Automated segmentation of caudate nucleus in MR brain images with voxel classification

Yulia Arzhaeva, Eva van Rikxoort, and Bram van Ginneken

Image Sciences Institute, University Medical Center Utrecht, the Netherlands
{yulia,eva,bram}@isi.uu.nl

Abstract. This paper presents a supervised voxel classification method for segmentation of the caudate nucleus from brain MRI images. Supervised voxel classification is a general pattern recognition technique. In this application general spatial and local structure features extracted from image voxels were used together with a k-nearest neighbor classifier. The trained classifier has been applied to different groups of test data. On test data that originated from the same population as the training images, the method yielded segmentations that correlated very well with human segmentations (Pearson correlation coefficient (PCC) of 0.82 and volumetric overlap (VO) of 74.2%). On data from a different source that exhibits intensity ranges similar to the training data, the method performed slightly worse (PCC of 0.52, VO of 64%), and the method failed on data with different intensity ranges.

1 Introduction

Magnetic resonance imaging (MRI) provides detailed three-dimensional (3D) images of living tissues and is widely used for brain studies. Quantitative analysis of brain MRI images often requires precise segmentation of specific neuroanatomical structures. One of these is the caudate nucleus (CN), a subcortical component of the basal ganglia that is involved in sensory-motor control, cognition, language, emotion and other important brain functions.

The caudate is a periventricular gray matter structure that shows up lighter compared to the majority of cortical gray structures in T1-weighted MRI images. The caudate has a rather homogeneous intensity but is difficult to segment because it is attached to other gray structures at multiple locations. It consists of a relatively large head, a cone-shaped body and a thin tail.

Manual 3D segmentation of the caudate is difficult and time consuming. For that reason various fully automated and computer-assisted segmentation methods have been developed including atlas-based registration techniques, deformable models, knowledge-driven and histogram-driven approaches, and statistical modelling (see for a survey [1] and a number of references from [2]). However, an automated or semi-automated method for CN segmentation has not yet become the technique of choice in cognitive and neuroscience laboratories. That suggests that an optimal automated technique is yet to be found.

T. Heimann, M. Styner, B. van Ginneken (Eds.): 3D Segmentation in The Clinic:
A Grand Challenge, pp. 65-72, 2007.

In this paper we present a fully automated segmentation method that relies on supervised voxel classification. This is a general segmentation technique exploiting methods originated from the pattern recognition and machine learning communities (see, e.g. [3] for an overview). It segments a structure by assigning a proper label (*a class*) to each voxel in an image. To make a differentiation between classes possible, image processing techniques are applied to extract discriminative features from a voxel. This segmentation approach has recently gained popularity and been successfully used to segmenting different structures from images of various modalities, e.g. see [4, 5].

The core of our method is assigning to each voxel in the brain a probability to belong to CN. Prior to voxel classification the brain is extracted from the whole MRI image of the head. A simple image processing technique – a combination of histogram analysis, region growing and morphology operations – is used to obtain the brain mask. The brain mask is exploited in two ways. Firstly, only voxels within the brain mask are classified since CN lies inside the brain. Secondly, for each voxel its relative position within the brain is estimated. Position features are very powerful features for differentiation between CN and adjacent gray matter structures. Other features characterize a voxel neighborhood and are calculated using a multi-scale filter bank of Gaussian derivatives. The k-nearest neighbor (*k-NN*) classifier is first trained with voxels from training images which labels are known from available manual CN segmentations and then applied to test images. After classification the posterior probabilities of the voxels are blurred and thresholded, resulting in a binary segmentation mask for the caudate. The obtained segmentations are compared with manual segmentations using several established comparison metrics.

2 Method

2.1 Voxel classification

The supervised voxel classification process can be broken into two phases: a training phase and a test phase. In the training phase a number of voxels are sampled from the training images, a set of features are calculated for each voxel, and a classifier is trained. This general scheme is adjusted to a particular medical imaging task through a proper choice of features and classifiers. In order to be able to train the classifier, a ground truth (a manual segmentation, for example) is required which assigns to each voxel its class label. In the test phase the same set of features are calculated for each voxel, and the trained classifier is applied yielding per voxel a class membership probability.

Feature extraction. Features should be meaningful and describe distinctive characteristics of a voxel. Spatial features are important because CN has a characteristic location within the brain. We employ the voxel cumulative position in the brain that indicates for each voxel in the brain which percentage of a brain volume is above, next or behind it in each direction. Formally, for discrete data, the cumulative position (x', y', z') of a point (x, y, z) inside the

object (the brain) is defined as follows: $x' = \frac{1}{V}\sum_{k=0}^{Z}\sum_{j=0}^{Y}\sum_{i=0}^{x} I(i,j,k), y' = \frac{1}{V}\sum_{k=0}^{Z}\sum_{j=0}^{y}\sum_{i=0}^{X} I(i,j,k), z' = \frac{1}{V}\sum_{k=0}^{z}\sum_{j=0}^{Y}\sum_{i=0}^{X} I(i,j,k)$, where V is the total brain volume, I is the segmentation mask which is 1 inside the brain and 0 elsewhere, and X, Y and Z are the dimensions of an image. In order to calculate these three features the brain mask has to be obtained. Subsection 2.2 describes our brain segmentation algorithm. Note that other brain segmentation algorithms could be used instead.

Additionally, the outputs of the Gaussian derivative filters of the first and second order ($L_x, L_y, L_z, L_{xx}, L_{yy}, L_{zz}, L_{xy}, L_{xz}, L_{yz}$) at three scales ($\sigma$ = 1,2,4 voxels) are used to characterize a local image structure. The gray value (L_0) in the original images, as well as the gray values in the images blurred with the Gaussian kernel at different scales (L, $\sigma = 1, 2, 4$), could be beneficial features on the condition that intensity distributions are similar for training and test data. Since MRI intensities do not have any absolute physical meaning and may differ a lot depending on scanning and post-processing parameters we prepared two feature sets, with and without the L and L_0 features. The feature set amounts either to 34 features, or 30 features excluding L's and L_0.

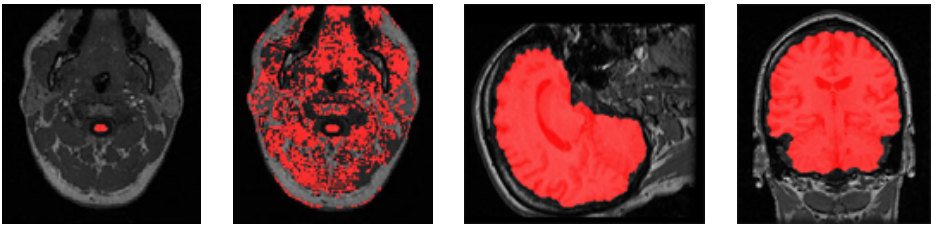
Classification. Only voxels within the brain are included in the training set. The caudate and background voxels are taken from subsampled training images (subsampling factors are 8 for background voxels and 2 for caudate voxels) to train the *k-NN* classifier (k =15). Different subsampling factors are used for selection of caudate and background voxels in order to balance the amount of samples of different classes in the training set. We have chosen the *k-NN* classifier because it is a non-parametric classifier that makes no assumption on the underlying distributions of features. The fast implementation of the *k-NN* classifier by Arya and Mount [6] is used in computations.

When the *k-NN* classifier is applied to a test image each voxel that lies within the brain is assigned a posterior probability of being a part of CN. This probability is estimated as a fraction of the caudate voxels among the test voxel's k nearest neighbors from the training set. To obtain a final segmentation, the posterior probabilities are first blurred with the Gaussian kernel ($\sigma = 1$), which has the effect of pooling local evidence and smoothing the borders. The resulting probabilities are thresholded to obtain a binary segmentation. Note that both left and right parts of the caudate are considered as the same class.

2.2 Extraction of the brain mask

The brain segmentation algorithm presented in this section is automatic but somewhat ad hoc; it is not intended to provide a very accurate segmentation since it plays only an auxiliary role in this application.

Firstly, for a given image, a range of intensitie values is found that encompasses the majority of the brain tissue. An optimal thresholding (as described in [7]) separates the majority of the head and brain tissue from the image background (perhaps excluding cerebrospinal fluid). Two major peaks from the histogram of image intensities higher than the optimal threshold characterize the gray and white matter (denoted as p_{gm} and p_{wm} respectively).



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

Fig. 1. An example of a leakage: the brain stem is correctly masked by 2D region growing (in 1(a)), but not after 3D region growing (in 1(b)). Figures 1(c) and 1(d) show an example of a final brain segmentation in a sagittal and coronal slice.

The brain is very roughly segmented by means of iterative 3D region growing that starts from a seed in the brain stem. The brain stem is found using a set of rules. It is searched as a connected component of a certain size within a particular range of intensities whose center of mass lies very close to $X/2$ (illustrated in Figure 1(a)). The range of likely sizes are estimated from the training images. The range of intensities are tentatively taken as $(p_{gm} - \delta, p_{wm} + \delta)$, where δ is an offset equal to the half of a distance between p_{gm} and the optimal threshold. The range of intensities allowed for region growing are initialized by the same interval. The region growing algorithm is iterative in the sense that the range of intensities are narrowed and the region growing is consequently repeated while *a leakage* occurs. The leakage is checked in a transversal slice on the level of the brain stem seed as an amount of overlap between the results of 2D and 3D region growing in this slice. In Figures 1(a) and 1(b) an example of the leakage (with an obviously low overlap between 2D and 3D segmentations) is demonstrated.

Sometimes it is not possible to get rid of the leakage in this way. Then 2D region growing is employed, first in a coronal and sagittal planes that go through the brain stem seed, then in all transversal slices using previously grown voxels as seeds. However, the second approach gives in average more occlusions of non-brain tissue than the first one.

In the final step, the brain segmentation is smoothed using 3D hole filling and morphological closing with a spherical structuring element to close gaps in the gray matter and cerebrospinal fluid and smooth the borders. All the brain masks include a part of the brain stem of various length and should be cut uniformly in transversal direction. The cut is done on the level where the cerebellum firstly appears in the mask which is easily recognizable as additional components appearing at the rear of the brain. The example of a final brain segmentation (sagittal and coronal slices) is given in Figures 1(c) and 1(d).

3 Experimental data

The training data set consisted of 15 brain MRI images and the corresponding manual caudate segmentations acquired from the Psychiatry Neuroimaging

Laboratory at the Brigham and Women's Hospital. Our method was tested on 4 distinct groups of data. The first group contained 14 MRI images acquired at the same site as the training data (further on denoted as BWH PNL test set). The second group consisted of 5 pediatric (patients no older than 2 years of age) MRI images acquired at the UNC Neuro Image Analysis Laboratory, Chapel Hill. 5 MRI images of elderly patients (above 55 years of age) with Parkinson disease acquired at the same place comprised the third group. The second and the third data sets are further referred as UNC Ped and UNC Eld respectively. Finally, the fourth group contained 10 MRI scans of the same young healthy person acquired within 60 days on 5 different scanners at the UNC Neuro Image Analysis Laboratory. For the first three groups the manual reference segmentations were available, and our automatic segmentations were compared against these. The fourth set was only used in testing the stability of the segmentation method.

The BWH PNL images from both training and test data sets were resized to 256 voxels in the sagittal direction and isotropic voxel size of 0.9375 mm. All UNC images originally had an isotropic voxel size of 1 mm and therefore were not resized.

4 Results and Discussion

Table 4 presents the results of the comparison metrics for BWH PNL, UNC Ped and UNC Eld data sets averaged for the left and right caudate segmentation. The metrics, listed from left to right in the table columns, are volumetric overlap error, relative volume difference, average symmetric absolute surface distance, symmetric RMS surface distance, and maximum symmetric absolute surface distance. Each measure is mapped into a score via linear scaling. A score of 100 points corresponds to a perfect segmentation, and a score of 90 points to an average human segmentation.

We used two different training sets to train the *k-NN* classifier. The classifier trained with the full set of features was applied to BWH PNL test images. The classifier trained with the feature set from which L and L_0 features were excluded was applied to images from UNC Ped and UNC Eld sets. The classification of one image required 35 to 50 minutes on a 3.2 GHz Pentium 4 PC. The computation time can be reduced by subsampling a test image and subsequent interpolation of posterior probabilities for neglected voxels. Another possibility to reduce computation time is to use a subset of features.

Our method demonstrates good average scores for BWH PNL and UNC Ped data sets, 73 and 74 respectively. In Figure 3 the first and last rows of images give examples of good correspondence between the manual and automatic segmentation for images from these data sets. The maximum symmetric absolute surface distance metric yields the worst score for BWH PNL data which confirms our observation that the tail of the caudate is often undersegmented by our method. UNC Ped images suffer from an overall undersegmentation that results in a relatively low volume difference score. On UNC Eld images our method largely fails. These particular images exhibit a much wider range of intensities

Correl	UNC Ped	UNC Eld	BWH PNL	Total
Left	0.3099	-0.2707	0.8558	0.2983
Right	0.7228	-0.0282	0.7915	0.4954
Average	0.5164	-0.1495	0.8236	0.3968

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	4353	4309	4078	3846	4281	4266	4175	4192	3923	4312	4174	173	4.1
Right	4278	4610	4235	4091	4619	4379	4446	4483	4067	4407	4362	193	4.4
Total	-										-	-	4.3

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.

than BWH PNL and UNC Ped images. Apparently, the exclusion of the gray value features did not overcome the influence of the difference in intensity resolution between training and test images. In a trial experiment we applied a linear intensity scaling to all test and training images in order to unify intensity ranges. The results were remarkably better for UNC Eld images (no failures at all), but seemed slightly worse for BWH PNL and UNC Ped images. It is likely that fine-tuning of transformation parameters, or a non-linear intensity transformation, would improve the segmentation of dissimilar data without deteriorating segmentations of similar images.

Another comparison measure is the Pearson correlation coefficient. This coefficient captures how well the volumetric measurements correlate with those of the reference segmentation, in a range of $[-1, 1]$. Manual segmentations of human experts for the caudate show in average a Pearson correlation of 0.71 as compared with each other. The results of our method for different test groups are presented in Table 1 showing a very good correlation for BWH PNL data, and a less stronger correlation for UNC Ped data.

Finally, repeatability measurements are presented in Table 2. Repeatability is measured as the coefficient of variation (= standard deviation / mean) of the volumetric measurements.Manual segmentation of human experts show in average a COV of 3.1%. Our method demonstrates slightly higher variation.

In conclusion, a voxel classification approach to caudate segmentation performed very well if the method was trained and validated on data from the same population. The method's performance could possibly be improved with an enhanced brain mask extraction method and a prior alignment of all cases, e.g. via registration. Our method was not applicable to images whose intensity resolutions differed a lot from the training scans.

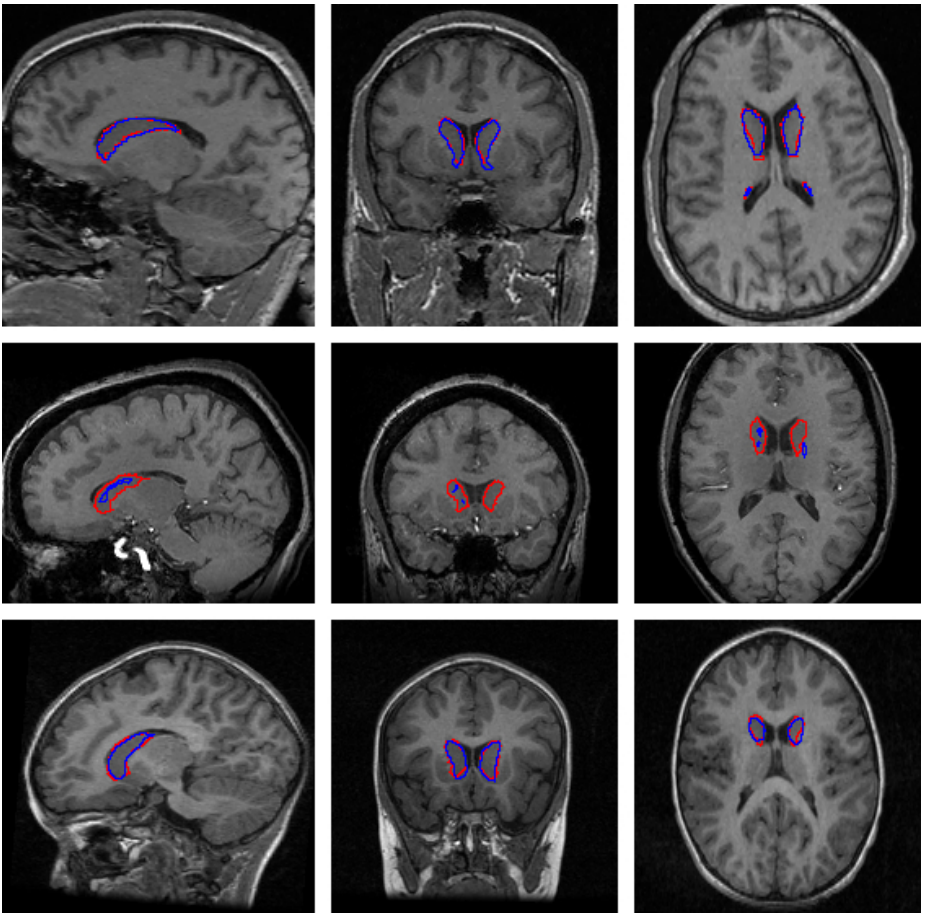


Fig. 2. From left to right, a sagittal, coronal and transversal slice from a a subject in the adults BWH group (top), one in the elderly UNC group (middle) and one in the pediatric UNC group (bottom). The outline of the reference standard segmentation is in red, the outline of the segmentation of the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

References

1. Clarke, L., Velthuizen, R., Camacho, M., Heine, J., Vaydianathan, M., Hall, L., Thatcher, R., Silbiger, M.: MRI segmentation: Methods and applications. Magnetic Resonance Imaging **13** (1995) 343–368
2. Xia, Y., Bettinger, K., Shen, L., Reiss, A.: Automatic segmentation of the caudate nucleus from human brain MR images. IEEE Transactions on Medical Imaging **26**(4) (2007) 509–517
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd edn. John Wiley and Sons, New York (2001)

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	39.7	75	-18.1	68	0.9	66	1.6	72	9.8	71	70
UNC Ped 14	41.3	74	-30.9	46	0.8	70	1.2	79	6.0	82	70
UNC Ped 15	26.2	84	-7.4	87	0.5	80	1.0	82	7.6	78	82
UNC Ped 19	40.5	74	-35.2	38	0.9	68	1.3	78	5.8	83	68
UNC Ped 30	32.9	79	-19.8	66	0.7	75	1.0	82	6.6	80	76
UNC Eld 01	100.0	0	-99.8	0	34.7	0	34.9	0	43.2	0	0
UNC Eld 12	97.5	20	-96.7	0	17.4	0	18.7	0	33.5	8	6
UNC Eld 13	91.6	42	-88.7	0	6.0	2	7.3	24	15.9	53	24
UNC Eld 20	99.8	0	-99.6	0	40.9	0	42.7	0	61.3	0	0
UNC Eld 26	100.0	0	-97.4	0	22.9	0	23.8	0	35.4	0	0
BWH PNL 16	27.3	83	-3.1	90	0.7	74	2.1	62	21.3	38	70
BWH PNL 17	25.3	84	-8.9	84	0.6	78	2.0	64	23.3	32	68
BWH PNL 18	26.6	83	-21.2	63	0.5	80	1.1	80	9.7	72	76
BWH PNL 19	25.2	84	-7.6	86	0.6	78	1.8	68	21.3	38	70
BWH PNL 20	26.1	84	-10.8	81	0.6	78	1.9	65	23.8	30	68
BWH PNL 21	29.9	81	-12.6	78	0.7	73	2.0	64	22.5	34	66
BWH PNL 22	30.0	81	-15.3	73	0.8	70	2.3	59	22.5	34	64
BWH PNL 23	21.3	87	1.4	96	0.4	87	0.8	86	8.2	76	86
BWH PNL 24	21.1	87	0.6	98	0.3	87	0.6	88	3.7	89	90
BWH PNL 25	24.3	84	-11.6	80	0.6	76	2.1	62	21.8	36	68
BWH PNL 26	23.8	85	4.6	92	0.4	86	0.8	86	7.4	78	86
BWH PNL 27	25.1	84	-10.5	82	0.7	74	2.0	64	19.1	44	70
BWH PNL 28	25.9	84	-20.1	64	0.7	76	2.1	62	19.9	42	66
BWH PNL 29	29.3	82	-18.1	68	0.5	82	0.8	86	4.0	88	82
Average All	42.9	68	-30.3	60	5.6	61	6.5	59	18.9	49	59
Average UNC Ped	36.1	77	-22.3	61	0.8	72	1.2	78	7.1	79	74
Average UNC Eld	97.8	12	-96.4	0	24.4	0	25.5	5	37.9	12	6
Average BWH PNL	25.8	84	-9.5	81	0.6	79	1.6	71	16.3	52	73

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.

4. van Ginneken, B., Stegmann, M., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical Image Analysis **10**(1) (2006) 19–40
5. Folkesson, J., Dam, E., Olsen, O., Pettersen, P., Christiansen, C.: Segmenting articular cartilage automatically using a voxel classification approach. IEEE Transactions on Medical Imaging **26**(1) (2007) 106–115
6. Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A.: An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. Journal of the ACM **45**(6) (1998) 891–923
7. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, analysis, and machine vision. 2nd edn. PWS Publishing, Pacific Grove, CA (1999)
8. Gerig, G. Jomier, M., Chakos, M.: Valmet: a new validation tool for assessing and improving 3D object segmentation. In: MICCAI 2001. Number 2208 in Lecture Notes in Computer Science, Springer, Berlin (2001) 516–523