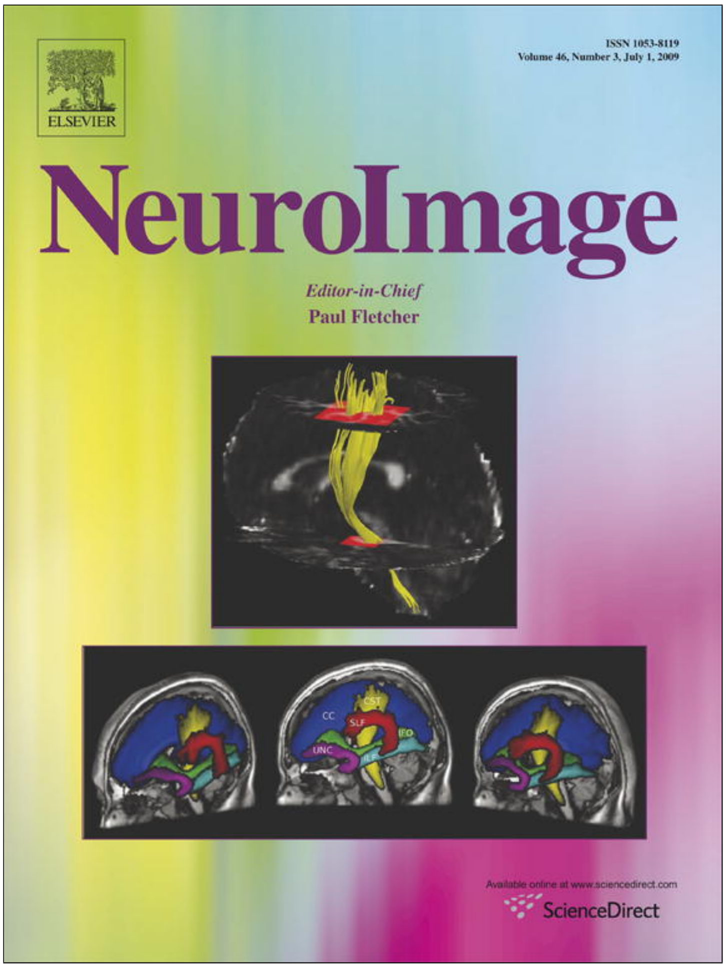
ISSN 1053-8119
Volume 46, Number 3, July 1, 2009
ELSEVIER
NeuroImage
Editor-in-Chief
Paul Fletcher
CST
CC
SLF
IFO
UNC
ILF
Available online at www.sciencedirect.com
ScienceDirect

NeuroImage 46 (2009) 726–738

Contents lists available at ScienceDirect

NeuroImage

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

Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy

P. Aljabar [a,*], R.A. Heckemann [b], A. Hammers [c], J.V. Hajnal [b], D. Rueckert [a]

[a] Visual Information Processing Group, Department of Computing, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK
[b] Imaging Sciences Department, MRC Clinical Sciences Centre, Imperial College London, UK
[c] Division of Neuroscience and Mental Health, MRC Clinical Sciences Centre, Imperial College London, UK

ARTICLE INFO

Article history:
Received 13 November 2008
Revised 22 January 2009
Accepted 7 February 2009
Available online 23 February 2009

Keywords:
MRI
Segmentation
Selection
Atlases
Registration

ABSTRACT

Quantitative research in neuroimaging often relies on anatomical segmentation of human brain MR images. Recent multi-atlas based approaches provide highly accurate structural segmentations of the brain by propagating manual delineations from multiple atlases in a database to a query subject and combining them. The atlas databases which can be used for these purposes are growing steadily. We present a framework to address the consequent problems of scale in multi-atlas segmentation. We show that selecting a custom subset of atlases for each query subject provides more accurate subcortical segmentations than those given by non-selective combination of random atlas subsets. Using a database of 275 atlases, we tested an image-based similarity criterion as well as a demographic criterion (age) in a leave-one-out cross-validation study. Using a custom ranking of the database for each subject, we combined a varying number *n* of atlases from the top of the ranked list. The resulting segmentations were compared with manual reference segmentations using Dice overlap. Image-based selection provided better segmentations than random subsets (mean Dice overlap 0.854 vs. 0.811 for the estimated optimal subset size, $n = 20$). Age-based selection resulted in a similar marked improvement. We conclude that selecting atlases from large databases for atlas-based brain image segmentation improves the accuracy of the segmentations achieved. We show that image similarity is a suitable selection criterion and give results based on selecting atlases by age that demonstrate the value of meta-information for selection.

Introduction

Magnetic resonance (MR) imaging of the brain has established itself as an essential diagnostic method in neurology research and clinical practice. Quantitative studies often rely on the capability to label or segment regions of the brain that have distinctive structural or functional properties. This enables comparisons within and between subjects for determining how such regions are affected by physiological and pathological processes as well as therapeutic measures. Such studies benefit from increasing numbers of MR images becoming publicly available for use in research. This availability has made the creation and maintenance of MR image databases incorporating structural segmentations (manual or otherwise) more feasible. Good examples are the Internet Brain Segmentation Repository[1] and the *LONI Probabilistic Brain Atlas* (Shattuck et al., 2008). An obvious application of this work is the use of expert annotations in the form of prior information to assist in providing automatic segmentations of query or unseen images.

An atlas, in the context of this work, is defined as the pairing of a structural MR scan and a corresponding manual segmentation. Given an atlas, a segmentation for an unseen query subject can be estimated using image registration. The atlas MR image can be registered to the query image, yielding a transformation which allows the atlas segmentation to be transformed and treated as a segmentation estimate for the query subject. Within this process, commonly called atlas-based segmentation, the atlas that is propagated can represent a single segmented individual (Iosifescu et al., 1997; Svarer et al., 2005; D'Haese et al., 2003). The propagation of the atlas might also form a step within a larger framework. For example, probabilistic or 'soft' atlases may be propagated and treated as priors in a Bayesian framework within a further segmentation step (Murgasova et al., 2006).

Sources of error in atlas-based segmentations include registration error, the possibility that the atlas used is anatomically unrepresentative of the query image to be segmented (for example if there are topological differences) or existence of labelling errors in the atlas segmentation, something that cannot be overcome by accurate registration.

If a database of atlases is available, multiple segmentations from a group of atlases can be propagated to the query. After propagation, they can be treated as separate classifiers and fused to form a single consensus segmentation estimate. The main benefit of the multi-atlas

* Corresponding author.
E-mail address: paul.aljabar@imperial.ac.uk (P. Aljabar).

[1] www.cma.mgh.harvard.edu/ibsr.

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.02.018

segmentation approach is that the effect of errors associated with any single atlas propagation can be reduced in the process of combination. Multi-atlas segmentation has been shown to be effective in comparison with other atlas-based approaches (Rohlfing et al., 2004a) and for the task of segmenting structures in the human brain (Heckemann et al., 2006a; Klein and Hirsch, 2005). Relevant work has also been carried out on the methods used for combining classifiers within a multi-atlas segmentation framework (Warfield et al., 2004; Rohlfing et al., 2004b).

Multi-atlas segmentation faces various issues, however, if the number of atlases in the database becomes large. On a practical level, if every atlas is registered with the query image, the computational cost of segmentation increases linearly with the size of the database. More importantly, it is possible that the population represented by the atlases is heterogeneous, for example in terms of age, morphology or pathology. In this case, for a given query, certain subjects in the database may be more appropriate as candidate segmentations than others. Propagating and combining only these subjects' atlases is likely to produce a better segmentation estimate than one that draws on the full atlas database.

These considerations provide a motivation for the selection of atlases that are appropriate for a given query image, and this work presents an investigation of a practical strategy for such a selection approach within the context of multi-atlas segmentation. Rohlfing et al. (2004a) and Wu et al. (2007) investigated the optimal selection of a single template during atlas-based segmentation. Our work contrasts with this in that we select multiple atlases for subsequent propagation and fusion. We present the results of a series of experiments to assess the performance of atlas selection using a database consisting of 275 MR images and accompanying manual subcortical segmentations. Automated segmentation is carried out based on ranking and selecting atlases from a database according to criteria that are expected to predict their suitability for segmenting a given target. To test this, the accuracy of the resulting segmentations is measured using leave-one-out cross-validation and compared with the accuracy of segmentations derived from combining random sets of atlases. We also investigate different criteria for ranking the atlases and the effect of selecting and combining increasing numbers of atlases from a ranked set.

Over the mainly subcortical structures studied, a mean Dice overlap of 0.854 was obtained using selection. This compares with a reference value of 0.811 obtained by fusing random sets of atlases. For individual structures, selection provides typical Dice accuracy gains of 0.02 to 0.05 over random sets with the biggest improvement of 0.12 being shown by segmentations of the caudate nucleus.

In this paper, methods for multi-atlas segmentation and selection are initially described. This is followed by descriptions of the experiments to assess the effectiveness of atlas selection and their results which are discussed in the final section. Part of the research presented in this study appeared previously in a conference paper (Aljabar et al., 2007).

Methods

We describe multi-atlas segmentation along with the motivation and possible strategies for atlas selection. These strategies can be based on image information within the atlases or on subject-specific meta-information.

Background: multi-atlas segmentation

Atlases within a database can be registered to a query image, and their segmentations can be transformed and subsequently fused or combined to provide a consensus segmentation estimate for the query. Sometimes described as classifier fusion or label fusion, this method is illustrated schematically in Fig. 1. This multi-atlas approach to segmentation reduces the effect of errors associated with individual propagated atlases. For example, a registration error for a particular propagated atlas is less likely to affect the final segmentation when combined with other atlases. The proportion of errors incurred during propagation that are independent are those that are averaged out when multiple atlases are combined (Heckemann et al., 2006a). As well as a gain in accuracy, Heckemann et al. (2006a) also demonstrate that precision improves as more atlases are combined.

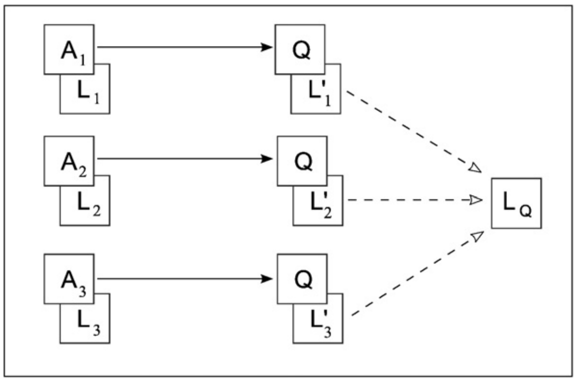

Fig. 1. Schematic illustration of multi-atlas segmentation. A set of atlas anatomical images A_i are registered to the query anatomy Q. The resulting transformations are used to transform the corresponding atlas segmentations L_i to the query. The transformed segmentations L'_i are then combined to create an estimate of the query segmentation L_Q.

The fusion of the propagated segmentations (or classifiers) takes place at the voxel level and can be achieved in different ways. In what is probably the simplest approach, the atlas segmentations are transformed using nearest-neighbour interpolation so that they each provide a discrete or 'hard' labelling for each voxel. The final label assigned to a voxel can then be decided by 'majority vote'.

More sophisticated methods for the combination or fusion of the segmentations are also available. For example it is possible to use a linear interpolator when transforming individual labels in order to obtain a probabilistic or 'soft' estimate for the label from each segmentation. This can be used to generate an array of values (p_{ij}) for a given voxel where p_{ij} represents the confidence level or probability of the voxel being assigned label i by the jth segmentation. A number of different rules can be used to generate a consensus estimate based on such data, and a good overview of these can be found in Kittler et al. (1998).

A notable example of producing consensus segmentations in the context of medical image processing is the STAPLE framework presented by Warfield et al. (2004). The STAPLE approach uses Expectation Maximisation to iterate between the estimation of the 'true' consensus segmentation and the estimation of reliability parameters for each of the raters (which in this work are represented by propagated segmentations). The reliability parameters are based on the sensitivity and specificity of each rater and are used to weight their contributions when generating the consensus estimate. The current consensus estimate can, in turn, be used to measure the reliability of the raters and this forms the basis of the EM iterations.

The use of majority voting for each voxel has, however, been shown to be effective in a number of contexts. Rohlfing et al. (2004a) used a database of images of bee brains to show that fusing segmentations using majority voting is robust and accurate compared with, for example, the propagation of an average shape atlas, or of an individual atlas, selected according to its similarity to the query image. The vote rule has also been shown to perform well relative to other fusion approaches in a more general pattern recognition context (Kittler et al., 1998).

In the context of human brain image segmentation, we have previously presented a series of experiments to investigate the precision and accuracy of structural multi-atlas segmentation using majority voting (Heckemann et al., 2006a,b). These show that multi-atlas segmentation performs at levels of accuracy approaching those of expert human raters with Dice overlap values (against a manual gold standard) increasing from an average of 0.754 for non-rigidly propagated single atlases to 0.836 for the fusion of 29 propagated label sets. The atlases in that work consisted of brain MR images for 30 subjects with corresponding manual segmentations created using the protocol published by Hammers et al. (2003).

A leave-one-out cross-validation approach can be used to assess the accuracy of a multi-atlas segmentation of a query image based on treating the manual segmentation available for that image as a gold standard. The precision of multi-atlas segmentation can be assessed by measuring the agreement of segmentations produced by different subsets of atlases. The agreement between an automated segmentation estimate and a manual segmentation or between a pair of automated segmentations can be assessed using overlap measures. A number of measures are available, with the Dice coefficient (Dice, 1945) being a popular option in the literature. The Dice coefficient is given by

$$d = \frac{2|A \cap B|}{|A| + |B|}$$

where A and B represent the regions of the label being compared and the volumes are measured by voxel counts.

Using the Dice coefficient as a measure, we have shown (Heckemann et al., 2006a) that, under reasonable assumptions, the overlap accuracy of a multi-atlas segmentation estimate, as a function of the number of segmentations fused n, is well modelled by

$$d(n) = a - \frac{b}{\sqrt{n},} \qquad (1)$$

where the constants to be determined, a and b, satisfy $0 \leq a \leq 1$ and $b > 0$.

Eq. (1) models monotonically increasing overlap as more segmentations are fused. This increase is limited by the asymptotic constant a while the constant b determines the rate at which the overlap increases. As more segmentations are fused, overlap accuracy increases because random errors, associated with individual atlas to query registrations and atlas variability, are increasingly cancelled out. There remains, however, a systematic bias that cannot be removed by simply using more segmentations. It is represented by the difference between the asymptotic value a and one. The systematic bias can arise for a number of reasons. For example, the anatomy of the query subject may represent a variation that is not, or not sufficiently, represented by the atlas database. There may also be a limit imposed by correspondence accuracy; the registrations may, for example, produce affine transformations that do not provide the local small scale alignment that may be necessary for accurate segmentation. In Heckemann et al. (2006a,b), it was shown that when the transformations used to propagate atlas segmentations are non-rigid and have a fine degree of local control (i.e. have a high-resolution), the asymptotic level of the Dice overlap achieved is higher and the rate of convergence is faster.

Atlas selection

Motivation and background

The size of the atlas database can affect various aspects of the process of multi-atlas segmentation as well as the quality of the final segmentation. As discussed in Background: multi-atlas segmentation, the average segmentation accuracy achieved by fusing random sets of atlases increases asymptotically as the number fused becomes large. This asymptotic increase in accuracy means that there are diminishing returns in using larger and larger numbers of atlases. For a large atlas database, however, the increased computational cost of registering large numbers of atlases to the query image is an immediate practical problem. A second difficulty relates to the way in which an anatomical structure might vary across the population. If a structure is represented by, say, two morphologically distinct variants in the population (and in the atlas database), then fusing a large number of atlases may give a shape that does not represent either variant very well. In such circumstances, for a given query subject, only a proper subset of the atlases in the database is appropriate to use — those sharing the variant represented in the query. Finally, in our experience, the fusion of a large number of atlases is more likely to create a smooth estimate of the structure being segmented and yet a shape which is less smooth may be a better estimate.

For such reasons, the selection of a limited number of atlases, appropriate for the query subject and prior to multi-atlas segmentation, would appear preferable to the fusion of an arbitrarily large number of atlases. Furthermore, it is natural to ask whether the asymptotic level of accuracy given by fusing large numbers of random atlases is the best that can be achieved, i.e. whether it is possible to equal or exceed this accuracy by fusing a smaller number of selected atlases.

In previous work on atlas-based segmentation, the term 'selection' has typically been used to describe the identification of the single best atlas for propagation to a query (see, for example, Rohlfing et al. (2004a) and Wang et al. (2005)). Han et al. (2008) compare the selection of a single atlas against the propagation and fusion of their entire atlas database. Our earlier work (Aljabar et al., 2007) and the work in this paper contrasts with these approaches as we apply selection to whole sets of atlases prior to propagation to the target and decision fusion. Klein et al. (2008) use a similar approach where multi-atlas segmentation is carried out using only atlases that reach a user-defined threshold of similarity with the target. For a given threshold, the number of atlases used for different targets may vary. Our work contrasts this by fixing (for each experiment) the number of atlases selected for different targets. This enables an assessment of the effect of selection on segmentation quality across a range of subjects.

The selection of atlases for segmenting a particular query image also has parallels with clustering problems. For a large atlas database, it is possible to search for clusters or modes of the population it represents. Given an unseen query image and an estimate of the cluster it belongs to, a better segmentation is expected using images from the same cluster rather than from different ones. See Blezek and Miller (2007) and Sabuncu et al. (2008) for examples of clustering approaches applied to brain MR images.

Proposed methods for selection

Given a large atlas database, a fixed subset size and a query image, it is theoretically possible to identify the optimal subset of atlases for generating a multi-atlas segmentation of the query. Leaving aside the question of how segmentations given by different subsets can be compared, exhaustively searching all possible subsets is clearly impractical for a large database.

We therefore present an alternative heuristic approach that uses a measure of the 'similarity' of each atlas to the query subject. Once assigned, this measure can be used to rank the atlases. Multi-atlas segmentation can then be carried out using a number of the top-ranked atlases. A simple approach is to interpret similarity as image similarity, i.e. to derive it from the intensities in the query and atlas images. Alternatively, similarity may be derived from meta-information relating to the subjects. The ranking of each atlas is then based on how closely the atlas subject matches the query in terms of a clinical variable, such as age, pathology, clinical history, genetics, gender, handedness, etc.

Selection using image similarity

A selection framework that relies on evaluating the similarity of a pair of images (an atlas and a query image) requires an estimate of the correspondence between them. It is possible to align all the atlases to each new query prior to making a selection. In order to avoid the computational burden of a large number of registrations direct to the query, we apply a selection framework that makes use of a standard space defined by a reference image.

This approach identifies an arbitrary reference image in advance and all atlases in the database are aligned to it. When a new query image is given, it is also aligned in the same way to the reference. The image similarity of the query image and each of the atlases can then be evaluated, since they are all aligned. These similarity values can then be used to assign ranks. The top-ranked atlases are selected and registered directly to the query in order to generate a multi-atlas segmentation estimate in the native space of the query image. This approach uses two types of registrations: those that align images to the reference prior to selection and those used to propagate atlases when generating the multi-atlas segmentations. This selection framework is illustrated schematically in Fig. 2.

Image similarity can be expressed using a variety of metrics, including sums of squared differences (SSD), cross-correlation (CC), mutual information (MI) (Collignon et al., 1995; Viola and Wells, 1995) and normalised mutual information (NMI) (Studholme et al., 1999). In the context of image registration, information theoretic measures such as MI and NMI are intended for aligning multi-modality images. In the context of selection, this makes such measures more appropriate for images with widely differing levels of contrast and appearance, for example if they were acquired on different MR scanners.

A related choice concerns the region over which to evaluate the similarity metric. The region of interest (ROI) where the similarity metric is evaluated can be the complete overlap of the atlas and query images, or it can be made more specific. For example, if a hippocampal segmentation is required, the ROI might represent a suitably located region that is large enough to be likely to encompass the target hippocampus and yet small enough to avoid evaluating the similarity metric over regions distant from the structure of interest that have little effect on its segmentation.

Another choice relating to image similarity selection concerns the transformations used during the registrations. These can be rigid, affine or non-rigid and, in the non-rigid case, the degree of local control (or flexibility) can be varied. As mentioned above, the spatially normalising transformations that are carried out prior to selection (Fig. 2 left) need to be distinguished from the transformations used to propagate atlas segmentations *after* selection and directly to the query image (Fig. 2 right). The principle we have adopted for spatially normalising transformations is that they should correct for gross differences in orientation and configuration but should not correct for small scale differences, as this will tend to make the atlases very similar to each other and harder to rank with respect to a given query.

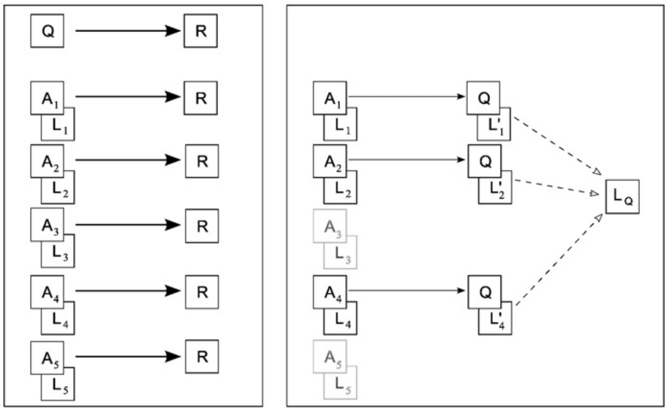

Fig. 2. Multi-atlas segmentation with image similarity selection. Left: All the atlas anatomies A_i and the query image Q are registered to the reference image R. Similarities between the spatially normalised query and each of the atlases are used to generate ranks. Right: Top-ranked atlases are selected and registered directly to the query image. The selected atlas segmentations L_i are propagated to the query giving the segmentations L'_i which are fused to generate the native space segmentation estimate L_Q for the query image.

By contrast, the propagation of atlases during multi-atlas segmentation uses transformations with a high degree of local control which, as discussed in Background: multi-atlas segmentation, have been shown to generate more accurate segmentations. The non-rigid registration method used to propagate atlas segmentations can be chosen from among a number of different approaches (see Zitová and Flusser (2003) for an overview). This work uses the free-form deformation (FFD) model of Rueckert et al. (1999) where displacements at a lattice of control points are blended using B-spline basis functions (De Boor, 1978).

Another distinction needs to be made between the use of a similarity metric for selection and its use for registration. During registration, the similarity metric is used as an optimisation objective function, whereas for selection, the similarity metric is evaluated once *post hoc* for the query and atlas images after alignment. We have used the same metric (NMI) for selection and for registrations (both pre- and post-selection) although there is no strict requirement that the same metric is used in all stages.

The type of spatial normalisation carried out has an effect on the selection process. If the atlases and the query are only rigidly aligned to the standard space, selection will favour atlases that are already very similar in size and configuration to the query while some atlases may be rejected that could have been useful for segmentation after, for example, a global change of scale.

In contrast, high-resolution non-rigid normalisation implies that the variation among atlas subjects is mainly represented in the normalising transformations rather than the aligned images and an image-based ranking of the atlases becomes harder to apply.

While it is possible to extract features from transformations as a basis for selection (see for example Commowick and Malandain (2007)), our focus in this work is on image similarity as a selection criterion and an intermediate level of spatial normalisation is used. These could be, for example, affine 12-parameter transformations or coarse non-rigid transformations that only correct for large scale configurational differences.

In terms of computation, most of the cost of multi-atlas segmentation with image similarity selection is incurred by registrations. If N atlases are in the database and S of the top-ranked atlases are propagated and fused, then the number of registrations for a given query is $1+N+S$. The N registrations spatially normalising the atlases prior to selection can, however, be carried out ‘off-line’ so that the bulk of the on-line computational cost is represented by the S fine-scale non-rigid registrations that propagate the segmentations to the query.

Selection using meta-information

Any meta-information collected from subjects at the time of scanning can also be used as a basis for atlas selection. If information such as gender, age, handedness, clinical status etc. is available for the atlas database as well as for the query subject, then atlases can be selected according to how well the corresponding subjects match the query subject on some aspect of this meta-information. For example, the atlas subjects who are closest in age to the query can be selected for multi-atlas segmentation. Selection using meta-information can be carried out independently of the image data, i.e. no pre-processing or alignment of the images is required. After selection by meta-information, the generation of the final segmentation is carried out in the manner described above and illustrated on the right hand side of Fig. 2.

Data and experiments

In order to assess the impact of the selection methods described above upon segmentation quality, a number of experiments were carried out which use selection prior to multi-atlas segmentation using vote rule decision fusion. The next section describes the data used and implementation choices for selection and multi-atlas segmentation. This is followed by experiments aiming to measure the effect of selection on segmentation accuracy. The objectives of these experiments were to estimate the accuracy possible using similarity selection, to assess the suitability of image similarity as a selection criterion, to describe the effect of the number of atlases selected after ranking and, finally, to compare image similarity selection with selection based on age.

Atlas database and choice of standard space

The data used in the experiments, consisting of T1-weighted MR brain images of 275 male and female subjects were made available by the Centre for Morphometric Analysis (CMA, Massachusetts General Hospital, Charlestown, MA). Ages were available for 224 subjects and ranged between 4 and 83 years with a mean of 27.9 ± 21.1 years. The images were acquired from multiple centres and various structures were manually delineated within each image according to a pre-defined protocol (Filipek et al., 1994). An example of an anatomical image and the corresponding manual segmentation are shown in Fig. 3.

The set of manually delineated structures that were available for all the images was mainly subcortical and consisted of the lateral ventricle, caudate nucleus, putamen, nucleus accumbens, pallidum, thalamus, amygdala, hippocampus and brainstem. In this work, using leave-one-out cross-validation, a subject can be treated as a query and its manually delineated structures can be treated as a gold standard for measuring the accuracy of segmentations obtained by multi-atlas segmentation (using the remaining atlases in the database).

The MNI simulated brain image provided by the Montreal Neurological Institute (MNI), McGill University, Quebec (Holmes et al., 1998) is used to define the standard space during image similarity selection.

Implementation

Given the available structures, as listed above, a region of interest (ROI) representing a mask of the subcortical region was used when evaluating image similarity for selection. This was generated by identifying a subcortical mask separately for each atlas image and spatially normalising all the masks using affine registrations between each of the CMA T1 images and the MNI reference. The union of all the aligned masks was then found and morphologically dilated twice (3 × 3 × 3 cubic structuring element) to generate a collective sub-cortical mask for the database that is uniformly applied during all experiments. The resulting mask is shown in Fig. 4.

The images for all subjects were transformed to standard space following affine registrations of the T1 scans with the reference. The similarity metric used for selection was NMI. Twenty atlases were selected from the ranked database for each leave-one-out experiment with the exception of the experiment described in Varying the number of atlases selected which investigates the effect of selecting varying numbers of atlases.

The non-rigid registrations from each set of selected atlases to the corresponding leave-one-out query were initialised with affine transformations and were modelled using FFDs. The FFD registrations

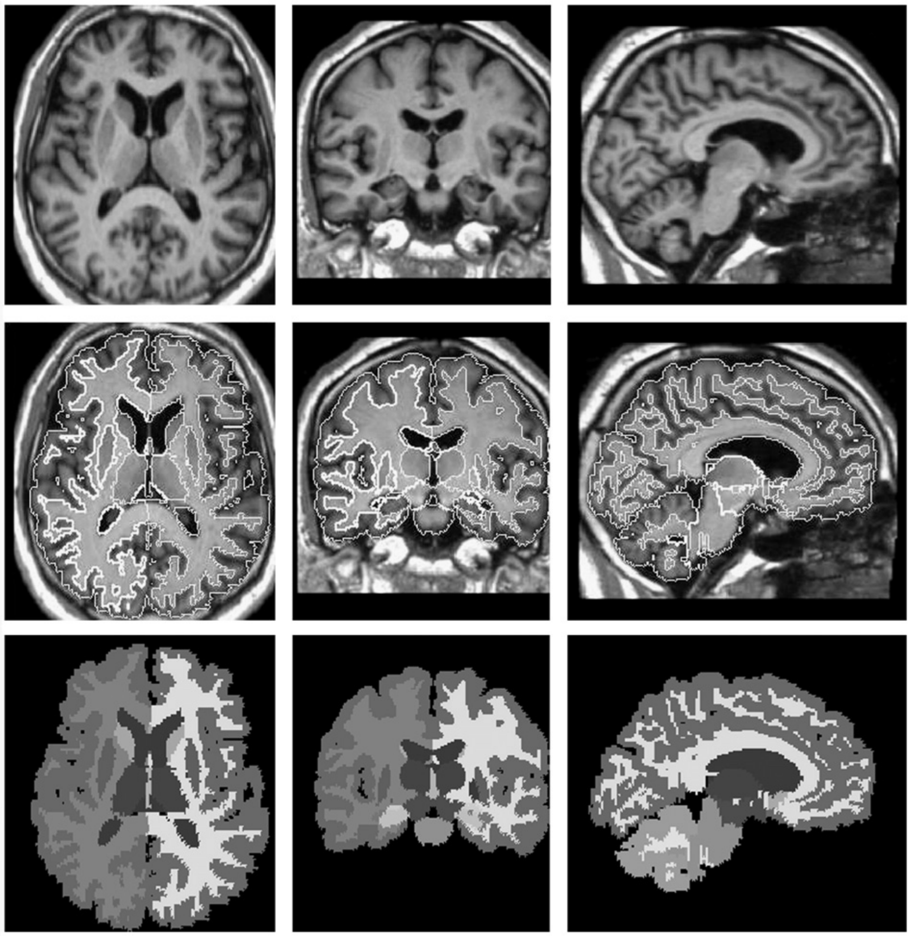

Fig. 3. An example dataset from the CMA data. Top to bottom: A T1-weighted MR image; the segmentation boundaries overlaid on the anatomy; the segmentation.

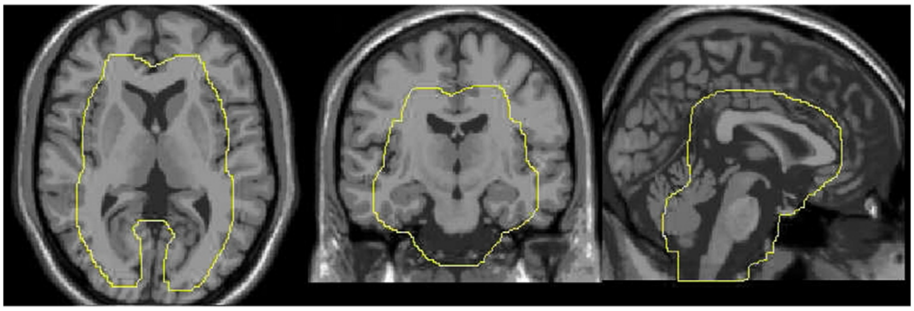

Fig. 4. The subcortical mask used for image similarity selection overlaid onto the image of the MNI simulated MR image.

were implemented using a hierarchical coarse-to-fine approach (Schnabel et al., 2001) with successive control point spacings of 20 mm, 10 mm and 5 mm. At each stage in the registration the control points with the current spacing were optimised and then used as a starting point for the next.

Where possible, all the subjects in the database were treated in turn as a test subject and the remaining subjects were used as a training set from which to draw atlases for segmentation. This was impractical, however, for experiments that use segmentations obtained from many different random sets of atlases and which require a large number of additional registrations. For such experiments, an exemplar group of three subjects was used as a test set. The three exemplar subjects were chosen to represent the adolescent, young adult and older subjects. The individuals chosen are referred to as Subjects 1, 2 and 3 and have ages at scan of 12, 29 and 79 years respectively. For the experiment that makes use of age data, 224 subjects for whom age data were available were each treated as test subjects.

Image similarity selection: assessing the accuracy obtained

Experiments were carried out using leave-one-out cross-validation to assess the accuracy of segmentations based on image similarity selection and multi-atlas segmentation. Summative results were obtained using all subjects in the database and more detailed results for separate structures within individual subjects were obtained using the exemplars as test subjects.

The summative data were obtained using segmentations for all 275 subjects in the database produced using similarity selection as described in Selection using image similarity. For each subject, 20 atlases were selected from the remainder of the database and registered to the subject. The accuracy of the resulting segmentation was measured using Dice overlap with the subject's manual segmentations.

The resulting average accuracy values for all 275 subjects using leave-one-out cross-validation are shown in Fig. 5. As a comparison, the Dice values for the fusion of random sets for the exemplar subjects (see below) were pooled and the resulting averages are also shown. This figure shows, for example, that the average combined left–right Dice accuracy for segmentations of the hippocampus was approximately 0.83 when using similarity selection and 0.81 based on fusing random sets. Separate left–right mean Dice values and standard deviations for each structure are also shown in Table 1. The mean Dice overlap over all structures was 0.854 using image similarity selection and 0.811 using random sets.

Table 1 also shows the averages based on the pooled random set data for comparison. Dice accuracy gains of 0.02 to 0.05 are typical when using similarity-based atlas selection instead of random sets. The most dramatic improvements are shown for the caudate nucleus with an improvement of approximately 0.12. The location of the caudate nucleus may mean that any improvement in its segmentation accuracy is related to improvement in the segmentation of the adjacent ventricles. Due to their much larger size, the increase in Dice for the ventricles (approximately 3–5 points assuming a scale of 0–100) is expected to be smaller than the increase for the caudate nucleus.

Experiments were also carried out for the three exemplars where accuracy was measured for each structure within each subject. The segmentations for each exemplar were generated using image similarity selection and the resulting Dice accuracy of each structure against the manual segmentation was subsequently compared against the accuracy of 1000 further segmentations, each obtained by fusing random sets of atlases. Comparing with the distribution of Dice obtained from random atlas sets helps to assess the significance of the accuracy obtained via atlas selection.

In order to enable the fusion of random sets of atlases for each of the exemplar subjects, all the remaining images in the database were non-rigidly aligned to each exemplar up to a 5 mm control point spacing. For each of the three exemplar subjects, random sets (20 atlases each) were drawn from among the remaining 274 non-rigidly aligned atlases. The segmentations within each random set were fused to generate an automated segmentation estimate and the Dice overlaps with the query's manual segmentation were calculated. This process was repeated 1000 times with different sets of random atlases in order to estimate the distribution of Dice overlaps for each structure within each of the exemplar subjects. Box plots of the estimated Dice overlap distributions are shown for the segmented structures in Fig. 6, which illustrates the results for Subject 2, the 29-year-old subject. The corresponding box plots for the younger and older subjects are shown in Figs. 7 and 8. As an example, focusing on the left pallidum in Fig. 6, the random overlap distribution has a median of approximately 0.84, the limits of the inter-quartile range

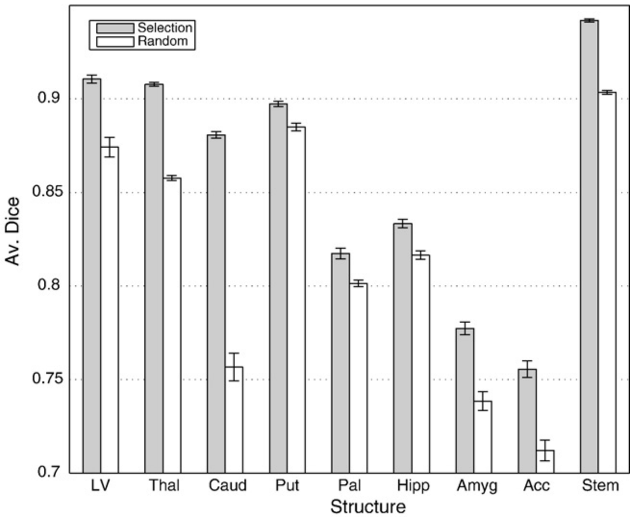

Fig. 5. Average Dice overlap values for all 275 subjects obtained by fusing atlases selected by image similarity (grey bars). For comparison, the average Dice values obtained by fusing random sets of atlases are also shown (white bars).

Table 1
Per structure results for the overlap accuracy achieved after image similarity selection.

Structure	L/R	Selection	Random
Lateral ventricle	L	0.914 (0.04)	0.866 (0.09)
	R	0.911 (0.04)	0.882 (0.05)
Thalamus	L	0.908 (0.02)	0.854 (0.02)
	R	0.909 (0.02)	0.862 (0.02)
Caudate	L	0.883 (0.03)	0.747 (0.11)
	R	0.879 (0.03)	0.766 (0.09)
Putamen	L	0.898 (0.02)	0.887 (0.03)
	R	0.898 (0.02)	0.882 (0.03)
Pallidum	L	0.819 (0.05)	0.803 (0.03)
	R	0.818 (0.05)	0.800 (0.02)
Hippocampus	L	0.832 (0.04)	0.808 (0.03)
	R	0.837 (0.04)	0.825 (0.03)
Amygdala	L	0.778 (0.06)	0.749 (0.05)
	R	0.776 (0.06)	0.728 (0.09)
Accumbens	L	0.765 (0.07)	0.726 (0.08)
	R	0.751 (0.07)	0.698 (0.07)
Brainstem	–	0.941 (0.01)	0.903 (0.02)

The middle column shows the average Dice value over 275 leave-one-out segmentations. For comparison, the final column shows the average overlap achieved by fusing 1000 random sets of 20 atlases each for the exemplar subjects. Left–right overlaps are shown on successive rows.

are 0.83 and 0.85 and the whiskers, representing the extremes of the distribution, are at 0.79 and 0.87. The Dice overlap of the single segmentation of the left pallidum derived using image similarity selection is plotted as a circle and is above the whisker at the upper end of the random distribution (≈0.89).

It can be seen in the figures that, in nearly every case, the segmentation accuracy obtained using selection exceeds the 75th percentile of the random Dice distribution. The order of structures by selection Dice accuracy and the spread of random Dice estimates differ markedly across the three exemplar brains. This might reflect local differences, or simply rater heterogeneity. For Subjects 1 and 3, the accuracy of caudate segmentation using selection exceeds the random Dice distributions by a large amount: the differences between the random distributions' upper quartile and selection accuracy are in the range of 9–12 points. For a number of other structures, the accuracy given by selection also exceeds the outlier values of the random distribution. A negligible percentage (<0.01%) of the fused random sets achieved overlap values outside the outlier limits defined by the whiskers of each box plot. These were not plotted for better readability. The outlier limits were set at 1.5× the inter-quartile range (IQR) above the upper quartile and 1.5× IQR below the lower quartile.[2]

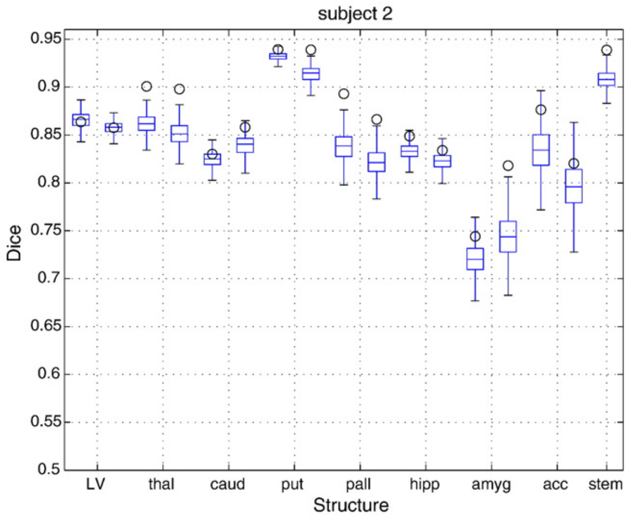

Fig. 6. Illustration for a 29-year-old subject (Subject 2 in the text) of similarity selection segmentation accuracy. This was assessed by Dice overlap with the subject's manual segmentation and is plotted as a circle. The boxplots indicate the Dice overlap distribution from 1000 fusions of random sets of 20 atlases each. The bounds of each box represent the 25th and 75th percentiles of the random distribution of Dice values obtained. The random overlap distributions provide a benchmark against which the overlaps from selection can be compared.

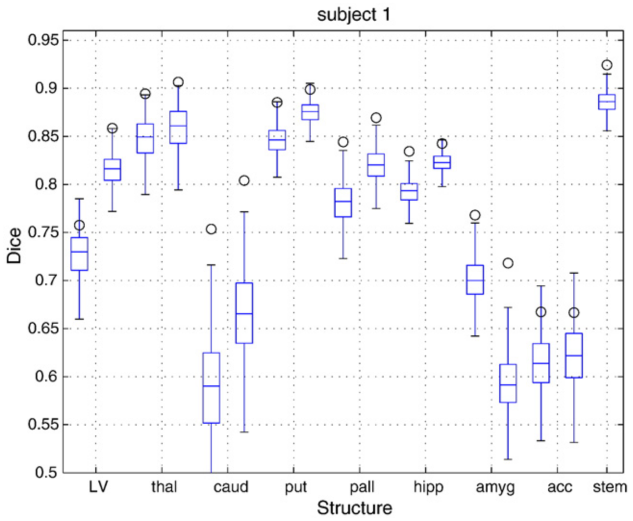

Fig. 7. Illustration of similarity selection segmentation accuracy for a 12-year-old subject (Subject 1 in the text). See Fig. 6 for further explanation.

The quality of similarity selection accuracy values can also be assessed by representing them as *z*-scores based on the random overlap distributions. The similarity selection Dice value for a particular subject and structure can be converted into a *z*-score by subtracting the mean of the random overlap distribution and dividing by the standard deviation. The *z*-score measures the signed difference between an overlap achieved using similarity selection and the random distribution mean expressed as a number of standard deviations. The use of *z*-scores assumes a Gaussian distribution for the Dice values which is theoretically not possible as Dice coefficients are strictly bounded between zero and one. We have, however, shown in previous work (Heckemann et al., 2006a) that the Gaussian assumption is reasonable when Dice values for a structure are sufficiently distant from these bounds.

The *z*-scores for Subjects 1–3 are shown in Table 2 (left–right values combined) and in Fig. 9 (left–right values separated). In general, these data show that similarity selection prior to multi-atlas segmentation performs much better than the fusion of random sets of atlases. The least improved structure across all three subjects was the nucleus accumbens, a small structure for which similarity selection gave accuracy values, on average, 1.6 standard deviations above the mean. The 29-year-old exemplar subject (number 2) showed a lower overall improvement for selection over random fusion (average *z*-score 2.12) than the older and younger subjects. This may relate to the age of exemplar Subject 2 being closer to the overall mean age (27.9 ± 21.1 years) of the database, i.e. that randomly chosen sets of atlases may be more likely to be similar to Subject 2. Therefore the accuracy after random atlas fusion may match more closely the accuracy given by selected atlases. The average *z*-score achieved overall was 2.91, which represents a cumulative percentage of 99.8% of a normal distribution, making the Dice overlaps achieved by selection significantly higher than those obtained by random fusion.

Image similarity as a selection criterion

We made an assessment of the extent to which image similarity represents a suitable selection criterion with respect to the ultimate

[2] see e.g. http://mathworld.wolfram.com/Outlier.html.

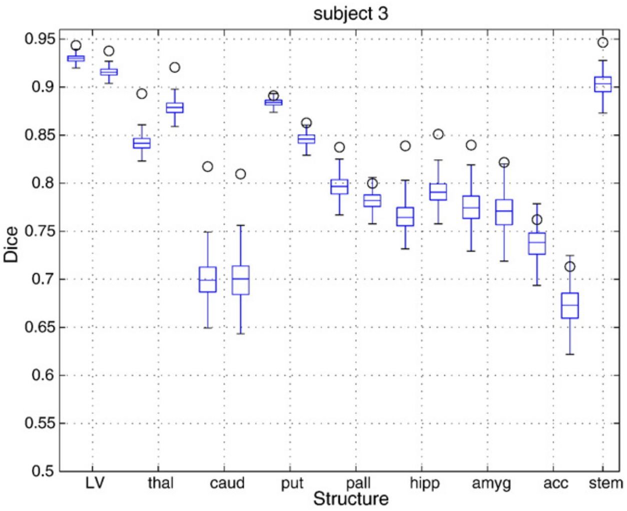

Fig. 8. Illustration of similarity selection segmentation accuracy for a 79-year-old subject (Subject 3 in the text). See Fig. 6 for further explanation.

aim of obtaining accurate segmentations. When registering two images, X and Y, the objective is to find a transformation T_θ with parameters θ that maximises the similarity, sim(X, $T_\theta(Y)$); the only images that are used during registration are X and Y. This contrasts with the use of similarity for selection where similarities between a query image and multiple other subject images are considered. If Q represents a query image and the atlas database images are represented by Y_i, $1 \le i \le n$, then the set of n values

$$\{\mathrm{sim}(Q, Y_1), \ldots, \mathrm{sim}(Q, Y_n)\}$$

are used to provide ranks for the database.

Given two atlases, A and B, it is desirable that similarity selection determines the better of A and B as potential segmentation atlas for Q. This means that if sim(Q, A)>sim(Q, B) then atlas-based segmentation using atlas A should generate a more accurate segmentation of Q. Such a desirable property may, however, be confounded, for example by contrast differences or varying quality of the atlases or scan protocol.

An experiment was therefore carried out to assess the suitability of image similarity as a selection criterion. This focuses on the relationship between the ranks of atlases as given by image similarity and their performance (as individual atlases) in segmenting the query image. Working in the affine normalised space, each of the atlas subjects was left out in turn and the similarity ranks of the remaining atlases were calculated. Additionally, the accuracy of the remaining images in segmenting a structure was estimated from the Dice overlap of their label for that structure with that of the left out subject. Repeating this process for all subjects allows the average accuracy to be calculated across the atlases at each rank. This procedure was carried out twice as the rankings are not symmetric: the rank of atlas A as a segmenter for query B does not, in general, equal the rank obtained by reversing their roles. This meant that 275×274 comparisons were carried out.

Table 2
z-scores for the Dice accuracy of combined left–right segmentations produced after image similarity selection.

	Subject			Average
	1	2	3	
Lateral ventricle	1.70	−0.05	4.21	1.96
Thalamus	2.21	3.87	6.31	4.13
Caudate	3.22	1.17	5.54	3.31
Putamen	2.48	2.38	2.37	2.41
Pallidum	2.83	3.31	2.89	3.01
Hippocampus	2.76	1.63	5.00	3.13
Amygdala	3.73	2.26	3.22	3.07
Accumbens	1.58	1.43	1.78	1.59
Brainstem	3.43	3.10	4.20	3.58
Average	2.66	2.12	3.95	2.91

The *z*-scores are based on the means and standard deviations of the random overlap distributions (see Image similarity selection: assessing the accuracy obtained).

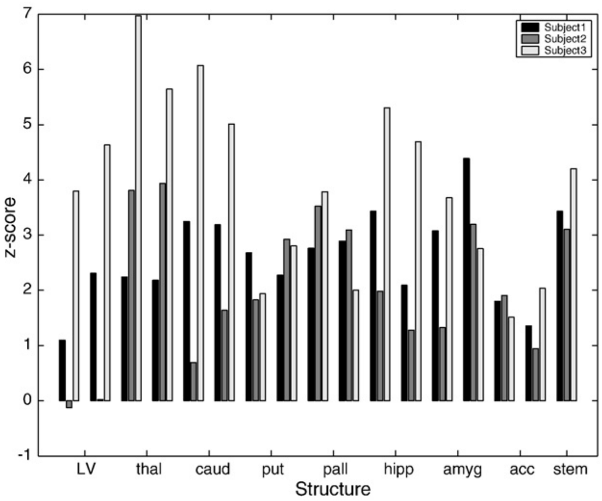

Fig. 9. *z*-scores for the Dice overlap accuracy of segmentations based on image similarity selection (see also Table 2). Separate bars are shown for paired left–right structures. Abbreviations: Lateral ventricle (LV), thalamus (thal), caudate nucleus (caud), putamen (put), pallidum (pall), hippocampus (hipp), amygdala (amyg), nucleus accumbens (accum), brainstem (stem).

The results are illustrated by plots of the average Dice overlap against the rank. These are shown for the hippocampus and the lateral ventricle in Fig. 10. Each point plotted shows the average Dice overlap for all atlases at a given rank in segmenting (as a single atlas) the query for which the rank is calculated. These plots indicate that high ranks are associated with a higher level of accuracy. Although no segmentations are fused in this experiment, the correlation between atlas rank, determined by similarity, and accuracy supports the use of similarity as a basis for selecting atlases prior to fusion (correlation coefficients: lateral ventricle, −0.95; hippocampus, −0.88).

Varying the number of atlases selected

When using similarity selection, the atlases are ranked according to their similarity with the query image. The number of atlases to be used for multi-atlas segmentation can be chosen by the user. This section describes a test of the effect of varying the number of atlases selected upon the final segmentation. Using each of the three exemplar subjects, the ranks of the remaining atlases were determined as described in Selection using image similarity. Increasing numbers are then selected from the *ordered* list of atlases. For each number selected, the corresponding segmentations were fused to provide separate segmentation estimates. The set of segmentations combined after selecting k atlases are the same used after selecting $k-1$ atlases with the inclusion of the kth atlas in the ordered list. The accuracy of each segmentation estimate was assessed using its Dice overlap with the query image's manual segmentation. This process was enabled by having all the atlases non-rigidly aligned to each of the exemplar subjects.

Graphs to illustrate how the resulting segmentation accuracy varies with the number of atlases used are shown in Fig. 11. These graphs show the average combined left–right Dice accuracies of various structures' segmentations for the three exemplar query subjects. The average accuracy over all structures is also shown in Fig. 11. In order to provide a comparison, for each number, k, of

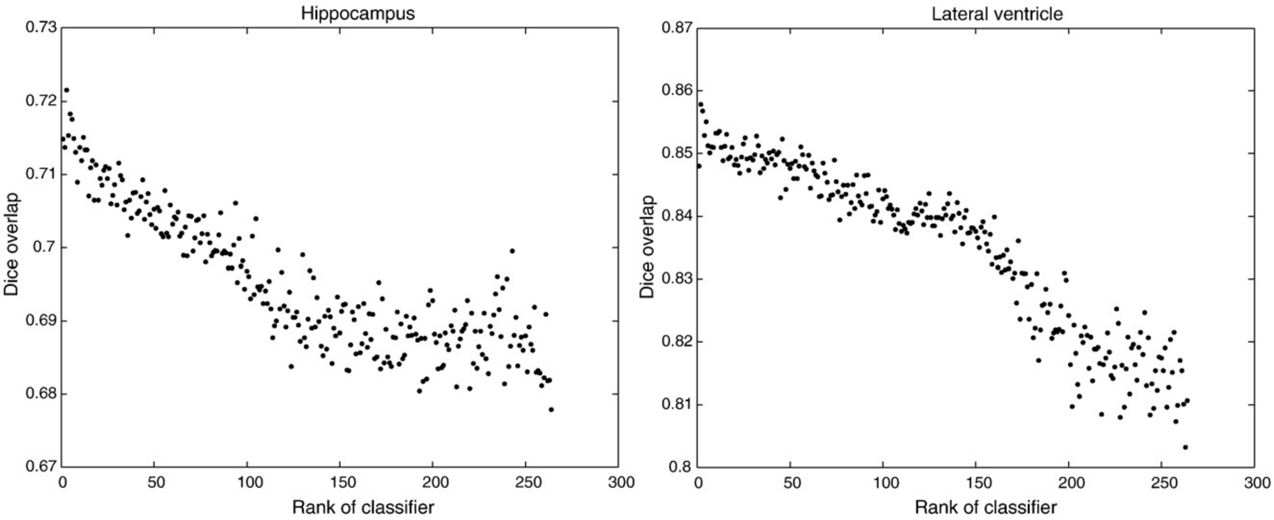

Fig. 10. Relationship between average Dice accuracy obtained by individual atlases for a given rank as determined by image similarity with the query ($N=275$). Results for the hippocampus are shown on the left, those for the lateral ventricle are shown on the right.

atlases selected from the ranked list, a number of random sets of k atlases each were also fused and their Dice overlaps were averaged. 50 random sets were fused for each value of k with each exemplar subject and the resulting averages are also shown in each chart as a dashed line.

The general pattern for the ranked atlas results shows a sharp initial increase in overlap accuracy up to a maximum level followed by a gradual decline. This contrasts with the overlaps achieved by fusing random sets of atlases; after an initial sharp rise, these show a continued slower and monotonic increase, always remaining below the accuracy obtained by selection. In Fig. 11, the shape of the curves obtained from the random atlas sets is predicted by Eq. (1). This equation does not apply to the curves for segmentations obtained from the ranked atlases.

As the number fused approaches the number of atlases in the database, the accuracy of ranked and random atlas fusion converges to the same level, the accuracy that would be obtained by combining the whole database to give a segmentation estimate. In all cases, this level of accuracy is exceeded by using a relatively small number of selected atlases.

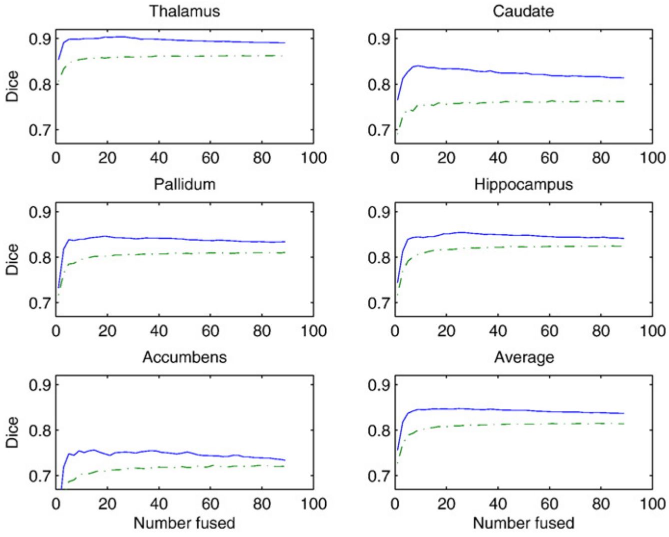

Fig. 11. Segmentation accuracy for various structures in the exemplar subjects after fusing increasing numbers of ranked atlases (Varying the number of atlases selected). The vertical axes show average Dice value obtained and the horizontal axes show the number fused. For comparison, the dashed lines show the average overlap obtained by fusing random sets of atlases of increasing size ($N=50$ for each point). The contiguous and dashed lines in each plot will converge as more atlases are used. Data for five structures (combined left–right) are shown. The average over all structures is shown in the plot at the bottom right.

The number of ranked atlases required for the highest accuracy varies for the different structures. The overlaps for the caudate nucleus reach a maximum for about 8 atlases, while the maximum for the hippocampus is reached after selection of the top 25 atlases on average. The average overlaps across all structures in the bottom right of the figure show a fairly flat section of the highest overlap values for between 15 and 25 atlases.

Comparing similarity- and age-based selection

An assessment was made of the value of meta-information by comparing the segmentation accuracy given by similarity-based selection and by age-based selection. This comparison was possible for 224 subjects in the database, the ages of the remaining subjects were not available. Similarity ranking and selection was carried out as described in Selection using image similarity. Age selection was

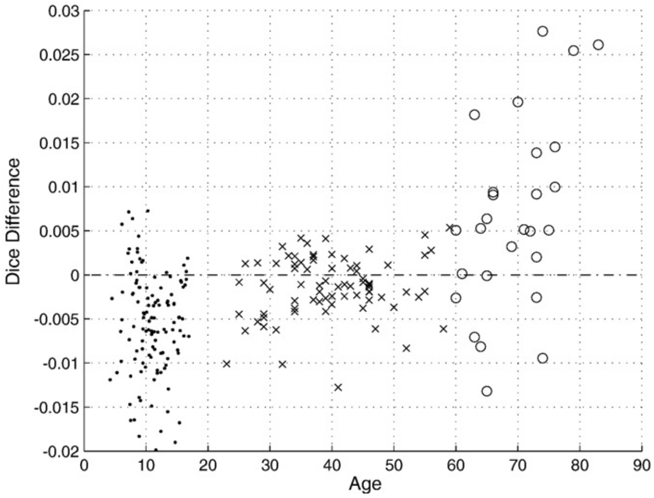

Fig. 12. Comparison of age-based selection and similarity-based selection for the 224 subjects with age data available. The horizontal axis shows the age. Subjects in different age groups are plotted with different symbols. The vertical axis shows the gain that age selection gives over similarity selection.

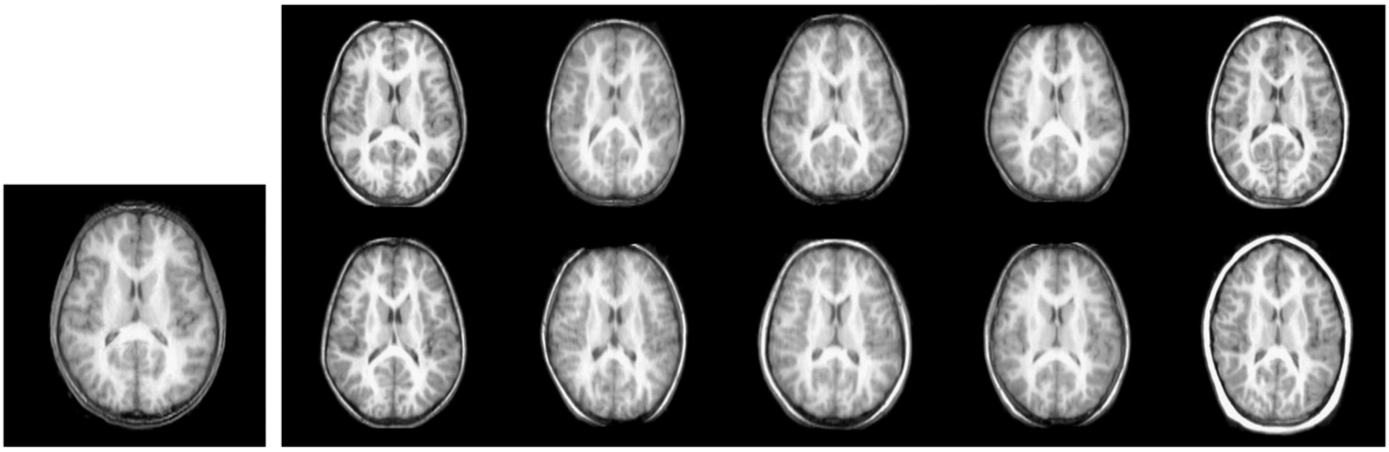

Fig. 13. Exemplar Subject 1 (age 12, left) and the top 10 atlases selected using image similarity (see also Table 3).

carried out by ranking the atlas subjects based on their age proximity to the query and identifying the twenty closest subjects in age.

The overlaps of the segmentations obtained using each selection method with the subjects' manual segmentations were then evaluated to allow the selection methods to be compared. The average overlap over the subcortical structures was used to give a summary measure for each selection criterion on each subject. Subtracting the means then gives a single figure per subject representing the gain obtained by age-based selection.

The resulting values are shown in Fig. 12, where points lying above the horizontal axis represent subjects for whom age selection represents an improvement over similarity selection. The subjects are divided into the three broad age groups and their results plotted with different markers.

Among the youngest subjects (represented by dots) there are some for whom age selection appears to produce a slightly worse result, although the differences are small. The mean gain through age selection for the adolescent subjects was -0.0055 ± 0.0057 and the maximum change was -0.0199. The results for the older group (represented by circles) were more varied than the younger groups, but contained the subjects for whom age selection made the most improvement. The mean Dice gain for the older subjects was 0.007 ± 0.011 with a maximum change of $+0.03$. The results for the young adult group (represented by crosses) are intermediate between those for the adolescents and those for the older groups.

To give a qualitative impression of the atlases selected, the top 10 atlases by image similarity rank were identified for the three exemplar subjects described in Implementation. Transverse sections through the frontal and occipital horns of both lateral ventricles from each of the MRIs associated with the top-ranked atlases for each subject are shown in Fig. 13 (age 12), 14 (age 29) and 15 (age 79).

The ages of the top ten atlases for each of these three subjects are shown in Table 3. The table shows that, although the top-ranked atlases were selected by image similarity, they match the age of the query subject well. This further demonstrates the level of agreement between age- and similarity-based selection and relates to their comparable segmentation accuracy as shown in Fig. 12.

Discussion

We have presented an investigation of strategies for atlas selection prior to multi-atlas segmentation. The accuracy obtained by atlas selection has been assessed against the levels of accuracy achieved by fusing random sets of atlases. The results show that the accuracy achieved by similarity selection is significantly higher than that achieved by the fusion of random sets of atlases. The results obtained from a large number of leave-one-out cross-validation experiments (Table 1) compare very well with the state-of-the-art (see for example Fischl et al. (2002), Klein and Hirsch (2005) or Chupin et al. (2007)) and are comparable with some previous manual segmentation

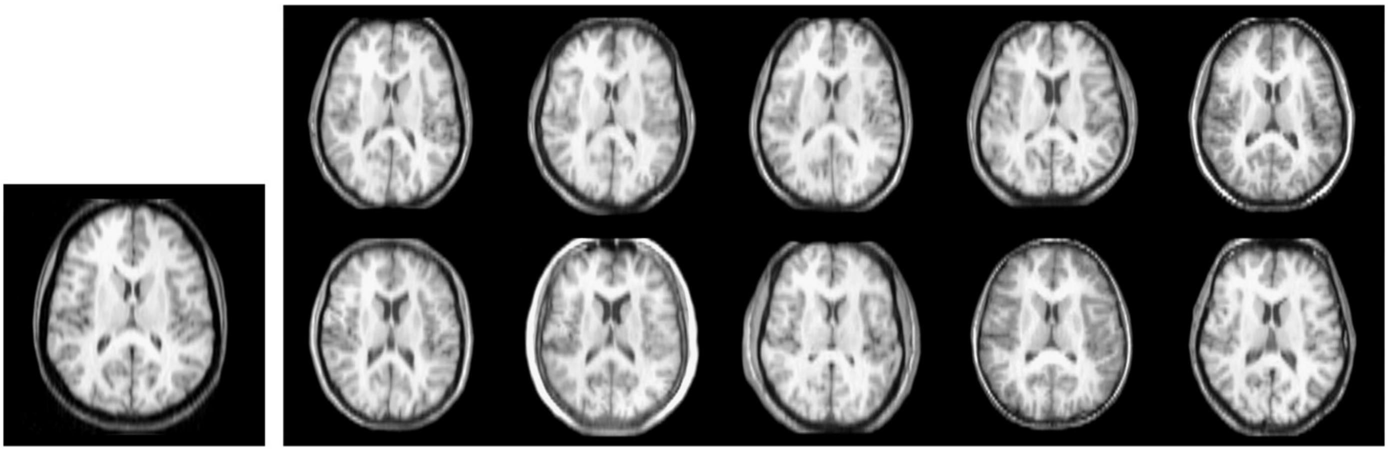

Fig. 14. Exemplar Subject 2 (age 29, left) and the top 10 atlases selected using image similarity (see also Table 3).

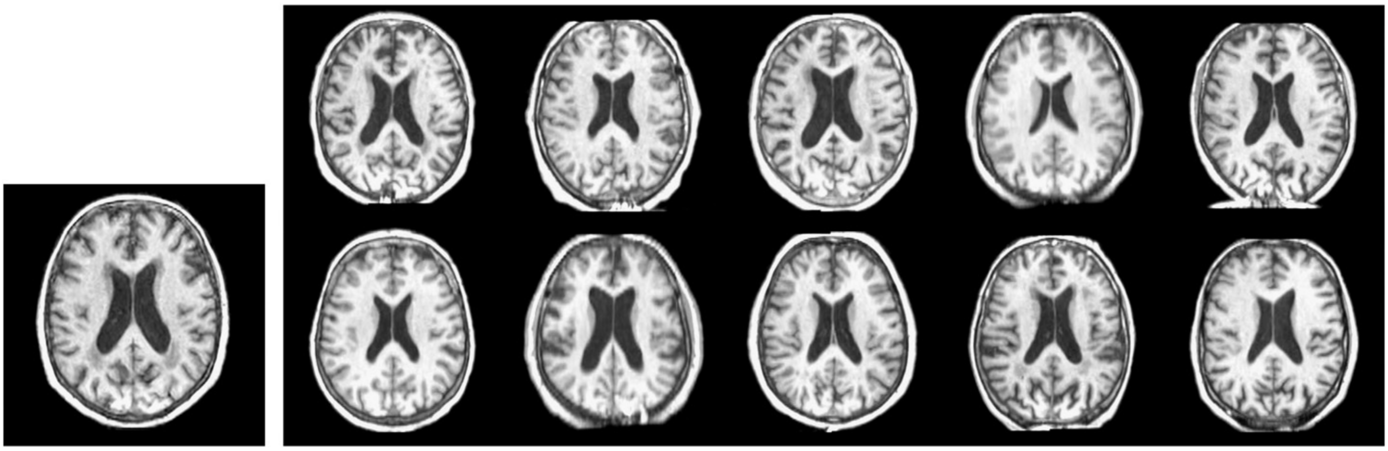

Fig. 15. Exemplar Subject 3 (age 79, left) and the top 10 atlases selected using image similarity (see also Table 3).

methods, for example Spinks et al. (2002). In particular, the results we obtained compare well with those presented in Heckemann et al. (2006a), which also use multi-atlas segmentation. It is worth noting, however, that our accuracy levels were achieved using a coarser FFD registration than those giving the best results in Heckemann et al. (2006a) (control point spacings of 5 mm vs.2.5 mm).

Age-based selection provides very similar levels of segmentation accuracy to similarity selection, indicating the potential that meta-information selection has when used with multi-atlas segmentation.

The framework for similarity selection presented in Selection using image similarity is reasonably simple and provides good results. Variations are clearly possible, for example over whether or not a standard space is used or in the choice of space in which similarity is measured. We have used the MNI simulated brain image as a reference for the selection step. We are aware that the choice of a single image may introduce a bias but feel that the results obtained justify the method as a whole and leave to future work an investigation of the effect of 'bias-free' or 'group-average' atlases at the selection stage. We have also varied the choice of similarity metric, again obtaining very similar results.

The number of exemplar subjects for which comparisons are made against random overlap distributions is limited by practical considerations, in particular CPU time, since many registrations are required. In spite of this limitation, it remains possible to obtain a reasonable impression of the gain in performance that can be obtained by similarity-based selection (see for example Table 1). In comparison with approaches that build explicit models, for example active appearance models or EM approaches, the computational time required for segmenting new images can be relatively high. The bulk of the computation required for our approach is taken up by atlas to query registrations. We ran these in parallel using the Condor distributed scheduling system[3] and the time required to segment an image varied, with typical times being 3 to 4 h. In contrast, methods that use explicit models typically require a significant time to build and train the model, although the application to new data is usually much faster.

Table 3
The ages of the top 10 atlas subjects selected using image similarity for three different query subjects.

Rank of atlas	Age of query subject		
	12	29	79
1	13.6	29	73
2	4.5	52	72
3	10.1	25	76
4	9	43	31
5	8.6	16.5	75
6	8	25	65
7	9.5	55	63
8	11.3	34	73
9	11.9	12.7	83
10	7.5	55	65
Mean (SD)	9.4 (2.55)	34.72 (15.74)	67.6 (14.20)

The three subjects were chosen to cover the age range of the data.
The atlas subjects for each exemplar in this table correspond to the transverse slices illustrated in order in Figs. 13–15.

We have assessed the suitability of image similarity as an atlas selection criterion. The results (Image similarity as a selection criterion) suggest that the similarity-based rank of an atlas correlates well with its accuracy as a potential segmenter of structures in the subcortical region for the query and provides justification for the use of image similarity in selection. Additionally, a visual inspection of the top-ranked atlases for three query subjects (Figs. 13–15) indicate that the anatomies of the image similarity selected atlases show a close match with those of the query subjects. It appears plausible that an individual atlas will perform better if the target subject is morphologically similar and our work provides evidence to support this intuition. The ROI used in this work encompassed a number of subcortical structures and the hippocampus. We are encouraged to carry out further work to assess the impact of ROIs that either target structures more specifically or are intended for use in segmenting cortical regions.

The accuracy figures achieved by the fusion of different numbers of ranked atlases presented in Varying the number of atlases selected indicate that simply using larger and larger numbers of atlases (after ranking and beyond approximately 20 atlases) leads to lower accuracy in the resulting segmentation. When applying image similarity selection, as the number of atlases fused increases from a low number, the accuracy achieved tends to rise quickly to a maximum and then gradually to decline. The decline in the overlap values as increasing numbers of atlases are used may be explained by the convergence or bias of the resulting fused segmentation towards the mean shape; this may not represent the query subject as well as a segmentation provided by fewer atlases selected specifically for the query.

The number of atlases required for the maximal accuracy varied according to structure, but the fusion of approximately 20 atlases produces near maximal accuracy when the results are averaged across structures. This suggests that the choice of 20 atlases after selection, as is carried out in the remaining experiments, is reasonable. If the

[3] www.cs.wisc.edu/condor.

segmentation of a particular structure is required, the number of atlases used from the ranked atlases could then be adapted specifically to suit the structure.

Another general implementation choice was that of the ROI used for making similarity comparisons. This represented a mask that covered all the subcortical regions for all subjects. A structure-specific implementation could easily incorporate a mask tailored to the target structure.

Comparing similarity- and age-based selection gives a comparison of the accuracy achieved after selection based on age and on similarity. The results suggest that the difference in their performance is generally negligible, with both approaches significantly outperforming random fusion (as shown by the results in Image similarity selection: assessing the accuracy obtained). For some subjects, selection on age produces a slightly worse result, although it should be stressed that the differences are very small. The subjects for whom age selection gave the largest improvement were among the oldest. This may be due to accelerating morphological change during old age and its impact on the appearance of the image data. This suggests that multi-atlas segmentation for older subjects may benefit more from the use of age-based selection instead of similarity-based selection. Interestingly, when comparing against the Dice overlaps generated by the fusion of random sets of atlases (Image similarity selection: assessing the accuracy obtained), the subject for whom the similarity selection gave the greatest improvement over random atlas fusion (Table 2) was the oldest subject (age 79).

There are a number of potential directions for future work on selection and multi-atlas segmentation. For example, adaptive weighting of the different atlases could be used during the fusion stage where, perhaps, a local measure of the similarity between atlas and query might be used to weight votes during fusion rather than using a simple majority vote. Future work can also be carried out to assess the use of STAPLE when fusing propagated atlases after selection. The simplest possible fusion method (vote rule) was used in this work in an effort to reduce computational cost. Alternatively, selection based on geometric features can be explored and compared with selection based on image similarity. Such an approach may draw upon work in deformation- or tensor-based morphometry and may make use of geometric features extracted from the transformations between the atlases and a reference. Such features can, in turn, be used as a basis for selection. Initial work in this area is presented by Commowick and Malandain (2007) where the displacement magnitudes of non-rigid query to atlas transformations are used to select the most similar atlas to the query image.

In a recent application, the methods for selection and structural segmentation presented in this work were successfully incorporated into a framework for assisting clinical diagnosis in cross-sectional studies. The framework also makes use of graph-theoretic spectral clustering techniques and was applied to image data acquired from subjects with early symptoms of dementia (Aljabar et al., 2008a,b). In further recent work, the accuracy of the selection and segmentation scheme presented in this work compared very favourably against other segmentation methods, based either on Active Appearance Models or on EM-based classification, which were applied to the same data (Babalola et al., 2008).

Acknowledgments

P. Aljabar acknowledges the support of the EPSRC Integrated Brain Image Modelling project (GR/S82503/01) and the MRC funding of the NeuroGrid project (www.neurogrid.ac.uk). R.A. Heckemann was supported by a grant from the Dunhill Medical Trust. The authors would like to thank David Kennedy and the Centre for Morphometric Analysis at MGH for making the image and segmentation data available.

References

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

Aljabar, P., Rueckert, D., Crum, W., 2008a. Automated morphological analysis of magnetic resonance brain imaging using spectral analysis. NeuroImage 43 (2), 225–235.

Aljabar, P., Rueckert, D., Crum, W., 2008b. Spectral clustering as a diagnostic tool in cross-sectional MR studies: an application to mild dementia. Eleventh Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '08). Vol. 5242 of Lecture Notes in Computer Science, pp. 442–449.

Babalola, K., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Jenkinson, T.C.M., Rueckert, D., 2008. Comparison and evaluation of segmentation techniques for subcortical structures in brain MRI. Eleventh Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '08). Vol. 5241 of Lecture Notes in Computer Science, pp. 409–416.

Blezek, D., Miller, J., 2007. Atlas stratification. Med. Image Anal. 11 (5), 443–457.

Chupin, M., Hammers, A., Bardinet, E., Colliot, O., Liu, R.S.N., Duncan, J.S., Garnero, L., Lemieux, L., 2007. Fully automatic segmentation of the hippocampus and the amygdala from MRI using hybrid prior knowledge. Tenth Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '07), pp. 875–882.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., Marchal, G., 1995. Automated multimodality image registration using information theory. In: Bizais, Y., Barillot, C., Paola, R.D. (Eds.), Information Processing in Medical Imaging. Kluwer Academic Publishers, Dordrecht, pp. 263–274.

Commowick, O., Malandain, G., 2007. Efficient selection of the most similar image in a database for critical structures segmentation. Tenth Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '07), pp. 203–210.

De Boor, C. (Ed.), 1978. A Practical Guide to Splines. Springer.

D'Haese, P., Duay, V., Merchant, T., Macq, B., Dawant, B., 2003. Atlas-based segmentation of the brain for 3-dimensional treatment planning in children with infratentorial ependymoma. Sixth Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '03), pp. 627–634.

Dice, L., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297.

Filipek, P., Richelme, C., Kennedy, D., Caviness, V., 1994. The young adult human brain: an MRI-based morphometric analysis. Cereb. Cortex 4 (4), 344–360.

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.

Hammers, A., Allom, R., Koepp, M., Free, S., Myers, R., Lemieux, L., Mitchell, T., Brooks, D., Duncan, J., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19, 224–247.

Han, X., Hoogeman, M., Levendag, P., Hibbard, L., Teguh, D., Voet, P., Cowen, A., Wolf, T., 2008. Atlas-based auto-segmentation of head and neck CT images. Eleventh Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '08). Vol. 5242 of Lecture Notes in Computer Science, pp. 434–441.

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

Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A., 2006b. Multiclassifier fusion in human brain MR segmentation: modelling convergence. Ninth Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '06). Vol. 4191, pp. 815–822.

Holmes, C., Hoge, R., Collins, L., Woods, R., Toga, A., Evans, A., 1998. Enhancement of MR images using registration for signal averaging. J. Comput. Assist. Tomogr. 22 (2), 324–333.

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.

Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20 (3), 226–239.

Klein, A., Hirsch, J., 2005. Mindboggle: a scatterbrained approach to automate brain labeling. NeuroImage 24 (2), 261–280.

Klein, S., van der Heide, U., Lips, I., van Vulpen, M., Staring, M., Pluim, J., 2008. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med. Phys. 35 (4), 1407–1417.

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.

Rohlfing, T., Brandt, R., Menzel, R., Maurer, C., 2004a. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21 (4), 1428–1442.

Rohlfing, T., Russakoff, D., Maurer, C., 2004b. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imag. 23 (8), 983–994.

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.

Sabuncu, M., Balci, S., Shenton, M., Golland, P., 2008. Discovering modes of an image population through mixture modeling. Eleventh Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '08). Vol. 5242 of Lecture Notes in Computer Science, pp. 381–389.

Schnabel, J., Rueckert, D., Quist, M., Blackall, J., Castellano-Smith, A., Hartkens, T., Penney, G., Hall, W., Liu, H., Truwit, C., Gerritsen, F., Hill, D., Hawkes, D., 2001. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In: Niessen, W.J., Viergever, M.A. (Eds.), Fourth Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI '01). Vol. 2208 of Lecture Notes in Computer Science, pp. 573–581.

Shattuck, D., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K., Poldrack, R., Bilder, R., Toga, A., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39 (3), 1064–1080.

Spinks, R., Magnotta, V., Andreasen, N., Albright, K., Ziebell, S., Nopoulos, P., Cassell, M., 2002. Manual and automated measurement of the whole thalamus and mediodorsal nucleus using magnetic resonance imaging. NeuroImage 17 (2), 631–642.

Studholme, C., Hill, D., Hawkes, D., 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recogn. 32 (1), 71–86.

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.

Viola, P., Wells, W., 1995. Alignment by maximization of mutual information. In: Grimson, W.E.L. (Ed.), Proc. 5th International Conference on Computer Vision (ICCV'95). IEEE Computer Society Press, pp. 16–23.

Wang, Q., Seghers, D., D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., Hammers, A., 2005. Construction and validation of mean shape atlas templates for atlas-based brain image segmentation. Information Processing in Medical Imaging: Proc. 19th International Conference (IPMI'05). Lecture Notes in Computer Science, pp. 689–700.

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.

Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C., Aizenstein, H., 2007. Optimum template selection for atlas-based segmentation. NeuroImage 34 (4), 1612–1618.

Zitová, B., Flusser, J., 2003. Image registration methods: a survey. Image Vis. Comput. 21 (11), 977–1000.