Comparison and Evaluation of Segmentation Techniques for Subcortical Structures in Brain MRI

K.O. Babalola[1], B. Patenaude[2], P. Aljabar[3], J. Schnabel[4], D. Kennedy[5], W. Crum[6], S. Smith[2], T.F. Cootes[1], M. Jenkinson[2], and D. Rueckert[3]

[1] Division of Imaging Science and Biomedical Engineering (ISBE), University of Manchester, UK
{kola.babalola,tim.cootes}@manchester.ac.uk
[2] FMRIB Centre, John Radcliffe Hospital, University of Oxford, OX3 9DU, UK
{mark,brian}@fmrib.ox.ac.uk
[3] Department of Computing, Imperial College London, SW7 2BZ, UK
{dr,pa}@doc.ic.ac.uk
[4] Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, UK
julia.schnabel@eng.ox.ac.uk
[5] MGH/MIT/HMS Athinoula A. Martinos Center for Biomedical Imaging, Building 149, 13th Street, Radiology/CNY149-Room 2301, Charlestown, MA 02129, USA
dave@cma.mgh.harvard.edu
[6] Institute of Psychiatry, Box P089, De Crespigny Park, London, UK, SE5 8AF
bill.crum@iop.kcl.ac.uk

Abstract. The automation of segmentation of medical images is an active research area. However, there has been criticism of the standard of evaluation of methods. We have comprehensively evaluated four novel methods of automatically segmenting subcortical structures using volumetric, spatial overlap and distance-based measures. Two of the methods are atlas-based – classifier fusion and labelling (CFL) and expectation-maximisation segmentation using a dynamic brain atlas (EMS), and two model-based – profile active appearance models (PAM) and Bayesian appearance models (BAM). Each method was applied to the segmentation of 18 subcortical structures in 270 subjects from a diverse pool varying in age, disease, sex and image acquisition parameters. Our results showed that all four methods perform on par with recently published methods. CFL performed significantly better than the other three methods according to all three classes of metrics.

1 Introduction

Functional and structural brain imaging are playing an expanding role in neuroscience and experimental medicine. The amount of data produced by imaging increasingly exceeds the capacity for expert visual analysis, resulting in a growing need for automated image analysis. In particular, accurate and reliable methods for segmentation (classifying image regions) are a key requirement for the extraction of qualitative or quantitative information from images.

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 409–416, 2008.

Image segmentation aims to separate an image into anatomically meaningful regions. The objective evaluation of image segmentation methods is crucially important in order to get automated image segmentation methods accepted in clinical practice. One of the key challenges for the evaluation of automated image segmentation methods is the lack of a gold standard against which to compare segmentation methods. In most cases expert manual segmentations are regarded as a gold standard. Given any gold standard segmentation there exist a large number of different methodologies that can be used to evaluate the quality of a given segmentation. These can be broadly divided into three groups. The first group, spatial overlap measures e.g. the Dice [1] coefficient and generalised overlap indicies [2], quantify the overlap between regions. Secondly, distance measures, e.g. signed surface distances, the mean absolute distance and the Hausdorff distance [3], quantify the distance between manually and automatically segmented surfaces. Thirdly, there are volumetric measures based solely on the volumes of the segmented regions e.g. the difference between volumes [4]. The selection of a class of metric depends on the clinical application of interest.

In this paper we compare four different state-of-the-art algorithms for automatic segmentation of subcortical structures in MR brain images. Two of these are model-based, and the other two are based on image registration. For objective comparisons and to address the limitations mentioned above, they have been evaluated using the same data and criteria. The dataset was large (270 subjects) and varied in age, disease and sex (see section 2). 18 subcortical structures were segmented from each subject and evaluated using a mixture of spatial overlap, distance and volumetric measures. We present qualitative and quantitative results of the evaluation of each method with respect to manually annotated data, and with respect to the other methods.

2 Materials and Method

The dataset consisted of 270 T1-weighted MR brain images which had been extensively labelled manually using methods similar to those described in [5]. For the purposes of this study, we used the a subset of the labels of 18 structures – the brain stem, the fourth ventricle and the left and right pairs of the accumbens, amygdala, caudate nucleus, hippocampus, lateral ventricles, pallidum, putamen and thalamus. The imaged cohorts included control subjects as well as subjects with Alzheimer's Disease, Schizophrenia, Attention Deficit Hyperactivity Disorder (ADHD), and prenatal drug exposure. Their ages ranged from 4.5yrs to 83yrs.

Four different methods have been used for segmentation. In each case the automatic segmentation, A has been compared to the manually labelled image which has been regarded as gold standard, G, by computing the metrics described below.

2.1 Segmentation Methods

Classifier Fusion and Labelling (CFL). [6] obtains segmentations by propagating labels from multiple atlases to the query subject and fusing them using a voting rule. When presented with a large repository of atlases it addresses problems of scale by ranking the atlases based on similarity with the query subject and choosing the best 20.

Profile Active Appearance Models (PAM). [7] are a variation of active appearance models which model the intensities along profiles normal to the boundary of a structure. A composite model of all structures and local models for each structure are coupled to perform a global search followed by more refined structure specific searches.

Bayesian Appearance Models (BAM). [8] Similar to the profile AAM, the BAM models texture along profiles normal to a surface. The BAM differs from the PAM mainly in that it models the relationship between shape and intensity via the conditional distribution of intensity given shape. Rather than synthesising intensities, BAM predicts intensity distributions and maximises the probability of the shape given the observed intensities.

Expectation-Maximisation-based segmentation using a dynamic brain atlas (EMS). [9] is a probabilistic approach combining a standard EM-based segmentation [10] with a dynamic brain atlas construction [11].

2.2 Evaluation Metrics

Dice coefficient. The Dice coefficient D [1] is one of a number of measures of the extent of spatial overlap between two binary images. It is commonly used in reporting performance of segmentation and gives more weighting to instances where the two images agree. Its values range between 0 (no overlap) and 1 (perfect agreement). In this paper the Dice values are expressed as percentages and obtained using Equation 1.

$$D = \frac{2(A \bigcap G)}{(A \bigcap G + A \bigcup G)} \times 100 \tag{1}$$

Hausdorff distance. The directed Hausdorff distance H_{ag}, between two sets of points A and G can be obtained in a two stage manner. First, for each point in A the minimum distance to all points in G is obtained. H_{ag} is the maximum of this set of minimum distances. In the present case, the minimum distance for the i^{th} surface voxel in A to the set of surface voxels in G is d_i^{ag}, therefore H_{ag} is the maximum value of the surface distance of all surface voxels in A (Equation 2). The Hausdorff distance, H, is the maximum of the directed form for $A \rightarrow G$ and $G \rightarrow A$ (Equation 3).

$$H_{ag} = max\{d_i^{ag}\}, i = \{1 \ldots n_a\} \tag{2}$$

$$H = max(H_{ag}, H_{ga}) \tag{3}$$

Volumes. For each individual segmentation result we find the volume, V, as the number of labelled voxels multiplied by the voxel dimensions. We then calculate the percentage absolute volumetric difference (AVD) as the ratio of the absolute difference between the original volume and the segmented volume, to the original volume (Equation 4). The absolute value is used to account for some segmentation results having a lower volume than the gold standard, and others having a higher volume.

$$AVD = \frac{|V_a - V_g|}{V_g} \times 100 \tag{4}$$

2.3 Experiments

The 270 subjects were randomly assigned into 27 groups of 10. Each of the methods described in section 2.1 was applied to segment each image in one of the groups of 10 using data from the other 26 groups. The results from each method were converted into binary voxel images with the same resolution as the input images. One binary file was obtained for each structure for each subject, and the measures described in section 2.2 applied to obtain quantitative results. Qualitative results were obtained by superimposing contours derived from the binary images onto the respective T1 images.

3 Results

Table 1 shows the results of applying the metrics of section 2.1 to the segmentation results of each structure using the manual labels as gold standards. The results are averages for each structure (left and right pairs combined) over the 27 sets of leave ten out experiments ($n = 540$ for all structures except brain stem and fourth ventricle where $n = 270$). summary box and whisker plots of the values of the Dice coefficient, Hausdorff distance and percentage absolute volumetric difference are given in Figure 1. The *p-values* of two-sample *t-tests* on the differences between the means were obtained at a structure by structure level and also over all structures pooled together. Differences were taken to be significant for *p-values* less than 0.05.

4 Discussion

Summary of results relative to gold standard
Table 1 contains results for each method relative to the gold standard as measured by spatial overlap, volumetric and distance-based metrics. The results are shown by structure, and a summary over all structures is given in Figure 1. Using the summary over all structures and the significance levels, the methods can be ranked in order of decreasing performance by a spatial overlap, a distance-based, and a volumetric metric as follows:

Table 1. Summary table for results of all methods applied on all structures over all measures with respect to manual annotations

Structure	Method	Dice Coefficient	Hausdorff Distance (mm)	Percent mean abs vol diff
accumbens	CFL	75.8 (7.2)	3.1 (1.0)	17.6 (16.9)
	PAM	67.7 (9.9)	3.8 (1.2)	18.2 (13.5)
	BAM	68.7 (7.9)	3.5 (1.0)	31.7 (29.1)
	EMS	67.9 (7.9)	4.3 (1.5)	28.6 (33.1)
amygdala	CFL	77.7 (5.8)	4.4 (1.6)	17.0 (15.8)
	PAM	66.9 (12.3)	5.3 (2.2)	20.9 (17.6)
	BAM	73.1 (6.9)	4.8 (2.2)	24.7 (22.8)
	EMS	70.8 (7.4)	5.4 (1.6)	22.1 (22.7)
brain stem	CFL	94.2 (1.4)	4.8 (1.5)	4.0 (3.0)
	PAM	87.8 (3.0)	6.0 (1.8)	6.8 (5.7)
	BAM	88.5 (2.0)	6.4 (2.1)	7.8 (5.8)
	EMS	82.9 (3.6)	7.7 (2.1)	21.1 (8.4)
caudate	CFL	88.1 (2.8)	4.1 (1.9)	7.7 (6.2)
	PAM	83.4 (5.1)	4.1 (2.0)	5.0 (5.5)
	BAM	85.6 (3.5)	4.6 (2.1)	13.2 (11.5)
	EMS	82.6 (5.7)	6.4 (3.2)	14.0 (12.9)
fourth ventricle	CFL	83.3 (4.7)	6.6 (2.9)	15.0 (11.5)
	PAM	70.6 (9.9)	7.7 (3.1)	15.4 (13.9)
	EMS	77.4 (8.6)	9.0 (4.2)	39.1 (34.2)
hippo-campus	CFL	83.5 (3.7)	4.5 (1.5)	9.2 (8.7)
	PAM	76.8 (6.2)	5.2 (1.6)	12.0 (7.4)
	BAM	79.12 (4.3)	5.0 (1.7)	22.1 (16.5)
	EMS	76.4 (5.9)	6.4 (2.0)	14.5 (13.3)
lateral ventricle	CFL	91.3 (3.7)	9.8 (7.3)	6.5 (6.3)
	PAM	80.9 (6.8)	14.0 (8.4)	7.3 (9.9)
	BAM	79.5 (9.6)	16.2 (9.5)	39.0 (34.7)
	EMS	82.9 (12.0)	10.5 (6.8)	39.4 (53.3)
pallidum	CFL	81.9 (4.8)	3.6 (1.1)	9.9 (7.1)
	PAM	79.3 (5.1)	3.8 (1.0)	9.4 (9.9)
	BAM	79.5 (4.3)	3.8 (1.0)	22.8 (15.7)
	EMS	80.5 (4.5)	3.9 (1.1)	13.8 (10.5)
putamen	CFL	89.8 (2.4)	3.6 (1.1)	6.9 (6.2)
	PAM	86.3 (2.8)	3.8 (1.1)	4.0 (4.2)
	BAM	86.4 (2.6)	4.4 (1.5)	14.6 (8.9)
	EMS	86.6 (2.5)	4.5 (1.2)	8.2 (6.6)
thalamus	CFL	90.8 (1.6)	4.0 (1.0)	4.8 (4.1)
	PAM	87.7 (2.8)	4.1 (1.0)	4.0 (3.3)
	BAM	87.6 (2.5)	4.3 (1.0)	13.7 (9.3)
	EMS	85.2 (2.1)	5.5 (1.4)	10.6 (6.5)

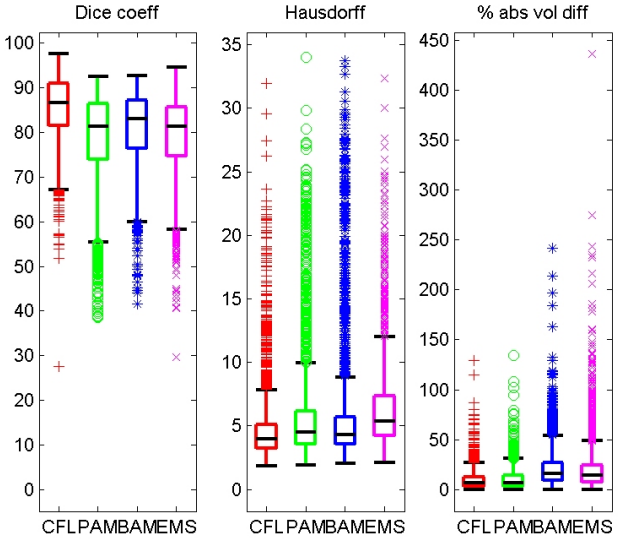

Fig. 1. Box and whisker plots of Dice coefficients, Hausdorff distance, and percentage absolute volumetric difference over all structures for each method. The whiskers are 1.5 × the inter-quartile range, and values outside these are plotted individually.

Spatial overlap (Dice): CFL → BAM → EMS → PAM
Distance-based (Hausdorff): CFL → BAM and PAM → EMS
Volumetric (AVD): CFL and PAM → EMS and BAM

The CFL method clearly gives the best overall performance with respect to the gold standard. When considering performance on a structure by structure basis the Dice coefficients of CFL are significantly better than those of the other three methods for all structures. The Dice values of BAM are either the same or better than those of PAM and EMS for all structures except the lateral ventricles (and pallidum for EMS). The Dice values of EMS were better than those for PAM for the amygdala, fourth ventricle, lateral ventricle and pallidum.

The Hausdorff distances of CFL at the structure level are better than those of all other methods for all structures except the caudate, pallidum and thalamus where it performs at least as well or better than the other methods. BAM and PAM have a mixture of better and same results relative to each other for this metric. EMS only performed better than BAM and PAM for the lateral ventricle.

CFL performed better than BAM and EMS for all structures, according to AVD. However, it didn't performed as well as PAM on the caudate, putamen and thalamus. PAM was better than EMS and BAM for all structures except the amygdala where it performed the same as EMS, and the brain stem where it was the same with BAM. EMS performed better than PAM on the hippocampus, pallidum, putamen and thalamus.

When comparing how well a method performed on the different structures, the order of best performance depends on the metric used to judge the performance.

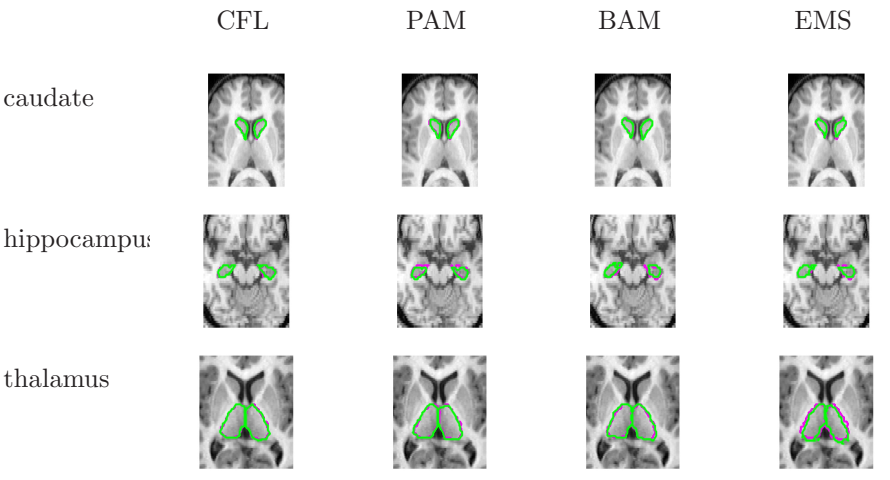

Fig. 2. Overlays of the segmentation results of all four methods for the subjects that gave the best Dice value for some structures for the CFL method. Pink is gold standard, green is result of method.

According to the spatial overlap metrics, the accumbens gives the worst overlap results for all methods, and the brain stem gives the best results. However the accumbens gives the best (lowest) Hausdorff distance and the lateral ventricles the worst for all methods. These differences are a combination of the fact that the accumbens are the smallest structures (hence small errors in overlap give high changes in the spatial overlap measures), and the shape of the lateral ventricles (in particular the length of the occipital horns coupled with partial volume effects along them) means that they are difficult to segment especially for methods relying on prior shape topology (BAM and PAM).

In terms of computing time and resources required, the CFL method is most expensive taking the order of a few hours to segment a 3D volumetric image. The model based methods perform faster in the order of tens of minutes and the EMS method falls in between these two.

5 Conclusion

We have presented a comprehensive comparison of four methods for fully automatic segmentation of 18 subcortical structures in the brain. These methods perform at least on par with currently available methods. The results will be of use to those needing to make a decision about a tool to use for applied research in brain imaging. The BAM method has been implemented as FIRST in the widely used FSL software suite[1], and the PAM method will be available for download over the internet from the University of Manchester[2] UK. The binaries and code

[1] www.fmrib.ox.ac.uk/fsl/first
[2] www.isbe.man.ac.uk/∼kob/ibim

to implement the CFL method on a database of images is also freely available from Imperial College, UK [3].

Acknowledgements

This work was funded by the EPSRC under the IBIM project. David Kennedy of the Center for Morphometric Analysis, Boston, provided the MR images used.

References

1. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26, 297–302 (1945)
2. Crum, W.R., Camara, O., Hill, D.L.G.: Generalised overlap measures for evaluation and validation in medical image analysis. IEEE Transanctions on Medical Imaging 25(11), 1451–1461 (2006)
3. Gerig, G., Jomier, M., Chakos, M.: Valmet: A new validation tool for assessing and improving 3D object segmentation. In: Niessen, W.J., Viergever, M.A. (eds.) MICCAI 2001. LNCS, vol. 2208, pp. 516–523. Springer, Heidelberg (2001)
4. Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C.: Automatic 3-D model-based neuroanatomical segmentation. Human Brain Mapping 3(3), 190–208 (1995)
5. Filipek, P., Richelme, C., Kennedy, D., Caviness, V.: The young adult human brain: An MRI-based morphometric analysis. Cereb. Cort. 4, 344–360 (1994)
6. Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D.: Classifier selection strategies for label fusion using large atlas databases. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 523–531. Springer, Heidelberg (2007)
7. Babalola, K.O., Petrovic, V., Cootes, T.F., Taylor, C.J., Twining, C.J., Williams, T.G., Mills, A.: Automated segmentation of the caudate nuclei using active appearance models. In: 3D Segmentation in the clinic: A grand challenge. Workshop Proceedings, MICCAI 2007, Brisbane, pp. 57–64 (2007)
8. Patenaude, B., Smith, S., Kennedy, D., Jenkinson, M.: Bayesian shape and appearance models, Technical report TR07BP1, FMRIB Centre - University of Oxford
9. Murgasova, M., Dyet, L., Edwards, A.D., Rutherford, M., Hajnal, J., Rueckert, D.: Segmentation of brain MRI in young children. Acad. Rad (in press, 2007)
10. Leemput, K.V., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE TMI 18(10), 897–908 (1999)
11. Hill, D.L.G., Hajnal, J.V., Rueckert, D., Smith, S.M., Hartkens, T., McLeish, K.: A dynamic brain atlas. In: Dohi, T., Kikinis, R. (eds.) MICCAI 2002. LNCS, vol. 2488, pp. 532–539. Springer, Heidelberg (2002)

[3] www.doc.ic.ac.uk/∼dr/software