Automatic Caudate Segmentation by Hybrid Generative/Discriminative Models

Zhuowen Tu, Mubeena Mirza, Ivo Dinov, and Arthur W. Toga

Lab of Neuro Imaging, School of Medicine
University of California, Los Angeles, USA
zhuowen.tu@loni.ucla.edu

Abstract. Segmenting sub-cortical structures from 3D brain images is of significant practical importance. This paper presents an experimental study for caudate segmentation in MRI images using a fully automatic algorithm based on [9], which is an extension of the hybrid model approach described in [1]. The method in [9] tackles multiple sub-cortical and cortical structures. In this study (the grand challenge competition), the focus is on segmenting the left and right caudate from MRI images. Given a set of training data, a classifier (PBT [9]) is used to learn a discriminative model for each voxel being on the left caudate, right caudate, or the background. In testing, a variational approach is used to perform segmentation by minimizing an energy of a hybrid model. This hybrid model is composed of a discriminative model term by the trained PBT classifier and a generative shape term. It takes, typically, 5 minutes to perform the segmentation on a modern PC.

1 Introduction

There has been considerable recent work on 3D segmentation in medical imaging and we cite several representative ones [11, 3, 8]. Existing approaches can be roughly categorized into two groups: (1) one which puts strong efforts into the shape priors [11, 8, 7]; (2) and the other which classifies pixels/voxels using various features [5, 6].

In [1, 9], a hybrid model approach for brain anatomical structure segmentation was presented. The systems adopted a probabilistic boosting tree (PBT) algorithm as multi-class classifier. It selects and combines hundreds of cues such as intensity, gradients, curvatures, and locations to model ambiguous appearance patterns.

2 Hybrid Discriminative/Generative Model

We give the basic problem formulation in this section. Brain imaging mostly deals with 3D images, which are referred to as volumes for the rest of the paper. For an input volume, $\mathbf{V}$, the task of brain segmentation is to obtain the full partition of each anatomical structure of interest. A solution W can be denoted

T. Heimann, M. Styner, B. van Ginneken (Eds.): 3D Segmentation in The Clinic:
A Grand Challenge, pp. 73-80, 2007.

as $W = \{R_k, k = 0, .., T\}$, where R_0 is the background region, and $R_k, k = 1, .., T$ denote the anatomical structures. We have $\bigcup_{k=0}^{T} R_k = \Lambda$ where Λ defines the 3D lattice of the input $\mathbf{V}$. $R_i \cap R_j = \emptyset, \forall i \neq j$. Let the optimal solution W^* to be the one which minimizes an energy

$$E(W, \mathbf{V}) = E_{AP}(W, \mathbf{V}) + \alpha_1 E_{SM}(W). \tag{1}$$

The first term, $E_{AP}(W, \mathbf{V})$, corresponds to a discriminative model

$$E_{AP}(W, \mathbf{V}) = -\sum_{k=0}^{T} \sum_{s \in R_k} \log p(y = k | \mathbf{V}(N(s))), \tag{2}$$

where $N(s)$ includes all the voxels in the sub-volume, and $y \in \{0, ..., T\}$ is the label/class for each voxel. $p(y = k|\mathbf{V}(N(s))$ essentially computes the classification probability of voxel s belonging to structure k.

$E_{SM}(W)$ represents the generative models about the shape prior $p(R)$ and it encourages the region boundaries to be smooth.

$$E_{SM} = \sum_{k=0}^{T} \int_{\partial R_k} dA. \tag{3}$$

α_1 is the weight balancing between how much we rely on the discriminative model and the local smoothness. α_1 is learned automatically and the details will be given in a longer version due to the space limit. Note that there are two other terms used previously, global shape priors and an edge model [9]. In this study, we do not include them as they are less important than $E_{AP}(W, \mathbf{V})$ and $E_{SM}(W)$.

3 Learning Discriminative and Generative Models

Learning $E_{AP}(W, \mathbf{V})$

To compute E_{AP}, our task is to learn and compute the discriminative model $p(y = k|\mathbf{V}(N(s)))$. Each input sample is a sub-volume and the output is the probability of the center voxel s being on region $R_k, k = 0..T$. This is not an easy task due to the complex appearance patterns of $\mathbf{V}(N(s))$.

We adopt a PBT.M2 algorithm [9] to learn a multi-class classifier. Given a set of training volumes where the left and right caudate are manually labeled as 1 and 2 respectively, we collect all the training samples and put them into 3 classes. Each sample in class 0 is a sub-volume (size $11 \times 11 \times 11$) in which the center voxel is labeled as the background. Likewise, each sample in class 1 or 2 is a sub-volume in which the center voxel is labeled as the left or right caudate respectively. There are millions of class 0 samples as majority of the volumes are the background and thousands of class 1 and class 2 samples. Given any sample (a sub-volume), we compute around 5,000 candidate features for selection, which include intensity, gradients, locations, curvatures and Haar at

various scales. PBT learns and computes an overall multi-class discriminative probability,

$$p(y|\mathbf{V}) = \sum_{l_1} \tilde{p}(y|l_1, \mathbf{V}) q(l_1|\mathbf{V}) = \sum_{l_1,..,l_n} \tilde{p}(y|l_n, ..., l_1), ..., q(l_2|l_1, \mathbf{V}) q(l_1|\mathbf{V}), \quad (4)$$

where l_i represents the ith layer in the tree, and $q(l_i)$ computes the discriminative probability by each boosting node in the tree. Fig. (1) gives a brief description of the learning procedure.

Given: Labeled training examples $S = \{(\mathbf{V}_a, y_a), a = 1..n\}$ with each $y_a \in \{0..T\}$

- Collect all the training samples and compute their candidate features.
- Train a boosting 3-class classifier to select and fuse a subset of the features.
- Split the training data using the trained classifier to expand the tree.
- Train the tree recursively until a certain stop criterion is met.

Fig. 1. Pseudo code for learning the PBT classifier.

The key of the PBT classifier is that it is capable of hierarchically fusing a set of informative features automatically selected from a large pool of candidate features. These features carry both intensity and local geometric properties of each voxel in a sub-volume.

Learning $E_{edg}(W, \mathbf{V})$

Another energy term, $E_{SM} = \sum_{k=0}^{T} \int_{\partial R_k} dA$, is added to encourage smooth surfaces. $\int_{\partial R_k} dA$ is the area of the surface of region R_k. When the total energy is being minimized in a variational approach, this term corresponds to the force that encourages each boundary point to have small mean curvature, resulting in smooth surfaces.

4 Segmenting 3D Brain Volumes

The goal of the segmentation stage is to find the optimal segmentation/solution which minimizes the energy in eqn. (1). In our problem, we apply a variational approach to perform boundary evolution to minimize eqn. (1). The details of the energy minimization algorithm can be found in a longer version. Basically, the boundary evolution is driven by two forces

$$\frac{dE_{AP}}{d\mathbf{M}} = -[\log \frac{p(y = i|\mathbf{V}(N(\mathbf{M}))}{p(y = j|\mathbf{V}(N(\mathbf{M}))}]\mathbf{N}, \quad (5)$$

and

$$\frac{dE_{SM}}{d\mathbf{M}} = -\alpha_1 H \mathbf{N}, \quad (6)$$

where H is the curvature at the surface point $\mathbf{M}$. Details can be found in [1].

5 Results

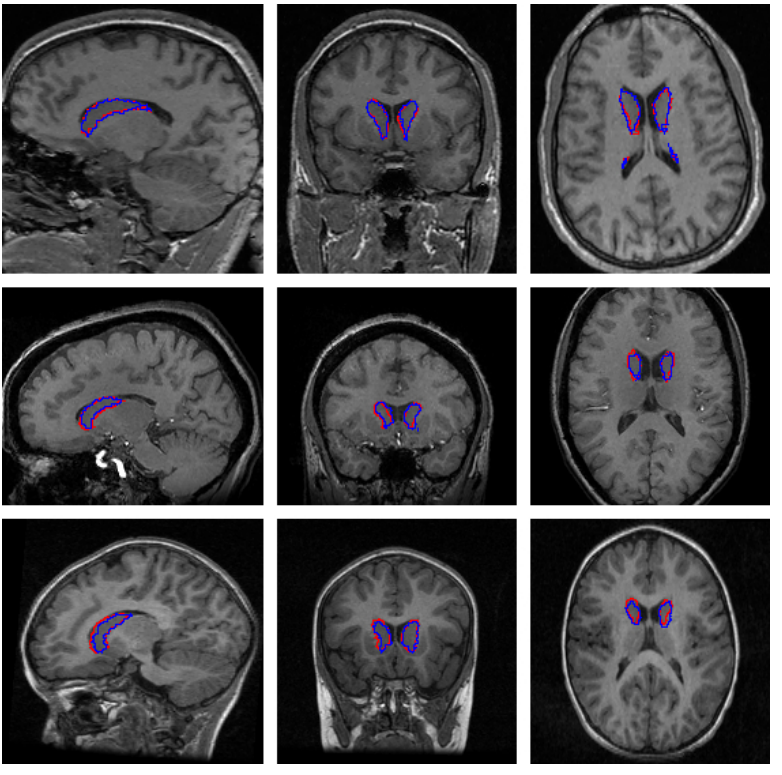

Fig. 2. From left to right, a sagittal, coronal and transversal slice from a a subject in the adults BWH group (top), one in the elderly UNC group (middle) and one in the pediatric UNC group (bottom). The outline of the reference standard segmentation is in red, the outline of the segmentation of the method described in this paper is in blue.

There are 4 sets of data provided in this grand challenge competition, 2 for training and 2 for testing. As described in the documents from the organizers: "All MRI images are scanned with an Inversion Recovery Prepped Spoiled Grass sequence on a variety of scanners (GE, Siemens, Philips, mostly 1.5 Tesla). Some datasets have been acquired in axial direction, whereas others in coronal direction. All datasets have been re-oriented to axial RAI-orientation, but have not been aligned in any fashion. All data is stored in Meta format containing an ASCII readable header and a separate raw image data file."

The 2 training sets are: (1) MRIs and structural segmentations from the internet brain segmentation repository (IBSR) at Mass General Hospital, Boston. (2) MRIs and caudate segmentations from the Psychiatry Neuroimaging Laboratory at the Brigham and Women's Hospital Boston (BWH). The 2 testing sets are: (1) MRIs from different disease study at the UNC Neuro Image Analysis

Correl	UNC Ped	UNC Eld	BWH PNL	Total
Left	-0.2990	0.2130	0.5403	0.1514
Right	0.5115	0.4429	0.2319	0.3955
Average	0.1063	0.3280	0.3861	0.2734

Table 1. Pearson correlation for the volume measurements in the three testing groups as well as in total. This coefficient captures how well the volumetric measurements correlate with those of the reference segmentations.

Test/Re-Test	UNC 03 [mm^3]	UNC 04 [mm^3]	UNC 09 [mm^3]	UNC 11 [mm^3]	UNC 17 [mm^3]	UNC 18 [mm^3]	UNC 21 [mm^3]	UNC 22 [mm^3]	UNC 24 [mm^3]	UNC 25 [mm^3]	Mean [mm^3]	Stdev [mm^3]	COV [%]
Left	3008	2896	2965	2862	2849	2917	2844	2836	3078	3041	2930	88	3.0
Right	2908	2980	2948	2806	2825	2971	2845	2889	3125	3059	2936	103	3.5
Total	-										-	-	3.3

Table 2. The volumetric measurements of the 10 data sets of the same young adult acquired on 5 different scanners within 60 days. The coefficient of variation (COV = standard deviation / average, last column) indicates the stability of the algorithm in a test/re-test situation including scanner variability.

Laboratory, Chapel Hill. (2) 14 MRIs from the Psychiatry Neuroimaging Laboratory at the Brigham and Womens Hospital, Boston. This data is from the same study as the BWH datasets in the training set.

Since the training sets BWH and IBSR are given as different forms, we trained two systems using BWH and IBSR data separately. One system was trained and tested using the BWH training and testing data respectively. Another system was trained using the IBSR data and tested on the UNC volumes. However, the training and testing procedures for the two systems are identical. It takes, typically, 5 minutes to perform the segmentation on a modern PC.

The Outline of The Algorithm

Training: (1) All the volumes were skull stripped using the BSE algorithm [2]. (2) We then used the first volume as the template to which the rest volumes are registered using the AIR program [10]. (3) Intensities were also normalized linearly by matching the peak in the histogram. (4) we trained a PBT classifier to compute $p(y|\mathbf{V}(N(s)))$ based on the MRI volumes from step (3). (5) Learn α_1 to combine the discriminative and generative models.

Testing: (1) Given a testing MRI volume, skull stripping was performed using the BSE algorithm [2]. (2) Register the data to the template using AIR [10]. (3) Perform intensity normalization. (4) Based on a normalized volume $\mathbf{V}$ from step (3), we computed $p(y|\mathbf{V}(N(s)))$ for each voxel and obtain a soft classification map. (5) Assign each voxel with the label of the highest probability in the soft classification map to obtain an initial segmentation. (6) Perform boundary evolution in minimizing the total energy E shown in eqn. (1).

Fig. (5) gives some slice views of the segmentation results with the ground truth. The segmentation results on the BWH and UNC testing volumes were sent to the workshop organizer to perform quantitative evaluation. Fig. (3,5,5) give detailed numeric measures for the 2 testing datasets.

All Dataset	Overlap Err		Volume Diff.		Abs. Dist.		RMS Dist.		Max. Dist.		Total
	[%]	Score	[%]	Score	[mm]	Score	[mm]	Score	[mm]	Score	Score
UNC Ped 10	44.9	72	-16.4	72	1.2	57	2.0	65	13.3	61	66
UNC Ped 14	46.1	71	-17.8	68	1.0	62	1.5	74	9.5	72	70
UNC Ped 15	37.7	76	-11.2	72	0.8	70	1.2	78	6.8	80	76
UNC Ped 19	47.6	70	-39.4	31	1.1	60	1.6	72	6.7	80	62
UNC Ped 30	44.1	72	-23.6	58	1.0	62	1.5	73	7.6	78	68
UNC Eld 01	67.4	58	-46.9	18	2.0	26	2.9	49	12.2	64	43
UNC Eld 12	55.3	65	-39.1	32	1.4	49	2.1	62	9.4	72	56
UNC Eld 13	34.2	78	-9.2	84	0.6	76	1.0	83	4.6	87	82
UNC Eld 20	41.5	74	-20.0	65	0.9	68	1.3	76	8.8	74	72
UNC Eld 26	37.2	76	-15.7	72	0.6	76	1.0	82	5.7	84	78
BWH PNL 16	34.3	78	-19.1	66	1.0	64	2.9	48	26.4	22	56
BWH PNL 17	31.7	80	-4.5	89	0.8	70	2.4	57	26.0	24	64
BWH PNL 18	40.9	74	-33.2	42	1.2	55	2.3	58	13.5	60	58
BWH PNL 19	43.8	72	-33.9	41	2.0	27	5.2	10	35.5	0	30
BWH PNL 20	35.6	78	-2.4	90	1.0	64	2.9	49	30.6	14	59
BWH PNL 21	52.5	67	-46.2	26	3.4	20	7.4	8	41.1	0	24
BWH PNL 22	33.8	78	-20.2	64	1.0	64	2.7	52	24.1	30	58
BWH PNL 23	28.0	82	7.3	88	0.5	80	1.1	80	9.4	72	80
BWH PNL 24	32.9	79	-13.8	76	0.6	78	1.0	82	6.8	80	79
BWH PNL 25	64.8	59	-57.2	5	3.9	0	7.4	0	39.5	0	12
BWH PNL 26	43.1	73	-30.3	46	1.6	42	3.8	32	25.3	26	44
BWH PNL 27	24.5	85	-3.6	94	0.7	75	2.1	62	20.5	40	71
BWH PNL 28	49.9	68	-42.8	24	2.5	14	6.3	1	35.4	0	22
BWH PNL 29	53.9	66	-28.3	50	1.9	31	3.3	41	18.3	46	47
Average All	42.7	73	-23.6	57	1.4	54	2.8	54	18.2	49	57
Average UNC Ped	44.1	72	-21.7	60	1.0	63	1.5	72	8.8	74	68
Average UNC Eld	47.1	70	-26.2	54	1.1	59	1.6	70	8.1	76	66
Average BWH PNL	40.7	74	-23.5	57	1.6	49	3.6	41	25.2	30	50

Table 3. Results of the comparison metrics and corresponding scores for all test cases averaged for the left and right segmentation. The summary rows at the end of the table display the overall average across all test cases, as well as grouped for the three testing groups.

The evaluation process was designed by the organizers of the MICCAI'07 "Segmentation Challenge" (http://mbi.dkfz-heidelberg.de/grand-challenge2007). The segmentation results are evaluated by assigning a score to each test case on a variety of metrics [4]. The minimum and maximum score for each metric are normalized to 0 and 100. The total score of a metric is obtained by averaging the individual scores of all test cases.

The score of each test case itself is the average of five metric, each also scaled from 0 to 100. These five metrics include: (1) Volumetric overlap. (2) Relative absolute volume difference, in percent. (3) Average symmetric absolute surface distance, in millimeters. (4) Symmetric RMS surface distance, in millimeters. (5) Maximum symmetric absolute surface distance, in millimeters. The detailed descriptions of how the measures are obtained can be found in the summary paper of this workshop.

The caudate segmentation challenge consists of multiple groups of subjects. One group of test cases are scans of the same subject, performed on different scanners. For these cases, no reference segmentations are available, and these cases will therefore not contribute to the total score. These cases are included to test if a method is reproducible. These results are reported separately in Fig. (5). The COV value for human experts is 3.1% and our result is 3.3%. This suggests that the output of our algorithm is as consistent as human experts.

6 Discussions

The proposed system is fully automatic, very general, and it can be used for segmenting a variety of structures in MRI. There is nearly no parameter to tune in the training stage and it usually takes one day on a modern PC to perform training. Once trained, the algorithm can quickly segment an MRI volume (5 minutes in this case and 20 minutes in [9] where 25 structures are segmented).

Table (3) suggests that our algorithm tends to under-segment the left and right caudate. This is due to two-fold: (1) There are many more background voxels than the object voxels. When learning the PBT classifier, we use a precision/recall criterion resulting a conservative probability measure for the foreground structure. Using the same measure to train our system may be helpful. (2) The smoothness term encourages the left and right caudate to be compact. In [9], we also used two additional terms, an edge term and a global PCA shape prior, which seem to improve this under-segmentation problem.

There are several future directions which might further improve the algorithm: (1) Adopting a better shape model, e.g. the m-rep [7], might guide the segmentation process with more informative high-level information. (2) Investigating key/invariant features is also important. (3) Learning better algorithm to reduce the complexity of the PBT. (4) Improving the hybrid model with more explicit context information.

References

1. Z. Tu, C. Narr, P. Dollar, P. Thompson, and A. Toga, "Brain Anatomical Structure Parsing by Hybrid Discriminative/Generative Models", (in press) IEEE Trans. on Medical Imaging, 2007.
2. B. Dogdas, D.W. Shattuck, and RM Leahy RM, "Segmentation of Skull and Scalp in 3D Human MRI Using Mathematical Morphology", *Human Brain Mapping*, 26(4):273-85.
3. B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, A.M. Dale, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain", *Neuron*, vol. 33, 341-355, 2002.
4. G. Gerig, M. Chakos, and M. Valmet, "A new validation tool for assessing and improving 3D object segmentation", *Proc of MICCAI 2001*, pp. 516-523, Berlin, 2001.

5. Z. Lao, D. Shen, A. Jawad, B. Karacali, D. Liu, E. Melhem, N. Bryan, C. Davatzikos, "Automated Segmentation of White Matter Lesions in 3D Brain MR Images, Using Multivariate Pattern Classification", *Proc. of 3rd IEEE In'l Symp. on Biomedical Imaging*, pp 307 310, April 6-9, 2006, Arlington, VA, USA.
6. T. Rohlfing, D. B. Russakoff, and C. R. Maurer, Jr., "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation", *IEEE Trans. on Medical Imaging*, vol. 23, no. 8, pp 983-994, August 2004.
7. S.M. Pizer, T. Fletcher, Y. Fridman, D.S. Fritsch, A.G. Gash, J.M. Glotzer, S. Joshi, A. Thall, G. Tracton, P. Yushkevich, and E.L. Chaney, "Deformable m-reps for 3D medical image segmentation", *Int'l. J. of Comp. Vis.*, 55(2/3), pp. 85C106 ,200.
8. K.M. Pohl, J. Fisher, R. Kikinis, W.E.L. Grimson, and W.M. Wells, "A Bayesian model for joint segmentation and registration", *NeuroImage*, 31, 228-239, 2006.
9. Z. Tu and A. W. Toga, "Towards whole brain segmentation by hybrid models", *Proc. of MICCAI*, 2007.
10. R. P. Woods, J. C. Mazziotta, S. R. Cherry, "MRI-PET registration with automated algorithm", *Journal of Computer Assisted Tomography*, 17:536-546, 1993.
11. J. Yang , L. H. Staib, and J. S. Duncan, "Neighbor-Constrained Segmentation with Level Set Based 3D Deformable Models", *IEEE Trans. on Medical Imaging*, vol. 23, no. 8, pp 940-948,August 2004.