0107 for the cortical and sub-cortical structures respectively. More detailed results are available in the attached Scoring Report.

4 Conclusion

The paper presents a validation of the STEPS algorithm on the *Neuromorphometrics, inc.* dataset. This algorithm incorporates a fast locally normalised cross correlation (LNCC) based ranking combined with a consensus based ROI selection and a new iterative MRF into a multi-label probabilistic formulation, resulting in highly accurate parcelations of key brain structures.

References

1. Warfield, S.K., Zou, K.H., Wells III, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7) (July 2004) 903–921
2. Cardoso, M.J., Modat, M., Keihaninejad, S., Cash, D., Ourselin, S.: Multi-STEPS: Multi-label Similarity and Truth Estimation for Propagated Segmentations. MMBIA (January 2012) 153–158
3. Ourselin, S., Roche, A., Prima, S., Ayache, N.: Block Matching: A General Framework to Improve Robustness of Rigid Registration of Medical Images. In: MICCAI, MICCAI 2000 (2000) 557–566
4. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine **98**(3) (June 2010) 278–284

Multi-Atlas Segmentation using Spatial STAPLE

Andrew J. Asman[1], Bennett A. Landman[1]

[1] Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235
{andrew.j.asman, bennett.landman}@vanderbilt.edu

Abstract. Multi-atlas segmentation provides a general purpose, fully automated class of techniques for transferring spatial information from an existing dataset ("atlases") to a previously unseen context ("target") through image registration. The method used to combine information after registration ("label fusion") has a substantial impact on the overall accuracy and robustness. We demonstrate the use of a recently proposed label fusion algorithm, Spatial STAPLE, for use in a general framework for multi-atlas segmentation. Unlike global approaches, Spatial STAPLE extends the traditional STAPLE framework to seamlessly account for spatially varying performance by extending the performance level parameters to a smooth, voxelwise performance level field that is unique to each rater (or registered atlas). In the end, we demonstrate a general multi-atlas segmentation framework that results in accurate and robust estimates for whole-brain multi-atlas multi-label segmentation.

Keywords: Simultaneous Truth And Performance Level Estimation (STAPLE), Spatial STAPLE, Multi-Atlas Segmentation

1 Introduction

The *de facto* standard baseline for large-scale, consistent, and robust segmentation is to perform a multi-atlas segmentation in which a collection of canonical atlases (with labels) are used to segment a target-of-interest [1, 2]. Herein, we focus on the problem of label fusion (i.e., resolving voxelwise label conflicts between the various registered atlases).

In general, there are two families of approaches for performing label fusion: (1) voting fusion (e.g., [3-5]) and (2) statistical fusion (i.e., approaches based upon Simultaneous Truth and Performance Level Estimation – STAPLE [6]). In this manuscript we use Spatial STAPLE [7]. Spatial STAPLE is a recently proposed algorithm that allows raters to exhibit spatially varying performance. For example, in multi-atlas segmentation the quality of a given registered atlas often varies across the image due to varying degrees of correspondence with the target. Through a reformulation of the traditional STAPLE performance level parameters, Spatial STAPLE provides a smooth voxelwise estimate of a given raters performance, and has been shown to be highly robust, particularly for problems where raters exhibit highly varying quality and reliability.

In this manuscript, we (1) outline our approach for performing whole-brain multi-atlas segmentation, (2) present the results of a leave-one-out cross-validation experiment on provided training data, and (3) provide brief concluding remarks.

2 Approach

All studies were run on a 64 bit quad-core 3.07GHz computer with 13GB of RAM running Ubuntu 11.04.

2.1 Data

The provided data consists of a collection 35 (15 training and 20 testing) atlases that are part of the Open Access Series of Imaging Studies (OASIS) [8] dataset. Each atlas was manually labeled by an expert anatomist (courtesy of Neuromorphometrics, Inc. Boston, MA). In total there were approximately 140 structures labeled on each atlas.

2.2 Registration

In a comparison between various deformable registration algorithms [9], the SyN registration algorithm ([10], *http://www.picsl.upenn.edu/ANTS/*) was consistently shown to be a top performer for multi-atlas segmentation. Here, we use the same parameters specified in [9] to perform all pairwise registrations between the targets and the atlases. Note that this registration algorithm incudes an initial affine registration, followed by highly deformable non-rigid registration. Syn required approximately 2 hours of runtime per registration.

2.3 Spatial STAPLE

For a full derivation and description of all of the parameters of the Spatial STAPLE fusion algorithm see [7]. Here, Spatial STAPLE was initialized with performance parameters equal to 0.95 along the diagonal and randomly setting the off-diagonal elements to fulfill the required constraints. For all presented results, the voxelwise label prior, $f(T_i = s)$, was initialized using the probabilities from a "weak" log-odds majority vote (i.e., decay coefficient set to 0.5) [4], Consensus voxels (voxels where $\max_s f(T_i = s) > 0.95$) were ignored during the estimation process. The window size governing each performance level parameter update was set to be 15% of the length of each dimension on the current region of interest. The number of windows was set using 70% overlap between windows, and linear interpolation was used to interpolate the voxelwise performance level parameters. The voxelwise performance level parameters were regularized using the global performance level parameters form a majority vote (with a bias value of unity). Lastly, convergence of the algorithm was detected when the average change in the trace of the performance level parameters fell below 10^{-6}. For all presented results Spatial STAPLE required approximately 5 hours of runtime per target volume to converge.

Due to the large number of labels and the required memory constraints of the current Spatial STAPLE implementation, the solution for each estimated label was calculated independently. For voxels at which multiple labels were estimated, ties were broken based upon a majority vote result. While it would be ideal to solve for all labels simultaneously, it is of note that, on average, less than 0.1% of the voxels in the target volume resulted in multiple estimated labels.

3 Results on Training Set

The results of a leave-one-out cross-validation experiment on the 15 provided training atlases can be seen in Figure 1. The mean Dice Similarity Coefficient (DSC) across the training subjects on all considered labels is 0.7646 ± 0.0107, on cortical labels is 0.7361 ± 0.0151 and on non-cortical labels is 0.8422 ± 0.0138.

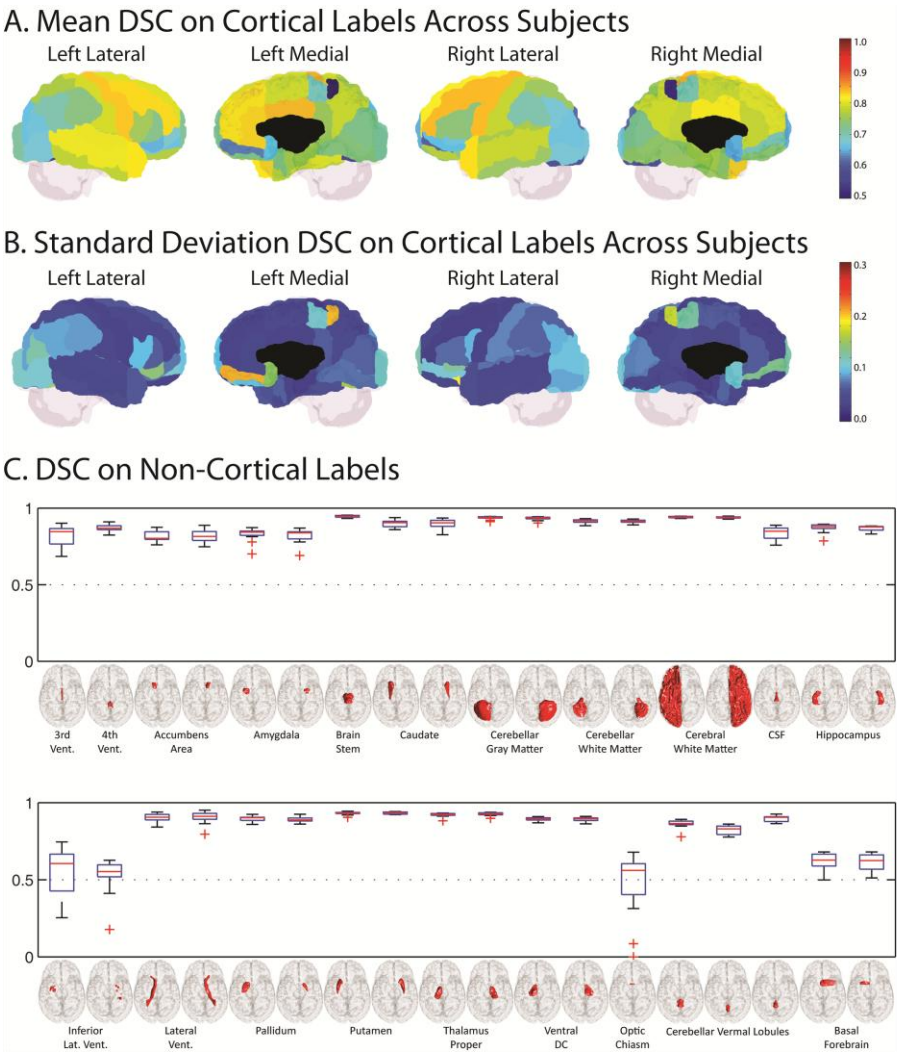


Fig. 1. Results on a leave-one-out cross-validation experiment on the provided training data. The top indicates the mean and standard deviation of the DSC on the cortical labels across the training subjects. The bottom indicates the accuracy on the non-cortical labels.

4 Discussion

Spatial STAPLE represents a promising statistical fusion algorithm that allows for smooth, voxelwise performance level parameters to be estimated for each rater. Here, we have presented highly promising results for whole brain segmentation using a fully general multi-atlas segmentation framework.

Nevertheless, several opportunities for future advancement remain. Recently, a myriad of advancements to the STAPLE framework have been suggested (e.g., [11, 12]). Incorporation of these advancements into the Spatial STAPLE fusion model presents fascinating areas of continuing research. Additionally, integration of intensity information into the Spatial STAPLE estimation process could provide important improvements in segmentation accuracy and algorithm performance (particularly for highly variable anatomy and low quality registration).

Acknowledgments: This work was supported in part by NIH/NINDS 1R01EB006136, 1R01EB006193, 1R03EB012461, and 1R21NS064534.

References

1. Heckemann, R.A., *et al.*: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115-126 (2006)
2. Rohlfing, T., *et al.*: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. Medical Imaging, IEEE Transactions on 23, 983-994 (2004)
3. Artaechevarria, X., *et al.*: Combination strategies in multi-atlas image segmentation: Application to brain MR data. Medical Imaging, IEEE Transactions on 28, 1266-1277 (2009)
4. Sabuncu, M.R., *et al.*: A generative model for image segmentation based on label fusion. Medical Imaging, IEEE Transactions on 29, 1714-1729 (2010)
5. Coupé, P., *et al.*: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54, 940-954 (2011)
6. Warfield, S.K., *et al.*: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. Medical Imaging, IEEE Transactions on 23, 903-921 (2004)
7. Asman, A.J., Landman, B.A.: Formulating Spatially Varying Performance in the Statistical Fusion Framework. IEEE transactions on medical imaging 31, 1326 - 1336 (2012)
8. Marcus, D.S., *et al.*: Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience 19, 1498-1507 (2007)
9. Klein, A., *et al.*: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786-802 (2009)
10. Avants, B., *et al.*: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26-41 (2008)
11. Asman, A., Landman, B.: Robust Statistical Label Fusion through Consensus Level, Labeler Accuracy and Truth Estimation (COLLATE). Medical Imaging, IEEE Transactions on 30, 1779-1794 (2011)
12. Asman, A.J., Landman, B.A.: Non-Local STAPLE: An Intensity-Driven Multi-Atlas Rater Model. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012, Nice, France (2012)

Multi-Atlas Based MRI Segmentation Framework with Atlas Selection for MICCAI 2012 Grand Challenge

Jiahui Wang[1], Zheng Fan[2], Yael Shiloh-Malawsky[2], Joe Kornegay[2,3,4], Martin Styner[1,5]

Departments of [1]Psychiatry, [2]Neurology, [3]Pathology and Laboratory Medeicine, and [5]Computer Science University of North Carolina, Chapel Hill, NC, USA
[4]Department of Veterinary Integrative Biosciences, Texas A&M University, TX, USA

Abstract. We applied a multi-atlas based MRI segmentation method with a novel atlas selection scheme to the MICCAI 2012 Grand Challenge contest dataset. We first pair-wise co-registered all atlas datasets and computed a directed graph with edge weights based on intensity similarity and transformation difference between atlases. Following co-registration of all atlas datasets to the subject MR image, the set of closest/neighboring atlases was selected via clustering of the graph information. Finally, a weighted majority-voting label fusion was employed to compute multi-atlas segmentation. The segmentation approach was evaluated on a testing dataset of 20 T1-weighted brain MRI scans. Mean Dice similarity coefficients of 0.74, 0.70, and 0.82 were obtained for the entire brain, cortical regions, and non-cortical regions, respectively.

Keywords: multi-atlas, atlas selection, label fusion, segmentation, MRI

1 Introduction

In this paper, we applied a multi-atlas segmentation method with a novel atlas selection technique [1] to the MICCAI 2012 Grand Challenge contest dataset. All atlases and target were aligned by a pair-wise deformable image registration. A fully connected graph was constructed by calculating the distances based on intensity and shape similarity between all pairs of registered images. We then clustered the graph by searching the shortest path between each atlas and target and selecting only those templates in each cluster that are closest to the target. The selected templates were fused to create the final segmentation via a weighted majority voting label fusion.

2 Method

2.1 Image Registration

Deformable registration plays an indispensable role in the multi-atlas based segmentation approaches. Researchers have proposed a variety of registration approaches with

different degrees of freedom, such as HAMMER [2], statistical parametric mapping [3], free-form deformations [4], and Thirion's Demons [5]. However, all these approaches are operated in the space of vector fields and do not necessarily preserve topology of the target. Avants et al. [6] proposed a symmetric diffeomorphic image registration approach (as part of the ANTS registration package) that preserves anatomical topology even with large deformation. The transformation is differentiable and guaranteed to be smooth and one-to-one, i.e., for every element in moving image, there is a single corresponding element in the fixed image.

In our segmentation scheme, we employed ANTS to register each atlas MR image to the target using a cross-correlation similarity metric. The cross-correlation has been widely used and shown to perform well in many image registration applications [6], where one requires robustness to unpredictable image noise and intensity inhomogeneity. The transformation field obtained from the registration was then applied to the manually segmented atlases with nearest neighbor interpolation. We also pair-wise registered all the atlas MR image pairs using the same approach.

2.2 Construction of Graph

We represented the registered dataset as a graph whose vertices correspond to the atlases and target. Every edge between two vertices was assigned a cost (e_{ij}), which is defined by a weighted sum of an intensity similarity term MS_{ij} (mean squared voxel-wise intensity difference) and harmonic energy HE_{ij} (harmonic energy) [Eq. (1)].

$$e_{ij} = w_1 \mathrm{MS}_{ij} + w_2 \mathrm{HE}_{ij} \quad (1)$$

where w_1 and w_2, represent the weighting factors for the intensity similarity term, and shape similarity terms, respectively. The mean squared intensity difference is defined by $\mathrm{MS}_{ij} = 1/N \sum_{m=1}^{N} \left(i_m - j_m\right)^2$, where i_m is the intensity of m-th voxel of a MRI scan I; j_m is the intensity of m-th voxel of another MRI scan J; N is the number of voxels in a MRI scan. The harmonic energy is defined as the mean Frobenius norm of the Jacobian of the deformation field [7].

2.3 Clustering-based Template Selection

From the graph constructed in the previous section, we can choose templates that are close to the target via an atlas clustering. On this graph, we clustered the atlas population into groups by searching the shortest path from each atlas to the target using the Floyd-Warshall algorithm. The atlases on the same shortest paths belong to the same cluster. We then selected the atlas that was closest to the target in each cluster as the neighboring template for the final segmentation.

2.4 Weighted Majority Voting Label Fusion

Weighted majority voting is an extension of the conventional majority voting [8]. The weighted majority voting technique assigns different weights to each atlas, i.e., as-

signing larger weights to the atlases more similar to the target image. In this study, for each selected neighboring template, we used one minus the cost [Eq. (1)] between a neighboring template and target on the graph as the weight. And then the final segmentation is determined by collecting weighted votes from all the segmentations of selected templates and assigning to each voxel the label that has the highest vote.

3 Experiments

Our segmentation approach is participated in the MICCAI 2012 Grand Challenge on Multi-Atlas Labeling. We applied our method to a contest dataset of 35 de-faced T1-weighted structural MRI scans. Fifteen scans (5 male and 10 female with an age range of 19 - 34) were selected as atlases and the remaining 20 scans (8 male and 12 female with an age range of 18 - 90) were used for testing. The testing MRI scans were segmented one-by-one using the 15 atlases. ANTS Symmetric Normalization (SyN) deformable registration with the cross-correlation similarity metric (windows radius 4), a Gaussian regularizer with $\sigma = 3$, and max-iterations of 100x50x25 was performed between all pairs of atlases and between all atlases and the target. The weighting factors for calculating the distance [Eq. (1)] between atlas pairs or between atlas and target were empirically determined as $w_1 = 0.5$ and $w_2 = 0.5$.

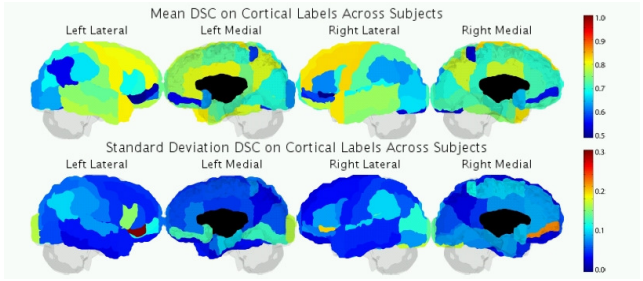


Fig. 1. Mean Dice similarity coefficients of cortical labels across subjects in testing dataset.

The Dice similarity coefficient (DSC) was used for evaluating the accuracy of the segmentation results on testing dataset. Our method achieved mean DSC of 0.74, 0.70, and 0.82 were obtained for the entire brain, cortical regions, and non-cortical regions, respectively. Figure 1 shows the mean DSC of cortical segmentations across the subjects. This result indicates that our method achieved consistent performance level on left Lateral/Medial and right Lateral/Medial. Figure 2 shows the mean DSC of non-cortical segmentations across the subjects. For most of the tissues our method achieved mean DSC higher than 0.80 (or very close to 0.80), except inferior lateral ventricle, optic chiasm, and basal forebrain.

The most time consuming step of our segmentation method is the ANTS deformable registration. The average computational time of the registration of one pair of images on a workstation with 2.6GHz CPU and 8GB RAM was 185 minutes. The remaining steps were performed on a workstation with 1.66GHz quad-core, 128GB RAM with an average computational time less than 5 minutes.

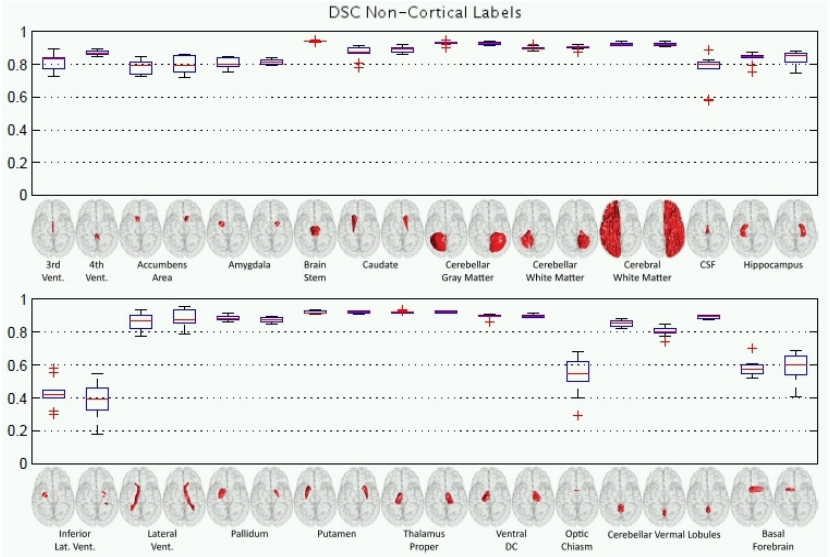


Fig. 2. Mean Dice similarity coefficients of non-cortical labels across subjects in testing dataset.

4 Conclusions

In the proposed method, a clustering technique was used to select neighboring templates that are close to the target on constructed graphs and determine weights of the selected templates for the label fusion procedure. This method provides the brain MRI studies an automated segmentation tool for efficient MR analysis.

References

1. Wang J., Fan Z., Shiloh-Malawsky Y., Kornegay J.N., Styner M.A.: Enhanced Atlas Selection for Multi-Atlas Segmentation with Application to Leg Muscle MRI. MICCAI 2012 Workshop on Multi-Atlas Labeling, (2012)
2. Shen, D., Davatzikos, C.: Hammer: Hierarchical attribute Matching Mechanism for Elastic Registration. IEEE Trans. Med. Imaging, 21, 1421 - 1439 (2002)
3. Ashburner, J., Friston, K.: Voxel-based morphometry-the methods. Neuroimage, 11, 805 - 821, (2000)
4. Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D.,. Nonrigid Registration Using Free-form Deformations: Application to Breast MR Images. IEEE Trans. Med. Imaging 18 (8), 712 – 721, (1999)
5. Thirion, J.P.: Image Matching as a Diffusion Process: an Analogy with Maxwell's Demons. Med. Image Anal. 2 (3), 243-260, (1998)
6. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain. Med. Image Anal., 12, 26-41, (2008)
7. Hamm, J., Ye, D.H., Verma, R., Davatzikos, C.: GRAM: A Framework for Geodesic Registration on Anatomical Manifolds. Med. Image Anal., 14, 633-642, (2010)
8. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic Anatomical Brain MRI Segmentation Combining Label Propagation and Decision Fusion. NeuroImage, 33, 115-126, (2006)

Joint generative model for segmentation and local atlas stratification

Annemie Ribbens, Frederik Maes, Dirk Vandermeulen, and Paul Suetens

Katholieke Universiteit Leuven, Department ESAT/PSI - Medical Image Computing
Medical Imaging Research Center - University Hospital Gasthuisberg,
Herestraat 49 bus 7003, B-3000 Leuven, Belgium
annemie.ribbens@uz.kuleuven.ac.be

Abstract. Multi-atlas-based methods for brain MR segmentation have been shown to be beneficial over single atlas segmentation approaches. However, the performance of (multi-)atlas-based segmentation methods is influenced by the available atlases, the way these are constructed and their (local) selection. We propose a method that combines segmentation of a set of brain MR images with probabilistic atlas construction and selection in a unified framework. The method is formulated for atlas-driven tissue segmentation. The atlas construction and selection is performed by a stratification approach, modeling the heterogeneity of the set of images. Validation on two publicly available sets of images shows that accurate segmentations are obtained and that the atlas stratification procedure is beneficial for the segmentation performance.

1 Introduction

Probabilistic atlas-driven segmentation of brain MR images typically involves atlas-to-image registration combined with a Gaussian mixture intensity model [1]. Probabilistic atlases are usually constructed by registration and averaging segmentations of multiple images. Many studies have emphasized the impact of the atlas construction procedure on segmentation accuracy, in particular the relation between atlas sharpness (or blurriness) and the flexibility of the atlas-to-image registration [2–5]. Ideally, registration flexibility should be identical during atlas construction and atlas-based segmentation such that the blurriness of the atlas is optimally adapted to the residual spatial mismatches after atlas-to-image registration. However, this is not guaranteed in practice as the atlas construction procedure is often not considered when a pre-existing atlas template is selected for atlas-based segmentation. Furthermore, brain tissue segmentation is often guided by a single atlas (e.g. [1]), although a single atlas is not sufficient to summarize the variability of the heterogeneous human population. In [6, 7], an effort is made to model the heterogeneity of a population by performing atlas stratification, i.e. one simultaneously searches for the major modes in the population, while creating a separate atlas for the subgroup of images corresponding to each specific mode. Several multi-atlas strategies have been proposed for segmentation of anatomical structures in MR images [8], including various approaches for

selecting the (locally) most appropriate atlases based on some measure of similarity between the atlases and the image to be segmented [9–12]. However, these methods assume that multiple atlases are already available. Moreover, most criteria for atlas selection are chosen heuristically and although intuitively the atlas selection criterion should be identical to the criterion used to select the images for atlas construction, this is again not guaranteed when atlas construction and atlas selection are treated separately.
In this work, we propose a Bayesian method for joint segmentation and atlas stratification. This work contributes to the current state-of-the-art by considering the problems of multi-atlas construction, atlas selection and atlas-guided image segmentation simultaneously. The approach improves the segmentation performance over using a single predefined atlas template because the subject-specific morphology is better modeled by using the multi-atlas stratification. Moreover, the method guarantees the same flexibility in the registration model and the same criteria for selecting the images and the atlases respectively in both atlas construction and atlas-guided segmentation.

2 Methods

The basis of our segmentation method is similar to [1]. Each image in a given data set of images is segmented using a Gaussian mixture model on the bias field corrected image intensities. The segmentation is guided by prior information in the form of a probabilistic atlas whose registration to the images is iteratively refined based on their segmentations. But instead of assuming the atlas to be given in advance as in [1], the atlas is estimated within the framework as explained below, by performing the segmentation for all images simultaneously. However, the data set can be heterogeneous containing different subgroups with a different morphology, e.g. healthy controls and diseased patients. Therefore, separate atlases need to be estimated for different subgroups (clusters) and the locally most appropriate atlas(es) need to be selected as prior for the segmentation model. This is performed using a local atlas stratification approach, i.e. locally similar images are assigned to the same subgroup (cluster) forming an atlas. The assignments (cluster memberships) are in turn used for locally selecting the most appropriate atlas for segmenting an image.

2.1 Framework

We denote the intensities of the images as $Y = \{y_{ij}\}$, their segmentations as $L = \{l_{ijk}\}$ and their voxelwise cluster memberships as $Z = \{z_{ijt}\}$, with $i = \{1 \dots, N_I\}$, $j = \{1 \dots, N_J\}$, $k = \{1 \dots, N_K\}$ and $t = \{1 \dots, N_T\}$ indexing the images, the voxels in each image, the segmentation labels (tissue classes) and the clusters respectively, with the number of clusters N_T and the number of tissue classes N_K specified by the user. Note that $l_{ijk} = 1$ if voxel j in image i has tissue class label k and $l_{ijk} = 0$ otherwise, while $z_{ijt} = 1$ if voxel j in image i belongs to cluster t and $z_{ijt} = 0$ otherwise. The cluster memberships are

determined voxelwise to allow local atlas selection, as discussed in more detail in the next section.

We define now a model that explains the observed variables, i.e. the image intensities $Y = \{y_{ij}\}$. We assume that the image intensities of tissue class k of an image i are generated from a Gaussian distribution with mean μ_{ik} and variance σ_{ik}^2. To account for MR field inhomogeneities, we include an additive bias field correction completely similar to [13] (by applying the model on the log-transformed MR intensities instead). We denote the bias field in voxel j as $B_j(C_i)$ with C_i the bias field parameters for image i. To guide the segmentations, a prior distribution is given for the segmentation labels l_{ijk} by the (within our model constructed) probabilistic atlas A_{kt} of cluster t deformed towards the image i using the registrations R_{it} with $R_{ijt} = R_{it}(j)$ (assuming that image i belongs to cluster t in voxel j). In addition to the diffeomorphic regularizer embedded in the used registration procedure (see further), we define a prior distribution on the atlas-to-image registrations to minimize the total amount of deformation. Therefore, the prior distribution on R_{ijt} is given as a Gaussian distribution with mean G_{jt} and variance ϵ_{jt}^2, avoiding that individual atlas-to-image registrations R_{ijt} drift away to far from a cluster-specific group mean G_{jt}. This group mean G_{jt} is the identify transform in case the atlas t is in minimal deformation space [14]. Finally, the prior distribution on the voxelwise cluster memberships z_{ijt} need to be defined to guide the clustering and therefore the selection of the appropriate cluster-specific atlases as prior for the segmentation. We here define the prior distribution on z_{ijt} as a uniform distribution over all voxels of the image i. Thus, the model assumptions can be summarized by the following distributions:

$$
\begin{aligned}
&P(y_{ij}|\mu_{ik}, \sigma_{ik}^2, C_i) = \mathcal{G}_{\sigma_{ik}^2}((y_{ij} - B_j(C_i)) - \mu_{ik}) \\
&P(R_{ijt}|G_{jt}, \epsilon_{jt}^2) = \mathcal{G}_{\epsilon_{jt}^2}(R_{ijt} - G_{jt}) \\
&P(l_{ijk}|A_{kt}, R_{ijt}) = A_{kt}(R_{it}(j)) \\
&P(z_{it}|\pi_{it}) = \pi_{it}
\end{aligned}
$$

with $\mathcal{G}_a(x - b)$ the Gaussian distribution on the variables x with mean b and variance a. From this model, it follows that:

$$P(Y, L, Z, \Upsilon) = P(Y|L, \mu, \sigma^2)P(L|Z, A, R)P(R|Z, G, \epsilon^2)P(Z|\pi) \qquad (1)$$

with $\Upsilon = \{\mu, \sigma, C, A, R, G, \epsilon, \pi\}$ denoting the model parameters. The model parameters Υ are estimated by optimizing their likelihood given the observed variables Y, resulting in the following Maximum A Posteriori (MAP) problem:

$$\hat{\Upsilon} = \arg\max_{\Upsilon} P(\Upsilon|Y) = \arg\max_{\Upsilon} \log P(\Upsilon|Y) = \arg\max_{\Upsilon} \log P(Y, \Upsilon) \qquad (2)$$

From equation (1) it becomes clear that solving this MAP problem is simplified by knowledge of the hidden (latent) variables, i.e. the segmentation L and the

clustering Z, as

$$\log P(Y,\Upsilon) \propto \log\left[\sum_{L,Z} P(Y,L,Z,\Upsilon)\right] = \tag{3}$$

$$= \sum_{i=1}^{N_I}\sum_{j=1}^{N_J} \log\left[\sum_{k=1}^{N_K}\sum_{t=1}^{N_T} P(y_{ij}|\mu_{ik},\sigma_{ik}^2,C_i)P(l_{ijk}|A_{kt},R_{ijt})P(z_{ijt}|\pi_{it})P(R_{ijt}|G_{jt},\epsilon_{jt}^2)\right]$$

To solve this problem, we compute a lower bound Q using Jensen's inequality:

$$Q(\Upsilon|\Upsilon^\eta) = E_{L,Z|Y,\Upsilon^\eta}\left[\log P(Y,L,Z,\Upsilon)\right] = \sum_{i=1}^{N_I}\sum_{j=1}^{N_J}\sum_{k=1}^{N_K}\sum_{t=1}^{N_T} P(l_{ijk},z_{ijt}|y_{ij},\Upsilon_{ijkt}^\eta)\cdot$$

$$\left[\log P(y_{ij}|\mu_{ik},\sigma_{ik}^2,C_i) + \log P(l_{ijk}|A_{kt},R_{ijt}) + \log P(z_{ijt}|\pi_{it}) + \log P(R_{ijt}|G_{jt},\epsilon_{jt}^2)\right]$$

which can be optimized iteratively using the Expectation Maximization (EM) algorithm with Υ^η indicating the parameters estimated in the previous EM iteration η. An expression for the posterior distribution, denoted by b_{ijkt}, can be found using Bayes' rule,

$$\begin{aligned} b_{ijkt} &= P(l_{ijk},z_{it}|y_{ij},\Upsilon_{ijkt}) \\ &= \frac{P(y_{ij}|\mu_{ik},\sigma_{ik}^2,C_i)P(l_{ijk}|A_{kt},R_{ijt})P(z_{ijt}|\pi_{it})P(R_{ijt}|G_{jt},\epsilon_{jt}^2)}{\sum_{t,k} P(y_{ij}|\mu_{ik},\sigma_{ik}^2,C_i)P(l_{ijk}|A_{kt},R_{ijt})P(z_{ijt}|\pi_{it})P(R_{ijt}|G_{jt},\epsilon_{jt}^2)} \end{aligned} \tag{4}$$

In the E-step, this posterior b_{ijkt} is updated using the previous estimate of the model parameters resulting in an estimate of the hidden variables (section 2.2). In the M-step, the parameters Υ are updated by maximization of the Q-function making use of the previous estimate of the posterior b_{ijkt}, i.e. hidden variables. Closed form analytical solutions are obtained for the parameters μ_{ik}, σ_{ik} and C_i of the intensity model, identical to [13]. For the atlases A_{jkt} and clustering prior π_{it}, we refer to sections 2.3 and 2.4. For the groupwise registrations G_{jt}, we find that they equal the weighted sum of the individual atlas-to-image registrations of the corresponding cluster t, with the weights determined by the voxelwise cluster memberships (see further). For the update of the individual registrations R_{ijt}, no closed form solution is found. Instead, the derivative of Q w.r.t. R_{ijt} for all j is interpreted as a force field that drives the registration. A physically acceptable deformation field is then obtained by imposing spatial smoothness using the diffeomorphic demons approach of [15]. In what follows, we focus on the update of the segmentations and the atlases. We denote our algorithm as CMACS, i.e. Combined Multi-Atlas Construction and Segmentation.

2.2 Local multi-atlas guided segmentation

The probabilistic segmentations of an image, denoted as p_{ijk}, are obtained from the posterior b_{ijkt} by summation over all clusters t. Hence, $p_{ijk} = \sum_t b_{ijkt} =$

$$\sum_t \left[\left(\sum_k b_{ijkt}\right)\cdot\left(\frac{P(y_{ij}|\theta_{ik},C_i)P(l_{ijk}|A_{kt},R_{ijt})}{\sum_k P(y_{ij}|\theta_{ik},C_i)P(l_{ijk}|A_{kt},R_{ijt})}\right)\right] = \sum_t \rho_{ijt}\cdot p_{ijkt} \tag{5}$$

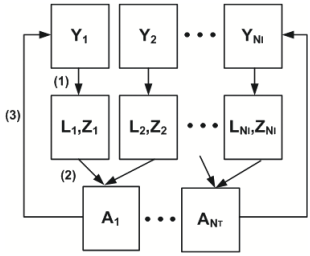


Fig. 1. Proposed scheme for joint atlas stratification and segmentation of a heterogeneous group of images. The images are segmented based on a Gaussian mixture model on the intensities Y guided by cluster-specific atlases A (E-step). The images are clustered based on the similarity between the segmentations L and the cluster-specific atlases A (E-step). The atlases are constructed from the segmentations L based on the cluster memberships Z (M-step).

Thus, p_{ijkt} are the probabilistic segmentations of image i obtained from a Gaussian mixture model on the intensities combined with the atlas of cluster t, transformed to the image (i.e. as in [1]). Such segmentations p_{ijkt} are computed for each cluster, i.e. using the cluster-specific atlas. The overall segmentation of image i equals thus the weighted sum of the cluster-specific segmentations p_{ijkt} for that image with the weights ρ_{ijt} defined voxelwise. Therefore, our algorithm yields a local-selection multi-atlas segmentation approach. The weights $\rho_{ijt} = \sum_k b_{ijkt}$ used for atlas selection are in fact the voxelwise cluster memberships, i.e. the probability that voxel j in image i belongs to cluster t. From equation (4), it follows that the weights are based on the amount of registration needed to transform the atlas to the image, on the similarity between the atlas and the intensity model of the image, and on the prior distribution on the cluster memberships.

2.3 Multi-atlas construction

Maximization of the Q-function with respect to the parameters A_{kt}, i.e. the atlases, subject to $\sum_k A_{kt} = 1$, results in the following expression:

$$A_{kt}(u) = \frac{\sum_i \rho_{it}(R_{it}^{-1}(u))|Jac(R_{it}^{-1}(u))|p_{ikt}(R_{it}^{-1}(u))}{\sum_i \rho_{it}(R_{it}^{-1}(u))|Jac(R_{it}^{-1}(u))|} = \sum_i \rho^*_{ij^it} \cdot p_{ij^ikt} \qquad (6)$$

with $j^i = R_{it}^{-1}(u)$ indicating voxel j in image i corresponding to position u in the atlas and with $|Jac|$ the determinant of the Jacobian. The atlases are thus constructed as a weighted sum of segmentations. These segmentations, denoted p_{ij^ikt}, are the cluster-specific segmentations p_{ijkt} (section 2.2) deformed towards the atlas space (using R_{it}^{-1}). The weights $\rho^*_{ij^it}$, are obtained from a normalization of the voxelwise cluster membership probabilities ρ_{ijt} in atlas space together with a modulation step, i.e. a multiplication with the Jacobian determinant to locally

preserve the volumes of the tissue classes. From equations (5) and (6), it follows that the segmentation and the atlas construction procedure use the same type of registrations and the same local selection criteria, i.e. the weights used for local atlas selection during segmentation are closely related to the weights used for multi-atlas construction.

2.4 Atlas stratification for a specific region

The model described above performs local atlas stratification based on voxelwise cluster memberships. However, the local clustering (stratification) is guided by a global uniform prior over the entire image. The parameter of the uniform prior distribution for image i and cluster t is estimated in the M-step of the EM algorithm as the average of the voxelwise cluster memberships of image i:

$$\frac{\partial Q}{\partial \pi_{it}} = 0 \text{ s.t. } \sum_t \pi_{it} = 1 \Rightarrow \pi_{it} = \frac{1}{N_J} \sum_j \rho_{ijt} \tag{7}$$

However, depending on the application, it could be desirable that the stratification is guided by a particular brain structure or region, for instance when studying specific regional differences between populations, e.g. hippocampal changes in Alzheimer disease. Thereto, we replace the uniform prior on the cluster memberships by a Markov Random Field (MRF) defined by the Gibbs distribution:

$$P(z_{ijt} = 1|\Phi_{z_{ijt}}) = Z(\Phi_{z_{ij}})^{-1} \exp\left[-U(z_{ijt} = 1|\Phi_{z_{ijt}})\right] \tag{8}$$

with $Z(\Phi_{z_{ij}})^{-1}$ a normalization constant and

$$U(z_{ijt} = 1|\Phi_{z_{ijt}}) = \sum_{t'} \left(\gamma \sum_{j' \in \Omega} z_{ijt} \alpha_{tt'} z_{ij't'} + (1-\gamma) \sum_{j' \notin \Omega} z_{ijt} \alpha_{tt'} z_{ij't'} \right)$$

where $\Phi_{z_{ijt}} = \{\alpha_{tt'}|\forall t'\}$ are the MRF parameters. The MRF gives a prior for the cluster membership of voxel j in image i, based on the cluster memberships of all other voxels j' in image i. The parameter γ determines the impact of voxels j' of a specific region Ω compared to voxels $j' \notin \Omega$ on the cluster membership of voxel j in image i. In other words, parameter γ determines the emphasis to perform the stratification on a specific region of interest Ω.

2.5 Implementation

The parameter ϵ_{jt}^2 describing the variance between the individual registrations R_{ijt} and the groupwise registration G_{jt}, needs to be determined in advance and is set to be 4 voxels in every voxel j and for every cluster t (heuristically chosen). Furthermore, in case atlas stratification for a specific region is performed, we need to determine the MRF parameters. We set $\alpha_{tt'} = 0$ in case $t = t'$ and $\alpha_{tt'} = 2$ in case $t \neq t'$ (heuristically chosen). The value 2 is the MRF field

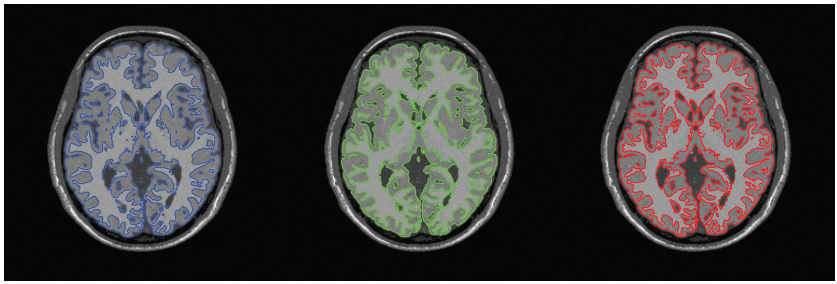
	WM	GM
CMACS	**93.48±0.35**	**92.84±0.52**
CMACS-a	93.47±0.36	92.30±0.63*
SPM8b	93.19±0.48*	92.20±0.85*
FSL4.1.9	93.44±0.36*	90.91±0.72*

Table 1. Experiment 1: Segmentation accuracy for the 20 simulated BrainWeb images in terms of % Dice overlap (mean ± standard deviation). Bold = highest values, * = values significantly different from CMACS (paired t-test with 5% significance level).

strength, i.e. a larger value implies a lower a priori probability for an individual voxel to have a different cluster membership than the other voxels of the same image. The calculation of the Q-function within the EM algorithm requires all possible realizations of the MRF, both in the expectation step (posterior) and maximization step (update of the prior distribution), which is computationally not feasible. The mean field approximation is used instead [16, 17].

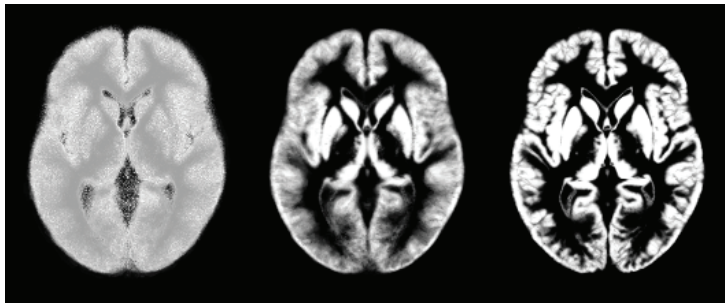
3 Experiments

Experiment 1: In this experiment, we demonstrate the benefits of simultaneous segmentation and atlas construction over doing both processes separately and sequentially. Therefore, we do not use the multi-atlas approach, but run our algorithm using one cluster ($N_T = 1$) such that a single atlas is constructed. The publicly available BrainWeb data set [18] is used which consists of 20 simulated MR images of healthy controls with ground truths for white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF). To investigate the benefits of combined segmentation and atlas construction, we compare CMACS with a basic version of our algorithm (denoted as CMACS-a) where the atlas itself is not updated, but a previously constructed atlas is used which is obtained by running the full version of CMACS on the same images (i.e. in advance). This comes down to sequential atlas construction and segmentation, whereby segmentation is performed by simultaneously estimating the Gaussian mixture model parameters and the atlas-to-image registrations (similar to [1]). For completeness, we compare our algorithm with two state-of-the-art segmentation methods, i.e. SPM5 (`http://www.fil.ion.ucl.ac.uk/spm/software/spm5/`, [1]) and FSL4.1.9 - FAST (`http://www.fmrib.ox.ac.uk/fsl/fsl/`).
The results are summarized in Table 1 in terms of the Dice overlap coefficient of the obtained segmentations with the BrainWeb ground truth. It is clear that our algorithm outperforms state-of-the-art methods. Furthermore, joint segmentation and atlas construction (CMACS) is superior to sequential atlas construction and segmentation (CMACS-a and SPM). A typical segmentation result is illustrated in Figure 2. The GM probabilistic maps of the atlases constructed in different iterations of CMACS are shown in Figure 3.



(a) Ground truth (b) CMACS (c) SPM

Fig. 2. Experiment 1: GM segmentation of a Brainweb image.



(a) Iteration 1 (b) Iteration 15 (c) Iteration 100

Fig. 3. Experiment 1: GM probabilistic atlas as estimated in different iterations of the algorithm.

Experiment 2: In this experiment, we focus on the atlas stratification when applying CMACS to a heterogeneous data set. Therefore, 90 images are selected from the publicly available ADNI data set, i.e. 45 images from patients suffering from Alzheimer disease (AD) (average age of 72.5 ± 3.1 years, range 65.6 - 77.6) and 45 age-matched normal controls (average age of 72.07 ± 2.14 years, range 65.1 - 75.1). Hippocampus segmentations are available from ADNI for all these subjects. As both groups are age-matched in this experiment, we expect the major mode of morphological variability in the group of 90 images to be caused by the disease patterns of Alzheimer and hence run our algorithm using 2 clusters. As AD is known to have an important impact on the hippocampal region, we want the atlas stratification to be determined by this region primarily. Therefore, we use the model described in section 2.4 with prior distribution as described in equation (8). The parameter γ in this experiment is chosen to be 0.9 and the hippocampal region, i.e. Ω in equation (8), is determined in advance using a test set of images with hippocampus segmentations (i.e. different from the data set

used in this experiment to make our method more general applicable).
To assess the outcome of the stratification with respect to the hippocampal region (i.e. the region of interest), the global clustering obtained by CMACS (i.e. defined as the average over the voxelwise cluster memberships) is visualized in Figure 4 in terms of the normalized hippocampus volumes (i.e. after affine registration of the images to MNI standard space). This shows that our framework is capable of discriminating the two subgroups and that the resulting clustering is indeed correlated with hippocampal differences. Our algorithm does not strictly follow the separation based on hippocampus volumes as the complete morphological appearance of this region, including also location and shape and not just volume, is taken into account, as well as, to a smaller extent (because of the weight γ), the morphological appearance in the rest of the image.
We also compare the cluster-specific atlases obtained from CMACS by stratification based on image features with the atlases formed based on clinical prior knowledge. To this end, we run CMACS again using one cluster ($N_T = 1$) on each of the two clinical subgroups of 45 images separately. We denote the algorithm on these clinical subsets as CMACS-c. The atlases and their difference maps obtained by CMACS and CMACS-c are shown in Figure 5. It is clear that the atlases constructed by stratification (CMACS) and by using clinical knowledge (CMACS-c) are very similar. This indicates again that our algorithm is capable of discriminating between different morphologies. Moreover, it shows that the obtained subgroups are clinically relevant and that the stratification is not restricted to the hippocampal region, i.e. other AD related patterns are exposed by CMACS such as enlarged ventricles, although the stratification was primarily guided by differences in the hippocampal region.
Furthermore, the stratified atlases (CMACS) seem to be sharper than those based on clinical prior knowledge (CMACS-c) and larger differences in the morphology of the hippocampal region are exposed by CMACS than by CMACS-c. These larger differences in the atlas stratification framework are also visible in other regions (e.g. ventricles and some cortical regions). This is an indication that the atlases obtained from stratification (CMACS) better capture the morphological variability in the population, than those obtained based on clinical knowledge (CMACS-c). Hence, it is expected that the atlases constructed by our method are also more optimally adapted to guide image segmentation, especially so for the hippocampal region as atlas stratification focused on this region primarily. For completeness, some segmented images obtained from CMACS are shown in Figure 6.

4 Discussion and conclusion

In this work, a Bayesian method is proposed for joint segmentation and atlas stratification of a heterogeneous data set. Images of the heterogeneous data set are assigned (voxelwise) to more homogeneous subgroups (clusters) in an unsupervised way. Per cluster a probabilistic atlas is then constructed. The voxelwise cluster memberships of an image are then used to locally select an atlas to guide

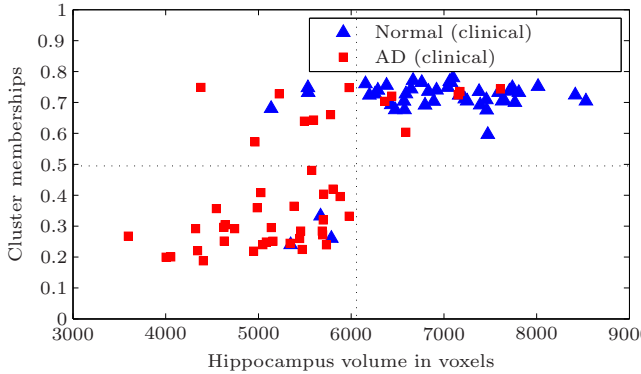


Fig. 4. Experiment 2: ADNI data set: Hippocampus volumes (after normalization to MNI space) in terms of the probabilistic cluster memberships (average over the voxel-wise cluster memberships) obtained from CMACS to belong to the normal cluster. Blue triangles indicate the subjects clinically diagnosed as healthy, while the red squares indicate the subjects clinically diagnosed with Alzheimer.

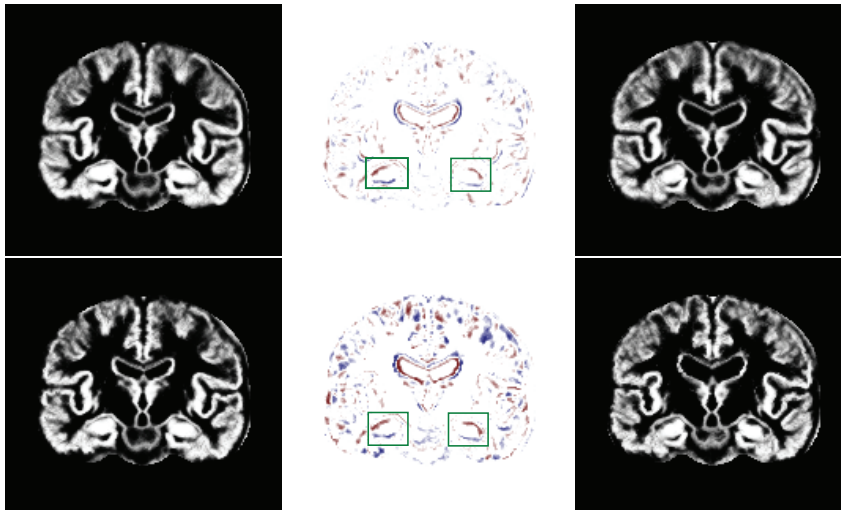
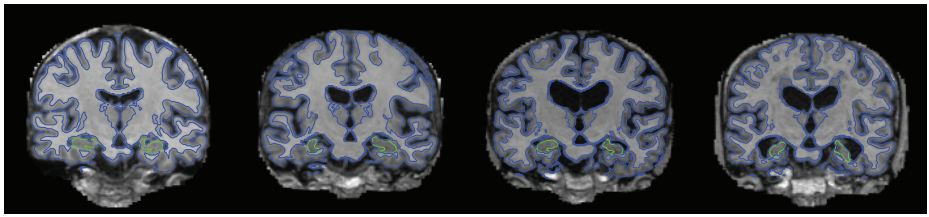


Fig. 5. Experiment 2: ADNI data set: GM probabilistic atlases obtained for cluster 1 ('healthy', left column) and cluster 2 ('AD', right column) by clustering based on clinical knowledge (CMACS-c, top row) and by image-based stratification (CMACS, bottom row). The difference between the GM probabilistic atlases of both clusters ('AD' - 'healthy', middle column) is scaled between -1 (red) and 1 (blue), and white indicates no differences (threshold on 0.2, i.e. the difference map is white in the interval [-0.2 0.2]). The green boxes are a rough indication of the hippocampal region.



(a) Normal (b) Normal (c) AD (d) AD

Fig. 6. Experiment 2: ADNI data with GM segmentations obtained from our algorithm (blue) and hippocampus segmentation obtained from ADNI (green): (a)-(b) 2 examples of healthy subjects, (c)-(d) 2 examples of AD subjects.

the tissue class segmentation.
The framework therefore guarantees that the same criterium is used for locally assigning an image to a cluster constructing multiple atlases, and for locally selecting an appropriate atlas for the image segmentation. The combination of both processes also assures that the same degree of flexibility of the registration is used in both atlas construction and segmentation.
Experiment 1 shows that simultaneously segmenting the images and constructing an atlas is beneficial over first constructing the atlas and subsequently performing the segmentation. Experiment 2 shows that the atlas stratification procedure is capable of discriminating between different morphologies. Therefore, the constructed atlases are more morphology specific than in case a single atlas was constructed for the entire heterogeneous population. The resulting atlases from our stratification follow the patterns observed in the atlases based on clinical knowledge, but seem to capture the heterogeneity of the population even better. This can be explained as the stratification is purely based on the brain morphology. This improved representation of the variability in brain morphology of the heterogeneous population, i.e. by the atlases, indicates that the atlases are formed in a more optimal way for segmentation purposes.
Future work could include an evaluation of the performance and the robustness of the method with respect to various parameters of the implementation. Moreover, we could perform a detailed analysis of the morphological differences between clusters exposed by our method. We could also analyze the method in case of morphological differences due to normal variability, e.g. differences between sulcal patterns. Another topic of interest is the segmentation of new unseen images based on the cluster-specific atlases.

References

1. Ashburner, J., Friston, K.: Unified segmentation. NeuroImage **26** (2005) 839–851
2. Zöllei, L., Shenton, M., Wells, W., Pohl, K.: The impact of atlas formation methods on atlas-guided brain segmentation. MICCAI: stat. reg. workshop **1** (2007) 39–46

3. Yeo, B.T., Sabuncu, M.R., Desikan, R., Fischl, B., Golland, P.: Effects of registration regularization and atlas sharpness on segmentation accuracy. MedIA **12** (2008) 603–615
4. Van Leemput, K.: Encoding probabilistic brain atlases using bayesian inference. IEEE Trans. on Med. Img. **28**(Pt 1) (2006) 822–837
5. Bhatia, K.K., Aljabar, P., Boardman, J.P., Srinivasan, L., Murgasova, M., Counsell, S.J., Rutherford, M.A., Hajnal, J., Edwards, A.D., Rueckert, D.: Groupwise combined segmentation and registration for atlas construction. MICCAI **10**(Pt 1) (2007) 532–540
6. Blezek, D., Miller, J.: Atlas stratification. MedIA **11** (2007) 443–457
7. Sabuncu, M.R., Balci, S.K., Shenton, M.E., Golland, P.: Image-driven population analysis through mixture modeling. IEEE Trans. on Med. Img. **28** (2009) 1473–1487
8. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage **33** (2006) 115–126
9. Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D.: Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage **46** (2009) 726–738
10. Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J.: Optimum template selection for atlas-based segmentation. NeuroImage **34** (2007) 1612–1618
11. van Rikxoort, E.M., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M.A., Pluim, J.P., van Ginneken, B.: Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. MedIA **14** (2010) 39–49
12. Artaechevarria, X., Munoz-Barrutia, A., de Solorzano, C.O.: Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Trans. on Med. Img. **28** (2009) 1266–1277
13. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based bias field correction of MR images of the brain. IEEE Trans. on Med. Img. **18(10)**(10) (1999) 885–896
14. Guimond, A., Meunier, J., Thirion, J.P.: Average brain models: A convergence study. Computer Vision and Image Understanding **77** (2000) 192–210
15. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image reigstration. NeuroImage **45** (2009) 61–72
16. Zhang, J.: The mean-field theory in EM procedures for Markov Random Fields. IEEE Trans. on Signal Processing **40** (1992) 2570–2583
17. Langan, D., Molnar, K., Modestino, J., Zhang, J.: Use of the mean-field approximation in an EM-based approach to unsupervised stochastic model-based image segmentation. Proc. ICASSP **3** (1992) 57–60
18. Aubert-Broche, B., Griffin, M., Pike, G., Evans, A., Collins, D.: Twenty new digital brain phantoms for creation of validation image data bases. IEEE Trans. on Med. Img. **25(11)** (2006. *http://www.bic.mni.mcgill.ca/brainweb/*) 1410–1416

Segmentation via Multi-atlas LDDMM

Xiaoying Tang[1], Susumu Mori[2], and Michael I. Miller[1*]

[1]Center for Imaging Science, Johns Hopkins University
[2]Department of Radiology, Johns Hopkins University

Abstract. In this paper, we examine the multi-atlas random orbit model in which imagery is modeled as conditional Gaussian random fields, conditioned on both the random atlas which generates it and the random diffeomorphism associated with the atlas's change of coordinates. The model is examined for segmenting T1-weighted MR imagery of the brain, in which an iterative algorithm is employed for simultaneously estimating the unknown atlas-diffeomorphism pair and generating the maximum-a-posteriori (MAP) estimator of the subject labels. Since the goal is to generate the MAP estimator of the segmentation labels, the iterative algorithm is a derivative of the EM algorithm thereby removing the conditioning on the unknown atlas labels and the diffeomorphism. The segmentation accuracy of our method is first evaluated in fourteen structures in eight MR scans, and then in whole brain segmentations in the fifteen training datasets provided by the organizer of the workshop using a leave-one-out test.

Keywords: Large Deformation Diffeomorphic Metric Mapping, Likelihood-fusion, FreeSurfer, FIRST, STAPLE, EM-algorithm

1 Introduction

Segmenting cortical and subcortical structures of the human brain is important in clinical neuroimaging studies. With the increase of the sample size of magnetic resonance (MR) imaging datasets, manual tracing of brain structures becomes prohibitive considering the time and cost of the labor-intensive process. There have been significant developments towards semi- or fully- automated segmentation methods [1] [2] [3]. The segmentation problem is usually handled in the Bayesian framework by solving a Maximum A Posteriori (MAP) estimation problem. There are typically two approaches, both defining appearance models (usually Gaussian appearance models). The first approach, which is more local in nature, models various features of the voxels, such as the intensity value, as Gaussian distributions and then performs MAP estimation combined with other techniques such as markov random fields [1] or level sets [4]. The second approach, which is more global in nature, tries to incorporate shape priors into the intensity models with a weighting matrix estimated from a training set [2].

Our method utilizes large deformation diffeomorphic metric mapping (LDDMM) [5] [6], in which a diffeomorphic deformation is obtained by mapping a

pre-labeled atlas T1 image globally to the target T1 image, and then propagating the labels of the atlas with the deformation. The aim of segmentation via LDDMM is to accurately cast the pre-labelled anatomical definitions onto a test image, but it is based on a global solution which could contain local errors when the shape difference between the template and the target is large. Using multiple atlases, it is possible to correct the local misclassifications from various atlases [7]. Various multi-atlas segmentation methods, based on label fusion methods such as STAPLE [8], have emerged in recent years which are elegant and effective.

The method developed here, which we term Multi-atlas LDDMM, is complementary in that we attempt to solve the actual Bayesian MAP problem for the unknown segmentation of the target given only the single observed MR image and the associated set of multiple atlases. We do not assume that we are given a set of segmentations for the same subject corresponding to different atlas interpretations. The problem we focus on is complementary of the one studied in STAPLE. In this Bayesian setup, given a set of pre-labeled T1-weighted atlas images, we model the to-be-segmented target image as a conditional Gaussian random field, conditioned on both the unknown atlas, which conceptually may have generated it, and the corresponding unknown diffeomorphism between that randomly conceived atlas and the target. The atlas selection is iteratively optimized using the expectation-maximization (EM) algorithm, which gives rise to the MAP estimation problem via a mixture of atlases. In this process, we never explicitly generate the set of segmentations associated with each of the single-atlas interpretations, and therefore no label fusion is performed. Instead, we generate the conditional likelihoods of the observed MR image to be associated to each of the atlas interpretations, and then perform likelihood fusion. This is essentially the E-step in the EM algorithm, from which the single set of segmentation labels of the target is generated each time in the M-step.

In this paper, we applied our Multi-atlas LDDMM segmentation method to eight MR datasets and evaluated them for fourteen structures (left and right amygdala, hippocampus, caudate, globus pallidus, putamen, thalamus, and lateral ventricle). We compared the segmentation results of our method with those from two state-of-art publicly available segmentation software tools, FreeSurfer [1] and FIRST [2], on the same datasets for subcortical structure segmentation. In addition, we compared our segmentations results to those obtained from STAPLE applied to the set of propagated segmentations from each single LDDMM. In the end, we validated our method for whole brain segmentation based on the fifteen training datasets provided by the organizer of the workshop using a leave-one-out technique.

2 Material and Method

2.1 Subject data

We collected high-resolution structural MR images from eight neurological healthy adults (all right-handed, four women, mean age = 23 years; age range = 19-

26 years). MR scans were acquired using a 3.0T scanner with 256 × 256 (1 × 1mm) in-plane resolution, 120 1-mm slices without gaps. Details about the fifteen training datasets can be found at the website of the workshop `https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Main_Page`. Briefly speaking, there are 10 females and 7 males. The age ranges from 19 to 34, with a mean age of 23.

2.2 Principles of LDDMM-image

Given an atlas T1-weighted image I_0 and a target T1-weighted image I_1, where I_0 and I_1 are functions defined on the image domain $\Omega \subseteq R^3$, the algorithm LDDMM-image [6] computes a diffeomorphic transformation $\varphi : \Omega \to \Omega$ as the end-point of the flow of an energy-minimizing velocity vector field $v_t : \Omega \to R^3, t \in [0,1]$. The velocity vector field is specified by the ordinary differential equation (ODE) $\dot{\phi}_t = v_t(\phi_t)$, which starts with $\phi_0 = Id$, where Id is the identity transformation such that $Id(x) = x, \forall x \in \Omega$. The diffeomorphic deformation φ is thus calculated as: $\varphi = \phi_1 = \int_0^1 v_t(\phi_t)\, dt$ with $\phi_0 = Id$. The optimal deformation is estimated by solving the variational problem:

$$\hat{v} = \underset{v:\dot{\phi}_t = v_t(\phi_t)}{\arg\min} \left(\int_0^1 \|Lv_t\|_{L^2}^2 \, dt + \frac{1}{\sigma^2} \left\| I_0 \circ \phi_1^{-1} - I_1 \right\|_{L^2}^2 \right) \tag{1}$$

To ensure that the solution lies in the space of diffeomorphisms [9], a sufficient amount of smoothness is achieved by defining the operator L as: $L = (-\alpha \nabla^p + \gamma)\, \mathrm{I}_{3\times 3}$, where $p \geq 1.5$ in 3-dimensions, γ is usually fixed to be 1, α affects the degree of smoothness of the deformation, and ∇ is the gradient operator.

2.3 Probabilistic Model

Let $A = 1, 2, ...$ be a set of T1-weighted atlas images, paired with its manual labels $\left(I^A, W^A\right)$, where I^A denotes the gray-scaled T1 image and W^A denotes its manual segmentations. Given a to-be-segmented subject with image intensity I_i at voxel x_i, we model it as a conditional Gaussian random field, conditioned on the unknown atlas and the corresponding unknown diffeomorphism $(A_i = a, \varphi_a)$. To accommodate locality, each voxel x_i can be interpreted or generated by a different atlas. The iterative algorithm for segmentation involves iterative atlas selection and diffeomorphism construction, which is a variant of the expectation-maximization (EM) method.

1. Initialize: for each voxel i of the target image, for each atlas $a \in A$, set the diffeomorphism to identity $\hat{\varphi}_a = Id$ and set initial weights to uniform conditional probability as:

$$\alpha^{old}(a) = \frac{1}{|A|}, \tag{2}$$

where $|A|$ denotes the total number of atlases.

2. For each voxel i, in terms of each atlas a, calculate:

$$\log p(I_i, W_i|a, \hat{\varphi}_a) = \log p(I_i|W_i, a, \hat{\varphi}_a) + \log p\left(W_i|a, \hat{\varphi}_a\right), \tag{3}$$

where

$$p(I_i|W_i, a, \hat{\varphi}_a) = \frac{1}{\sqrt{2\pi}\sigma(W^a \circ \hat{\varphi}_a^{-1})} e^{-\frac{\left|I_i - \mu(W^a \circ \hat{\varphi}_a^{-1})\right|^2}{2\sigma(W^a \circ \hat{\varphi}_a^{-1})^2}}, \tag{4}$$

with

$$\mu(W^a)_i = \frac{\sum\limits_{j \in structure} I_j^{(a)}}{\sum\limits_{j \in structure} 1}, (\sigma(W^a)_i)^2 = \frac{\sum\limits_{j \in structure} \left(I_j^{(a)} - \mu(W^a)_i\right)^2}{\sum\limits_{j \in structure} 1}, \tag{5}$$

where i indexes different structures. The quantity $p\left(W_i|a, \hat{\varphi}_a\right)$ is calculated by performing trilinear interpolation when transferring the manual labels $W^{(a)}$ of the atlases under the action of diffeomorphism $\hat{\varphi}_a(\cdot)$ – composition with $\hat{\varphi}_a^{-1}$.

3. Update the label classification of each voxel in the target via:

$$W_i^{new} = \arg\max_{W_i} \sum_j \sum_a \alpha_j^{new}(a) log p(I_j, W_j|a, \hat{\varphi}_a), \tag{6}$$

where i indexes voxels.

4. Update segmentation $W^{old} \leftarrow W^{new}$, and compute optimum diffeomorphism for each $A_i = a$ via:

$$\hat{\varphi}_a = \arg\max_{\varphi} p(a, \varphi|W^{old}, I) \tag{7}$$

$$= \arg\max_{\varphi} \log p(W^{old}|a, \varphi, I) + \log \pi(a, \varphi) \tag{8}$$

where $\pi(a, \varphi)$ is the prior probability of the atlas a and its diffeomorphism to the subject. We use the metric distance in LDDMM [5] to estimate this prior probability.

5. Update $\alpha_i^{old}(a) \leftarrow \frac{p(a, \hat{\varphi}_a|W^{old}, I_i)}{\sum\limits_a p(a, \hat{\varphi}_a|W^{old}, I_i)}$ for each $A_i = a$, go to 2.

2.4 Comparison Metrics

In order to validate the segmenting accuracy of our method, we compare our automated segmentation results with the manual segmentations of the same datasets in three respects:

– Kappa Score κ
The Kappa Score [10] is defined by:

$$\kappa = \frac{p_{agree} - p_{random}}{1 - p_{random}}, \tag{9}$$

where p_{agree} is the fraction of voxels in which the given segmentation agrees with the manual segmentation, and p_{random} is the fraction you would expect by random chance (based only on the volumes of foreground and background). κ is biased by the volume of the structure. Generally, the bigger the structure, the higher the kappa score. For applications involving brain structures, a value of $\kappa = 0.8$ is considered quite good.

– Volume Difference (VD)
Given that many studies are only interested in quantifying the structural volumetric changes, another metric, which quantifies volume difference between too labels, was defined in [11]:

$$VD\left(L_A, L_M\right) = \frac{\left|V\left(L_A\right) - V\left(L_M\right)\right|}{\left(V\left(L_A\right) + V\left(L_M\right)\right)/2} \tag{10}$$

where $V\left(L_A\right)$ is the volume size of the automated segmentation, $V\left(L_M\right)$ is the volume size of the manual labeling.

– In the end, we compare the volume size of the automated segmentation averaged across subjects with that of the manual segmentation in mm^3.

3 Results

3.1 Comparison to FreeSurfer and FIRST

In the first experiment, accuracy and reliability of Multi-atlas LDDMM is compared with that of the segmentation module FIRST [2] in FSL and that of FreeSurfer [1] in segmenting fourteen structures in eight T1-weighted images. FIRST and FreeSurfer are chosen for the comparison because they both provide state-of-art segmenting accuracy for most subcortical structures among the most widely used segmentation algorithms. The results are given in Figs. 1- 3, which illustrate the comparisons of the three methods in terms of kappa score (Fig. 1), volume difference (Fig. 2), and mean volume sizes of different structures (Fig. 3).

3.2 Comparisons with STAPLE

In the second experiment, we compare the segmentation accuracy of Multi-atlas LDDMM with that of STAPLE in terms of Kappa overlaps. The kappa overlap results of the two methods for each structure are tabulated in Table 1. Both Multi-atlas LDDMM and STAPLE achieved high accuracy. To compare

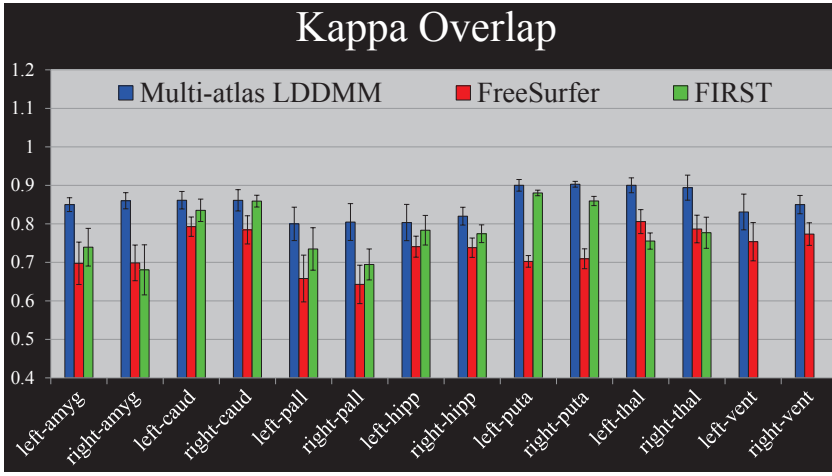


Fig. 1. Demonstration of mean and standard deviations of kappa scores of each segmented structure for Multi-atlas LDDMM (blue), FreeSurfer (red), and FIRST (green) of the 8 T1 images.

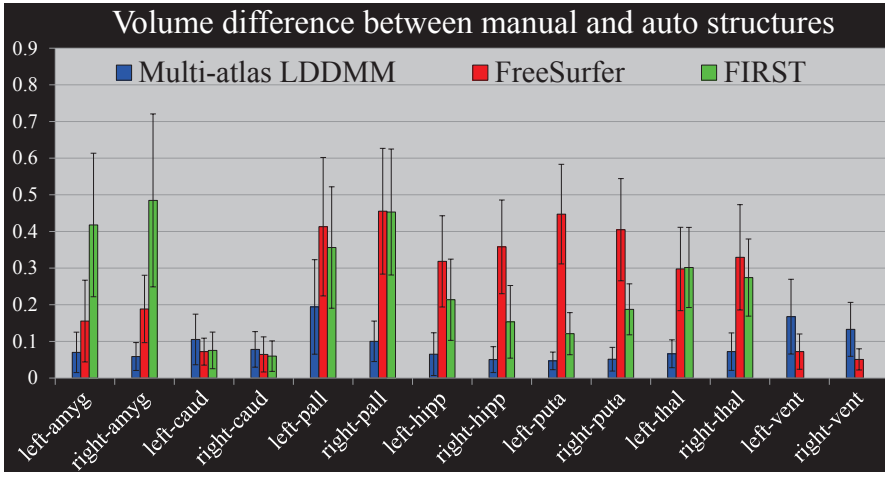


Fig. 2. A comparison of Multi-atlas LDDMM (blue), FreeSurfer (red), and FIRST (green) in terms of volume difference (mean value and standard deviations) of the 14 structures from the 8 subjects.

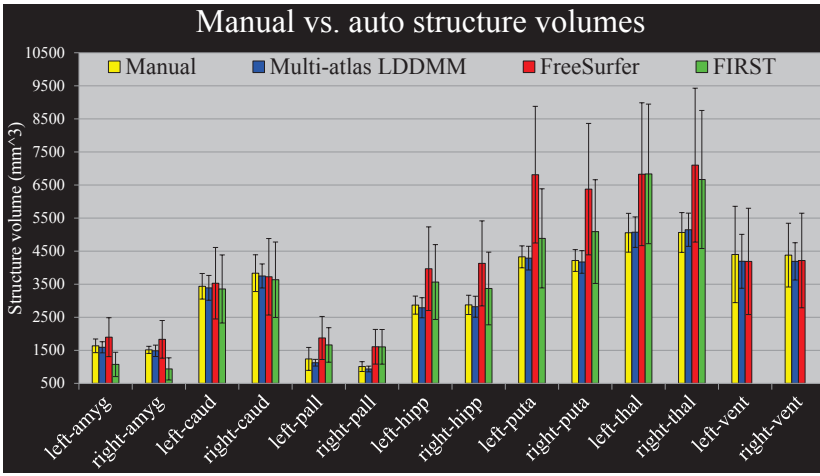


Fig. 3. Comparison of structure volumes in terms of the mean value and the standard deviation computed from manual labelings (yellow) and automated segmentations of Multi-atlas LDDMM (blue), FreeSurfer (red), as well as FIRST (green) for all the 14 structures studied in the 8 T1 images.

the two methods statistically, we performed 10,000 permutation tests to get the correct p-values for each comparison between the two samples of results for each structure. Permutation test results revealed that the kappa overlap ratios of left amygdala, right hippocampus, right putamen, and ventricles in both hemispheres are significantly higher than those obtained from STAPLE ($p < 0.05$). For other structures, Multi-atlas LDDMM is comparable to STAPLE but seemingly shows slight improvements.

3.3 Whole brain segmentation of Multi-atlas LDDMM

In the third experiment, we extend the segmentation from subcortical structures to the whole brain using the training datasets provided by the organizer of the workshop using a leave-one-out test. Based on the manual labelings of the training datasets, we segment the whole brain into 136 cortical and subcortical regions. The automated results have been compared with those of manually labeling the same datasets. Due to the space limitations, we only list the kappa scores of 41 regions including all the subcortical structures, ventricular structures, and some of the cortical and white matter regions. The kappa score results of subcortical regions on these 15 MR datasets agree with what we obtained in the eight subjects, which shows that our method is robust, and insensitive to the datasets it is applied to. According to the results shown in Fig. 4, Multi-atlas LDDMM is capable of achieving a kappa score higher than 0.8 for most brain structures.

The automated whole brain segmentations from Multi-atlas LDDMM of two representative subjects from the 15 MR datasets are depicted in Fig. 5, including

Table 1. Mean and standard deviation of Kappa overlap ratio computed across the 8 subjects for all the 14 structures for comparison of STAPLE and Multi-atlas LDDMM. Bold typesetting indicates that Kappa overlap ratio of Multi-atlas LDDMM is statistically significantly higher than that of STAPLE.

	STAPLE	Multi-atlas LDDMM
left-amyg	0.817 (0.0252)	**0.846 (0.0178)**
right-amyg	0.838 (0.0193)	0.859 (0.0209)
left-caud	0.825 (0.0440)	0.862 (0.0227)
right-caud	0.830 (0.0387)	0.862 (0.0276)
left-pall	0.781 (0.0490)	0.800 (0.0434)
right-pall	0.781 (0.0367)	0.805 (0.0478)
left-hipp	0.758 (0.0414)	0.804 (0.0468)
right-hipp	0.775 (0.0257)	**0.816 (0.0231)**
left-puta	0.885 (0.0198)	0.901 (0.0152)
right-puta	0.890 (0.0119)	**0.903 (0.0074)**
left-thal	0.883 (0.0258)	0.901 (0.0193)
right-thal	0.876 (0.0423)	0.895 (0.0324)
left-vent	0.783 (0.0917)	**0.831 (0.0464)**
right-vent	0.795 (0.0586)	**0.846 (0.0237)**

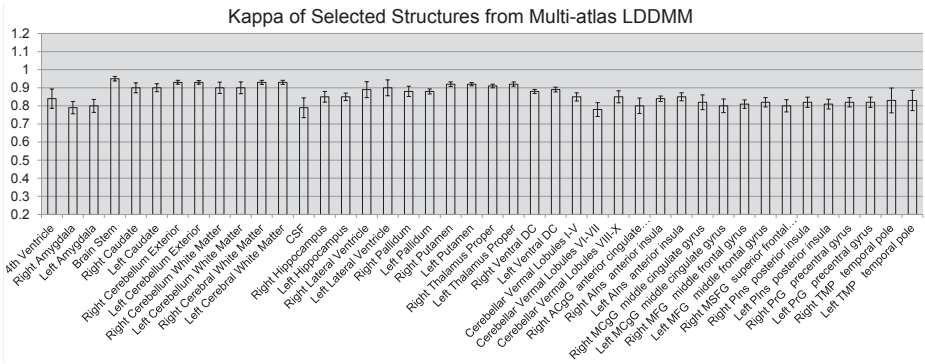


Fig. 4. The averaged kappa scores and the standard deviations of the 15 training subjects provided by the workshop organizer for 41 different brain structures obtained from Multi-atlas LDDMM.

a total of 136 structures. The names of the structures can be found at the website of the workshop.

4 Conclusion

We present an automated whole brain segmentation method based on multiple atlases which we term Multi-atlas LDDMM. Our method was shown to attain automated segmentations for subcortical and ventricular structures with kappa scores ranging from 0.8 to 0.9. For white matter segmentation, our method is capable of obtaining automated segmentation with kappa score larger than 0.91. For example, for cerebellum white matter in both hemispheres, the mean kappa score obtained from 15 subjects is 0.91, and 0.96 for the cerebral white matter. Our method has also been shown to achieve good segmentation results for cortical regions. For example, the mean kappa score of the automated segmentations for middle frontal gyrus is 0.82, 0.82 for precentral gyrus, 0.82 for posterior insula, and 0.83 for anterior insula.

We have included a brief comparison with STAPLE which is based on labeled fusion and demonstrated seemingly similar or more accurate results. The crucial difference between Multi-atlas LDDMM and STAPLE is that our Eq. (5) 'fuses likelihoods' thereby only generating a single segmentation, rather than generating multiple segmentations and performing label fusion on them.

5 Acknowledgments

The work presented here was supported by grants: NIH R01 EB000975, NIH P41 EB015909 and NIH R01 EB008171. The authors would like to acknowledge Dr. Tilak Ratnanather and Dr. Steven Yantis for contributions to the data collection and Timothy Brown and Huong Trinh for assistance in manual delineation of several subcortical structures.

References

1. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.: Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. Neuron 33, 341–355 (2002)
2. Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M.: A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 907–922 (2011)
3. Pohl, K.M., Fisher, J., Grimson, W.E.L., Kikinis, R., Wells, W.M.: A Bayesian model for joint segmentation and registration. NeuroImage 31, 228–239 (2006)
4. Baillard, C., Hellier, P., Barillot, C.: Segmentation of brain 3D MR images using level sets and dense registration. Med. Image Anal. 5, 185–194 (2001)
5. Miller, M.I., Trouv, A., Younes, L.: On the metrics and Euler−Lagrange equations of computational anatomy. Annu. Rev. Biomed. Eng. 4, 375–405 (2002)

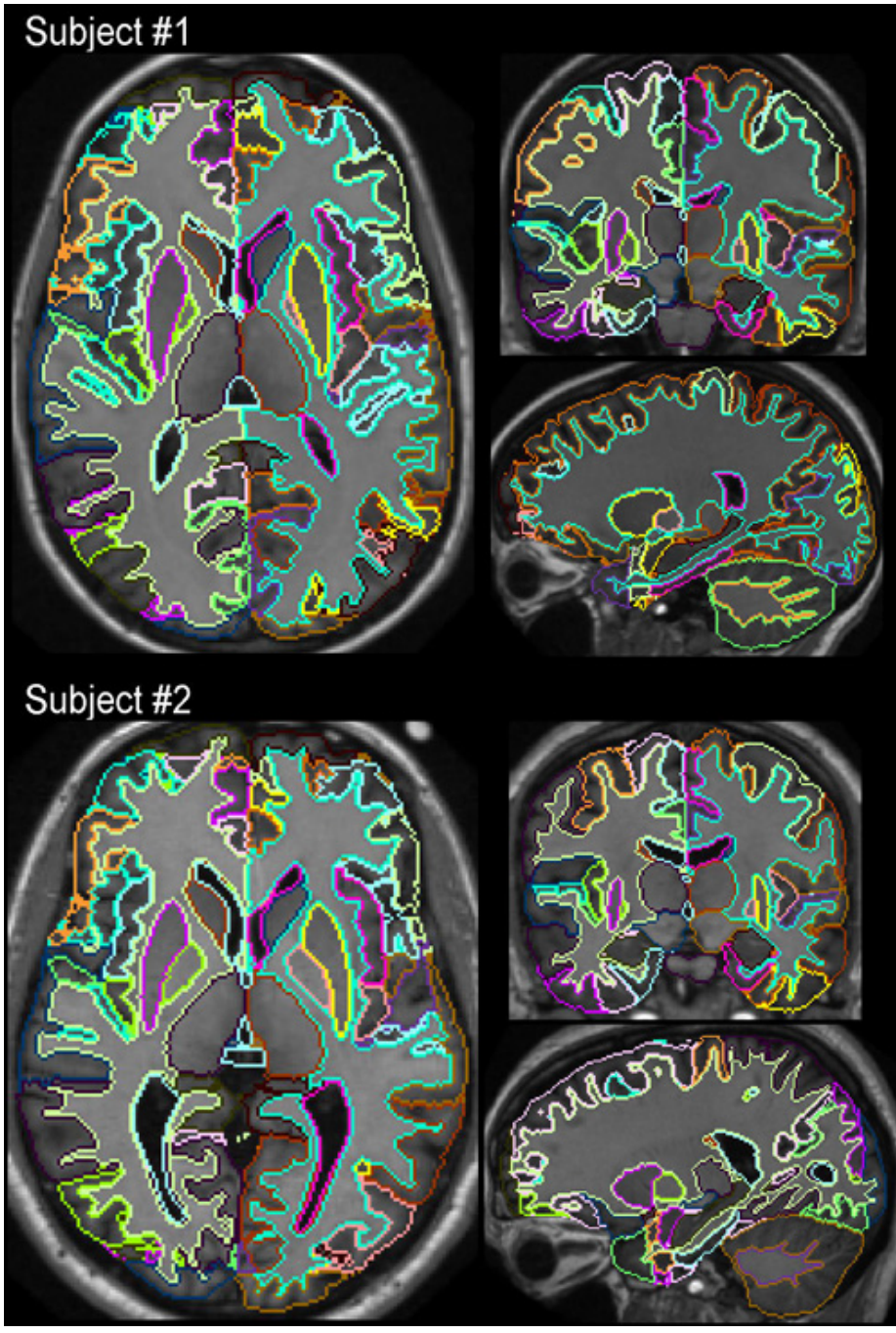


Fig. 5. Figure shows whole brain segmentations of two representative T1-weighted images among the 15 training subjects in three views: Axial (left), Coronal (top right), and Saggital (bottom right).

6. Beg, M.F., Miller, M.I., Trouv, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. 61, 139–157 (2005)
7. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126 (2006)
8. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903-921 (2004)
9. Dupuis, P., Grenander, U., Miller, M.I.: Variational problems on flows of diffeomorphisms for image matching. Quaterly of Applied Math 56, 587-600 (1998)
10. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 159–174(1977)
11. Collins, D.L., Holms, C.J., Peters, T.M., Evans, A.C.: Automatic 3-D model-based neuroanatomical segmentation. Hum. Brain Mapp. 3, 190–208(1995)

Multi Atlas Segmentation applied to in vivo mouse brain MRI

Ma Da[1,2], M. Jorge Cardoso[1], Marc Modat[1], Nick Powell[1,2], Holly Holmes[2], Mark Lythgoe[2], Sébastien Ourselin[1,3]

[1] Centre for Medical Imaging Computing, Department of Medical Physics and Bioengineering, University College London, UK,
[2] Centre for Advanced Biomedical Imaging, University College London, UK,
[3] Dementia Research Centre, Institute of Neurology, WC1N 3BG, University College London, UK.

Abstract. Multi-atlas segmentation propagation has evolved quickly in recent years, becoming a state-of-the-art method for automatic structural parcellation for brain MRI. However, few studies have applied these methods to preclinical research. In this study, we present a fully automatic multi-atlas segmentation pipeline for mouse brain MRI tissue parcellation. The pipeline adopts the Multi-STEPS multi-atlas segmentation algorithm, which utilises a locally normalised cross correlation (LNCC) similarity metric for atlas selection and an extended STAPLE framework for multi-label fusion. The segmentation accuracy of the pipeline was evaluated using an *in vivo* mouse brain atlas with pre-segmented manual labels as gold standard, and optimised parameters were obtained. Results show a mean Dice similarity coefficient of 0.839 over all the structures and for all the samples in the database, significantly higher than in a single atlas propagation strategy, and also generally higher than STAPLE strategy, although the improvement is not significant.

1 Introduction

Mice share more than 80% of genomes with human, making it a good animal for preclinical study of human brain diseases, such as Alzheimers disease and Downs syndrome. Preclinical studies normally require relatively large sample size, thus an accurate, robust and reproducible method for quantitative analysis of preclinical MRI images is necessary for high-throughput studies. More specifically, structural parcellation enables the study of volume, shape and morphological characteristics of key brain structures. Despite the labour intensive and expert-dependent nature of the task, manual labelling of anatomical structures is still standard practice in mouse brain MRI studies [1, 2]. Various automated labelling algorithms have thus been developed to address these limitations [3, 4]. Segmentation propagation is a method where a template, i.e. an accurate manual delineation of anatomical structures that follows a well-defined segmentation protocol, is propagated to a query image through the process of image registration. Although the accuracy and efficiency of image registration algorithms has been constantly improving, the segmentation performance is still

limited by inaccuracies in the registration process, especially between morphologically dissimilar subjects. This problem can be greatly ameliorated by propagating multiple image templates and subsequently fusing them into a consensus segmentation through a process known as label fusion [5–7]. A great deal of effort has been put in exploring the structural parcellation of human brain MRI [8–11]. However, in preclinical research (e.g. mouse model), a study about structural parcellation techniques is still lacking. Maheswaran *et al.* compared a single atlas segmentation propagation with deformation based morphormetry (DBM) [12], and concluded that atlas-base method can identify longitudinal and cross-sectional group difference, but is less sensitive to much smaller regional changes compared to DBM. Artaechevarria *et al.* [13] adopted a weighted majority voting label fusion using an ex vivo mouse brain MRI atlas containing 10 individual samples with 20 manually labelled structures [14]. In cases where only one single template is available, Chakravarty *et al.* [15] proposed to first propagate the template to a set of unlabelled images with traditional single-atlas segmentation propagation and then propagate the resultant set of segmentations to another new image using majority voting label fusion, which resulted in an increase in performance. However, with the improvement of hardware and scanning protocols, mouse studies are moving from *ex vivo* to *in vivo* imaging, leading to much lower contrast to noise ratios (CNR) and signal to noise ratios (SNR). Bai *et al.* [3] have recently published a study using majority voting and STAPLE (Simultaneous Truth and Performance Level Estimation) multi-atlas label fusion to in vivo mouse brain MRI, and compare the improvement gained with that of the non-rigid image registration. In this paper, we use a new multi-atlas based structural parcellation method, Multi-STEPS (Multi-label Similarity and Truth Estimation for Propagated Segmentations) [16], on *in vivo* mouse brain MRI images. We developed a fully automated pipeline for brain parcellation and optimised the fusion strategy parameters using a leave-one-out cross validation.

2 Methods

In this section we will describe the steps used for multi-atlas structural parcellation. A schematic diagram of the pipeline is shown in Figure 1.

2.1 Brain extraction

Starting from a set of template images with associated tissue parcellations and brain masks, the first step of the pipeline was to create a brain mask for the query image. This goal was achieved by propagating the brain masks defined on the template images using the Multi-Atlas Propagation and Segmentation (MAPS) strategy developed by Leung *et al.* [17].

2.2 Bias field correction

MR images are corrupted by intensity non-uniformity, or bias, cause by the inhomogeneity of the RF excitation field, the spatially nonuniform receiver coil

sensitivity profiles, the induced currents and standing wave affects [18]. Intensity non-uniformity leads to misalignment in the registration process due to corrupted intensity profile. We thus used the N3 intensity non-uniformity correction algorithm developed by Sled *et al.* [18] to correct the bias field.

2.3 Template registration

After correction of the intensity non-uniformity, we first globally and then non-linearly registered the masked template images to the query image. The global registration was performed using a block-matching approach [19]. A parametric approach based on a cubic B-Spline parameterisation [20] was used for non-linear registration. We used the efficient implementation proposed by Modat *et al.* [21]. The resulting deformation fields obtained from the registrations were then used to resample the labels from the template image spaces to the query image. Nearest-neighbour interpolation was used to preserve the integer nature of the original labels.

2.4 Multi-atlas label fusion

The label fusion was conducted using the Multi-STEPS algorithm developed by Cardoso *et al.* [16]. Multi-STEPS is an extension of the original STAPLE algorithm [5, 6] with several improvements. Firstly, it includes a Markov Random Field (MRF) used in an iterative manner to maintain spatial consistency. It also incorporates a template selection step using a ranking strategy based on the locally normalised cross correlation (LNCC) over a local Gaussian window. This fusion strategy has two main user-defined parameters: the width of the Gaussian kernel for image comparison and the number of labels to fuse after ranking. The optimisation of these parameters is described in section 3.2.

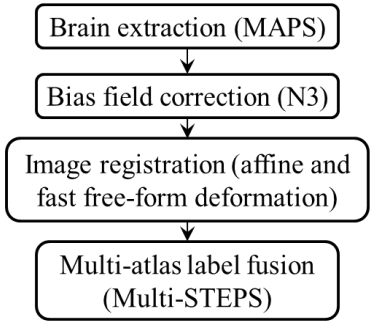


Fig. 1. Proposed multi-atlas segmentation propagation pipeline.

3 Validation and results

This section will present the optimisation of the fusion strategy parameters as well as the segmentation performance evaluation.

3.1 Data

To evaluate the performance of the method and optimise the parameters of the proposed pipeline, we used a previously described *in vivo* mouse brain MRI database containing 12 individual brain T2* MRI of 12-14 weeks old C57BL/6J mouse. Each MRI brain image is associated with 20 manually delineated structures. Detailed scanning parameters are described in [1]. Due to missing labels in 3 of the 12 available templates, only 9 images and associated parcellations were included in this study.

3.2 Parameter optimisation

The Multi-STEPS label fusion performance depends mostly on the width of the Gaussian kernel and the number of top ranked templates used for fusion. We will thus focus on optimising these parameters.

A leave-one-out cross validation was performed to assess the segmentation accuracy as well as to optimise the parameters of Multi-STEPS. For each of the 9 samples, the remaining 8 samples were used as template sets for multi-atlas segmentation. The average Dice similarity coefficient between the automatic segmentation and the manual segmentation of all the structures for all the individual sample images was calculated. We ran the leave-one-out validation for each combination of parameters, with the Gaussian kernel standard deviation varying from 1 to 6 (step of 0.5) and the number of templates used from 3 to 8. The parameter combination that gave the highest Dice similarity coefficient was selected and regarded as the optimal combination.

The results of the Multi-STEPS parameter optimisation are shown in Figure 2. The best combination of parameters was: number of local templates used equal to 6, with a Gaussian kernel with a standard deviation of 4. The corresponding average Dice similarity coefficient between automatic and manual segmentation, for all structures and templates, was 0.839 with standard deviation of 0.025.

Figure 2 shows that there is a large plateau zone (i.e. a small variation in Dice similarity coefficient) close to the optimal model parameters, indicating high stability of the pipeline with regards to the selection of parameters. Another possible explanation for the segmentation stability could be the smaller inter-template morphological variation for mice when compared to humans, thus making the fusion less dependent on the parameter selection. Example images of segmentation results and the correspond-ing manual labelling are presented in Figure 3.

We also compared the average Dice similarity coefficient of our pipeline result with the single template-based segmentation propagation and STAPLE. For single template-based segmentation, we propagated all templates and averaged the

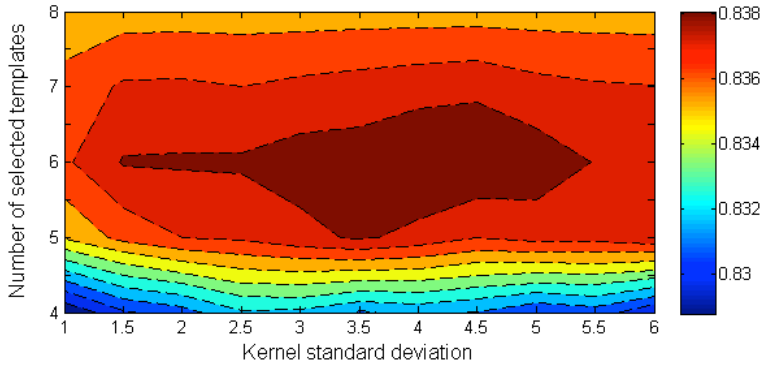


Fig. 2. Average Dice similarity coefficients for different combinations of Gaussian kernel standard deviation and number of selected templates in Multi-STEPS algorithm. The optimal parameter were found to be: number of templates = 6 and Gaussian kernel standard deviation = 4 (Dice = 0.839).

Dice similarity coefficients. The results are shown in Figure 4. The average Dice similarity coefficient of our pipeline was in general higher than both the single-atlas method and STAPLE for most of the structures. When compared with single-atlas method, significant improvements were found in External Capsule, Ant Commissure, Internal Capsule, Ventricles and Fimbria. The improvements compared to STAPLE were not statistically significant. We believe the low samples number of the manual segmentations is the main cause of this effect. Also, the fact that all the manual labels come from the same atlas in leave-one-out cross validation may effectively result in relatively high Dice similarity coefficient for STAPLE, which may not be the case for newly acquired images. Further research will explore and characterise these limitations.

3.3 Pipeline robustness testing

In order to test the ability to segment new unseen datasets, we acquired *in vivo* images of mouse brains and applied the pipeline to obtain the corresponding anatomical labels. These scans were obtained using a Varian VNMRS 9.4 Tesla MRI system (Agilent Technologies Inc. Palo Alto CA, USA). A 72-mm volume coil (RAPID Biomedical GmbH, Würzburg, Germany) was used for excitation and a quadrature mouse brain surface coil (Bruker Biospin GmbH, Ettlingen, Germany) was used for signal detection. T1 weighted contrast was achieved using an efficient fast spin echo (FSE) sequence. The data was acquired with the parameters $\mathrm{TR/TE}_{eff}$ = 2500/12.7 ms, ETl = 4, 1 average, field of view = $24.6 \times 16.8 \times 12.0$, with spatial resolution of $150 \times 150 \times 150$ isotropic. The total *in vivo* imaging time was approximately 1 hour and 30 minutes. An example of the segmentation results in one of the scanned subjects is shown in Figure 5. Due to the lack of available manual segmentations, quantitative analysis could

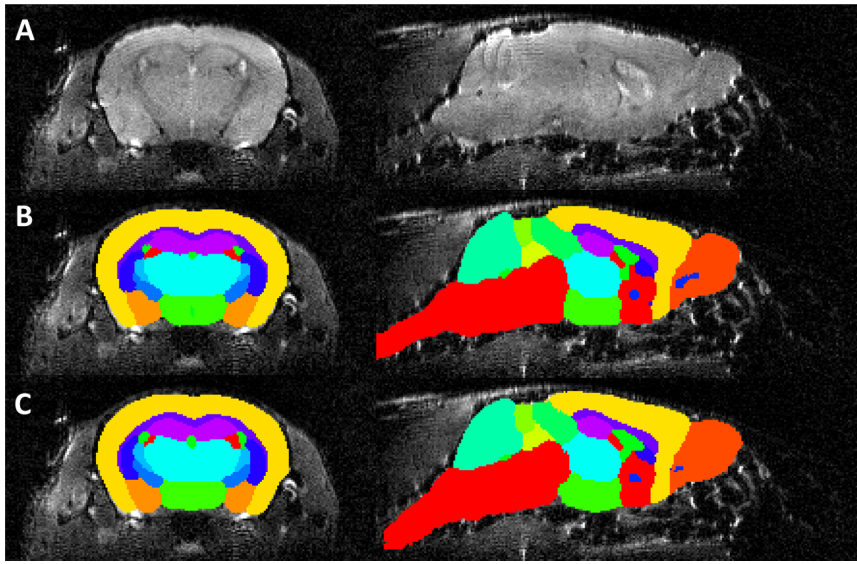


Fig. 3. Sample images showing the coronal view (left) and sagittal view (right) of the template image (A), overlaid with the multi-atlas segmentation results (B) and the manual labels (C).

not be performed on these new datasets. However, visual inspection has shown good segmentation accuracy.

4 Discussion and conclusion

The proposed work utilises the state-of-the-art multi-atlas segmentation propagation method Multi-STEPS, along with the fast free-form deformation registration algorithm and other pre-processing techniques such as brain extraction (MAPS) and intensity non-uniformity correction (N3), to create an integrated and fully automated pipeline for brain segmentation. This paper presents the successful application of advanced multi-atlas segmentation techniques for *in vivo* mouse brain parcellation.

The optimised Multi-STEPS parameters were chosen based on the average Dice similarity coefficient over all the structures and for all samples in the template data-base. However, one should note that the Dice similarity coefficient intrinsically favours large structures. For example, small structures (e.g. external capsule, anterior commissure) show much worse performance than larger structures (e.g. hippocampus, neocortex) due to local registration errors, inter-template morphological variability and human segmentation consistency. Contrast between structures can also have a detrimental effect on segmentation performance. The lack of contrast between some neighbouring anatomical regions can lead to decreased performance as the registration algorithm will have

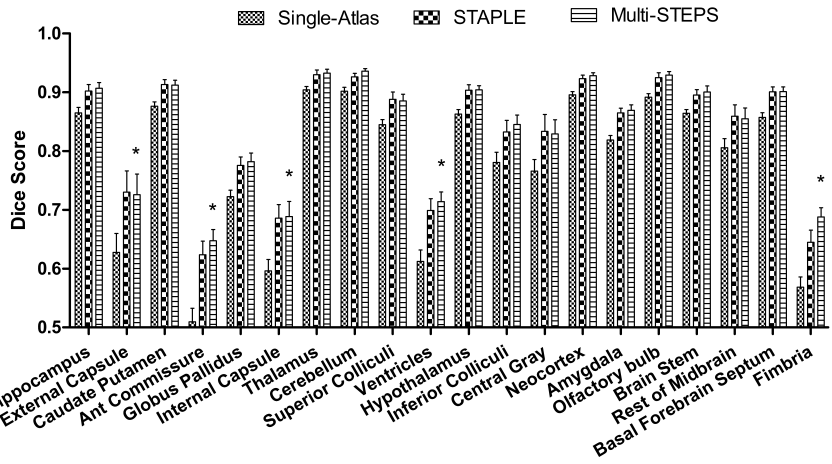


Fig. 4. Average Dice similarity coefficient comparison between traditional single-atlas segmentation propagation, STAPLE and the Multi-STEPS method. Two-way ANOVA statistical test was performed. Significant differences were found for some structures (*) between single-atlas segmentation propagation and Multi-STEPS method (*: $p < 0.001$). The improvement of STEPS compared to STAPLE does not reach statistically significance.

to rely on the regularisation term (rather than on image features) for accurate structural matching. Conversely, the nature of most measure of similarity will also lead to a registration algorithm that is governed by high contrast edges, possibly reducing the propagation accuracy in low-contrast areas.

Compared to human brain MRI segmentation studies, the availability of mouse templates and the amount of information about the segmentation protocols is very limited. Subsequently, label fusion techniques are limited in performance by several different factors. First, the templates used for the presented work are limited in number and are defined only on T2* images, impeding their direct application to other imaging modalities. While certain similarity measures for image registration can deal with multi-modal images, the lack of contrast between certain anatomical structures in other modalities will reduce the accuracy of the parcellation algorithm. Second, the lack of anatomical standardisation and vague definition of the segmentation protocol reduces the consistency between human raters. Finally, intra- and inter-rater labelling variability has not been assessed in mice. Since the manual segmentations are used for comparison, as they are considered as gold standard, the information about intra- and inter-rater labelling variability is of critical importance because it represents the theoretical performance upper limit for automated methods.

Lastly, the estimated optimal parameters and segmentation performance were only assessed within the template database. Although the application to new testing data has good visual assessed segmentation accuracy, further validation is still necessary in order to enable the unsupervised use of this algorithm

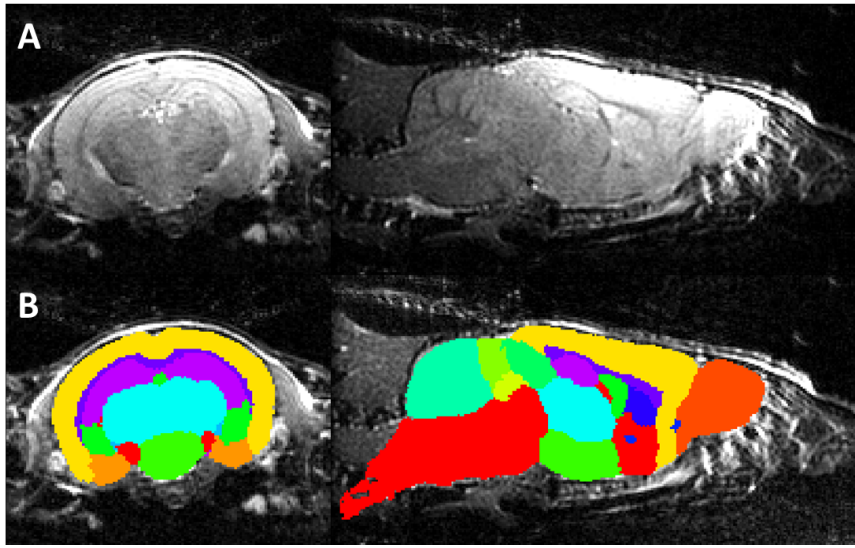


Fig. 5. Sample images showing the coronal view (left) and sagittal view (right) of the test image data (A), overlaid with the multi-atlas segmentation results (B).

in a pre-clinical setting and for different mouse models. Future work will also include the optimisation of the registration parameters, which are here considered as fixed.

Acknowledgement

This work was undertaken at UCL which received a proportion of funding from Faculty of Engineering funding scheme. This project is also supported by CBRC grant 168. The author would like to thank Y. Ma *et al.* who publically release the in vivo mouse MRI brain atlas [1]. Without their atlas, this project would not be possible.

References

1. Ma, Y., Smith, D., Hof, P.R., Foerster, B., Hamilton, S., Blackband, S.J., Yu, M., Benveniste, H.: In vivo 3D digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. Frontier Neuroanatomy **2** (Apr 2008) 1–10
2. Richards, K., Watson, C., Buckley, R.F., Kurniawan, N.D., Yang, Z., Keller, M.D., Beare, R., Bartlett, P.F., Egan, G.F., Galloway, G.J., Paxinos, G., Petrou, S., Reutens, D.C.: Segmentation of the mouse hippocampal formation in magnetic resonance images. NeuroImage **58**(3) (Oct 2011) 732–740

3. Bai, J., Trinh, T.L.H., Chuang, K.H., Qiu, A.: Atlas-based automatic mouse brain image segmentation revisited: model complexity vs. image registration. Magnetic Resonance Imaging **30**(6) (Jul 2012) 789–798
4. Lee, J., Jomier, J., Aylward, S., Tyszka, M., Moy, S., Lauder, J., Styner, M.: Evaluation of atlas based mouse brain segmentation. In: Proceedings of SPIE. (2009) 725943–725949
5. Rohlfing, T., Russakoff, D.B., Maurer, C.R.: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Transactions on Medical Imaging **23**(8) (Aug 2004) 983–994
6. Warfield, S.K., Zou, K.H., Wells III, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7) (Jul 2004) 903–921
7. Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J.V., Rueckert, D.: Classifier selection strategies for label fusion using large atlas databases. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Volume 10. (2007) 523–531
8. Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., Hammers, A., Initiative, A.D.N.: Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. NeuroImage **51**(1) (May 2010) 221–227
9. Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S.: Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Human Brain Mapping **19**(4) (Aug 2003) 224–247
10. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage **33**(1) (Oct 2006) 115–126
11. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron **33**(3) (Jan 2002) 341–355
12. Maheswaran, S., Barjat, H., Bate, S.T., Aljabar, P., Hill, D.L.G., Tilling, L., Upton, N., James, M.F., Hajnal, J.V., Rueckert, D.: Analysis of serial magnetic resonance images of mouse brains using image registration. NeuroImage **44**(3) (Feb 2009) 692–700
13. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C.: Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Transactions on Medical Imaging **28**(8) (Aug 2009) 1266–1277
14. Ma, Y., Hof, P.R., Grant, S.C., Blackband, S.J., Bennett, R., Slatest, L., McGuigan, M.D., Benveniste, H.: A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. Neuroscience **135**(4) (Sep 2005) 1203–1215
15. Chakravarty, M.M., van Eede, M.C., Lerch, J.P.: Improved segmentation of mouse MRI data using multiple automatically generated templates. In: International Society for Magnetic Resonance in Medicine. Volume 15. (2011) 134
16. Cardoso, M.J., Modat, M., Keihaninejad, S., Cash, D., Ourselin, S.: Multi-STEPS: Multi-label similarity and truth estimation for propagated segmentations. In: Mathematical Methods in Biomedical Image Analysis, 2012 IEEE Workshop on. (2012) 153–158

17. Leung, K.K., Barnes, J., Modat, M., Ridgway, G.R., Bartlett, J.W., Fox, N.C., Ourselin, S., Initiative, A.D.N.: Brain MAPS: an automated, accurate and robust brain extraction technique using a template library. NeuroImage **55**(3) (Apr 2011) 1091–1108
18. Sled, J., Zijdenbos, A., Evans, A.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging **17**(1) (Feb 1998) 87–97
19. Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N.: Reconstructing a 3D structure from serial histological sections. Image and Vision Computing **19**(1-2) (Jan 2001) 25–31
20. Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D.: Nonrigid registration using free-form deformations: Application to breast MR images. IEEE Transactions on Medical Imaging **18**(8) (Aug 1999) 712–721
21. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine **98**(3) (Jun 2010) 278–84

Parametric Images: An Image Representation that Preserves Edge Strength in Registration and Atlasing

Blake C. Lucas[1], Yoshito Otake[1], Mehran Armand[1,2], and Russell H. Taylor[1]

[1] Johns Hopkins University, Baltimore, MD, USA
[2] Johns Hopkins Applied Physics Laboratory, Laurel, MD, USA
blake@cs.jhu.edu, otake@jhu.edu, mehran.armand@jhuapl.edu, rht@jhu.edu

Abstract. Deformable registration is commonly used to construct atlases of the human body from medical images. During registration, warping images with deformation fields can stretch or compress edges; and after registering images to a template, any misalignment of anatomical boundaries will likely produce a blurry image when images are averaged together. This work introduces a novel parametric image (p-image) representation that preserves edge strength when averaging or deforming images, regardless of the registration algorithm's performance. The key idea is to decouple shape and intensity information, manipulate them independently, and then synthesize images. We will describe how to warp and form linear combinations of parametric images. Applications to whole brain and whole body atlas construction will be presented, but the method is not particular to these applications nor does it require any specific registration algorithm.

Keywords: registration, segmentation, atlas, active appearance models, spring level sets.

1 Introduction

One way to construct an anatomical atlas is to combine multiple 3D medical images (e.g. CT or MR) of the human body. The atlas should ideally capture variations in geometry and intensity for different tissue types. A common approach is to use deformable registration to create a dense 3D-to-3D mapping between template and subject [1, 2]. The template image can be deformed into the subject's image space or vice versa to construct an atlas.

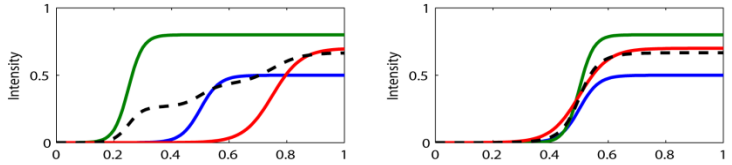
One way to construct an image intensity atlas is to register all subjects to a template image and then average those warped subject images [2, 3]. A known problem with this approach is that the average image is blurry and has worse contrast than any individual image. In fact, the sharpness of the average image is commonly used to assess the performance of deformable registration algorithms [2]. Blurriness is an indicator of misaligned anatomical boundaries. If boundaries are misaligned, then the average intensity at a boundary voxel is computed across different structures, which creates ambiguity as to which structure the average intensity corresponds.

The atlas construction method just described characterizes differences in intensity, not geometry. To capture variations in geometry, the template image is

deformably registered to all other subjects. After which, any geometric structures identified in the template image can be warped into the subject's image space. There are several approaches to analyze geometric differences. One approach is to create mesh segmentations of each anatomical structure, warp them into each subject's image space, and analyze them with Principal Component Analysis (PCA) [3]. Alternatively, one can transform image segmentations into a LogOdds representation and analyze them with PCA [4].

Active Appearance Models (AAM) [5, 6] extend the PCA method to model shape and intensity information simultaneously. Active Appearance Models decompose an image domain into elements (i.e. tetrahedra in 3D) and attach a texture map to each element. Intensity information is analyzed and combined in shape-normalized space (texture space) instead of image space to factor out variability due to differences in geometry. PCA is conducted on both shape and intensity information simultaneously to model variability in both. Similar to the deformable registration approach, any shape variability not accounted for by the registration algorithm can blur boundaries when images are averaged together [6]. Also, warping images with either deformable image registration or AAMs can stretch or compress edges, which contributes to a change in sharpness.

To illustrate the importance of preserving edge strength, consider a 1D example in which two structures are differentiated by one edge that is slightly misaligned in three images. Fig. 1a shows that averaging images (edges located at x={0.25,0.5,0.75}) blurs the edge because the average intensity is computed across different structures. Fig. 1b shows that if we first average the location of the edge, shift the intensity profile for each edge to the average edge position, and then average the image intensities, edge shape and strength are maintained. Similarly, warping an image can stretch and compress edges (Fig. 2b), but if we warp the geometry and then shift the intensity profiles, edge shape is maintained (Fig. 2c). We would like to extend this procedure to 3D; but before doing so, we must have explicit knowledge of edge locations and an intensity representation that shifts along with an edge's location.



(a) (b)

Fig. 1. (a) 1D images (solid lines) and average image (dotted line). (b) 1D images after first averaging edge locations (solid lines) and then averaging image intensities (dotted line).

(a) (b) (c)

Fig. 2. (a) original image, (b) warped image, (c) warped parametric image. Notice in (b) that the edges around "NICE 2012" are stretched and edges around "MICCAI" are compressed.

Previous work [1-3] has focused on improving boundary alignment in registration algorithms so that the scenario in Fig 1b. is more likely. However, even if boundaries are perfectly aligned by the registration algorithm, warping an image with a deformation field can stretch or compress edges (Fig. 2). This work presents a parametric image representation that explicitly aligns edges before combining intensities so that the scenario in Fig. 1b. is always the case. The key idea is to decouple shape and intensity information, manipulate them independently, and then synthesize images. We will make the assumption that all edges are step edges (i.e. no texture, shading, or thin lines), and edge sharpness varies as a function of distance to a region's boundary. To demonstrate the utility of such a representation, we construct shape and intensity atlases from MR and CT images of the human body.

2 Method

2.1 Representation

Geometric Model. To extend the procedure in Fig. 1 and 2 to 3D, a multi-object boundary representation is required to represent edge locations, and a distance field is required to encode intensity information. We use Multi-object Spring Level Sets (MUSCLE) [7, 8] because it provides a memory efficient, sub-voxel precision framework for modeling this type of information.

For image domain $\Omega \subset \mathcal{R}^3$ containing labeled object regions $\mathcal{L} = \{1, \cdots, L\}$, the MUSCLE representation consists of a label mask $\chi: \Omega \to \mathcal{L}$, distance field $\psi: \Omega \to \mathcal{R}$, and triangle mesh with vertices $q_n \in \mathcal{R}^3$. Each object has a corresponding triangle mesh that represents its boundary. To identify objects in the aggregate mesh, a label $l_n \epsilon \mathcal{L}$ is associated with each triangle vertex q_n. Meshes are coupled with signed distance fields $\varphi_l: \Omega \mapsto \mathcal{R}$ that provide a redundant, implicit representation of each object l. Since an image may have hundreds of different regions, level sets are compressed into a "distance field + label mask" representation. The distance field image measures the unsigned distance to the closest object boundary at each voxel (i.e. $\psi(\boldsymbol{x}) = \min_l |\varphi_l(\boldsymbol{x})|$), and the label mask image indicates to which object a voxel belongs. The smallest label is used in the event that a voxel is inside more than one object (i.e. $\chi(\boldsymbol{x}) = \min_{\{l|\varphi_l(\boldsymbol{x})<0\}} l$). This geometric representation is sufficient to reconstruct the boundaries of each object with sub-voxel precision. It is formally defined as the set of data structures $\mathcal{M}_g = \{\psi, \chi, \{q_1, \cdots, q_N\}, \{l_1, \cdots, l_N\}\}$.

In the MUSCLE framework, deformations are applied to a triangle mesh; after which, the distance field and label mask are evolved to track the moving mesh with Multi-Object Geodesic Active Contours (MOGAC) [9]. MUSCLE is an efficient way to parametrically represent and evolve multiple objects that guarantees no overlaps, self-intersections, or air-gaps between adjacent structures. It is also trivial to determine whether a voxel is inside a particular object with $\chi(\boldsymbol{x})$ and determine the distance from each voxel to the nearest boundary with $\psi(\boldsymbol{x})$. These two properties will be used to attach intensity information to $\mathcal{M}_g$ that accounts for mixtures (partial volumes) of tissue classes that occur near the boundary of object regions.

Intensity Model. To model edge intensity profiles analogous to the step edges in Fig. 1b, we compute the average intensity for voxels within distance ranges $Q = \{[0,1),[1,2),[3,4),[4,\infty)\}$ to the closest object boundary and store them in the function $f(l,d)$, where l is the object label and d is the distance to the closest object boundary. This model accounts for some of the inhomogeneity in image intensities that occur near the boundary of object regions as a function of distance. However, the model assumes only two objects are involved in the mixture.

We would like to attach more intensity information to $\mathcal{M}_g$ without introducing more geometric data structures. One useful extension is to model intensity mixtures for pairs of objects along object boundaries. To do so, voxels $\boldsymbol{y} \in \Lambda$ at object boundaries $\Lambda = \{\boldsymbol{y} | \exists \boldsymbol{x} \in \mathcal{N}(\boldsymbol{y}) \text{ s.t. } \chi(\boldsymbol{x}) \neq \chi(\boldsymbol{y})\}$ are located where $\mathcal{N}(\boldsymbol{x})$ is the 18-connected neighborhood around $\boldsymbol{x}$. For each location $\boldsymbol{x}$, we identify location $\boldsymbol{x}^* \epsilon\, \mathcal{N}(\boldsymbol{x})$, which is the point on the closest neighboring object. The average intensity in image $I: \Omega \to \boldsymbol{\mathcal{R}}$ for each label pair (a,b) is $g(a,b) = \frac{1}{|\mathcal{H}|}\sum_{\boldsymbol{x}\epsilon\mathcal{H}} I(\boldsymbol{x})$ where $\mathcal{H}$= $\{\boldsymbol{x} | a = \chi(\boldsymbol{x}) \text{ and } b = \chi(\boldsymbol{x}^*)\}$. Given a label image, distance field, and model $\mathcal{M}_I = \{f, g\}$, the intensity of a pixel is,

$$\hat{I}(\boldsymbol{x}) = \begin{cases} g\big(\chi(\boldsymbol{x}), \chi(\boldsymbol{x}^*)\big), \boldsymbol{x} \in \Lambda \\ f\big(\chi(\boldsymbol{x}), \psi(\boldsymbol{x})\big)\ \boldsymbol{x} \notin \Lambda \end{cases}. \quad \textbf{(1)}$$

This model assumes objects enclose homogeneous regions of intensity with the possibility that intensities near the boundary are a mixture from at most two objects. The model is capable of recovering all the intensity information in the image as $L \to |\Omega|$, but we would like to keep L small so that the representation is compact.

Linear Transforms. Before parametric images (p-images) can be combined, we must ascribe a linear structure to the space of image models $\mathcal{M}_p = \big(\mathcal{M}_g, \mathcal{M}_I\big)$. Given two parametric images $A, B \in \mathcal{M}_p$ and scalars $\alpha, \beta \in \mathcal{R}$, the parametric image $C = \alpha A + \beta B$ is computed by independently transforming $\mathcal{M}_g$ and $\mathcal{M}_I$. Since point correspondences are maintained in the MUSCLE framework when re-sampling is disabled, mesh vertices for C are computed by $q_n^C = \alpha q_n^A + \beta q_n^B$ and vertex labels computed by $l_n^C = l_n^A$. After determining mesh vertices and vertex labels, the distance field and label mask are computed by rasterizing each triangle mesh to a level set φ_l and then combining level sets into $\psi_C(\boldsymbol{x}) = \min_l |\varphi_l(\boldsymbol{x})|$ and $\chi_C(\boldsymbol{x}) = \min_{\{l|\varphi_l(\boldsymbol{x})<0\}} l$. The computational bottleneck for p-images is in rasterizing triangle meshes to level sets. Rasterization is implemented on the CPU in this work, but there are GPU accelerated methods for performing this task [10]. The final geometric model is $\mathcal{M}_g^C = \big\{\psi_C, \chi_C, \{q_1^C, \cdots, q_N^C\}, \{l_1^C, \cdots, l_N^C\}\big\}$. It is straightforward to combine intensity information since g: $\mathcal{L} \times \mathcal{L} \to \mathcal{R}$ and $f: \mathcal{L} \times \mathcal{R} \to \mathcal{R}$. The final intensity model is $\mathcal{M}_I^C = \{\alpha f_A + \beta f_B, \alpha g_A + \beta g_B\}$. Equipped with a linear structure on the space of parametric image models, we can form linear combinations of p-images; and because intensity varies as a function of distance to a region boundary, it is simple to shift intensity profiles along with edge locations as in Fig. 1 and 2. $\mathfrak{L}_p$ will be used in subsequent sections to analyze shape and intensity information with PCA.

2.2 Parametric Image Construction

There are several ways to construct parametric images. To construct parametric images (p-images) of the brain, we process MR images with Topology-preserving, Anatomy Driven Segmentation (TOADS) to produce a hard classification image consisting of 10 tissue classes [11].The hard classification is then converted to a "distance field + label mask" representation, and MOGAC is used to smooth object boundaries with mean-curvature flow [12]. It is straightforward to compute $\mathcal{M}_I$ from the MR image once the distance field and label mask are aligned with mesh boundaries (see the Intensity Model sub-section in 2.1). As shown in Fig. 3a-c, the final model captures most of the shape and intensity information present in the MR image. Another approach is to use an existing geometric phantom (e.g. mesh, level set, or region mask) with a corresponding synthesized medical image (like the XCAT [13]) and re-express it as a p-image (Fig. 6a-b).

2.3 Parametric Image Warping

As a prerequisite for atlas construction, the following describes how to warp parametric images using displacement fields produced by existing registration algorithms. Although it is likely beneficial to incorporate p-images into the registration process, we do not want to restrict p-image usage to any particular registration method. Instead, affine transformations and displacement fields are applied to p-images with techniques developed in the MUSCLE framework [7]. Given an affine transformation $\boldsymbol{A}$ produced by a global registration algorithm, such as FLIRT [14], the transformation is first applied to the triangle mesh, label mask, and distance field. The image re-sampling process required to transform the label mask and distance field introduces distortion in the level set functions φ_l implied by the "label mask + distance field" representation. This distortion is corrected by evolving level sets to minimize the distance from their iso-surfaces to the triangle mesh. To do so, the image $\omega: \Omega \to \mathcal{R}$ is first computed, which measures the clamped unsigned distance to the triangle mesh:

$$\omega(\boldsymbol{x}) = \min\{2d_{max}, d_1(\boldsymbol{x}) \quad \dots \quad d_N(\boldsymbol{x})\} \tag{2}$$

where $d_k(\boldsymbol{x})$ is the distance from location $\boldsymbol{x}$ to triangle k, and $d_{max} = 0.5$ is the clamped distance. Level sets φ_l are then evolved to minimize the following objective function:

$$E = \sum_l \int \left(\tfrac{1}{2}\big(\omega(\boldsymbol{x})\big)^2 + \lambda|\nabla\varphi_l(\boldsymbol{x})|\right)\delta(\varphi_l(\boldsymbol{x}))d\boldsymbol{x} \tag{3}$$

where λ is a regularization weight that controls the model's smoothness. Instead of evolving $\varphi_l{}'s$ directly, the label mask and distance field are evolved with MOGAC [9]. Global registration is followed by deformable registration, which produces a displacement field $\vec{v}: \Omega \mapsto \mathcal{R}^3$ describing where locations in the source image map to in the target image. The geometry of a p-image is warped with the displacement field by incrementally advecting mesh vertices q_n from source to target with $q_n(t) = q_n(0) + t\vec{v}\big(q_n(0)\big)$ where $t \in [0,1]$. After each displacement step k s.t. $t = k\Delta t$ for $k = 0, 1, \dots, \lceil 1/\Delta t \rceil$, the label mask and distance field are evolved to track the moving

mesh. The step size is chosen to be $\Delta t \leq d_{max}/\max_n \|\vec{v}(q_n)\|$ so that the level sets stay within the capture range of the clamped distance field ω.

One alternative to level set tracking is to simply apply the displacement field to triangle mesh, rasterize the mesh to a collection of level sets, and compress the level sets into a label mask and distance field. The computational complexity of rasterizing L level sets to an $M \times M \times M$ image is $O(LM^3)$, whereas the step complexity for evolving L level sets with MOGAC is $O(M^2)$. Level set evolution is usually faster because the number of steps $K = \lceil 1/\Delta t \rceil \ll M$. However, rasterization is necessary when synthesizing a p-image from a Point Distribution Model (PDM) [15], which we will describe in the subsequent section.

2.4 Atlas Construction

Given a template image and collection of subject images, the atlas construction pipeline begins by constructing a p-image for the template with the methods presented in section 2.2. The template is affine registered to each subject [14] and then deformably registered with the Mutual Information based method from Rohde et al. [16]. The affine transform and displacement field are applied to the template p-image with the method described in section 2.3. A Point Distribution Model [15] is then constructed for the warped meshes. The mean triangle mesh is rasterized to a label mask and distance field to construct an average geometry model $\mathcal{M}_g$. Since the warped p-images have the same intensity model as the template, the average intensity model $\mathcal{M}_I$ is the intensity model for the template. Another option is to perform PCA on both shape and intensity models for each subject to construct something similar to an Active Appearance Model (AAM) [5, 6]. We will leave AAMs for future work.

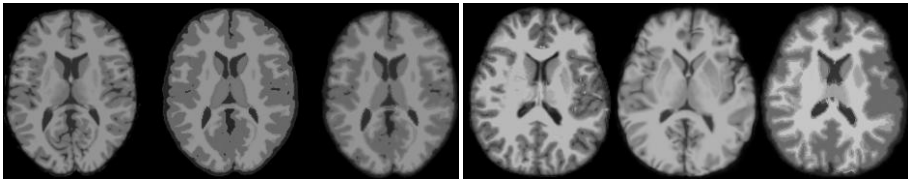
3 Results and Discussion

3.1 Human Brain Atlas

A human brain atlas was constructed from 19 normal subjects chosen at random from the OASIS MRI database [17]. The BrainWeb phantom [18] with no-noise or inhomogeneity was used for the template image. Note that the template is a simulation of a T1 weighted SPGR sequence and the subjects were imaged with a T1 weighted MPRAGE sequence. Therefore, their intensities follow slightly different statistical distributions. An advantage of p-images is that either intensity model (SPGR or MPRAGE) can be attached to the geometric model. Table 1 summarizes all registration results. MR image and p-image registration results look slightly different (Fig. 3d-f) because p-images and MR images use different displacement fields. P-image deformation is a Lagrangian technique that uses the forward source-to-target displacement field whereas MR image deformation is a semi-Lagrangian technique that uses the target-to-source displacement field. Displacement fields produced by target-to-source and source-to-target registration are usually not inverses of each other. Looking at MR registration to the template and p-image registration to the subject, both which use the source-to-target displacement field, Normalized Mutual Information (NMI) is the same in both cases (Table 1). Looking at MR registration to

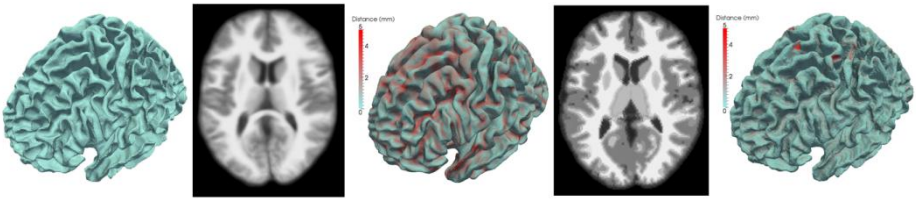
the subject and p-image registration to the template, both which use the target-to-source displacement field, the NMI is slightly higher for p-images.

Fig. 4 shows a mean intensity atlas using either MR Images or P-Images. Although both atlases are similar (Correlation Coefficient of 0.95), the p-image atlas is noticeably sharper. To quantitatively assess image quality, we ran the CRUISE cortical reconstruction pipeline [19] on the BrainWeb phantom, MR Image atlas, and p-image atlas. The cortical reconstruction produced by segmenting the p-image atlas is closer (mean surface distance of 0.37±0.35 mm) to the BrainWeb White Matter (WM)/Gray Matter (GM) surface than the reconstruction from the MR image atlas (surface distance of 0.88±0.71 mm). To synthesize images that deviate from the mean p-image, mode weights were specified to generate a mesh. After which, the mesh was rasterized to a label mask and distance field. Fig. 5 shows both the geometry and synthesized images for the first PCA mode with the mean intensity model. Synthesized images from the p-image atlas are consistently sharp and represent both shape and intensity information. These results suggest p-images could easily substitute PDMs in existing registration and segmentation algorithms.



(a) (b) (c) (d) (e) (f)

Fig. 3. (a) BrainWeb template image, (b) p-image with just distance intensity model $\mathcal{M}_I = \{f\}$, and (c) p-image with distance and edge intensity model $\mathcal{M}_I = \{f, g\}$. The Correlation Coefficients between (a)/(b) and (a)/(c) are 0.91 and 0.93 respectively. (d) target image, (e) registered MR Image, (f) registered p-image.

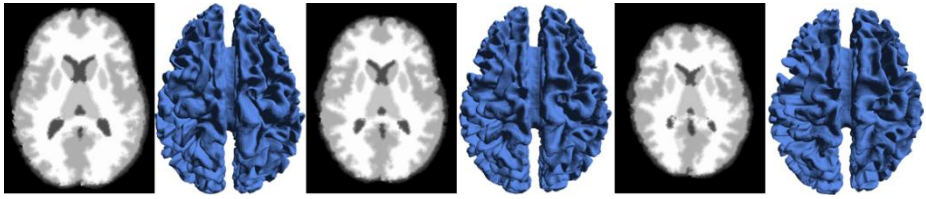


(a) (b) (c) (d) (e)

Fig. 4. (a) BrainWeb WM/GM surface, (b) mean of registered MR images, (c) WM/GM surface reconstructed from (b), (d) mean intensity p-image, (e) WM/GM surface reconstructed from (d). Mean error in (c) is 0.88±0.71 mm. Mean error in (e) is 0.37±0.35 mm. Correlation Coefficient between (b) and (d) is 0.95.

Table 1. Registration results for different image representations and registration targets. The table reports Normalized Mutual Information (NMI) between the registered image and target. Regions outside the brain were excluded in these measurements.

Representation	Template (SPGR)	Subject (MPRAGE)
MR Image	1.12±0.01	1.08±0.01
P-Image (SPGR)	1.10±0.02	1.12±0.01
P-Image (MPRAGE)	1.10±0.02	1.12±0.01

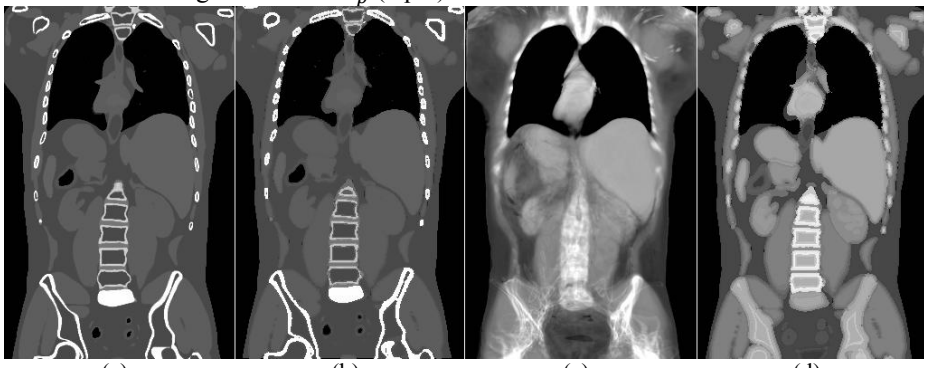


(a) (b) (c) (d) (e) (f)

Fig. 5. P-image atlas (SPGR intensities) and WM/GM surface for (a/b) mean minus the largest model, (c/d) mean, (e/f) mean plus the largest mode.

3.2 Human Body Atlas

A human body atlas was constructed from 20 whole-body CT images (12 adult males, 8 adult females) using the XCAT phantom [13] as a template. The XCAT is a whole-body NURB surface representation from which a CT image is simulated (Fig. 6a). We used an instance of the phantom for the 50 percentile male consisting of 50 structures. NURB surfaces for each structure were converted to triangle meshes and then rasterized to a label mask and distance field. After which, an image intensity model was constructed to form a p-image representation for the XCAT (Fig. 6b). Each subject was registered to the XCAT's CT image using the same procedure and methods for the human brain. Then, the p-image template was warped into the space of each subject. PCA analysis was conducted on the p-images. Fig 6. compares the mean intensity image achieved by warping CT images and warping p-images. One can surmise from Fig. 6c that the registration algorithm did not align bones well because these regions are blurry and bear little resemblance to bone geometry. However, averaging p-image intensity models (Fig. 6d) preserves bone geometry in these regions. Fig. 7. shows the mean and plus / minus the first (largest) shape mode. The MUSCLE representation captures variation in all 50 structures simultaneously and because the structures do not overlap or have air-gaps, it is straightforward to simulate a CT image based on $\mathcal{M}_p$ (eq. 1).



(a) (b) (c) (d)

Fig. 6. (a) XCAT CT image, (b) p-image constructed from XCAT CT image, (c) mean CT image after registration, (d) mean intensity p-image after registration. The correlation coefficient between (a)/(b) is 0.96 and between (c)/(d) is 0.97. Image size: $256 \times 256 \times 512$.

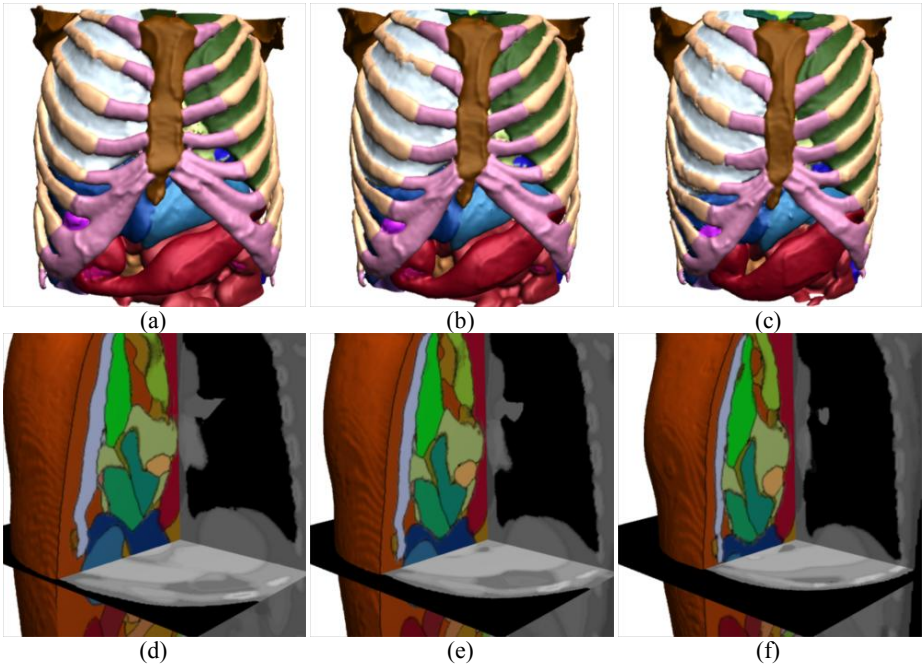


Fig. 7. P-image renderings showing the geometry (top) and synthesized CT images (bottom) for (a/d) mean minus one std. dev., (b/e) mean, (c/f) mean plus one std. dev. Computation time for a $256 \times 256 \times 256$ image and 2M triangles was 30 sec on a PC with dual Intel E5630s.

Conclusion

This work has presented a new data structure for representing and manipulating images that preserves edge strength. To demonstrate, we applied p-images to deformable registration and construction of an atlas for the human brain and human body. Parametric images assume all edges are step edges and structures do not contain texture, shading, or thin lines. These assumptions do not hold in general, but for MR images of the human brain (after inhomogeneity correction) and CT images of the human body, p-images are highly correlated with the real images (Correlation coefficient of 0.93 for the human brain and 0.96 for the human body).

One future direction would also be interesting to incorporate a p-image atlas into registration and segmentation pipelines to see if there is any improvement in performance. An interesting feature of p-images is that they can model aritculated structures because the MUSCLE representation allows surfaces to slide on one another. For this reason, a p-image atlas may be advantageous when segmenting regions of the human body that contain bone joints or other articualted structures.

Acknowledgments. This research was supported in part by a graduate student fellowship from Johns Hopkins University Applied Physics Laboratory (JHU-APL), internal funds from JHU, and the United States Army Natick Soldier Research Development and Engineering Center through a subcontract with JHU-APL. We

thank George Fung and Ben Tsui at the Johns Hopkins Outpatient Center (JHOC) for use of the XCAT phantom and John Carrino (department of radiology, JHOC) for assistance in obtaining CT data and Catherine Carneal and Andrew Merkle (JHU-APL Program Management) for managing Human body atlas research.

References

1. J. P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Medical image analysis*, vol. 2, pp. 243-260, 1998.
2. D. Shen and C. Davatzikos, "HAMMER: hierarchical attribute matching mechanism for elastic registration," *Medical Imaging, IEEE Trans. on*, vol. 21, pp. 1421-1439, 2002.
3. D. Rueckert, A. F. Frangi, and J. A. Schnabel, "Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration," *Medical Imaging, IEEE Transactions on*, vol. 22, pp. 1014-1025, 2003.
4. K. Pohl, J. Fisher, M. Shenton, R. McCarley, W. Grimson, R. Kikinis, and W. Wells, "Logarithm odds maps for shape representation," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pp. 955-963, 2006.
5. T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 681-685, 2001.
6. T. Cootes, C. Beeston, G. Edwards, and C. Taylor, "A unified framework for atlas matching using active appearance models," 1999, pp. 322-333.
7. B. C. Lucas, M. Kazhdan, and R. H. Taylor, "Multi-object Spring Level Sets (MUSCLE)," in *MICCAI*, Nice (to appear), 2012.
8. B. C. Lucas, M. Kazhdan, and R. H. Taylor, "SpringLS: A Deformable Model Representation to provide Interoperability between Meshes and Level Sets," in *MICCAI*, Toronto, 2011.
9. B. C. Lucas, M. Kazhdan, and R. H. Taylor, "Multi-Object Geodesic Active Contours (MOGAC)," in *MICCAI*, Nice, 2012.
10. H. Nguyen, *Gpu gems 3*: Addison-Wesley, 2007.
11. P. L. Bazin and D. L. Pham, "TOADS: topology-preserving, anatomy-driven segmentation," in *Biomedical Imaging: Nano to Macro, IEEE Int. Symposium on*, 2006, pp. 327-330.
12. J. Sethian, *Level set methods and fast marching methods*: Cambridge Univ Pr, 1999.
13. W. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. Tsui, "4D XCAT phantom for multimodality imaging research," *Medical physics*, vol. 37, p. 4902, 2010.
14. M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical image analysis*, vol. 5, pp. 143-156, 2001.
15. T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, pp. 38-59, 1995.
16. G. K. Rohde, A. Aldroubi, and B. M. Dawant, "The adaptive bases algorithm for intensity-based nonrigid image registration," *Medical Imaging, IEEE Trans. on*, vol. 22, pp. 1470-1479, 2003.
17. D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS)," *J of Cognitive Neuroscience*, vol. 19, pp. 1498-1507, 2007.
18. C. A. Cocosco, V. Kollokian, K. S. K. Remi, G. B. Pike, and A. C. Evans, "Brainweb: Online interface to a 3D MRI simulated brain database," *NeuroImage*, vol. 5, p. 425, 1997.
19. X. Han, D. L. Pham, D. Tosun, M. E. Rettmann, C. Xu, and J. L. Prince, "CRUISE: cortical reconstruction using implicit surface evolution," *Neuroimage*, vol. 23, pp. 997-1012, 2004.

Enhanced Atlas Selection for Multi-Atlas Segmentation with Application to Leg Muscle MRI

Jiahui Wang[1], Zheng Fan[2], Yael Shiloh-Malawsky[2], Joe Kornegay[2,3,4], Martin Styner[1,5]

[1]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA
[2]Department of Neurology, University of North Carolina, Chapel Hill, NC, USA
[3]Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC, USA
[4] Department of Veterinary Integrative Biosciences, Texas A&M University, TX, USA
[5]Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA

Abstract. Accurate segmentation of proximal leg muscles in magnet resonance imaging (MRI) studies of a golden retriever model of Duchenne muscular dystrophy (GRMD) is an important and difficult task for muscular dystrophy studies. In this study, we developed a multi-atlas based muscle segmentation method with a novel atlas selection scheme. For the atlas selection, we first pair-wise co-registered all atlas datasets and computed a directed graph with edge weights based on intensity and shape similarity between atlases. Following co-registration of all atlas datasets to the subject MR image, the set of closest/neighboring atlases was selected via clustering of the graph information. Finally, a weighted majority-voting label fusion was employed to compute multi-atlas segmentation. The mean Dice coefficient of 73.9% was obtained when performing leave-one-out cross validation in a longitudinal GRMD dataset. As leg muscles are of elongated, thin shape, this performance is within intra-rater performance range. The proposed method performed better than the conventional multi-atlas segmentation approaches without atlas selection and provides enhanced segmentation for individual leg muscle from MRI.

Keywords: multi-atlas, atlas selection, label fusion, segmentation, clustering, MRI, muscle, golden retriever muscular dystrophy, Duchenne muscular dystrophy

1 Introduction

Duchenne muscular dystrophy (DMD) is a fatal X-linked muscle disorder characterized by progressive degeneration of skeletal and cardiac muscles. Currently, no therapy halts or reverses progression of DMD. Although cellular and gene therapies are promising, key questions must first be addressed in relevant animal models. Because golden retriever muscular dystrophy (GRMD) dogs develop progressive and fatal disease strikingly similar to the human condition, this model has increasingly been used in preclinical trials [1-3]. Magnetic resonance imaging (MRI) has been used as

to provide data on disease progression in both natural history and treatment trials for GRMD [4-6]. However, to date, manual muscle segmentation in MRI is still the norm for most muscular dystrophy imaging studies [4,5,7]. Manual muscle segmentation is tedious, time-consuming, subject to rater errors, and impractical for large studies. Thus, an automated muscle segmentation method is highly demanded.

However, automated muscle segmentation in MRI is challenging, since the contrast between different muscles is low in the MR images. Furthermore, the muscle tissues affected by muscular dystrophy will generate many shape and intensity variations, which will cause difficulty in the muscle segmentation. Atlas-based segmentation method is a simple approach for automated segmentation that involves performing non-rigid registration between a labeled atlas image and a target, propagating the labels from the atlas to the target to generate a labeling for the target [8]. The accuracy of the resulting segmentation thus depends on the ability of the registration to find accurate and meaningful correspondence between the images, which is inherently related to the anatomical similarity of the two images. However, because of the inherent disease related variability in MRI appearance and muscle shape in muscular dystrophy muscles, it is difficult to segment muscles using a single muscle atlas. As an alternative, multi-atlas segmentation attempts to resolve this problem by using a number of different subjects as atlas images, performing multiple registrations from all the atlases to the target and fusing the results to generate the target segmentation [9,10]. However, in our experiments even multi-atlas based segmentation approaches have difficulty to obtain accurate segmentation results between atlases and targets with large shape and intensity variability. Because dissimilar atlases will more likely lead to poor segmentations, such atlases should be weighted less during the label fusion step. Therefore, an appropriate atlas selection technique should improve the final segmentation accuracy.

One example of atlas selection is the use of atlas-target registration accuracy estimators to weight the influence of a given atlas [11-14]. Similarly, methods that employ image similarity metrics, such as mutual information, to select atlases [15] are also examples of atlas selection, which presume that choosing those atlases whose registered images are similar to the target will result in more accurate segmentations. However these approaches could not handle the registration with large initial dissimilarity in shape between atlases and target. This can lead to inappropriately high weights in cases of initially large shape differences resulting in incorrect image correspondences established by the atlas registration..

Recently, several segmentation method using graph-based [16,17] or tree-based [18] intermediate templates guided registration methods have been demonstrated to be effective in the segmentation of brain MR images. The key concept of these methods is to decompose a large deformation into several small deformations that can be estimated with higher reliability. And then each atlas was warped through the intermediate templates one by one on the path towards the target. However, the major problem of the above strategy is that the quality of the warped atlas will be affected by the accumulated registration errors. Similar to these approaches, Langerak et al.[19] proposed a multi-atlas segmentation method with pre-registration atlas selection. The atlas set was clustered [20] and exemplars for each cluster were selected to generate a preliminary segmentation of the target using a majority voting label fusion.

The cluster with the highest similarity to the preliminary segmentation was selected to create the final segmentation of target. While this method is somewhat close to the proposed method here, it assumes that the difference between preliminary segmentation and true segmentation is minor, which, however, is not necessarily satisfied. Furthermore, this method ignores the appearance information in the target image and the atlas images. Finally, any sample bias in the multi-atlas population that could bias a subsequent segmentation is further aggregated by employing only the closest/best cluster. In contrast, our method proposes the use of all clusters with each one only contributing a single exemplar atlas, the one closest to the target image.

In this paper, we proposed a multi-atlas based muscle segmentation method with a graph based atlas selection. The muscles were manually segmented by expert raters using semi-automated segmentation scheme in a longitudinal GRMD MRI dataset and used as atlases for muscle segmentation. We then performed a pair-wise deformable image registration to align all atlases and target. A fully connected graph was constructed by calculating the distances based on intensity similarity and shape similarity between all pairs of registered images. We then clustered the graph by searching the shortest path between each atlas and target and selecting only those templates in each cluster that are closest to the target image. The selected templates were fused to create the final segmentation via a standard weighted majority voting label fusion.

2 Method

2.1 Materials

This study was approved by the Institutional Animal Care and Use Committee at the University of North Carolina at Chapel Hill (UNC-CH). The dataset included 5 GRMD and 6 normal dogs. The proximal legs of 3 GRMD dogs were longitudinally scanned at approximately 3, 6, and 9 months of age; the other GRMD and normal dogs were longitudinally scanned at approximately 3 and 6 months of age. All the dogs were produced through a GRMD colony maintained at UNC-CH (PI, Kornegay). Dogs were scanned on Siemens 3T Allegra Head-Only MRI scanner with standard CP head coil or Siemens 3T Tim Trio Whole-Body MRI scanner with 32-channel body coil at the UNC Biomedical Research Imaging Center. A T2-weighted image (T_2w) sequence was acquired using a turbo spin echo (TSE) sequence with the following parameters: repetition time (TR) 3000 ms / echo time (TE) 406-409 ms, 256 mm field of view (FOV), slice thickness ranged from 0.80 to 1.00 mm, matrix size of 256 x 256 pixels, and pixel size ranged from 0.60 to 1.00 mm.

2.2 Reference Muscle Segmentation

In this study, we focused on the segmentation of six proximal leg muscles of GRMD dogs: adductor magnus, biceps femoris, cranial sartorius, gracilis, rectus femoris, and semitendinosus. Because legs are bilaterally symmetrical, we conducted the segmen-

tation on left leg in this study; the muscle segmentation of the right leg can be obtained by mirroring the left leg atlas images to the right leg and applying the same multi-atlas segmentation approach. We conducted an interpolation-based semi-automated muscle segmentation to obtain reference segmentation of proximal leg muscles. These reference segmentations will be used as atlas images as well as used for assessing the segmentation results. For this method, expert raters first manually delineated the outline of each muscle in one out of five slices. The remaining slices were then automatically interpolated via a straightforward linear interpolation scheme computed independently for each muscle. The interpolated segmentations were thresholded at 50%. Figure 1 shows the reference segmentation of the six muscles of a GRMD dog in a transverse T_2w slice.

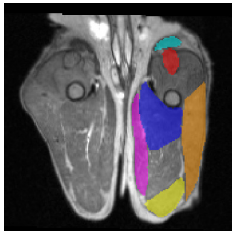


Fig. 1. Reference segmentation of adductor magnus (blue), biceps femoris (brown), cranial sartorius (light blue), gracilis (pink), rectus femoris (red), and semitendinosus (yellow) in a transverse view.

2.3 Image Registration

Deformable registration plays an indispensable role in the multi-atlas based segmentation approaches; researchers have proposed a variety of registration approaches with different degrees of freedom, such as HAMMER [21], statistical parametric mapping [22], free-form deformations [23], and Thirion's Demons [24]. However, all these approaches are operated in the space of vector fields and do not necessarily preserve topology of the target. Avants et al. [25] proposed a symmetric diffeomorphic image registration approach (as part of the ANTS registration package) that preserves anatomical topology even with large deformation. The transformation is differentiable and guaranteed to be smooth and one-to-one, i.e., for every element in moving image, there is a single corresponding element in the fixed image.

In this study we employed ANTS to register each atlas T_2w MR image to the target T_2w image using a cross-correlation similarity metric. The cross-correlation has been widely used and shown to perform well in many image registration applications [8,25], where one requires robustness to unpredictable image noise and intensity inhomogeneity. The transformation field obtained from the registration was then applied to the muscle segmentations with nearest neighbor interpolation. We also pairwise registered all the atlas MR image pairs using the same approach.

2.4 Construction of Graph

We represented the registered dataset as a graph (Fig. 2) whose vertices correspond to the atlases and target. Every edge between two vertices was assigned a cost (e_{ij}), which is defined by a weighted sum of an intensity similarity term MS_{ij} (mean squared voxel-wise intensity difference) and two shape similarity terms SS_{ij} and HE_{ij} (harmonic energy) [Eq. (1)].

$$e_{ij} = w_1 MS_{ij} + w_2 SS_{ij} + w_3 HE_{ij} \quad (1)$$

where w_1, w_2, and w_3 represent the weighting factors for the intensity similarity term and shape similarity terms, respectively. We empirically determined a combination of $w_1 = 0.1$, $w_2 = 0.8$, $w_3 = 0.1$ for the weighting factors.

The mean squared intensity difference is defined by

$$MS_{ij} = \frac{1}{N}\sum_{m=1}^{N}(i_m - j_m)^2$$

where i_m is the intensity of m-th voxel of a MRI scan I; j_m is the intensity of m-th voxel of another MRI scan J, N is the number of voxels in a MRI scan.

The first shape similarity term is defined by one minus the absolute difference of circularities (2D) at the mid-point of the proximal leg in two MRI scans [Eq. (2)].

$$SS_{ij} = 1 - |C_i - C_j| \quad (2)$$

where C_i and C_j are circularities of image I and image J, respectively. The circularity is defined by

$$C_i = 4\pi * (S_i / P_i^2)$$

where S_i and P_i are the area and perimeter of proximal leg at mid-point in the MR image, respectively. The second shape similarity term is defined as the harmonic energy, which is the mean Frobenius norm of the Jacobian of the deformation field [16].

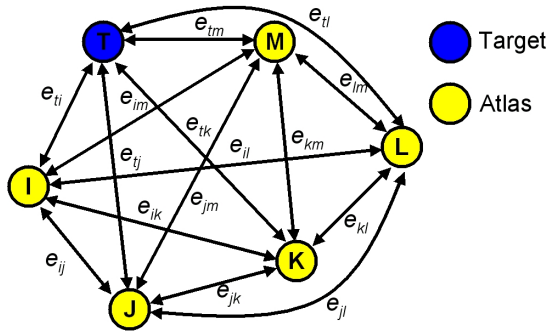


Fig. 2. Example of a graph with the target T and atlas I, J, K, L and M. The graph is constructed based on the similarity measurements between image pairs.

2.5 Clustering-based Template Selection

From the graph constructed in the previous section, we can choose templates that are

close to the target via an atlas clustering. On this graph, we clustered the atlas population into groups by searching the shortest path from each atlas to the target using the Floyd-Warshall algorithm. The atlases on the same shortest paths belong to the same cluster. We then selected the atlas that was closest to the target in each cluster as the neighboring template for the final muscle segmentation. An example of the clustering from a graph is illustrated in Fig. 3 to demonstrate the framework of the atlas selection. In this example, the atlases were partitioned into three clusters. Three neighboring temples were selected for creating the final segmentation of target.

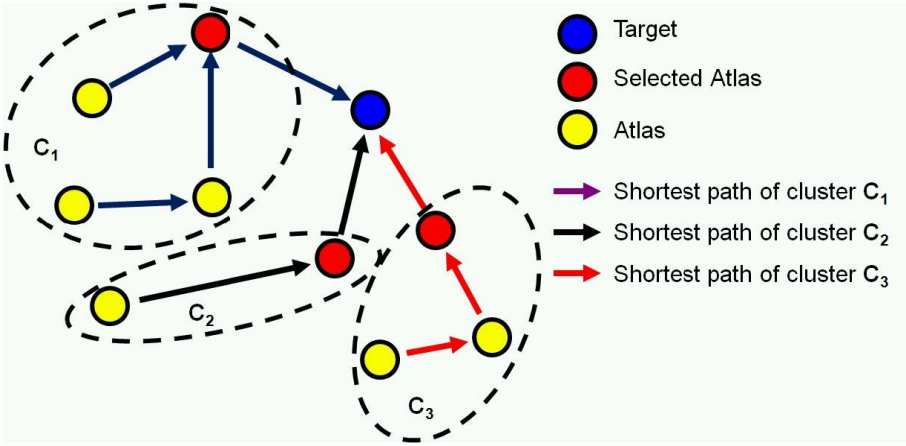


Fig. 3. Clustering-based atlas selection framework

2.6 Weighted Majority Voting Label Fusion

Majority voting is the most widely used label fusion algorithm for multi-atlas based segmentation approaches. This algorithm weights each candidate segmentation equally and assigns to each voxel the label that most segmentations agree on [10]. However, the assigned label by this simple majority rule dose not necessarily imply a correct segmentation in applications with large variation in size, shape, and appearance, such as for DMD muscles. This issue can be solved by a weighted majority voting approach, i.e., assigning larger weights to the atlases more similar to the target image. For each selected neighboring template, we used one minus the cost between a neighboring template and target on the graph as its weight. And then the final segmentation for each muscle is determined by collecting weighted votes from all the segmentations of selected templates and assigning to each voxel the label that has the highest vote.

2.7 Segmentation Performance Assessment

We assessed the performance of our proposed segmentation method by evaluating how close the resulting segmentation is to the corresponding reference segmentation.

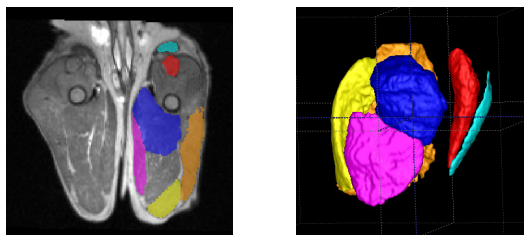
The most commonly used metrics is the overlap between the segmentations. In this study, we used the Dice similarity coefficient (DSC), also referred to as the mean overlap or the similarity index, which is computed between two segmentations as:

$$DSC = 2 \times \frac{V_{auto} \cap V_{ref}}{V_{auto} + V_{ref}} \times 100\%$$

where V_{auto} and V_{ref} are the volume of automated segmentation result and the volume of reference segmentation, respectively. A DSC of 1 indicates complete volumetric overlap, and 0 indicates no overlap at all.

3 Results

To validate our method, we applied the proposed method in a leave-one-out experiment for all 25 MRI scans in our dataset, resulting in 24 atlases to be used for atlas selection and 17 neighboring templates were selected in average to segment each target image. Figure 4 shows that the proposed method achieved satisfactory segmentation results of a GRMD dog in 2D and 3D views. We also quantitatively evaluated the proposed segmentation method using DSC and compared the performance levels to other segmentation methods, i.e., standard majority voting and standard weighted majority voting using all the atlases. For the standard weighted majority voting, we used all the available atlases in our dataset without atlas selection for label fusion. Table 1 shows the mean values, standard deviations, and range of the DSC for majority voting, standard weighted majority voting and our proposed method, respectively. Overall, the average DSCs of majority voting, weighted majority voting and our proposed method were 69.9%, 72.5%, and 73.9%, respectively. By use of paired t-test, we found that overall the segmentation accuracy of proposed method is significantly better than the standard majority voting ($p < 0.0001$) and standard weighted majority voting ($p < 0.0001$).



(A) (B)

Fig. 4. Multi-atlas based segmentation result of proximal leg muscles of a GRMD dog in (A) 2D view and (B) 3D view.

Furthermore, the segmentation accuracy of the proposed method is significantly better than the majority voting for each individual muscle: adductor magnus ($p < 0.0001$), biceps femoris ($p < 0.0001$), cranial sartorius ($p < 0.0001$), gracilis ($p < 0.0001$), rectus femoris ($p < 0.0001$), and semitendinosus ($p < 0.0001$).

The segmentation accuracy of the proposed method is also significantly better than the weighted majority voting for most of the muscles: biceps femoris ($p < 0.0001$), cranial sartorius ($p < 0.0001$), gracilis ($p < 0.0001$), rectus femoris ($p = 0.004$), and semitendinosus ($p = 0.003$). The only muscle not to show significant differences was the adductor magnus ($p = 0.65$).

Table 1. The mean, standard deviation (SD), and range [minimum (Min) - maximum (Max)] of the Dice similarity coefficient for the segmentation of GRMD proximal leg muscles in MRI by standard majority voting, standard weighted majority voting without atlas selection, and our proposed multi-atlas based segmentation method.

	Majority Voting			Weighted Majority Voting			Weighted Majority Voting with atlas selection		
	Mean	SD	Range (Min - Max)	Mean	SD	Range (Min - Max)	Mean	SD	Range (Min - Max)
AD	83.8%	4.0%	73.9% - 90.2%	84.7%	3.8%	77.3% - 90.9%	84.8%	4.2%	75.6% - 91.4%
BF	79.2%	10.9%	50.0% - 90.0%	80.7%	9.5%	55.8% - 89.5%	82.1%	8.5%	60.7% - 91.2%
CS	48.3%	18.8%	2.6% - 79.8%	53.0%	16.0%	18.4% - 80.5%	56.0%	14.5%	29.3% - 79.5%
GR	71.0%	12.7%	39.1% - 88.0%	73.5%	13.4%	31.5% - 88.0%	75.2%	12.6%	36.0% - 87.6%
RF	61.3%	19.1%	0% - 87.6%	65.0%	18.1%	2.6% - 88.0%	66.1%	17.1%	7.9% - 87.8%
ST	76.1%	13.3%	37.6% - 91.1%	77.8%	12.1%	40.2% - 91.7%	79.2%	11.1%	43.2% - 92.1%

4 Discussion

Because of the large shape and intensity variations of the muscle tissues caused by muscular dystrophy in GRMD MRI scans we need a database whose size is large enough to represent the variations of the data. However, due to the difficulty of collecting either human DMD or DMD animal models, the number of available images in DMD studies generally is low. Although the current amount of data (25 MRI scans) in our experiment is not large, the dataset we used has been shown to provide improved segmentation results compared to conventional multi-atlas segmentation schemes without atlas selection.

Different from brain applications, the proximal leg muscles of GRMD dogs show large shape variations on MRI images. The variations were caused by the disease progression of muscular dystrophy or the location and pose of legs in the MR coil. In order to overcome this issue, the incorporation of shape similarity between images was needed for optimal segmentation. Our shape similarity measurement is based on a premise that the muscles should show similar circularity at the mid-point of the leg as well as should show reduced harmonic energy.

We selected the neighboring templates via an atlas clustering technique. As shown in Fig. 3, there would be some overlap between different paths between atlases and target image. Because of this overlap the number of selected neighboring templates would be vary for different targets. However, because the similarity to the target of atlases along a shortest path is incremental and we chose the atlas closest to the

target on the path, it ensures that we always selected the atlases from the population with high similarity to the target for label fusion.

5 Conclusions

In conclusion, we have introduced a multi-atlas based GRMD proximal leg muscle segmentation method. A clustering technique was used to select neighboring templates that are close to the target on constructed graphs and determine weights of the selected templates for the label fusion procedure. We validated this method on a longitudinal GRMD MRI dataset. The results had shown that the proposed method improved the overall accuracy of muscle segmentation in GRMD MRI and could provide enhanced segmentation for individual leg muscle. This method also provides the field of muscle MRI and DMD with an automated muscle segmentation tool for efficient muscle MR analyses.

Acknowledgments

This work was supported by the National Institutes of Health Grant Nos. R42 NS059095-03 (NINDS) (Styner), and 1U24NS059696-01A1 (NINDS) (Kornegay), the Muscular Dystrophy Association (Kornegay), North Carolina Translational and Clinical Sciences Institute 50K Grant No. 50KR71104 (Fan) and UNC Intellectual and Developmental Disabilities Research Center (P30-HD003110-41). The authors thank Weili Lin, and Kathleen Wilber for their help in acquisition of MRI scans and their helpful discussions and Janet and Dan Bogan and Jennifer Dow for technical assistance in managing the dogs.

References

1. Coulton, G.R., Morgan, J.E., Partridge, T.A., Sloper, J.C.: The mdx Mouse Skeletal Muscle Myopathy: I. A Histological, Morphometric and Biochemical Investigation. Neuropathol. Appl. Neurobiol. 14, 53-70, (1988)
2. Kornegay J.N., Tuler S.M., Miller D.M., Levesque D.C.: Muscular Dystrophy in a Litter of Golden Retriever Dogs. Muscle Nerve. 11, 1056-1064, (1988)
3. Shelton G.D., Engvall E.: Canine and Feline Models of Human Inherited Muscle Diseases. Neuromuscular Disord. 15, 127-138, (2005)
4. Thibaud, J.L., Monnet, A., Bertoldi, D., Barthelemy, I., Blot, S., Carlier, P.G.: Characterization of Dystrophic Muscle in Golden Retriever Muscular Dystrophy Dogs by Nuclear Magnetic Resonance Imaging. Neuromuscul. Disord.,17, 575-584, (2007)
5. Kobayashi, M., Nakamura, A., Hasegawa, D., Fujita, M., Orima, H., Takeda, S.: Evaluation of Dystrophic Dog Pathology by Fat-suppressed T2-weighted Imaging. Muscle Nerve., 40, 815-826, (2009)
6. Yokota, T., Lu, Q.L., Partridge, T, Kobayashi, M., Nakamura, A., Takeda, S., Hoffman, E.: Efficacy of Systemic Morpholino Exon-skipping in Duchenne Dystrophy Dogs. Ann. Neurol., 65, 667-676, (2009)

7. Mathur, S., Lott, D.J., Senesac, C.: Age-related Differences in Lower-limb Muscle Cross-Sectional Area and Torque Production in Boys with Duchenne Muscular Dystrophy. Arch. Phys. Med. Rehabil., 1051-1058, (2010)
8. Bajcsy, R., Lieberson, R., Reivich, M.: A Computerized System for the Elastic Matching of Deformed Radiographic Images to Idealized Atlas Images. J. Comput. Assist. Tomogr., 5, 618-625, (1983)
9. Rohlfing, T., Brandt, R., Menzel, R., Maurer, C.R.: Evaluation of Atlas Selection Strategies for Atlas-based Image Segmentation with Application to Confocal Microscopy Images of Bee Brains. NeuroImage, 21, 1428–1442, (2004)
10. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic Anatomical Brain MRI Segmentation Combining Label Propagation and Decision Fusion. NeuroImage, 33, 115-126, (2006)
11. Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M.A., van Ginneken, B.: Multi-atlas-based Segmentation with Local Decision Fusion - Application to Cardiac and Aortic Segmentation in CT Scans. IEEE Trans. Med. Imaging, 28,1000-1009, (2009)
12. Artaechevarria, X., Munoz-Barrutia, A., de Solorzano, C.O.: Combination Strategies in Multi-atlas Image Segmentation: Application to Brain MR Data. IEEE Trans. Med. Imaging, 28, 1266-1277, (2009)
13. van Rikxoort, E.M., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M.A., Pluim, J. P.W., van Ginneken, B.: Adaptive Local Multi-atlas Segmentation: Application to the Heart and the Caudate Nucleus. Med. Image Anal. 14 (1), 39-49, (2010).
14. Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J.: Optimum Template Selection for Atlas-based Segmentation. Neuroimage, 34, 1612-1618, (2007)
15. Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D.: Multi-atlas Based Segmentation of Brain Images: Atlas Selection and Its Effect on Accuracy. Neuroimage, 46, 726-738, (2009)
16. Hamm, J., Ye, D.H., Verma, R., Davatzikos, C.: GRAM: A Framework for Geodesic Registration on Anatomical Manifolds. Med. Image Anal., 14, 633-642, (2010)
17. Jia, H., Wu, G., Wang, Q., Wang Y., Kim M., Shen, D.: Directed Graph Based Image Registration. Comput. Med. Imaging Graph., 36, 139-151, (2012)
18. Jia, H., Yap, P.T., Shen, D.: Iterative Multi-atlas-based Multi-image Segmentation with Tree-based Registration. Neuroimage., 59, 422-430, (2012)
19. Langerak T.R., Van der Heide U.A., Kotte A.N.T.J.,Berendsen F.F., Pluim J. P. W.: Multi-atlas-based segmentation with pre-registration atlas selection. MICCAI 2011 Workshop on Multi-Atlas Labeling and Statistical Fusion, (2011)
20. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science, Vol. 315, pp. 972--976, (2007).
21. Shen, D., Davatzikos, C.: Hammer: Hierarchical attribute Matching Mechanism for Elastic Registration. IEEE Trans. Med. Imaging, 21, 1421 - 1439 (2002)
22. Ashburner, J., Friston, K.: Voxel-based morphometry-the methods. Neuroimage, 11, 805 - 821, (2000)
23. Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D.,. Nonrigid Registration Using Free-form Deformations: Application to Breast MR Images. IEEE Trans. Med. Imaging 18 (8), 712 – 721, (1999)
24. Thirion, J.P.: Image Matching as a Diffusion Process: an Analogy with Maxwell's Demons. Med. Image Anal. 2 (3), 243-260, (1998)
25. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain. Med. Image Anal., 12, 26-41, (2008)