NeuroImage 43 (2008) 225–235

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Automated morphological analysis of magnetic resonance brain imaging using spectral analysis

P. Aljabar [a], D. Rueckert [a], W.R. Crum [b,*]

[a] Department of Computing, Imperial College London, Visual Information Processing Group, 180 Queen's Gate, London, SW7 2AZ, UK
[b] Centre for NeuroImaging Sciences (P089), Institute of Psychiatry, Kings College London, De Crespigny Park, London, SE5 8AF, UK

ARTICLE INFO

Article history:
Received 25 February 2008
Revised 24 July 2008
Accepted 31 July 2008
Available online 9 August 2008

Keywords:
Image segmentation
Spectral analysis
Generalized overlap
Label fusion
Dementia

ABSTRACT

Analysis of structural neuroimaging studies often relies on volume or shape comparisons of labeled neuroanatomical structures in two or more clinical groups. Such studies have common elements involving segmentation, morphological feature extraction for comparison, and subject and group discrimination. We combine two state-of-the-art analysis approaches, namely automated segmentation using label fusion and classification via spectral analysis to explore the relationship between the morphology of neuroanatomical structures and clinical diagnosis in dementia. We apply this framework to a cohort of normal controls and patients with mild dementia where accurate diagnosis is notoriously difficult. We compare and contrast our ability to discriminate normal and abnormal groups on the basis of structural morphology with (supervised) and without (unsupervised) knowledge of each individual's diagnosis. We test the hypothesis that morphological features resulting from Alzheimer disease processes are the strongest discriminator between groups.

Introduction

Magnetic resonance imaging (MRI) of the brain has become an indispensable tool for diagnosis and research in neuroimaging. Segmentation of brain regions of structural or functional interest via labeling is a requirement for quantitative studies of morphology as it provides a neuroanatomical context to subsequent measurements or forms the basis of those measurements. The classic structural neuroimaging experiment seeks morphological measures which discriminate two sets of subjects grouped on the basis of other information (such as genetics, neuro-psychology, medication, etc). A related experiment first discovers such discriminators from training data and then applies them to classify new subjects. This can form the basis of a diagnostic system (e.g. Klöppel et al., 2008) Techniques employed range from simple manual volumetry (Jack et al., 1997) to sophisticated shape-based measurement and classification techniques (Wang et al., 2007). The alternative framework of "hypothesis-free" analysis exemplified by Voxel Based Morphometry (VBM) (Ashburner and Friston, 2000) is concerned with the detection and significance of local tissue density differences rather than an analysis of their morphological structure. More recent developments such as the incorporation of local measures of volume change into VBM as well as so-called DBM (Deformation-Based-Morphometry) (Ashburner et al., 1998) and TBM (Tensor-Based-Morphometry) (Studholme et al., 2004) have blurred the operational distinction between traditional morphological analysis and voxel-wise methods. While there is on-going debate about the reliability and interpretation of hypothesis-free techniques (Bookstein, 2001; Davatzikos, 2004), morphological analysis of individual structures, identified either manually or with computer-assistance, can be regarded as a practical gold-standard.

Manual segmentation methods requiring expert neuroanatomical knowledge or at least a protocol derived from expert knowledge, have been used for many years, and retain particular importance in the case of structures which are challenging for automatic segmentation techniques such as the hippocampus (Jack et al., 1997; Pruessner et al., 2000) and the entorhinal cortex (Du et al., 2001). Such methods are time-consuming and suffer from errors which are a function of a range of human factors (e.g. inter- and intra-observer variation, practice and temporal drift effects), segmentation protocol details and acquisition details (scan signal and contrast characteristics, patient motion and other artifacts, other scanner calibration and performance issues etc). In parallel there has been a huge amount of research effort devoted to automation, from techniques which simply separate brain from non-brain (Smith, 2002) to those which provide detailed gyral and sulcal labeling (Mangin et al., 2004). Automated techniques have improved immensely but can be computationally demanding, complex, and sensitive to image acquisition details and the presence of abnormal anatomy (Duncan and Ayache, 2000). Nevertheless, the identification of brain structures and/or tissue-classes is a necessary prerequisite to virtually all morphological analyses. The simplest and most common analysis which depends on neuroanatomical labeling is a cross-sectional (single time-point) volumetric comparison. Many authors have investigated higher order measures of shape (Csernansky et al.,

* Corresponding author. Fax: +44 20 3228 2116.
E-mail address: bill.crum@iop.kcl.ac.uk (W.R. Crum).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.07.055

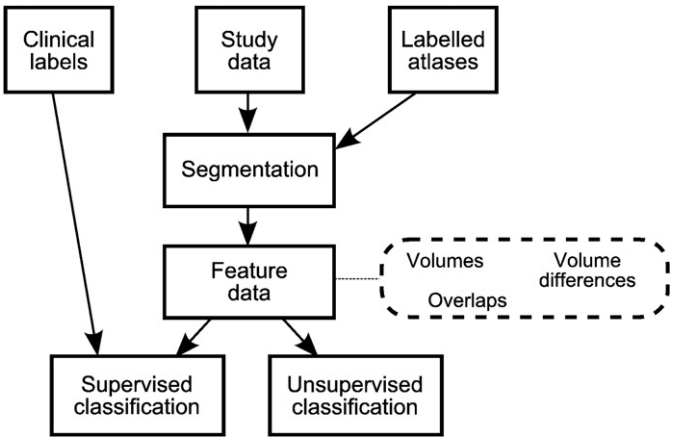

Fig. 1. A schematic illustration of the different components of the analysis pipeline.

1998; Kim et al., 2005; Wang et al., 2006) with varied success and interpretation of results and reproducibility on large cohorts remains difficult.

We have two goals in this work: The first is to move beyond simple volumetry but avoid some of the drawbacks, including computational cost and interpretability, of traditional higher order shape analysis without resorting to intensive manual techniques. The second goal is to partition a group of subjects purely on the basis of observed morphology, i.e. an unsupervised classification approach without prior knowledge of clinical status, and compare the associated discriminators with those derived from a supervised approach. We focus on achieving high-quality structural segmentation using state-of-the-art automated label fusion based segmentation techniques (Aljabar et al, 2007). These techniques select candidate segmentation atlases from a pre-existing database and, by appropriate combination of candidate labels at the voxel level, become robust to many sources of random error including unavoidable anatomical variation, registration error and random labeling errors in the atlas population. The subsequent analysis step uses the overlap of labeled structures as the simplest possible generic indicator of shape similarity beyond volumetric measures. We summarise group morphology by constructing a complete graph where each subject is represented by a node and pairs of nodes are connected with edge-weights that are a function of the morphological similarity (e.g. label overlap) of one or more structures. We apply spectral analysis techniques (von Luxburg, 2007) to the graph to generate indicator vectors which can be used to partition the graph, and therefore the subjects, on the basis of morphological similarity. The resulting unsupervised morphological classification is compared with a supervised linear discriminant analysis which seeks the morphological measure which best separates groups when the clinical status of each subject is known. The analysis framework is generic in that we are at liberty to choose both the methods for generating morphological features and the manner in which we compare those features between subjects.

In this paper we focus on an exemplar application in dementia where departures from normal anatomy are gradual and progressive and where previous studies suggest that label fusion and volumetric and overlap-based similarity measures should be able to describe the morphology present in the cohort. There has been an immense amount of work on characterising the appearance of dementia in structural MRI (Chetelat and Baron, 2003) and thereby measuring disease progression (Fox et al., 2001), detecting early disease (Jack et al., 1997) and distinguishing disease processes from normal ageing (Laakso et al., 1998). There is evidence of subtle pre-clinical global changes in brain morphology exemplified by the work in Fox et al. (1996,1999). The reliable automated morphological analysis of structural changes associated with Alzheimer's disease will add to our understanding of the structural consequences of pathology in this group of diseases. With the advent of treatments which provide symptomatic relief and the prospect of disease-modifying agents, it is increasingly important to characterise and detect Alzheimer's disease by its effect on brain morphology.

Methods

The analysis pipeline has several stages: First, label fusion of registered atlases is used to obtain high-quality segmentations of neuroanatomical structures for each subject. After this, all subjects in the analysis cohort are spatially normalised to a standard reference space for group analysis. Feature data are then extracted from the spatially normalised segmentations for use in either a supervised or an unsupervised classification step. The data extracted are either raw

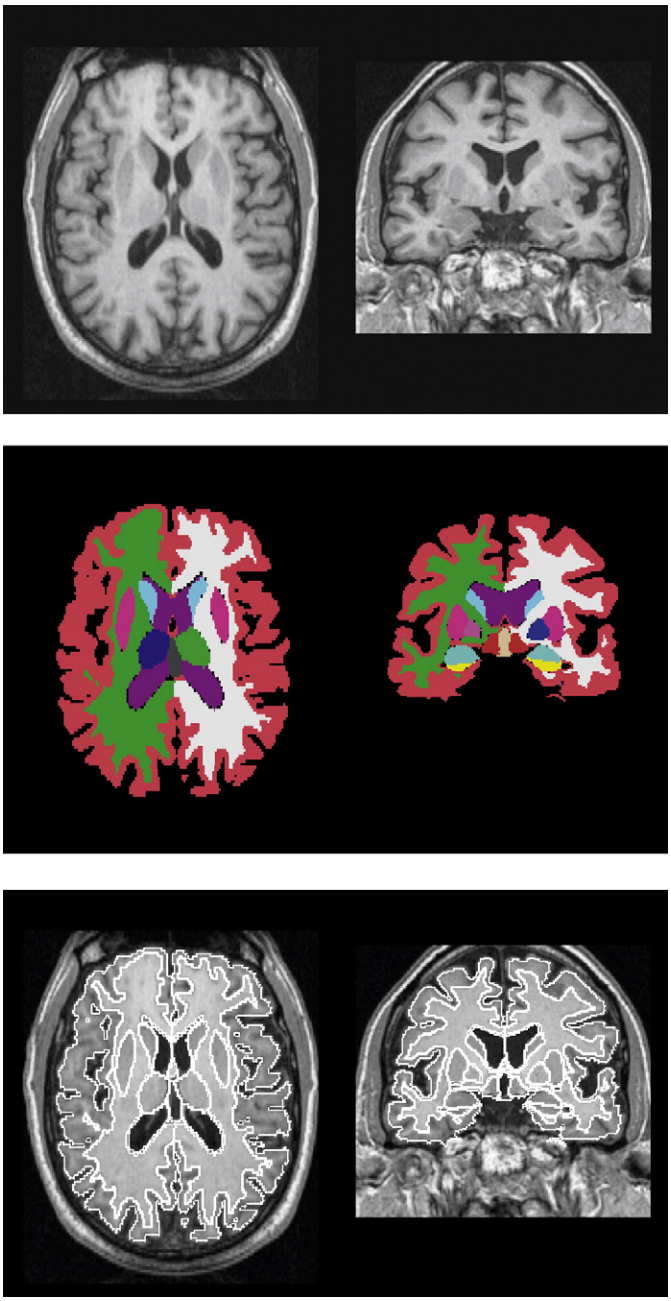

Fig. 2. An example of a label fusion segmentation of a control subject. Top: original scan; Middle: label fusion result; Bottom: label fusion overlay on the original scan.

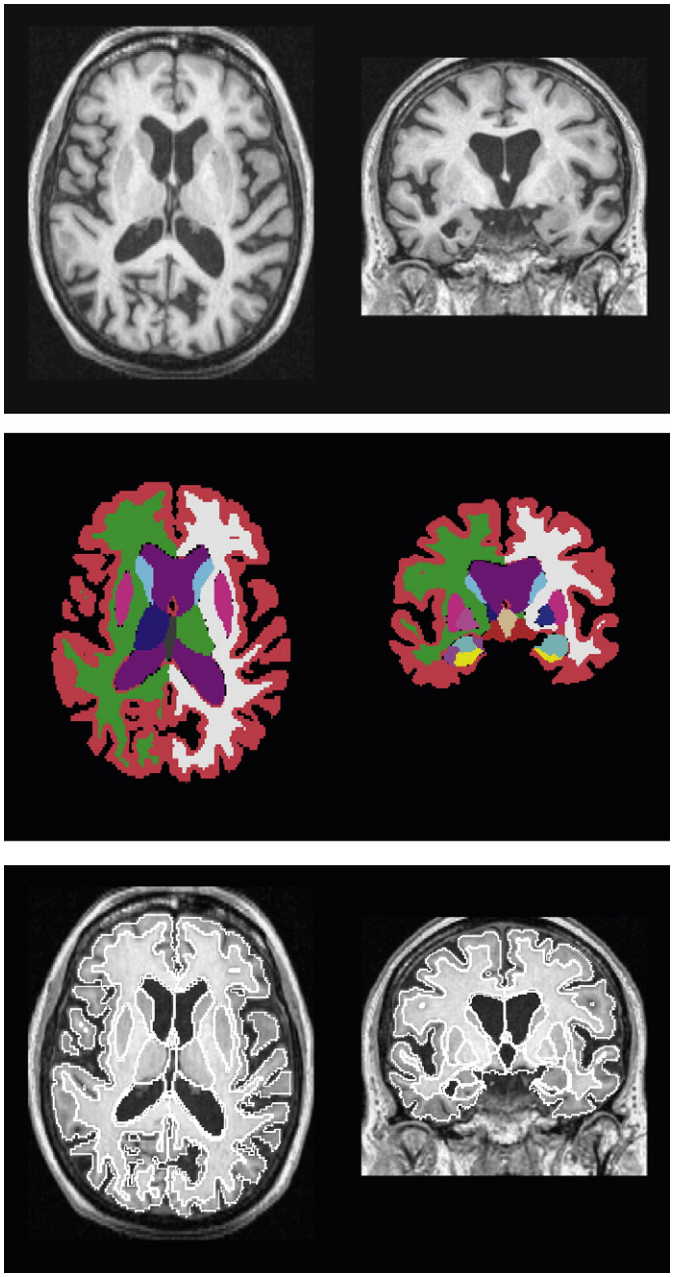

Fig. 3. An example of a label fusion segmentation of a probable Alzheimer's subject. Top: original scan; Middle: label fusion result; Bottom: label fusion overlay on the original scan.

volumes of structures for each subject or low-order pairwise measures of morphological similarity between subjects. The pairwise similarity measures are derived from either the differences in volume between corresponding structures for each pair of subjects or the overlaps for corresponding structures (or groups of structures). The raw volumes of structures already represent per-subject feature data and can be used directly for classification. Spectral analysis techniques are used to convert pairwise measures of similarity between subjects into per-subject features that can be classified using standard clustering techniques. A schematic of the analysis framework is shown in Fig. 1.

Label fusion for automated segmentation

Label fusion is a generalisation of atlas-based segmentation (Iosifescu et al., 1997; Svarer et al., 2005) which uses image registration to obtain multiple candidate segmentations of a target from an atlas repository. The final labeling is obtained by applying a voting rule to the set of candidate labels at each voxel. We applied a simple per-voxel majority voting rule, which produces results which compare well with other atlas-based methods (Rohlfing et al., 2004). Label fusion approaches are less susceptible to errors from registration, atlas labeling or anatomical variation than single atlas methods. When applied to the segmentation of MR images of the human brain, label fusion has been shown to be robust and accurate, achieving levels of accuracy comparable with expert manual raters (Heckemann et al., 2006). Large numbers of atlases, however, represent a significant computational burden and the resulting segmentation can tend towards the population mean rather than the target anatomy. This motivates the use of a scheme for selecting the most appropriate atlases from a larger repository (Aljabar et al., 2007). Therefore, in this work we adopted the following strategy:

- Atlas selection
  - Affinely register the atlas and target images to a common reference image.
  - Rank the atlas images based on their similarity with the target.
  - Select the n top-ranked atlases.
- Segmentation
  - Non-rigidly register the selected atlases with the target.
  - Propagate labels to target.
  - Fuse using the vote rule.

The reference space was defined by the MNI single subject atlas (Cocosco et al., 1997). Normalised mutual information (Studholme et al., 1999) was used to assess the similarity of atlases with the target over a region of interest encompassing the labels. The top 20 atlases were selected for the label fusion step. We found that label fusion could underestimate the size of internal CSF spaces for subjects with large ventricles or increased parahippocampal CSF. Therefore a post-fusion correction step based on tissue classification was also applied. Specifically, an expectation maximisation (EM) based tissue classification (Leemput et al., 1999; Murgasova et al., 2006) was used to generate tissue probability maps for grey and white matter and for cerebro-spinal fluid (CSF). Regions identified as brain-tissue by label fusion that were subsequently assigned a high probability (>0.75) of being CSF by the EM approach were re-labeled as CSF. This is particularly important for applications in dementia. Examples of segmentations generated using the full fusion procedure for an AD and a control subject are shown in Figs. 2 and 3.

Inter subject measures of morphological similarity

Automated morphological analysis of groups requires measures which quantify the morphological similarity of corresponding structures between pairs of subjects. The low-order morphological similarity measures used in this work were derived from measures of overlap of corresponding structures. We also tested similarity measures based on simple volumetric differences of corresponding labeled structures to determine whether label overlap resulted in more powerful classifiers.

Label volume difference

We define a normalised similarity measure between subjects from the difference in volume of corresponding structures. Raw volume differences must be transformed to normalised similarities before use in a spectral analysis step. The volumes of each structure over N subjects after affine alignment, $s_1, \ldots, s_N$, are first transformed to z-scores, $s'_1, \ldots, s'_N$, by subtracting the mean and dividing by the standard deviation. The normalised measure of volumetric similarity between subjects i and j is then

$$v_{ij} = \frac{1}{c}\exp\left(-\frac{(s'_i - s'_j)^2}{c^2}\right)$$

where $c=2$ specifies the constant kernel width. An equivalent formulation would use the raw volumes in the similarity measure with the kernel width, c, varying in order to reflect structure-specific variation in volume distributions. A general description on the use of the Gaussian form as a neighbourhood function can be found in von Luxburg (2007).

Label overlap

Overlap measures are one of the simplest measures of morphological similarity between paired instances of structures. Because they incorporate location information they provide a higher level of morphological description than volume differences. In this work the Dice overlap coefficient was used. If $N(A)$,$N(B)$ and $N(A\cap B)$, and represent the volumes of two labels and their intersection, then the Dice coefficient is defined as:

$$d = \frac{2N(A \cap B)}{N(A) + N(B)}$$

Simple Dice overlaps compare a single pair of labels. When comparing two brains, the overlaps between several different labeled structures may be a more sensitive indicator than comparing each individual structure in turn. Generalized overlap measures which summarise the agreements of multiple labels in terms of the total intersection and total mean volume were defined in Crum et al. (2006). The generalized Dice coefficient is given by

$$d = \frac{2\sum_i \alpha_i N(A_i \cap B_i)}{\sum_i \alpha_i (N(A_i) + N(B_i))}$$

The weights, α_i, control the relative impact of small versus large labels. Choosing α_i as the inverse square of the average volumes A_i and B_i (Crum et al., 2006) makes the label pair contribute to the overall overlap in inverse proportion to its volume. Simple and generalized overlaps both represent pairwise measures of similarity between subjects and can therefore be used directly within a spectral analysis step. Overlap measures are computed once confounding translational, scaling and rotational factors have been removed. When brain images are normalised using globally rigid or affine transformations, residual local translational or rotational effects may masquerade as morphological differences. A local normalisation can remove these effects; in this work we applied a low-resolution B-spline normalisation to remove confounding effects while preserving genuine morphological differences.

Spectral approaches for classification

Spectral analysis is used to convert the pairwise similarity measures described in the Inter subject measures of morphological similarity section to per-subject features for use in classification. First a complete, undirected, weighted graph which summarises the morphological similarity between all pairwise combinations of N subjects is constructed. Each node represents a subject and the edge weight connecting two nodes is a measure of similarity between those two subjects. Spectral analysis generates indicator (feature) vectors from the eigenvectors of the Laplacian matrix associated with the graph. These features summarise the group similarity structure and are used to partition the cohort into two sub-groups. Technical details of the spectral analysis approach are in the Appendix. The features are clustered using conventional algorithms to assign group membership to each subject. We can choose the dimension of generated features, k. A standard choice for binary partitioning problems is $k=1$ which generates "Fiedler" scalar features. When we consider classification using multiple structures we generate Fiedler features for each structure and concatenate them to form a multi-structure feature vector. Alternatively, overlap data from multiple structures can be combined using generalized overlaps to generate a single Laplacian for all structures.

Supervised and unsupervised classification

Classification methods may be broadly divided into supervised (trained) and unsupervised (untrained) approaches. Both approaches require a feature or vector of features associated with each subject in the study group. Supervised classification derives a classification rule from training data so that variation in morphological features in relation to the known group assignments is learned. We used the Fisher Linear Discriminant (FLD) (Fisher, 1936) method which maximises inter-group variance and minimises intra-group variance. Performance was assessed using leave-one-out cross-validation. Supervised classification is used in two parts of this work. First, to assess which structures are best at discriminating normal controls from subjects diagnosed with dementia so that these structures can be further investigated. Second, as a benchmark for unsupervised classification based on spectral analysis.

For unsupervised classification following spectral analysis, we used the fuzzy c-means clustering technique (Dunn, 1973; Bezdek, 1981) which generalizes the well-known k-means clustering (MacQueen, 1967). This assigns each subject to one of two possible clusters which do not themselves have a clinical label. Therefore we label the clusters as "normal" or "AD" respectively based on the majority clinical label of subjects present in each cluster. We can then compute sensitivity and specificity in the standard way for comparison with the supervised case.

Atlas and application data

Two sets of images were used in this work, a pre-labeled atlas pool for label fusion, and an application study group. We used data from 275 anonymised subjects provided by David Kennedy of the Centre for Morphometric Analysis for the atlas pool. A subset of these images are publicly available as part of the Internet Brain Segmentation Repository[1]. This database was constructed from cohorts used in previous clinical research studies and included male and female subjects, of varying ages, left and right-handed, and with varying numbers designated "normal", "Alzheimer", "schizophrenic", "cocaine-user", "ADHD" and "psychotic". Each subject had sub-cortical manual labels of the following structures: Lateral ventricle, thalamus, caudate, putamen, pallidum, hippocampus, amygdala, accumbens, hippocampus, and brainstem.

Our study group comprised 38 subjects diagnosed with probable Alzheimer's disease and 19 age-matched controls. The subject selection criteria were age >55 with Mini-Mental State Exam (Folstein et al., 1975) ≥27 for controls or in the range 13–26 inclusive for probable AD. The gender match was (women/men+women) 23/38 (AD) and 10/19 (controls). The group ages were (mean±standard deviation) AD 69.8±7 years and controls 69.3±7 years. The AD/control MMSE scores were 19.5±4.0 and 29.5±0.7 respectively. More detail about the cohort can be found in Schott et al. (2005).

Software implementation

Individual registrations for label fusion were performed using software from the Image Registration Tool Kit[2] (IRTK) (Rueckert et al., 1999). The spectral analysis and Fisher discriminant software were implemented in MatLab[3] (The Mathworks, Inc. Natick, 01760-2098, MA, USA). All other significant software was written by the authors.

[1] http://www.cma.mgh.harvard.edu/ibsr.
[2] http://wwwhomes.doc.ic.ac.uk/~dr/software.
[3] http://www.mathworks.com.

Experiments

Investigation of age-related bias in automated segmentation

The study group demographic is not typical of the atlas pool. In order to ensure that label fusion was not biased for or against this group, a separate experiment was carried out using only the atlas pool. The 275 subjects were divided into older subjects, some of whom were AD patients, and other subjects. There were 248 subjects aged below 60 years and 27 subjects aged 60 years and above. Of the older group four subjects were diagnosed with AD and a further four subjects were diagnosed with mild cognitive impairment (MCI) converting to AD at time of scan. Label fusion segmentations were generated for all subjects on a leave-one-out basis. The Dice overlap of the automatically obtained structural labels with each subject's pre-existing manual labels allowed a comparison across the different age groups.

Application experiments

Label fusion segmentation was applied to each subject using the labels selected from the atlas pool. Normalised Laplacian matrices were constructed based on the subjects' pairwise volume and label similarities. The Fiedler (second) eigenvector, $\boldsymbol{v}_2$, of the Laplacian was used to generate a scalar feature indicator for each subject. Three sets of classification features were used:

1. Label volumes after affine alignment of the subjects.
2. Feature data from a spectral analysis of similarities derived from volume differences.
3. Feature data from a spectral analysis of label overlaps.

Group discrimination associated with each structure

The ability of each individual structure to separate the AD group from the control group was assessed on a leave-one-out basis using *t*-tests applied for each structure and each classification feature above. The five most discriminating structures (those which resulted in the largest absolute t-values in a supervised classification experiment) were identified and ranked for each cohort of $n-1$ subjects and used to classify the nth subject to establish a benchmark supervised classification rate. To allow comparison between the supervised and unsupervised approaches in subsequent experiments we defined the population-wide five most discriminating structures as those which had the minimum summed rank over all n leave-one-out classifications. We compared classification performance in the experiments below using either (a) all structures or (b) the five population-wide most discriminating structures.

In a separate experiment to examine the effect of combining features, generalized overlap measures were calculated by aggregating overlaps from multiple structures. The aggregated overlaps were used to generate a Laplacian matrix and t-tests were applied to the resulting Fiedler vector components. This was repeated twice, once using all the structures and once using the five population-wide top-ranked structures.

Supervised and unsupervised classification

Classification clusters each subject into one of two groups using either supervised (clinical status known) or unsupervised (clinical status unknown) methods. We further explored classification using label overlaps depending on whether (a) overlaps were computed individually and the resulting scalar Fiedler features combined during the classification stage or (b) generalized (combined) overlaps were computed and the resulting scalar Fiedler features from spectral analysis classified directly. The significance (p values) of the supervised classification rates was estimated using the permutation-based approach described by Golland and Fischl (2003). The clinical labels for the subjects were permuted and a full leave-one-out cross-validation was carried out for the supervised classifier to give the classification rate associated with the permutation. This was repeated 10,000 times to estimate the distribution of the classification rate which is, in turn, used to estimate the significance of the observed classification rate (without permutation). Separate permutation tests were carried out for each supervised classification experiment.

A standard permutation-based significance test cannot be applied to unsupervised classification as the group membership is determined

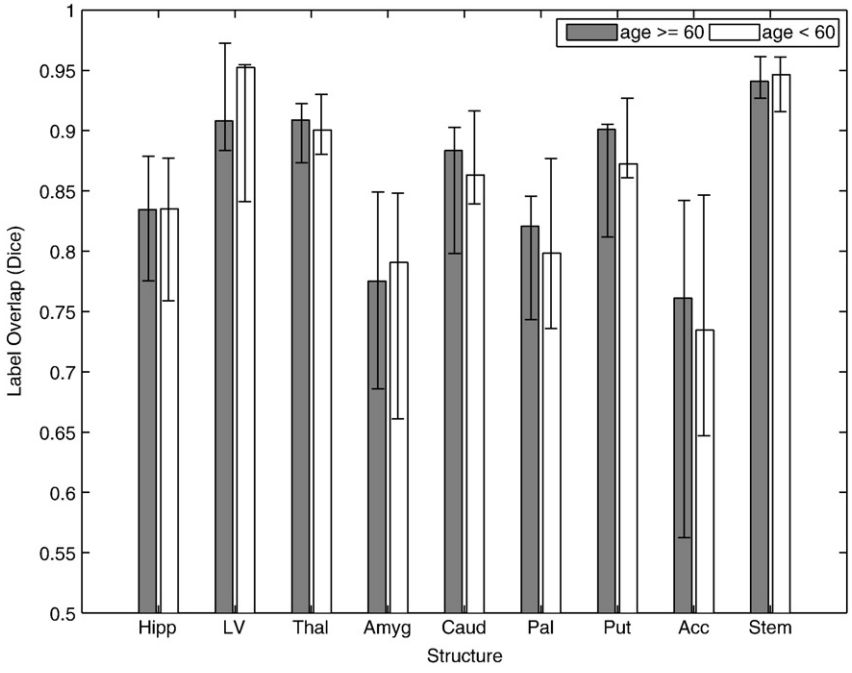

Fig. 4. Label fusion segmentation accuracy in 275 subjects split into older (including those with Alzheimer's disease) and younger sub-groups. Bars indicate the distributions of the Dice values for each combination of structure and age group—this is indicated by the 5th and 95th percentiles of the corresponding Dice values. Key: LV: lateral ventricle; Thal: thalamus; Caud: caudate; Put: putamen; Pal: pallidum; Hipp: hippocampus; Amyg: amygdala; Acc: accumbens; Stem: brainstem. Left right structures are averaged.

Table 1
The group separation performance of individual structures and corresponding t-statistic ranges computed during leave-one-out testing

Label	Volumes			Differences			Overlaps		
	Rank(S)	t_{min}	t_{max}	Rank(S)	t_{min}	t_{max}	Rank(S)	t_{min}	t_{max}
Hipp-L	4(266)	2.80	3.77	1(60)	3.37	3.89	1(58)	5.08	5.80
Hipp-R	2(125)	3.27	3.93	3(168)	2.70	3.19	2(166)	4.57	5.27
LV-R	8(466)	2.26	2.70	14(828)	0.33	0.72	3(174)	3.96	4.57
Thal-L	1(58)	3.75	4.67	2(133)	2.89	3.37	4(255)	3.37	3.89
Amyg-R	6(340)	2.47	3.20	9(486)	1.63	2.05	5(285)	3.19	3.72
Thal-R	3(172)	3.02	4.07	13(734)	0.65	1.04	6(330)	3.19	3.72
LV-L	10(578)	1.89	2.33	5(303)	2.20	2.69	7(437)	2.30	2.77
Acc-R	9(528)	1.88	2.76	4(220)	2.45	2.90	10(566)	2.25	2.70
Pall-R	5(269)	2.80	3.54	7(449)	1.63	2.05	11(622)	2.05	2.50

The features used were raw volumes (left) or Fiedler vectors derived from volume differences (middle) or overlaps (right). The overall top five ranked structures for each feature are shown with their ranking and total summed rank over all leave-one-out experiments as described in the section on group discrimination. Structures are presented in order of their overall ranking for experiments using overlap data. Abbreviations are: Hipp: hippocampus; LV: lateral ventricle; Thal: thalamus; Amyg: amygdala; Acc: accumbens; Pal: pallidum. L = left and R = right.

by the algorithm. There is no standard method for computing significance in this case. With larger study groups one could design a boot-strap procedure based on repeatedly resampling a cohort for unsupervised classification from a larger pool. We leave this to future work.

Results

Investigation of age-related bias in automated segmentation

The differences in segmentation accuracy between older subjects and the remaining subjects in the atlas pool are shown in Fig. 4. The mean Dice accuracy value for each combination of structure and age group is shown along with the 5th and 95th percentiles. Between groups there is good agreement in the distributions of Dice values; systematic bias with respect to either age group is not apparent.

Application experiments

Group discrimination associated with each structure

Table 1 summarises the overall top-ranked structures after leave-one-out classification testing for each feature and the range of associated t-statistic values. The discriminatory ability of structures depends to some extent on the features used for classification. However, as might be expected for this cohort, the hippocampi are consistently ranked highly as discriminators and result in the largest *t*-statistic values (range [5.08–5.80]) when overlap-derived features are used.

The leave-one-out classification rates computed for features which combine the individual top five ranked structures were 0.74, 0.71, and 0.72 for volumes, volume differences and overlaps respectively. These rates should be compared with those obtained using the population-wide top-ranked structures in the Supervised and unsupervised classification section.

Table 2
T-statistics based on Fiedler vector components derived from aggregated overlap Laplacian matrices

Structures	*T*-statistic	*p*-value
All	2.8998	0.005
Selection	7.2256	$<10^{-4}$

The significances (values) were estimated using permutation tests (see text). Two aggregated label cases were analysed: (i) aggregation of all labels (ii) aggregation of the best individually discriminating labels. (left hippocampus, right hippocampus, right lateral ventricle, left thalamus, right amygdala). See Table 1 final column and Fig. 5.

Table 3
Sensitivity, specificity and classification rate when using feature vectors representing volumes (*V*) or cluster indicator variables derived from volume differences (*D*) or from overlaps (O)

Combination	Specificity			Sensitivity			Rate		
	V	*D*	*O*	*V*	*D*	*O*	*V*	*D*	*O*
sup-all	0.74	0.58	0.89	0.72	0.69	0.77	0.72	0.66	0.81
sup-sel	0.79	0.74	0.89	0.74	0.74	0.82	0.76	0.74	0.84
unsup-all	0.79	0.68	0.84	0.79	0.82	0.74	0.79	0.78	0.78
unsup-sel	0.79	0.79	0.89	0.77	0.85	0.82	0.78	0.83	0.84

Experiments are ordered according to whether the classifier used was supervised (sup) or unsupervised (unsup) and whether all (all) or a selection (sel) of structures were used. See also Fig. 6.

Table 2 shows the discrimination results obtained using features derived from aggregated overlaps. The significances were evaluated using 10,000 random group membership permutations. (see e.g. Benjamini and Yekutieli, 2001) under the null hypothesis that group membership did not affect discrimination. It can be seen that aggregating selected structures improves both the power and the significance of the discrimination.

Supervised and unsupervised classification

The feature vectors for different structures were calculated based on either (a) volumes or (b) Fiedler vectors derived from volume differences or overlaps. These feature vectors either used all the available structures ($k=17$) or used a selection of structures based on the top five structures with respect to group separation performance ($k=5$) (See the Group discrimination associated with each structure section above and Table 1). The resulting sensitivity, specificity and overall classification rate for the different combinations of data type, number of structures used, and type of classifier are shown in Table 3 and in Fig. 6. Significance levels for the supervised case are shown in Table 4. These values should be corrected for multiple (6) comparisons giving a corrected per-comparison *p*-value of ~0.0085. All the significances in Table 4 are within this value with the exception of those using volume-difference features computed over all structures.

Classification based on overlap-derived Fiedler features produced comparable or higher classification rates than volume and volume-difference based classification. The highest classification rates of are for the cases where overlaps from selected structures are used with either

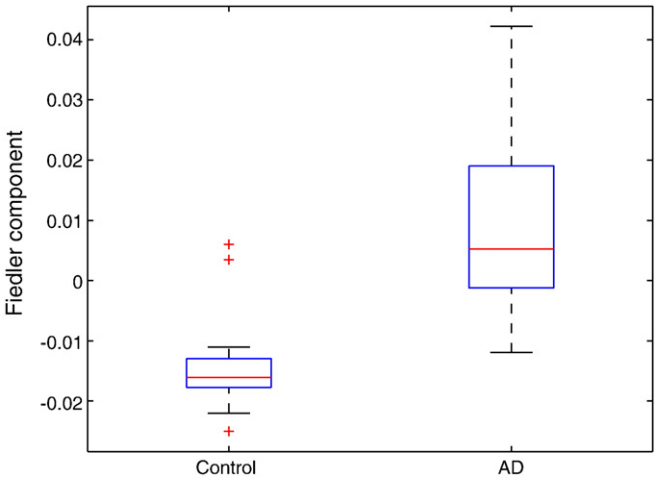

Fig. 5. The group separation obtained by the Fiedler vector components for the control and AD subjects. The components were obtained from a Laplacian matrix derived from generalized Dice overlaps based on a group of structures. The structures selected were those that best separated the groups individually (left hippocampus, right hippocampus, right lateral ventricle, left thalamus, right thalamus, see Tables 1 and 2 and the Group discrimination associated with each structure section).

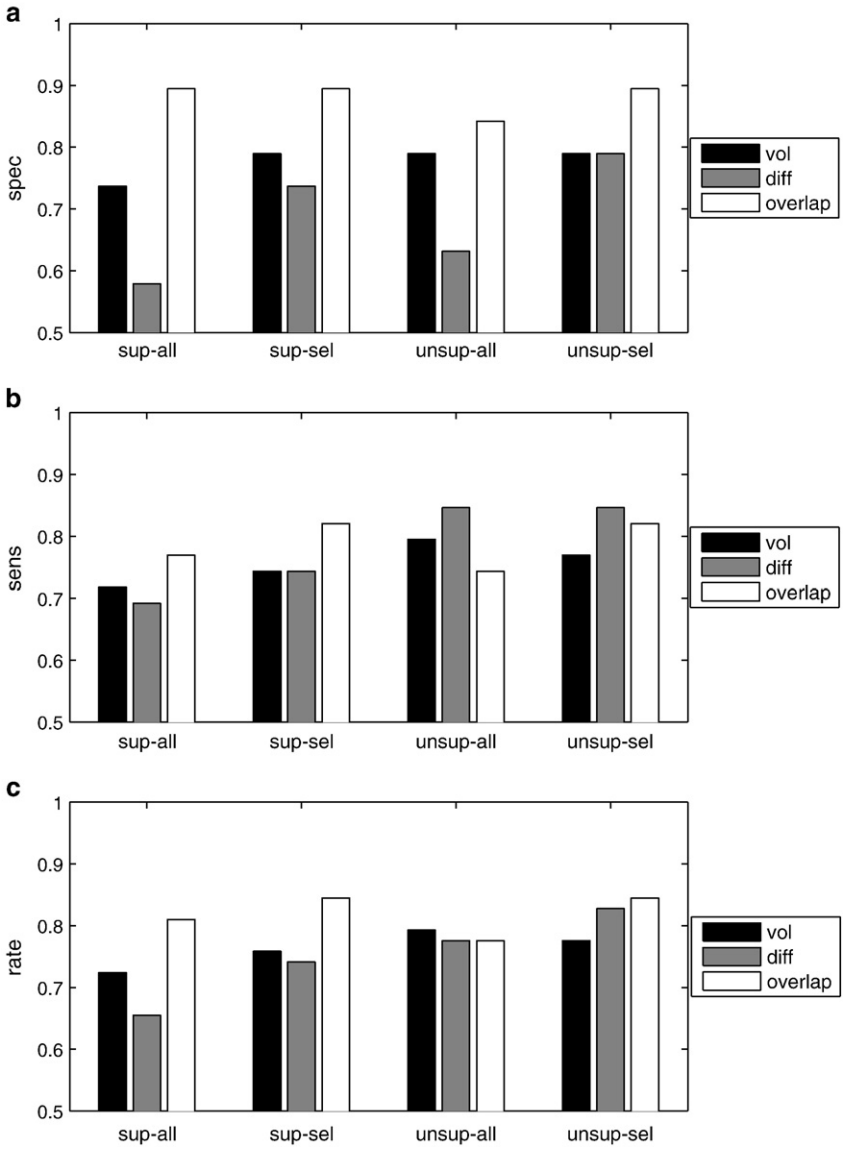

Fig. 6. Classification results distinguishing normal controls from probable Alzheimer subjects. (a) Specificity, (b) Sensitivity, (c) overall classification rates. The shading of each bar indicates the data used for classification: raw volumes, volume differences and overlaps. The volume differences and overlaps required a spectral analysis step. The bars are grouped according to whether the classifier used was supervised (sup) or unsupervised (unsup) and whether all (all) structures or a selection (sel) were used. See also Table 3.

a supervised or an unsupervised classifier ('*sup-sel*' and '*unsup-sel*' in the terms of Table 3). This suggests that most of the important morphology for distinguishing groups has been correctly identified by the initial group separation analysis. We further analysed the incorrectly classified subjects. In both cases two subjects were consistently false positive (normal classified as dementia) and five subjects were consistently false negative (dementia classified as normal) with two additional different false negative subjects for each classification method. We looked at the volumes of the selected structures in each case and found no consistent pattern that differentiated the shared misclassified subjects with the exception of one of the false positive subjects. For this individual, the left lateral ventricle volume was within one standard deviation of the control group mean while the right lateral ventricle volume was more than three standard deviations above the mean. No other evidence from imaging or clinical follow-up was found to explain the mis-classification. In summary, the use of overlaps outperforms volumes and volume differences in these classification tasks; this is consistent with the view that overlaps provide higher order morphological descriptions.

Table 4
Values of p estimating the significance of the classification rates obtained by the supervised classifiers shown in Table 3

Combination	V	D	O
sup-all	0.0036	0.0389	0.0001
sup-sel	0.0004	0.0033	$<10^{-4}$

The p values were obtained by permuting the subject labels 10,000 times in order to estimate the distribution of the classification rate in each case (see text).

Table 5
The classification performance based on a cluster indicator variable taken from a single aggregated overlap Laplacian is compared with the performance of vectors derived from separate Laplacians for the top five structures with respect to group separation

Data	Classifier	Specificity	Sensitivity	Rate
Separate structure overlaps	Supervised	0.89	0.82	0.84
	Unsupervised	0.89	0.82	0.84
Aggregated overlaps	Supervised	0.89	0.69	0.76
	Unsupervised	0.89	0.69	0.76

The figures in the top two rows of the table are taken from the overlaps (O) sup-sel and unsup-sel cases in Table 3.

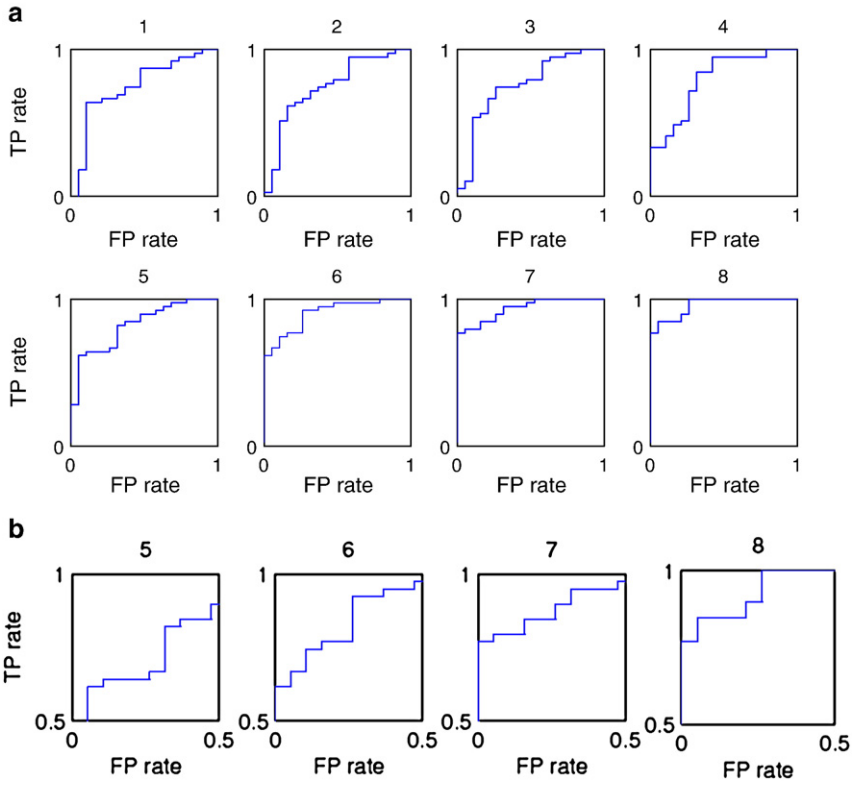

Fig. 7. (a) ROC curves assessing the discrimination ability of varying numbers of eigenvectors selected from the aggregated overlap Laplacian. The number of eigenvectors used (starting with the Fiedler vector) is shown above each chart. (b) Enlarged versions of the upper right quarter of the corresponding plots in (a).

We repeated the experiments above using Fiedler features derived from a spectral analysis of generalized overlap of structures. The results shown in Table 5 indicate that the single aggregated overlap Laplacian (correct classification rate=0.76) does not perform as well as the vectors of indicator variables from per-structure Laplacians reported above. This result is, perhaps, not surprising given that the classification based on aggregated overlaps is derived from a single eigenvector from a single Laplacian in contrast to the use of separate structure overlaps, where the classification is derived from Fiedler vectors taken from multiple Laplacian matrices. However, the similarity structure of the cohort can be more completely represented if further eigenvectors from the aggregated overlaps Laplacian are used to form feature vectors. This motivated an investigation of the effect of using different numbers of ordered eigenvectors for clustering.

The effect of the number of Laplacian eigenvectors

We used a combination of eigenvectors $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$, to generate a vector feature indicator for each subject. The data represented by these feature vectors were projected onto a single dimension using the matrix generated by the Fisher linear discriminant model. The resulting univariate data was then assessed for its group discrimination ability using a receiver operator characteristic (ROC) curve. The ROC curves based on the selection of up to eight eigenvectors ($k=2,\ldots,9$) are shown in Fig. 7. The discrimination performance of feature data derived from the Laplacian generally improves as the number of eigenvectors used increases. The area under the ROC curve ranges from 0.75 (1-eigenvector) to 0.96 (8-eigenvectors). Table 6 shows that the use of eight eigenvectors taken from a Laplacian based on aggregating overlaps for the top five structures (based on t-statistics) gives classification rates of up to 0.92. These findings confirm there is benefit for classification accuracy in using higher order features from the aggregated overlap Laplacian.

Table 6
The classification performance based on the use of 8 eigenvectors taken from a single Laplacian derived from the aggregated overlaps

Classifier	Sensitivity	Specificity	Rate
Supervised	0.89	0.84	0.86
Unsupervised	0.89	0.92	0.92

These overlaps were obtained using the top 5 structures with respect to group separation (see Table 1). A significance estimate $p<10^{-4}$ of was obtained for the classification rate of the supervised classifier using a permutation test.

Discussion

The principal motivation for this work was to determine whether a relatively simple label fusion approach to automated brain labeling could be applied in conjunction with supervised and unsupervised classification to extract meaningful relationships between clinically normal and abnormal subjects imaged as part of a dementia study. It is clear from the results above that the combination of label fusion and spectral analysis has significant value in this application.

The analysis pipeline presented in this paper is intended to be generic. One specific requirement for it to be applied to a new population is that the automated segmentation via label fusion be validated on the new data. In our case we drew atlas images from a varied population and carried out a validation study and automatically corrected for the occasional mis-classification of CSF. For investigations of conditions which result in subtle changes in neuroanatomy we believe a generic atlas population should suffice. However bias and error could easily be introduced if unsuitable generic pools are applied to subject groups who lie outside or at the extremes of the demographic covered by the pool. To reduce error, particularly where significant departures from normality are expected, either large atlas pools coupled with selection strategies which include other clinical information (e.g. age, sex, MMSE, other medical factors), or more specific pools that focus on specific clinical, ethnic, or age-banded populations could be used. Either strategy might have removed the need for our explicit CSF correction step. In an analogous way to the use of custom-templates in SPM, it may also be

possible to boot-strap more accurate label fusion by iterating the registration and label fusion steps to create a customised atlas population. With the use of a selective atlas strategy, the approach is eminently scalable to large studies. The principle computation cost apart from assembling the pool of candidate atlases is to non-rigidly register each selected atlas to each image in the study. Established non-rigid registration algorithms produce results of sufficient accuracy in a few minutes to a few hours and the combination of constantly evolving computing power and optimised registration algorithms means that analysis of studies of the order of hundreds of subjects is feasible with resources available today. We note that registering atlases to images is trivially parallelizable and there are no significant computational challenges in applying this technique to large cohorts.

Label fusion has previously been shown to have high accuracy compared with other automated segmentation approaches (Heckemann et al., 2006). Dawant et al. (1999) used a registration and atlas approach to automated segmentation and quoted Dice overlaps ~0.84 for the caudate. Fischl et al. (2002) described an automated labeling approach based on anisotropic non-stationary Markov Random Fields and compute Dice overlaps between manual and automated segmentations of 7 subjects on a leave-one-out basis for evaluation. Our charted overlaps in Fig. 4 compare well with Fig. 2 in Fischl et al. (2002) and therefore also in terms of expected manual labeling error. Zhou and Rajapakse (2005) apply a fuzzy template approach driven by image registration to automate segmentation of sub-cortical structures. The method models intensity variation, location and spatial relationship across structures to introduce population anatomical context. Quantitative evaluation used 18 publicly available brain images from the Internet Brain Segmentation Repository, (which form a subset of our much larger training population). A Dice-like overlap measure resulted in mean overlaps of 0.81 (caudate), 0.83 (putamen), 0.84 (thalamus), 0.71 (hippocampus) and 0.65 (amygdala). Although the experiments cannot be compared directly, our overlap results in Fig. 4 computed in a similar way on similar data are higher for each structure. In this work, we have used a simple vote rule to fuse labels and generate segmentations. It is possible to use other more sophisticated schemes, for example STAPLE (Warfield et al., 2004). We plan to make a comparison of vote rule fusion and STAPLE in future work.

While spectral analysis techniques are generic and powerful it is known that their performance can be sensitive to the manner in which the edge-weights (inter-subject similarities) are defined, and the connectivity of the Laplacian. At present there is little theoretical guidance on the optimal way to apply spectral analysis to a new dataset. We followed the advice of von Luxburg (2007) by using the normalised Laplacian, a Gaussian form for the volume similarity, a normalised measure of label overlap, and a fully connected graph. This strategy remains computationally practicable for studies of several thousand subjects; the largest computational cost is in achieving the labeling of each subject rather than the subsequent classification of those subjects. We solved a simple two-way partitioning problem where subjects were simply classified as “normal” or “abnormal”. Where further sub-groups are expected (e.g. normal, MCI and AD or normal AD and VaD) some consideration of the distribution of similarities across groups and the appropriate degree of connectedness in the connectivity graph may be necessary. It is possible in principle to determine the appropriate number of clusters post-hoc from an analysis of the eigenvalue distribution of the Laplacian; we leave this for future work. Spectral analysis was well suited to our application but there are other techniques which make use of pairwise similarity comparisons which could be used. One example is Hierarchical Clustering where pairwise comparisons of points are iterated to construct hierarchies of connected components. See Jain et al. (1999) for a survey of this and other techniques. A detailed examination of both theory and practical application of competing techniques in this application is required but is beyond the scope of the present work.

Initial analysis of individual structure labeling showed that the most sensitive structures for distinguishing controls from AD by volume were the lateral ventricles, the hippocampus and the thalamus. Volumetric change in the lateral ventricles is a well-reported and powerful marker of atrophy (Schott et al., 2005). The hippocampi are thought to be implicated with early disease processes and hippocampal atrophy is also a well-reported sign associated with dementia. Volumes of the ventricles, hippocampi and thalamus obtained by automated labeling were also shown to distinguish control subjects from AD (Fischl et al., 2002). It is thought that both Alzheimer's disease and vascular dementia carry an increased incidence of vascular lesions in sub-cortical regions and increased atrophy in the thalamus in dementia has recently been reported by two deformation based automated techniques (Teipel et al., 2007; Cardenas et al., 2007). Fischl et al., (2002) also found that amygdala volumes could distinguish groups with the right amygdala producing better classifications. Our volumetric results agreed with this trend but other structures (e.g. the pallidum) were better classifiers. However, when overlaps were considered the classification rate of both amygdala increased markedly with the right amygdala comparable with the thalamus.

The comparison of supervised versus unsupervised classification requires some careful interpretation. First, the unsupervised classification is at liberty to classify on the basis of any morphological trends it can discover in the data. Our hypothesis is that structural changes due to dementia provide the strongest basis for classification in the cohort. We compared classification performance with features derived from a generic label set and labels found to be individually good discriminators. We found that unsupervised classification rates were highest and equal to supervised classification results when the overlap of these selected structures was used for classification (Table 3). This demonstrated that overlap added significant power beyond simple volumetry and that for this cohort at least, knowledge of the clinical status of each subject was not necessary for good classification provided an appropriate structural basis for classification was chosen. In this study we chose the basis empirically, however these labels could have been selected on the basis of previous reports in the literature as they are consistent with known characteristics of Alzheimer's disease. The unsupervised classification rate determined from volumes, volume differences and overlaps was high and comparable when all structures were used for classification. This suggests that disease-related changes were the most important factors overall for classification. However morphological information from label overlaps compared with volumetric measures did not improve classification when all structures were included in the classification. Therefore, at least in this cohort, morphological variation unrelated to disease processes was confounding the classification and a hypothesis for selecting structures whose morphology is most likely to reflect pathological processes is necessary to obtain maximum classification power. We also found (Table 6) that the highest classification rate of all was obtained by (a) including further spectral analysis components in the classification to capture more of the structural relationships in the data and (b) by applying spectral analysis to a generalized label overlap. Again these results were obtained using the pre-selected structures. It is interesting to compare our classification rates with two recent studies using Support Vector Machines for supervised classification. Klöppel et al. (2008) analysed normalised Grey Matter segments and achieved a supervised classification rate in a cohort of normal controls and post-mortem confirmed Alzheimer's disease subjects. This is comparable with our best unsupervised classification rates but we note in passing that our Alzheimer subjects had slightly higher MMSE, were 10 years younger on average and were not post-mortem confirmed cases. Fan et al. (2008) achieve equally impressive supervised classification rates (~94%) of control and AD subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI).

Both these studies use voxel-wise techniques which can be considered complementary to our use of distinct morphological features.

It is likely to remain the case that longitudinal studies are the most powerful way to study disease-related morphology in the individual (Fox and Freeborough, 1997; Smith et al., 2002). However longitudinal studies are expensive, prone to drop-out and for the individual, stressful and inconvenient. For diagnostic purposes, the inter-scan interval remains a compromise between size of measurable morphological effects, normal physiological variation, and a desire to obtain a result quickly. Therefore it is likely that single time-point exams be they MR-related or via some other measurement technique will remain important. We have shown that new techniques in automated segmentation and classification, which are readily available and scalable, can be applied to these problems and show potential for use in novel analysis pipelines.

Acknowledgments

P. Aljabar acknowledges support of the EPSRC GR/S82503/01, Integrated Brain Image Modelling project. W.R. Crum acknowledges the support of the Medical Images and Signals IRC (EPSRC GR/N14248/01 and UK Medical Research Council Grant No. D2025/31). W.R. Crum was previously at Centre for Medical Image Computing, University College London, London, UK. The authors thank Professor Nick Fox, Dementia Research Centre, Institute of Neurology, University College London, London UK for the application data and David Kennedy and the Centre for Morphometric Analysis at MGH for the labeled training data. Brian Patenaude of The Oxford Centre for Functional Magnetic Resonance Imaging of the Brain wrote the Fisher linear discriminant software.

Appendix. Normalised spectral analysis

A brief description of the essential steps of the specific normalised spectral analysis approach adopted follows; see Ng et al. (2002) for more detail. For N subjects, a $N\times N$ matrix $\mathbf{W}$ of pairwise similarity values is evaluated, where $\mathbf{W}=(w_{ij}), i,j=1,\ldots,N$ and w_{ij} represents the similarity of subjects i and j. These pairwise similarities are assumed to correspond to weights on the edges of a complete graph in which the subjects are represented by the nodes. The diagonal degree matrix $\mathbf{D}$, which measures the total similarity between each subject and all others, is constructed from $\mathbf{W}$ by summing the edge-weights along each row, $D_{ii}=\sum_{j=1}^{N} w_{ij}$. $\mathbf{D}$ and $\mathbf{W}$ are used to construct the normalised Laplacian (Chung, 1997) $\mathbf{L}$, where $\mathbf{L}=\mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-\frac{1}{2}}$, which contains the information required to cluster the subjects. $\mathbf{L}$ is symmetric positive semi-definite and therefore has real non-negative eigenvalues. From the definition of $\mathbf{D}$, it can be shown that the vector $\mathbf{D}^{-\frac{1}{2}}1$ is an eigenvector of $\mathbf{L}$ with eigenvalue zero. It can also be shown that the remaining eigenvalues are all positive (Chung, 2007) and therefore provide an ordering for the corresponding eigenvectors. Let $\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$ represent an ordered selection of eigenvectors starting with the eigenvector corresponding to the first non-zero eigenvalue (i.e. the second eigenvalue). A feature matrix is constructed by taking $\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$ as columns and normalising its rows to one. The rows of this matrix correspond to the original subjects and are used as feature vectors in a clustering algorithm. The features become scalar cluster indicator variables if only the first of these eigenvectors–the 'Fiedler vector' (Chung, 2007)–is used. The reasons why the components of the Laplacian eigenvectors can be used in this way are non-trivial and we refer to von Luxburg (2007) who presents multiple interpretations of this process including one that adopts a graph-cuts perspective.

References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2007. Classifier selection strategies for label fusion using large atlas databases. In: Tenth Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '07). vol. 4791 of Lecture Notes in Computer Science. pp. 523–531.

Ashburner, J., Friston, K., 2000. Voxel-based morphometry — the methods. NeuroImage 11 (6), 805–821.

Ashburner, J., Hutton, C., Frackowiak, R., Johnsrude, I., Price, C., Friston, K., 1998. Identifying global anatomical differences: deformation-based morphometry. Hum. Brain Mapp. 6, 638–657.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29 (4), 1165–1188.

Bezdek, J., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press.

Bookstein, F., 2001. "Voxel-based morphometry" should not be used with imperfectly registered images. NeuroImage 14 (6), 1452–1462.

Cardenas, V., Boxer, A., Chao, L., Gorno-Tempini, M., Miller, B., Weiner, M., Studholme, C., 2007. Deformation-based morphometry reveals brain atrophy in frontotemporal dementia. Arch. Neurol. 64 (6), 873–877.

Chetelat, G., Baron, J.-C., 2003. Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. NeuroImage 18 (2), 525–541.

Chung, F.R.K., 1997. Spectral Graph Theory. American Mathematical Society.

Cocosco, C., Kollokian, V., Kwan, R.-S., Evans, A., 1997. BrainWeb: online interface to a 3D MRI simulated brain database. NeuroImage 5 (4), S425.

Crum, W., Camara, O., Hill, D., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imag. 25 (11), 1451–1461.

Csernansky, J., Joshi, S., Wang, L., Haller, J., Gado, M., Miller, J., Grenander, U., Miller, M., 1998. Hippocampal morphometry in schizophrenia by high dimensional brain mapping. Proc. Natl. Acad. Sci. 95 (19), 11406–11411.

Davatzikos, C., 2004. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. NeuroImage 23, 17–20.

Dawant, B., Hartmann, S., Thirion, J.P., Maes, F., Vandermeulen, D., Demaerel, P., 1999. Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations: Part I, methodology and validation on normal subjects. IEEE Trans. Med. Imag. 18 (10), 909–916.

Du, A., Schuff, N., Amend, D., et al., 2001. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J. Neurol. Neurosurg. Psychiatry 71 (4), 441–447.

Duncan, J., Ayache, N., 2000. Medical image analysis: progress over two decades and the challenges ahead. IEEE Trans. Pattern Anal. Mach. Intell. 22 (1), 85–106.

Dunn, J., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. 3, 32–57.

Fan, Y., Batmanghelich, N., Clark, C., Davatzikos, C., 2008. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39, 1731–1743.

Fischl, B., Salat, D., 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., 2002. Whole brain segmentation: automated labeling of neuroanatomical structure in the human brain. Neuron 33 (3), 341–355.

Fisher, N., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7 (II), 179–188.

Folstein, M., Folstein, S., McHugh, P., 1975. "Mini-mental state". a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12 (3), 189–198.

Fox, N., Freeborough, P., 1997. Brain atrophy progression measured from registered serial MRI: validation and application to Alzheimer's disease. J. Magn. Reson. Imag. 7, 1069–1075.

Fox, N., Warrington, E., Rossor, M., 1999. Serial magnetic resonance imaging of cerebral atrophy in preclinical Alzheimer's disease. Lancet 353, 2125.

Fox, N., Warrington, E., Freeborough, P., Hartikainen, P., Kennedy, A., Stevens, J., Rossor, M., 1996. Presymptomatic hippocampal atrophy in Alzheimer's disease: a longitudinal MRI study. Brain 119, 2001–2007.

Fox, N., Crum, W., Scahill, R., Stevens, J., Janssen, J., Rossor, M., 2001. Imaging of onset and progression of Alzheimer's disease with voxel-compression mapping of serial magnetic resonance images. Lancet 358, 201–205.

Golland, P., Fischl, B., 2003. Permutation tests for classification: towards statistical significance in image-based studies. In: Information Processing in Medical Imaging: Proc. 18th International Conference (IPMI'03). Lecture Notes in Computer Science, pp. 330–341.

Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33 (1), 115–126.

Iosifescu, D., Shenton, M., Warfield, S., Kikinis, R., Dengler, J., Jolesz, F., McCarley, R., 1997. An automated registration algorithm for measuring MRI subcortical brain structures. NeuroImage 6 (1), 13–25.

Jack Jr., C., Petersen, R., Xu, Y., Waring, S., O'Brien, P., Tangalos, E., Smith, G., Ivnik, R., Kokmen, E., 1997. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. Neurology 49, 786–794.

Jain, A., Murty, M., Flynn, P., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.

Kim, J., Kim, Y., Choi, S., Kim, M., 2005. Morphometry of the hippocampus based on a deformable model and support vector machines. Artificial Intelligence in Medicine. Vol. 3581 of Lecture Notes in Computer Science, pp. 353–362.

Klöppel, S., Stonnington, C., Chu, C., Draganski, B., Schaill, R., Rohrer, J., Fox, N., Jack Jr, C., Ashburner, J., Frakowiak, R., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689.

Laakso, M., Soininen, H., Partanen, K., Lehtovirta, M., Hallikainen, M., Hanninen, T., Helkala, E.-L., Vainio, P., Riekkinen, S., 1998. MRI of the hippocampus in Alzheimer's disease: sensitivity, specificity and analysis of the incorrectly classified subjects. Neurobiol. Aging 19 (1), 22–31.

Leemput, K.V., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imag. 18 (10), 897–908.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297.
Mangin, J., Riviere, D., Cachia, A., et al., 2004. Object-based morphometry of the cerebral cortex. IEEE Trans. Med. Imag. 23 (8), 968–982.
Murgasova, M., Dyet, L., Edwards, A., Rutherford, M., Hajnal, J., Rueckert, D., 2006. Segmentation of brain MRI in young children. Ninth Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '06). vol. 4190 of Lecture Notes in Computer Science, pp. 687–694.
Ng, A., Jordan, M., Weiss, Y., 2002. On spectral clustering: analysis and an algorithm. Adv. Neural Inf. Process. Syst. 14, 849–856.
Pruessner, J., Li, M., Serles, W., et al., 2000. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb. Cortex 10 (4), 433–442.
Rohlfing, T., Brandt, R., Menzel Jr, R., C. M., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21 (4), 1428–1442.
Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D., 1999. Non-rigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imag. 18 (8), 712–721.
Schott, J., Price, S., Frost, C., Whitwell, J., Rossor, M., Fox, N., 2005. Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months. Neurology 65 (1), 119–124.
Smith, S., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.
Smith, S., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P., Federico, A., Stefano, N., 2002. Accurate, robust, and automated longitudinal and crosssectional brain change analysis. NeuroImage 17 (1), 479–489.
Studholme, C., Hill, D., Hawkes, D., 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recogn. 32 (1), 71–86.
Studholme, C., Cardenas, V., Blumenfeld, R., Schuff, N., Rosen, H., Miller, B., Weiner, M., 2004. A deformation tensor morphometry study of semantic dementia with quantitative validation. NeuroImage 21 (4), 1387–1398.
Svarer, C., Madsen, K., Hasselbalch, S., Pinborg, L., Haugbol, S., Frokjaer, V., Holm, S., Paulson, O., Knudsen, G., 2005. MR-based automatic delineation of volumes of interest in human brain PET images using probability maps. NeuroImage 24 (4), 969–979.
Teipel, S., Born, C., Ewers, M., Bokde, A., Reiser, M., Moller, H., Hampel, H., 2007. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. NeuroImage 38 (1), 13–24.
von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17 (4), 395–416.
Wang, L., Miller, J., Gado, M., McKeel, D., Rothermich, M., Miller, M., Morris, J., Csernansky, J., 2006. Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. NeuroImage 30 (1), 52–60.
Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J., Csernansky, J., Miller, M., 2007. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. IEEE Trans. Med. Imag. 26 (4), 462–470.
Warfield, S., Zou, K., Wells, W., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23 (7), 903–921.
Zhou, J., Rajapakse, J., 2005. Segmentation of subcortical brain structures using fuzzy templates. NeuroImage 28 (4), 915–924.