NeuroImage 49 (2010) 2352–2365

Contents lists available at ScienceDirect

NeuroImage

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

Fast and robust multi-atlas segmentation of brain magnetic resonance images

Jyrki MP. Lötjönen [a,*], Robin Wolz [b], Juha R. Koikkalainen [a], Lennart Thurfjell [c], Gunhild Waldemar [d], Hilkka Soininen [e], Daniel Rueckert [b] and The Alzheimer's Disease Neuroimaging Initiative [1]

[a] Knowledge Intensive Services, VTT Technical Research Centre of Finland, P.O. Box 1300 (street address Tekniikankatu 1), FIN-33101 Tampere, Finland
[b] Department of Computing, Imperial College London, London, UK
[c] Medical Diagnostics R and D, GE Healthcare, Uppsala, Sweden
[d] Memory Disorders Research Group, Department of Neurology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
[e] Department of Neurology, University of Kuopio, Kuopio, Finland

ARTICLE INFO

Article history:
Received 12 July 2009
Revised 9 October 2009
Accepted 10 October 2009
Available online 24 October 2009

Keywords:
MRI
Segmentation
Atlases
Registration
Hippocampus

ABSTRACT

We introduce an optimised pipeline for multi-atlas brain MRI segmentation. Both accuracy and speed of segmentation are considered. We study different similarity measures used in non-rigid registration. We show that intensity differences for intensity normalised images can be used instead of standard normalised mutual information in registration without compromising the accuracy but leading to threefold decrease in the computation time. We study and validate also different methods for atlas selection. Finally, we propose two new approaches for combining multi-atlas segmentation and intensity modelling based on segmentation using expectation maximisation (EM) and optimisation via graph cuts. The segmentation pipeline is evaluated with two data cohorts: IBSR data ($N=18$, six subcortial structures: thalamus, caudate, putamen, pallidum, hippocampus, amygdala) and ADNI data ($N=60$, hippocampus). The average similarity index between automatically and manually generated volumes was 0.849 (IBSR, six subcortical structures) and 0.880 (ADNI, hippocampus). The correlation coefficient for hippocampal volumes was 0.95 with the ADNI data. The computation time using a standard multicore PC computer was about 3–4 min. Our results compare favourably with other recently published results.

Introduction

Brain MR imaging is playing an important role in neuroscience. Neurodegenerative brain diseases mark the brain with morphological signatures; detection of these signs may be useful to improve diagnosis, particularly in diseases for which there are few other diagnostic tools. For example, early and significant hippocampal atrophy in people who have memory complaints points to a diagnosis of Alzheimer's disease. Quantitative analysis and objective interpretation of images usually require segmentation of various structures from images. Reliable and accurate segmentation is a prerequisite for comprehensive analysis of images. Current state-of-the-art brain segmentation algorithms can be classified into algorithms that label voxels (a) into brain/non-brain (Ségonne et al., 2004; Smith, 2002); (b) into different tissue types such as white matter (WM), grey matter (GM), or cerebral spinal fluid (CSF) (Ashburner and Friston, 2005; Bazin and Pham, 2007; Pham and Prince, 1999; Scherrer et al., 2008; van Leemput et al., 1999; Zhang et al., 2001); or (c) algorithms that identify anatomical areas, e.g., hippocampus, thalamus, putamen, caudate, amygdala, and corpus callosum (Bazin and Pham, 2007; Chupin et al., 2009; Corso et al., 2007; Desikan et al., 2006; Fischl et al., 2002; Heckemann et al., 2006; Klein et al., 2005; Morra et al., 2008; Scherrer et al., 2008).

Atlas-based segmentation is a commonly used technique to segment image data. In atlas-based segmentation, an intensity template is registered non-rigidly to a target image and the resulting transformation is used to propagate the tissue class or anatomical structure labels of the template into the space of the target image. Many different approaches have been published using registration-based segmentation, for example, for segmenting subcortical structures (Avants et al., 2008; Bhattacharjee et al., 2008; Han and Fischl, 2007; Pohl et al., 2006). A comparison of different atlas-based segmentation algorithms was recently published by Klein et al. (2009). A review of registration techniques is presented in Gholipour et al. (2007).

The segmentation accuracy can be improved considerably by combining basic atlas-based segmentation with techniques from machine learning, e.g., classifier fusion (Heckemann et al., 2006;

* Corresponding author. Fax: +358 20 722 3499.
E-mail address: jyrki.lotjonen@vtt.fi (J.M.P. Lötjönen).

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu\ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf.

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.10.026

Klein et al., 2005; Rohlfing et al., 2004; Warfield et al., 2004). In this approach, several atlases from different subjects are registered to target data. The label that the majority of all warped labels predict for each voxel is used for the final segmentation of the target image. Babalola et al. (2008) compared in a recent study different algorithms for the segmentation of subcortical structures. They found that multi-atlas segmentation produced the best accuracy from the algorithms tested. However, the major drawback of multi-atlas segmentation is that it is computationally expensive, limiting its every day use in clinical practice.

Several factors affect the segmentation accuracy and computation time in multi-atlas segmentation (Fig. 1). First, all atlases are non-rigidly registered to the target (patient) image. During the non-rigid registration, an atlas is deformed in such a way that a similarity measure between the atlas and the target data is maximised. The selection of the similarity measure and the deformation model are central components in optimising the performance of non-rigid registration. A prolific number of solutions are available for similarity measures and for ways to deform the atlas. In this work, we study similarity measures although the deformation model also plays an important role. Second, when the majority voting is applied after non-rigid registration, the objective is to keep the number of atlases as low as possible because the computation time increases correspondingly. As shown in Heckemann et al. (2006), segmentation accuracy increases in a logarithmic way when new atlases are included, i.e., first rapidly and finally very slowly when the number of atlases is high. For these reasons, a compromise must be made when selecting the number of atlases. On the other hand, not only the number of atlases matters but also their quality. If an atlas is very similar to the target data, the inclusion of this atlas probably increases the segmentation accuracy more than less similar atlases. Appropriately implemented atlas selection improves the accuracy of multi-atlas segmentation (Aljabar et al., 2009). Third, the standard multi-atlas segmentation does not model and utilise the statistical distributions of intensities in different structures although this information could be highly valuable in improving the segmentation accuracy. Combining multi-atlas segmentation and intensity modelling as a post-processing step improves the segmentation accuracy (van der Lijn et al., 2008). This work investigates these three factors in more detail.

The ultimate objective of this study is to develop a segmentation method for the clinical practice. This means that we aim (1) to search methods to further improve the segmentation accuracy and (2) to speed up processing without compromising segmentation accuracy, in the context of multi-atlas segmentation. To be clinically feasible, the automatic segmentation algorithm should produce accuracy comparable with manual segmentation made by an expert, and require only a few minutes computation time in a stand-alone PC workstation. The major contribution of this work is the optimisation of the whole multi-atlas segmentation pipeline. We develop and compare different (1) similarity measures in non-rigid registration, (2) atlas-selection methods, and (3) methods to combine multi-atlas segmentation and intensity modelling.

In this article, methods for non-rigid registration, atlas-selection, and combination of multi-atlas segmentation and intensity modelling are first described. This is followed by describing the experiments to assess the multi-atlas segmentation pipeline. Finally, results for two data cohorts are shown and discussed. Part of the research presented in this work appeared previously in conference articles (Lötjönen et al., 2009; Wolz et al., 2009).

Materials and methods

In this section, the whole pipeline for multi-atlas segmentation is described: pre-processing, non-rigid registration, atlas selection, and combination of multi-atlas segmentation and intensity modelling as a post-processing step.

Pre-processing

Intensity normalisation of atlases

The intensity values of CSF, GM, and WM in the atlases were first normalised; the mean intensity values of CSF, GM, and WM were computed and mapped to pre-defined intensity values (see details in Intensity difference as a similarity measure section).

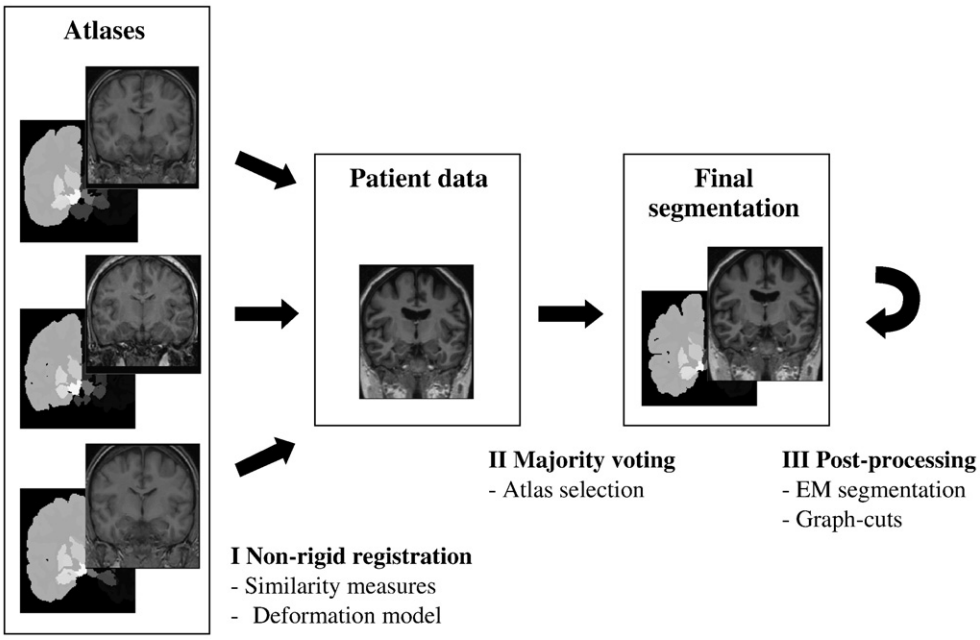

Fig. 1. Steps of multi-atlas segmentation: (I) non-rigid registration used to register all atlases to patient data, (II) classifier fusion using majority voting for producing class labels for all voxels, and (III) post-processing of multi-atlas segmentation result by various algorithms taking into account intensity distributions of different structures.

Affine registration

Atlases and target images were registered using 9-parameter affine transformation. Normalised mutual information (NMI) (Studholme et al., 1997) was maximised between images using a gradient-descent algorithm. NMI was used because intensities were not yet normalised between atlases and target image making the use of intensity differences unstable as a similarity measure.

Inhomogeneity correction

Intensity inhomogeneities were removed from the images using the algorithm proposed by Studholme et al. (2004). The bias field was obtained by dividing intensity values of a low-pass filtered image by intensity values of a low-pass filtered template, which had been registered non-rigidly to the image. As the images were, at this point, only affinely registered, the bias field was estimated within the white matter region after two morphological erosion operations had been performed. In addition, the mean and standard deviation were computed for the bias field, which was modelled as a multiplicative term. Values exceeding 95% confidence interval were excluded. Finally, the bias field in other regions was extrapolated and low-pass filtered. In addition to inhomogeneity correction, the algorithm performs intensity normalisation as intensities in the WM region become approximately equal.

Non-rigid registration

Background

Normalised mutual information (NMI) (Studholme et al., 1997) is a widely used similarity measure for atlas-to-image registration (Heckemann et al., 2006; van der Lijn et al., 2008). In this work, we study atlas-to-image registration techniques that replace NMI by a much simpler and faster intensity difference similarity measure. Intensity differences have been used as a similarity measure for a long-time but comparison with NMI and requirements for using it in atlas-to-image registration need clarification.

The challenge in using the intensity difference is that the intensity of a specific tissue type can vary across different magnetic resonance (MR) images even if the same imaging parameters were used. Therefore, some form of intensity normalisation is needed. Several approaches have been reported. One strategy is to align the intensity histograms of images; Nyúl et al. (2000) defined several landmarks (percentiles and modes) from histograms and matched the landmarks. Hellier (2003) estimated a mixture of Gaussians that approximates a histogram, and matched the mean intensities of the histogram peaks between images. Jäger and Hornegger (2009) proposed recently a method for normalisation of multi-spectral images. They computed joint histograms for multi-spectral images and aligned histograms using non-rigid registration. Spatial tissue correlations between images can also be used to normalise intensities. Guimond et al. (2001) estimated the intensity mapping between two images by a high-dimensional polynomial. The polynomial minimised the difference between the images in the least square sense. Their algorithm alternates between intensity and spatial normalisations. Schmidt (2005) defined, for intensities of a template image, a scaling factor that minimised the absolute value of the difference between the template and target images. The images were assumed to be aligned non-rigidly before the normalisation. The difference was computed only for regions that were well aligned. The neighbourhood of each pixel is considered to be aligned if the local intensity distributions are similar, measured by their joint entropy. In addition, the computation is only performed in the region of interest, e.g., in the brain region. In this work, we propose a technique for intensity normalisation based on spatial tissue correlations using ideas similar to Guimond et al. (2001) and Schmidt (2005). We demonstrate two techniques where spatial and intensity normalisations are done iteratively during the registration.

Framework for non-rigid registration

In atlas-based segmentation, an atlas image $A=A(x, y, z)$ is mapped to a target image, $I=I(x, y, z)$. In the following, the intensity value of the voxel p at location (x, y, z) is denoted by A_p and I_p. The transformation that maps the atlas to the target image is denoted by a vector field $\boldsymbol{T}=\mathbf{T}(x, y, z)$. In addition to the intensity values, each voxel p in the atlas includes a label f_p, which defines the tissue class for the voxel. The segmentation of the target image is produced by transforming the labels f_p by the transformation $\boldsymbol{T}$.

Non-rigid registration is often formulated as a maximisation or minimisation problem of the cost function:

$$E = E_{\text{data}} + \gamma E_{model,} \tag{1}$$

where E_{data} represents similarity or dissimilarity between atlas and target image, and E_{model} is a regularisation term that constrains the transformation $\boldsymbol{T}$ to be smooth. We constrained the curvature of the transformation as defined in Rueckert et al. (1999). The parameter γ is a user-defined weight that determines the trade-off between both terms.

Normalised mutual information is one of the most widely used similarity measures allowing fully automatic registration even of multi-modal images such as MR and PET. NMI is defined as:

$$E_{data} = \frac{H(A) + H(I)}{H(A, I)}, \tag{2}$$

where $H(A)$ and $H(I)$ are marginal entropies and $H(A,I)$ is a joint entropy of the images. In this work, the computation of NMI was implemented as described in Maes et al. (1997).

The spatial transformations were defined using our in-house proprietary VolumeWarp registration software package (http://volumewarp.vtt.fi). The software is based on local registrations and the multi-resolution framework, an approach very similar to the method proposed in Andronache et al. (2008). The floating image, i.e., in our case the atlas, is divided to sub-images, and the similarity of each sub-image and the target image is maximised by a rigid registration stage. Linear interpolation is applied to transformation parameters between sub-images to guarantee a continuous transformation. One major reason for the improved speed is the careful optimisation of various components of the registration. The optimisation of registration includes approximations and simplifications of different routines, e.g., replacing NMI by intensity difference, and the maximised usage of the cache memory.

Intensity difference as a similarity measure

When intensity difference is used as a similarity measure, the following measure is maximised:

$$E_{data} = \sum_{p \in A \cap I} -\| T \circ A_p^{'} - I_p \|, \tag{3}$$

where $A_p'=A_p'(x, y, z)$ is an intensity normalised image at voxel p, and $\boldsymbol{T} \circ A_p'$ denotes a spatially transformed image.

In this work, intensity normalisation was implemented via a piecewise linear function, $m=m(g)$, which transforms intensity g to intensity $m(g)$ (Fig. 2). For brain MRI, the mapping function was determined by defining values for $m(g_{\text{CSF}})$, $m(g_{\text{GM}})$, and $m(g_{\text{WM}})$ where g_{CSF} , $g_{\text{GM,}}$ and g_{WM} are mean intensity values of CSF, GM, and WM, respectively. As the segmentations of these structures were included in the atlas, the intensities can be computed easily. If segmentation is not available, it can be computed using an auto-

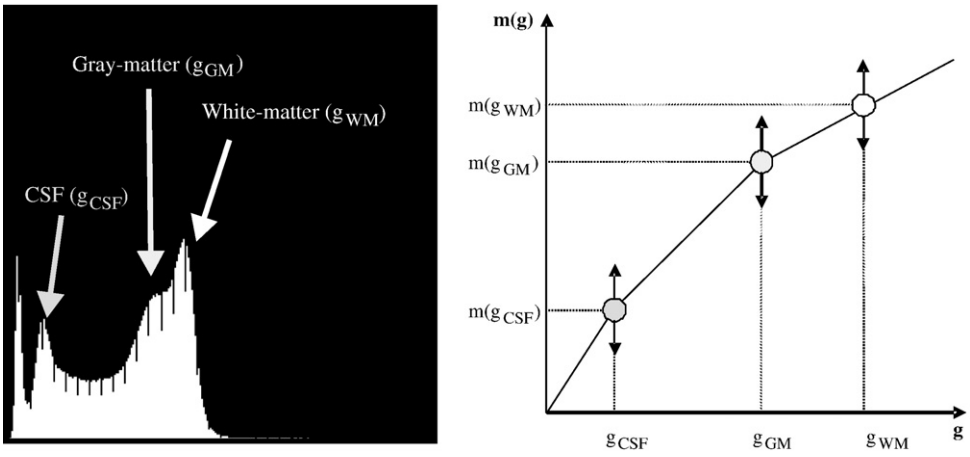

Fig. 2. Intensity normalisation via a piecewise linear mapping function, $m = m(g)$. Intensity values are first defined for the CSF (g_{CSF}), gray-matter (g_{GM}) and white matter (g_{WM}), indicated in the gray-scale histogram of the atlas (on left). The values of the mapping function $m(g_{CSF})$, $m(g_{GM})$, and $m(g_{WM})$ are optimised (demonstrated by arrows, on right) in such a way that the absolute value of the difference between the target image and intensity normalised atlas is minimised.

matic tissue classifier, e.g., proposed by van Leemput et al. (1999) or Pham and Prince (1999). Alternatively, these three values can be specified manually by the user.

The following iterative algorithm was used during the non-rigid registration:

1. Optimise the spatial transformation $\boldsymbol{T} = \mathbf{T}(x, y, z)$ while keeping $m = m(g)$ constant
2. Optimise the intensity mapping $m = m(g)$ while keeping $\boldsymbol{T} = \mathbf{T}$ (x, y, z) constant.
3. Go to step 1 if the maximum number of iterations has not been reached, otherwise stop.

Two approaches were tested for producing the piecewise linear intensity mapping.

Minimise intensity difference (MIN). The intensity mapping was optimised by an exhaustive search for the function values $m(g_{CSF})$, $m(g_{GM})$, and $m(g_{WM})$. We searched for the optimal combination of these three values by maximising Eq. (3). The mapping function was modified only gradually during each iteration; the search range for each value was $[m(g) - \Delta, m(g) + \Delta]$. In this work, we used $\Delta = 9$ but other small values could be used as well. The mapping was unity, $g = m(g)$, in the beginning. Schmidt (2005) used local intensity distributions to exclude regions where the alignment of images was not good. In this work, we formed a histogram from differences and excluded upper quartile (75% percentile) from the summation. This approach rejects voxels on the borders where the differences can be high due to misalignments. Alternatively, the differences were weighted by the square root of distances from the closest borders, computed from distance maps. However, neither of these strategies improved the segmentation accuracy and was not used in computing the final results.

Direct evaluation (DE). In this approach, the values g_{CSF}, g_{GM}, and g_{WM} for an atlas were estimated by averaging all voxel values under corresponding structures weighted by the square root of the distances from the closest border. The values $m(g_{CSF})$, $m(g_{GM})$, and $m(g_{WM})$ for the target image were estimated in a similar way. Because segmentations of CSF, GM, and WM for the target image were not available, the segmentations of the atlas were used.

The spatial transformations were defined using data from the whole head but only the brain region was used for the intensity normalisation. Intensity normalisation was performed only at the highest resolution level of the multi-resolution registration.

Atlas selection

Background

In the simplest form, atlases can be selected either randomly or using all the atlases available. However, in Aljabar et al. (2009), it was shown that the best multi-atlas segmentation accuracy is obtained by optimally selecting a subset of the atlases (about 10–20) instead of using all the atlases. There are several ways how to intelligently select atlases (Aljabar et al., 2009; Rohlfing et al., 2004; Wu et al., 2007). Most often, the selection is done based on an intensity-based similarity measure computed for the atlases and the target image. In addition, the magnitude of the deformations from the atlases to the target image (Rohlfing et al., 2004) and demographic data (Aljabar et al., 2009) have been proposed for atlas selection. Aljabar et al. (2009) performed atlas selection in a template space. All atlases and a target image were registered using a 12-affine transformation to a separate template. This reduced the computational load significantly, as compared to atlas selection in a target space, i.e., all the atlases registered to the target image. Artaechevarria et al. (2009) demonstrated recently an alternative approach where atlas selection was not performed but a weigh factor, based on a similarity measure, was defined for each atlas. They showed that defining the weights locally produces better results than global weighting. In STAPLE (Warfield et al., 2004), the performance level of each atlas is estimated using expectation maximisation (EM) algorithm, and the individual segmentations are combined by weighting the atlases based on their performance level.

Atlas selection methods studied

Several methods to select atlases for majority voting were tested. The simplest way is to randomly select atlases from a database, i.e., to select n atlases randomly from a set of N atlases, where $n < N$, providing a baseline for the selection strategies.

Intensity-based selection methods. In the previous atlas selection studies using intensity-based measures (Aljabar et al., 2009; Rohlfing et al., 2004; Wu et al., 2007), normalised mutual information (NMI) has proven to be the best choice and was chosen also for this study. The NMI value was computed from the structures of interest by dilating the binary segmentations of the structures three times and using the resulting binary image as a mask for the NMI computation. The dilatation was used for including the borders of the structures and their small surrounding into the mask.

Five methods were used for atlas selection:

AS1. The atlases and the target image were affinely registered to a single template (in template space), the NMI between each atlas and target image was computed, and the n atlases with the highest NMI values were selected for the multi-atlas segmentation (Aljabar et al., 2009).

AS2. After the affine registration, the atlases and the target image were non-rigidly registered to a template MRI image. Then, the NMI values between the deformed target image and deformed atlases were computed and the n atlases with the highest NMI values were selected.

AS3. In the third method, multiple templates were used. Three templates were chosen to represent different subgroups of the datasets: Alzheimer's disease, mild cognitive impairment and control subgroups for the ADNI data and three age groups for the IBSR data (see Image data section below). The target image and all atlases were non-rigidly registered to each template and the NMIs were computed between the target image and atlases. Finally, the atlases were ranked based on the maximum value of the three NMI values. Our hypothesis is that if the template used is similar to the atlas and the target image, registration errors are smaller and the similarity value becomes higher than using a dissimilar template. All methods 1–3 are fast, requiring only a small number of affine and non-rigid registrations of the target image. These registrations can be efficiently computed simultaneously with multi-processor computers.

AS4. For comparison, all the atlases were first affinely and then non-rigidly registered to the target image, requiring notably more computation time than methods 1–3.

AS5. A well-known STAPLE algorithm presented by Warfield et al. (2004) was used as a reference method. The STAPLE was applied to the images that were non-rigidly registered to the target image (same as in AS4). The binary STAPLE was applied for each structure of the IBSR data to decrease computation time as compared to the multi-class method proposed by Warfield et al. (2004).

Non-image-based selection methods. Atlas selection from non-image-based data is a tempting option as it does not require any image registrations. In this study, we tested the utilisation of demographic information in selection. The information used was age and MMSE (Mini-Mental State Examination) score. The differences of the values of the target subject and the atlas subjects were computed, and the atlases were ranked based on the absolute values of the differences. The combination of intensity-based and non-image-based measures was also tested by using weighted sum of the measures. Different values were tested for the weights and the values producing the highest segmentation accuracy were chosen.

In addition, we studied the segmentation accuracy if the optimal set of atlases was selected. This accuracy was then compared to the results obtained with the atlas selection methods presented above. In this study, the optimal set of atlases was obtained as follows. The atlases were added to the multi-atlas segmentation one by one. The combination that produced the best segmentation accuracy was determined in each iteration. This was continued until all the atlases were used. This was repeated for all target subjects, and the segmentation errors were averaged.

Combined multi-atlas segmentation and intensity modelling

Background

If intensity difference is used as a similarity measure, the registration algorithms implicitly expect that the intensity distributions of different structures in an atlas and a target image are fairly similar. This assumption is not strictly valid in practice. Modelling of intensity distributions of different structures or tissue types provides data for classifying voxels; in principle the intensity of each voxel is compared with the intensity distributions and the most probable class is chosen using, for example, a Bayesian framework (Han and Fischl, 2007; van Leemput et al., 1999). In many cases, a probabilistic atlas is used as a priori information to constrain the segmentation.

van der Lijn et al. (2008) recently proposed a technique to further improve the accuracy of multi-atlas segmentation taking into account this intensity modelling aspect. Their method uses graph cuts to optimise an energy function based on the following terms: a statistical intensity model, a spatial prior derived from multi-atlas registrations, and a regularisation term based on Markov Random Field (MRF).

The use of the graph cuts for optimisation is attractive as it provides the global minimum or maximum of an energy functional. However, the segmentation produced is optimal only if the energy function is able to separate perfectly the structure from the background, i.e., all assumptions are valid and all necessary parameters can be estimated correctly. One limitation of the method proposed by van der Lijn et al. (2008) is the reliance on an intensity model derived from manual training, which restricts its application to images acquired with the same MRI sequence.

We used two alternative methods for the problem (1) based on a modification of the graph cuts approach presented in van der Lijn et al. (2008) and 2) based on the well-known expectation maximisation (EM) algorithm (van Leemput et al., 1999).

Graph cuts approach

In Wolz et al. (2009) , we proposed a modified version of van der Lijns graph cuts approach that does not rely on manual training and that can be applied to more than one structure of interest. A Markov Random Field (MRF) is defined for the segmentation of an unseen image with graph cuts. As in van der Lijn et al. (2008), the a priori probability of a voxel being in foreground or background of a structure of interest is determined from a subject-specific probabilistic atlas obtained from multi-atlas segmentation. This spatial prior is combined with an intensity model for foreground and background that is directly estimated from the target image. This generalised intensity model makes the approach more robust to a variation in grey-level intensities resulting from different MRI sequences and therefore the method applicable to a broader range of images. For more details of this approach, see Appendix A or Wolz et al. (2009).

Expectation maximisation (EM) approach

We used energy terms similar to those of van der Lijn et al. (2008). As in the graph cuts approach, the intensity model was computed directly from the target volume. The details of the energy terms are described in Appendix B.

The classification algorithm used was as follows:

1. Estimate model parameters mean μ and standard deviation σ (maximisation step of the EM algorithm, M-step).
2. For each voxel $v_p \in V$, define classes C in the 6-neighborhood including also voxel v_p.
3. Classify voxel v_p to a class from C according to the maximum a posterior probability (expectation step of the EM algorithm, E-step).
4. Iterate until the segmentation does not change.

The motivation for defining the classes (C) in the neighbourhood of each voxel is that the number of possible classes becomes small, in most cases only two. As only voxels on the borders are processed during each iteration (number of classes in $C>1$), the object is updated in a similar way to the well-known region-growing

approach. This reduces the need for the regularity prior. Because we noticed that the use of the regularity prior does not improve the segmentation accuracy, it was not used in computing the final results.

The algorithm requires that segmentations of objects surrounding the object of interest are available. Otherwise, the classes C in the neighbourhood of each voxel cannot be defined. As only hippocampus segmentations were available in ADNI data (see below), the following procedure was adopted: WM, GM, and CSF were segmented from the atlases using the method by van Leemput et al. (1999), and propagated to the target space using the deformations obtained. Then, the surroundings of the hippocampus segmentation were replaced by this tissue segmentation. In this case, the surroundings of the object of interest did not contain segmentations of anatomical structures but only tissue classes.

Image data

The experimental validation of the developed algorithms was performed using data from two publicly available datasets containing manual segmentations.

IBSR data

T1 weighted MR image volumes from 18 subjects (4 females and 14 males) with age between 7 and 71 years were used (Fig. 3). The size of the volumes were 256 × 256 × 128 voxels with the voxel size from 0.8 × 0.8 × 1.5 mm to 1.0 × 1.0 × 1.5 mm. The images were spatially normalised into the Talairach orientation (rotation only). In addition to intensity images, the data contained two separate segmentations: one with a tissue classification into CSF, GM, and WM and another for 34 different structures. In multi-atlas segmentation, cross-validation was used, that is, the case to be segmented was left out from the set of atlases which contained therefore 17 atlases. The MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/.

ADNI data

The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The principle investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California—San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 adults, aged 55 to 90 years, to participate in the

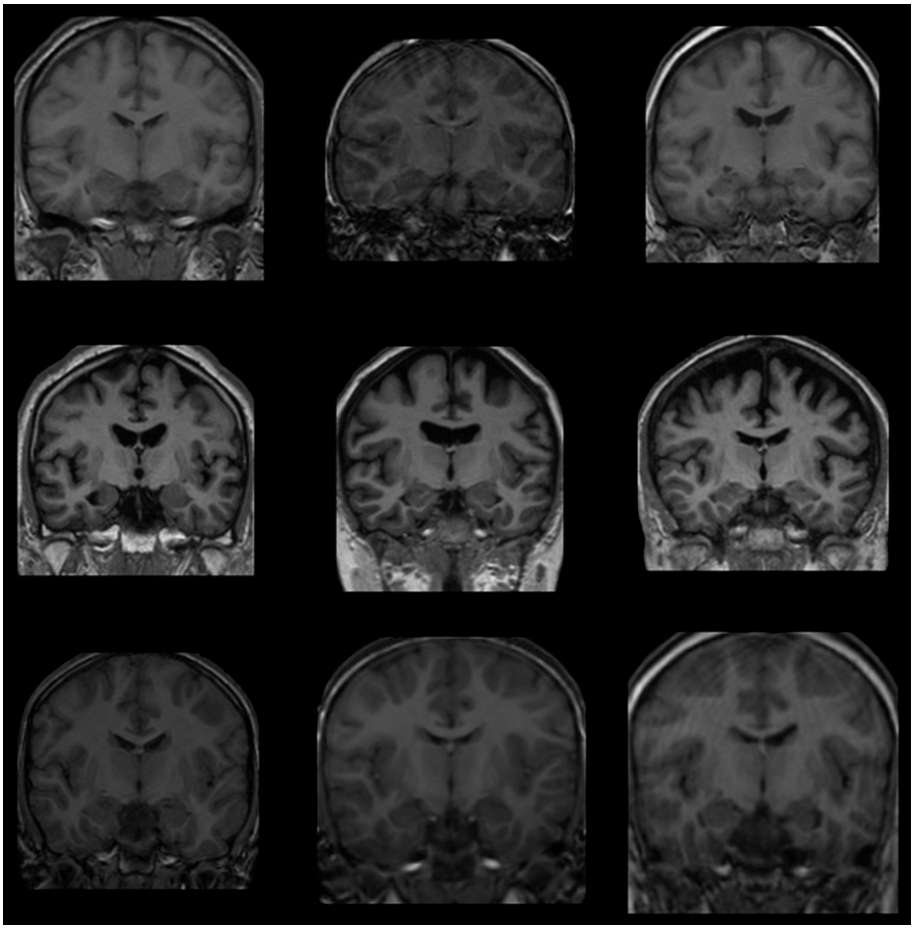

Fig. 3. Coronal slices from nine IBSR cases.

research—approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years.

T1-weighted 1.5-T MR images were studied from 60 subjects in the ADNI database, http://www.loni.ucla.edu/ADNI (Fig. 4). The ADNI consortium has classified data into three groups: Alzheimer's patients (AD), mild cognitive impairment (MCI) and control subjects (controls). In this study, 20 subjects, having manual segmentations available, were chosen randomly from each group (Table 1). The images were acquired using MRI scanners from three different manufacturers (General Electric Healthcare (GE), Siemens Medical Solutions, Philips Medical Systems) and using a standardised acquisition protocol. Acquisition parameters on the SIEMENS scanner (parameters for other manufacturers differ slightly) were echo time (TE) of 3.924 ms, repetition time (TR) of 8.916 ms, inversion time (TI) of 1000 ms, flip angle 8°, to obtain 166 slices of 1.2-mm thickness with a 256×256 matrix. The size of the volumes were from 192 × 192 × 160 to 256 × 256 × 180 voxels with the voxel size from 0.9 × 0.9 × 1.2 mm to 1.3 × 1.3 × 1.2 mm.

Table 1
Demographic data and clinical scores for 60 ADNI cases used in the study; the mean value and standard deviation are shown.

	Number	Male/Female	Age	MMSE
Control	20	6/14	76.5 ± 6.3 [63–88]	28.8 ± 1.3 [25–30]
MCI	20	14/6	75.9 ± 8.0 [61–88]	26.6 ± 2.4 [21–30]
AD	20	8/12	75.5 ± 8.1 [57–89]	22.3 ± 3.4 [10–26]

The minimum and maximum values are in brackets. The abbreviations used are MCI, mild cognitive impairment; AD, Alzheimer's disease; MMSE, Mini-Mental State Examination.

For each image, a manual segmentation of the hippocampus was provided by ADNI. The set of atlases used in multi-atlas segmentation consisted of 30 ADNI images, different from the 60 cases used for evaluation. The atlas contained cases from AD, MCI and controls, 10 from each.

As manual segmentations were available only for hippocampus in ADNI, a subvolume was automatically extracted containing left and right hippocampus. This was done to speed up the computation. The size of the subvolume used was 100 × 100 × 100 voxels.

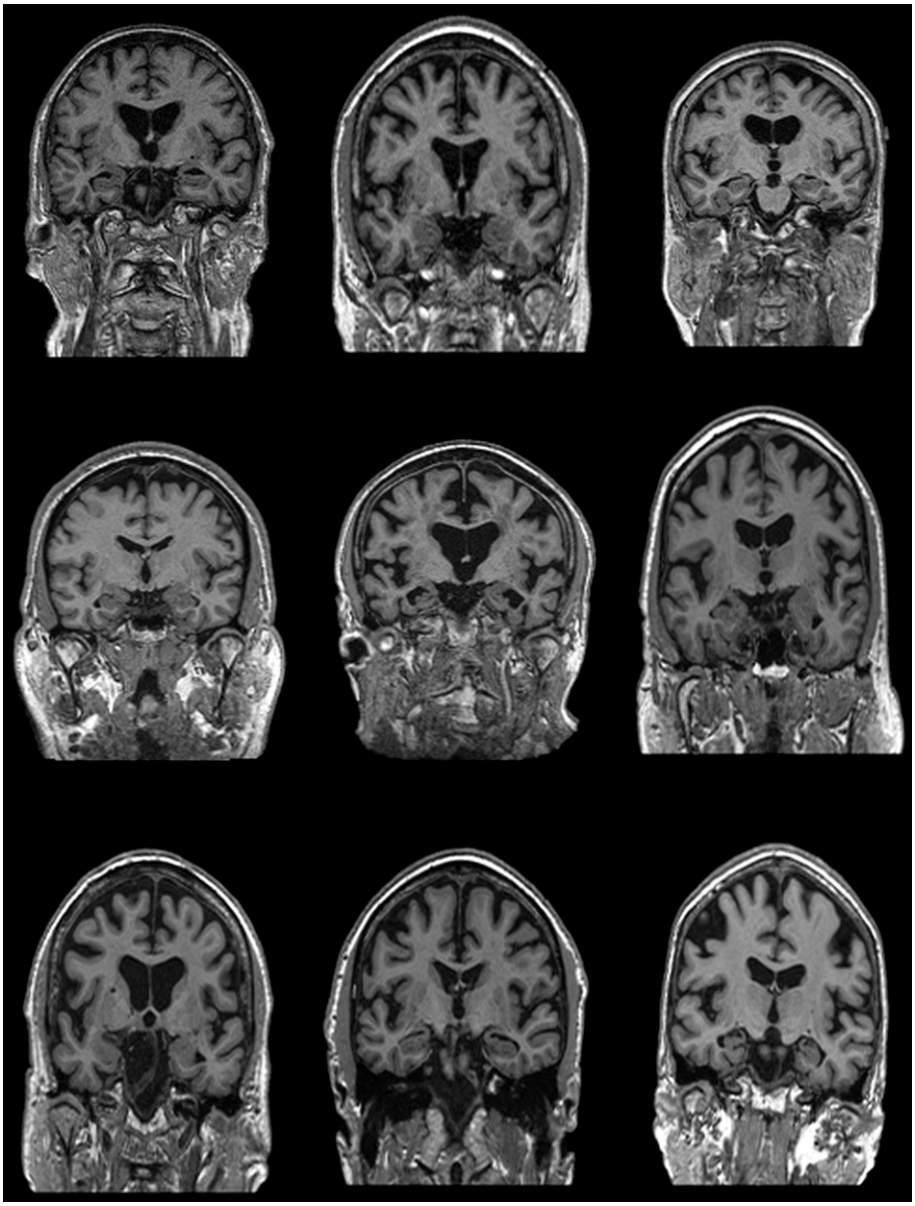

Fig. 4. Coronal slices from nine ADNI cases.

Evaluation tools

Because the Dice similarity index (SI) is one of the most widely used measures in assessing the performance of segmentation, it was a basis for the comparison. In addition, we report some other commonly known measures when summarising the results:

- Similarity index (SI) $= 2\frac{A\cap B}{A+B}$
- Precision$=\frac{A\cap B}{B}$
- Recall$=\frac{A\cap B}{A}$
- Distance$=[d(A\rightarrow B)+d(B\rightarrow A))]/2$

where A and B represent automatically and manually generated segmentations and $d(A\rightarrow B)$ is the distance of the surface A from the surface B.

The statistically significant differences between groups were studied by Wilcoxon Rank Sum test for paired samples (SPSS 14.0 For Windows, Chicago, USA). The difference between similarity indices was considered statistically significant if $p<0.05$.

In addition, correlation coefficients between hippocampus volumes based on automatic and manual segmentations were computed for the ADNI data.

Results

Intensity difference as a similarity measure

The similarity indices produced after applying different intensity normalisation methods are shown in Table 2 for different subcortical structures using IBSR data. In the single-atlas case, the values are averages over all atlases ($N=17$). Each atlas from the database was used separately to segment the data. In multi-atlas segmentation, all available cases were used in the voting ($N=17$), i.e., no atlas selection was performed.

The results indicate an expected finding that intensity normalisation is needed if intensity difference is used as a similarity measure. Two previously published methods for intensity normalisation, published by Nyúl et al (2000) and Hellier (2003), were tested but the average of similarity values was lower than obtained with NMI-based segmentation (difference statistically significant, except for 'ID Hellier (2003)' in the Multi-atlas approach). When the intensity normalisation methods developed in this work were applied and the AVG column was analysed, no difference compared with the NMI-based method was observed except for the direct evaluation (DE) in the single-atlas approach. No statistically significant difference was identified between minimise intensity difference (MIN) and direct evaluation methods.

The results show also the well-known result that the multi-atlas method performs better than the single-atlas method (sub-tables 'Single-atlas' vs. 'Multi-atlas'; difference statistically significant for the averages of structures). In addition, the combination of multi-atlas segmentation and intensity modelling improves the accuracy compared with the situation when only multi-atlas segmentation is used (sub-tables 'Multi-atlas' vs. 'Multi-atlas+EM'; difference statistically significant for all intensity normalisation methods when computed for the AVG column).

In Lötjönen et al. (2009), we showed that a combination of NMI and image gradient-based features increases the segmentation accuracy. The row 'ID MIN+' shows for comparison results when this gradient term was used in addition to regulating the curvature of the transformation, i.e., using the term E_{model} in Eq. (1). Statistically significant differences compared with the row 'NMI' are shown in the table. If 'ID MIN+' is compared with 'ID MIN', the

Table 2
Similarity index produced after applying different intensity normalisation methods using single-atlas, multi-atlas, and combined multi-atlas and intensity modelling approaches (EM approach).

	Thalamus	Caudate	Putamen	Pallidum	Hippoc	Amygdala	AVG
Single atlas							
NMI	0.830	0.748	0.815	0.693	0.689	0.591	0.728
ID NO	0.812*	0.729*	0.743*	0.607*	0.680	0.555*	0.688*
ID Hellier (2003)	0.813	0.728	0.733*	0.619*	0.674	0.534*	0.683*
ID Nyúl et al. (2000)	0.809*	0.728*	0.762*	0.602*	0.696	0.574	0.695*
ID MIN	0.838*	0.754	0.792*	0.645*	0.695	0.578	0.718
ID DE	0.838*	0.753	0.793*	0.644*	0.695	0.580	0.717*
ID MIN+	0.849*	0.764*	0.845*	0.760*	0.724*	0.659*	0.767*
Bhattacharjee et al. (2008)	0.820	0.750	0.840	0.760	0.660	0.610	0.740
Multi-atlas							
NMI	0.882	0.836	0.881	0.785	0.802	0.726	0.819
ID NO	0.872	0.824	0.847*	0.740	0.793	0.699*	0.796*
ID Hellier (2003)	0.860	0.800	0.818*	0.739	0.790	0.663*	0.778
ID Nyúl et al. (2000)	0.878	0.836	0.860*	0.721*	0.804	0.716	0.803*
ID MIN	0.890*	0.841	0.876	0.757*	0.805	0.720	0.815
ID DE	0.891*	0.843	0.876	0.749*	0.805	0.719	0.814
ID MIN+	0.888*	0.847	0.898*	0.833*	0.804	0.752*	0.837*
Multi-atlas + EM							
NMI	0.889	0.853	0.896	0.803	0.818	0.737	0.833
ID NO	0.871*	0.843	0.857*	0.756*	0.811	0.725	0.811*
ID Hellier (2003)	0.861*	0.827	0.840*	0.767*	0.810	0.703	0.801*
ID Nyúl et al. (2000)	0.888	0.855	0.881*	0.753*	0.817	0.731	0.821*
ID MIN	0.899*	0.865*	0.890	0.780*	0.819	0.740	0.832
ID DE	0.898*	0.864*	0.888*	0.775*	0.818	0.738	0.830
ID MIN+	0.896*	0.866*	0.905*	0.844*	0.814	0.767	0.849*
Han and Fischl (2007)	0.88	0.84	0.85	0.76	0.83	0.75	0.818
Heckemann et al. (2006)	0.90	0.90	0.90	0.80	0.81	0.80	0.852
Artaechevarria et al. (2009)	0.88	0.83	0.87	0.81	0.75	0.72	0.810

Results are reported for six subcortical structures. Statistically significant differences are indicated by asterisk (*) when compared with the values on the NMI row. Abbreviations used: NMI , normalised mutual information; ID , intensity difference; NO , no intensity normalisation; Hellier , method presented in Hellier (2003); Nyul , method presented in Nyúl et al. (2000); MIN , minimise ID; DE , direct evaluation (Section 3.2.1); MIN+ , as MIN but gradient features (Lötjönen et al., 2009) and the regularisation of transformation is also used; AVG , average over six subcortical structures. For comparison, results from four other publications are given.

Table 3
Computation times in seconds for normalised mutual information (NMI) and intensity difference (ID) as a similarity measure in non-rigid registration.

	Non-rigid (1 Core)	Total (8 Core)
NMI	126 s	416 s
ID	44 s	266 s

The first column shows the computation time for registering non-rigidly single atlas to a target image using 1 Core. The second column shows the total computation time of multi-atlas segmentation including also pre- and post-processing steps and using 14 atlases in a standard 2 processor 4 Core PC computer.

average of all structures is higher in 'ID MIN+' (difference statistically significant, not shown in Table 2). The difference is also statistically significant separately for all structures in the single-atlas approach but not in the multi-atlas approaches. For example, the segmentation accuracy of the hippocampus does not improve in the multi-atlas segmentation by using image gradients and curvature regularisation.

For comparison, results from four other publications are shown. When compared with the results from (Bhattacharjee et al., 2008) using also IBSR data and single-atlas approach, the accuracy is comparable. However, Bhattacharjee et al. (2008) reported that 8 min were needed for registration, which is slower than our algorithm (44 s). Results are also comparable when the accuracy of multi-atlas segmentation is compared with three other previously published methods (Artaechevarria et al., 2009; Han and Fischl, 2007; Heckemann et al., 2006). Detailed comparisons with Han and Fischl (2007) and Heckemann et al. (2006) are not possible as the data used in those two publications are not from the IBSR database. However, Artaechevarri et al. (2009) used IBSR data.

The computation times are presented in Table 3. The results show that NMI-based registration is three times slower than non-rigid registration based on intensity differences. The value is only indicative as the actual implementation of the measures affects the results. However, we have tried to optimise also the computation of NMI but further optimisation might still be possible. The relative difference in the total computation time is not as dramatic because time needed for pre- and post-processing operations is equal in both cases.

Similarity index, precision, recall, and the distance between surfaces are reported in Table 4 for the segmentations produced using the gradient component and the curvature regularisation (the row 'ID MIN+' in Table 2).

Atlas selection

The mean similarity indices for IBSR and ADNI data are shown in Figs. 5a and b, respectively. The number of non-rigid registrations needed for atlas selection using different methods studied is listed in Table 5. In addition, the affine registration of a target image to the template space and the non-rigid registrations of the atlases selected to the target image need to be taken into account when considering the total computation time.

Both datasets showed similar behaviour. With the optimal atlas selection, the best segmentation accuracy was obtained with relatively few atlases, about 8–15. After this, the segmentation accuracy worsened when more atlases were added. Consequently, the possible saving in the computation time obtainable with atlas selection is remarkable, especially for large datasets of atlases.

From the atlas selection methods studied, one based on NMI after non-rigidly aligning atlases to a target image (AS4) turned out to be best. Non-rigid registration of atlases to a single template space (AS2) gave slightly worse results for the IBSR data but almost identical results for the ADNI data, but this strategy required only one non-rigid registration. Utilisation of three templates instead of just one (AS3) improved the results close to the results of the NMI in the target space for the IBSR data. All these selection methods performed better than the STAPLE algorithm (AS5) on the ADNI data (difference statistically significant), whereas on the IBSR data, the STAPLE algorithm outperformed atlas selection. However, the STAPLE was applied separately for each structure, but atlas selection and voting were applied simultaneously for each structure when using the IBSR data. When atlas selection and voting were performed separately for each structure, the similarity index of AS4 increased from 0.805 to 0.814, which was close to the accuracy of the STAPLE (0.815, difference not statistically significant). Affine registration to template space (AS1) gave clearly worse results but still better than the results of random selection. The selection based on age was better than NMI after affine registration in the case of the IBSR dataset, and the combination of these two still improved the results. On the other hand, the selection using demographic data did not give as good results for the ADNI dataset (not shown in the Fig. 6b for clarity). This may be due to the smaller age range of the ADNI dataset and the different structures to be segmented. In addition, we tested the performance of the Mini-Mental State Examination (MMSE) score in atlas selection. However, no increase in the accuracy was obtained when combined with the method where the data were registered non-rigidly to the template space (curve 'non-rigid, template space' in Fig. 5).

Segmentation of hippocampus from ADNI data

Both left and right hippocampi were segmented from 60 ADNI cases. The results for multi-atlas segmentation with all 30 atlases and for 13 atlases (maximum in Fig. 5b) selected either in target space or template space are shown in Table 6. In addition, the graph cuts and EM approaches have been applied to all the segmentation results. The similarity indices are shown for each

Table 4
Similarity index, precision, recall, and the average distance in millimetres between surfaces and its standard deviation for segmentation of six subcortical structures from IBSR data ($N=18$).

	Thalamus	Caudate	Putamen	Pallidum	Hippoc	Amygdala	AVG
Single atlas							
Similarity index	0.849	0.764	0.845	0.760	0.724	0.659	0.767
Precision	0.818	0.763	0.817	0.741	0.692	0.646	0.746
Recall	0.890	0.790	0.879	0.787	0.766	0.701	0.802
Distance (average)	0.96	0.86	0.70	0.84	0.96	1.21	0.92
Distance (SD)	0.38	0.34	0.11	0.14	0.22	0.33	0.17
Multi-atlas + EM							
Similarity index	0.896	0.866	0.905	0.844	0.814	0.767	0.849
Precision	0.872	0.863	0.889	0.824	0.763	0.722	0.822
Recall	0.926	0.876	0.924	0.871	0.878	0.829	0.884
Distance (average)	0.74	0.57	0.50	0.64	0.74	0.93	0.69
Distance (SD)	0.32	0.22	0.06	0.11	0.18	0.27	0.13

The values are for the segmentations produced by the 'ID MIN+' configuration in Table 2.

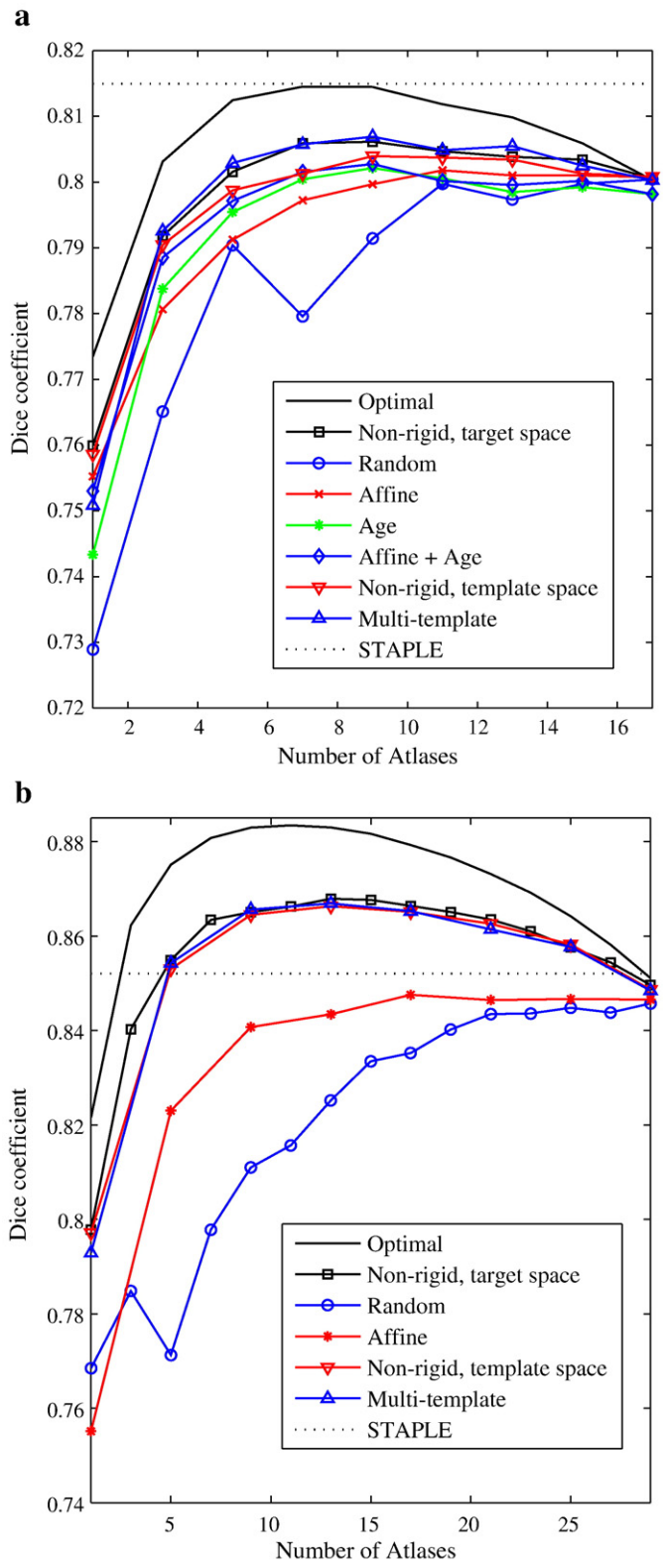

Fig. 5. Similarity indices for different number of atlases and for different atlas selection methods for (a) IBSR dataset and (b) ADNI dataset.

case (120 segmentations) in Fig. 6 using atlas selection in template space and the graph cuts approach. The best average SI of the methods studied was 0.885. Morra et al. (2009) studied the similarity index between two human raters using ADNI data ($N = 21$). They obtained the value SI = 0.853. These values indicate that our segmentation pipeline produced results comparable to the accuracy of manual segmentation.

Table 5
Number of non-rigid registrations needed for atlas selection.

Random	0
Affine, template space	0
Non-rigid, template space	1
Multi-template	Number of templates
Non-rigid, target space	Number of atlases
Demographics	0

In Wolz et al. (2009), 60 images and 30 atlases from ADNI were used for multi-atlas segmentation with the well-known registration algorithm by Rueckert et al. (1999). The cases were the same as we used in this work. In their article, an average overlap of SI = 0.86 without atlas selection was reported. Applying the same strategy with our algorithm leads to an average overlap of SI = 0.87. Using atlas selection (13 atlases) with registrations produced using the algorithm by Rueckert et al. leads to an average overlap of SI = 0.88, which is identical to the value reported in this work. In Wolz et al. (2009), the computation time of one registration was around 1 h on a multi-core PC computer.

We studied also the correlation of the automatically and manually computed volumes of hippocampus. Fig. 7 shows a scatter plot of the hippocampus volumes when 13 atlases were selected in template space and the graph cuts approach was used. The correlation coefficient was 0.95 ($R^2 = 0.9037$). The value 0.854 was reported in Morra et al. (2009) for two human raters.

The total computation time for segmenting one case including also pre- and post-processing steps and using 13 atlases selected in template space was about 3 min using a standard 2 processor 4 Core PC computer. For comparison, van der Lijn et al. (2008) reported that non-rigid registration required 5 to 8 h for each 19 atlases using a single core computer. The time needed for non-rigid registration using only a single core in our system was 17 s for a subvolume of 100 × 100 × 100 voxels and 2 min 20 s for the original volume of 256 × 256 × 166 voxels. When compared with the IBSR results (Table 3), much more iterations were performed with ADNI data to maximise the segmentation accuracy.

Discussion

In this work, different steps of multi-atlas segmentation were studied: non-rigid registration, atlas selection, and post-processing steps. All these factors have an important role in multi-atlas segmentation. We demonstrated that the segmentation accuracy can be clearly improved when optimising these factors. The results of automatic segmentation showed a good overlap with manual segmentations: the average SI was 0.849 for six subcortical structures (IBSR data) and 0.885 for the hippocampus (ADNI data). The correlation coefficient for hippocampal volumes in ADNI was high, 0.95.

Intensity normalisation is a prerequisite for using intensity difference as a similarity measure. We proposed two methods that produced piecewise linear transformation for intensities. Intensities of CSF, GM, and WM were matched between images. We demonstrated that using intensity difference as a similarity measure produced equal segmentation accuracy compared with standard NMI-based segmentation. The computation time needed for non-rigid registration was, however, decreased by a factor of 3. With IBSR data, the registration time of an atlas to target image reduced from 126 to 44 s (Table 3). This finding makes multi-atlas segmentation more attractive to clinical practice where computation time plays a crucial role. The major limitation in using intensity differences is that images to be segmented and atlases used should be acquired with

Table 6
Similarity index, precision, recall, distance in millimetres between surfaces and correlation coefficients for segmentation of hippocampus from ADNI data ($N = 60$).

ADNI data, hippocampus ($N = 60$)	Similarity index	Precision	Recall	Distance (average ± SD) [mm]	Correlation of volumes
Multi-atlas (30 atlases)	0.846	0.872	0.833	0.54 ± 0.14	0.66
Multi-atlas (30 atlases)+ GC	0.869	0.894	0.851	0.48 ± 0.09	0.89
Multi-atlas (30 atlases)+ EM	0.866	0.863	0.880	0.50 ± 0.16	0.71
AS non-rigid target space (13 atlases)	0.868	0.873	0.867	0.48 ± 0.10	0.91
AS non-rigid target space (13 atlases) + GC	0.882	0.887	0.879	0.44 ± 0.07	0.95
AS non-rigid target space (13 atlases) + EM	0.883	0.870	0.902	0.45 ± 0.09	0.94
AS non-rigid template space (13 atlases)	0.866	0.890	0.849	0.48 ± 0.08	0.93
AS non-rigid template space (13 atlases) + GC	0.880	0.899	0.864	0.45 ± 0.06	0.95
AS non-rigid template space (13 atlases) + EM	0.885	0.884	0.890	0.44 ± 0.07	0.94
Morra et al. (2009), ADNI data ($N = 21$) AUT	0.856	0.845	0.875	0.005	0.71
Morra et al. (2009), ADNI data ($N = 21$), MAN	0.854	0.877	0.836	0.004	0.71
van der Lijn et al. (2008)Lijn et al (2008), Rotterdam study ($N = 20$), AUT	0.858			0.38 ± 0.08	0.81
van der Lijn et al. (2008)Lijn et al (2008), Rotterdam study ($N = 20$), MAN	0.858			0.33 ± 0.08	0.83

For comparison, the corresponding values from two other recent studies are shown. Abbreviations: GC , graph cuts; AS , atlas selection; AUT , automatic segmentation; MAN , manual segmentation for van der Lijn et al. (2008) for intra-rater reliability (the values are averages for the left and right hippocampus) and for Morra et al. (2009) for inter-rater reliability.

approximately similar imaging parameters, for example, both images should be T1-weighted MRI images.

Different atlas selection methods were compared. The method based on the similarity of the atlas and the target image in the template space after non-rigid registration (AS2) provided a good compromise between the accuracy and computation time. The accuracy based on non-rigid registrations was clearly better than using only affine registrations (AS1), as done in Aljabar et al. (2009). Although the atlas selection methods evaluated gave better segmentation results than the random selection, there was still a clear difference to the results of the optimal atlas selection. This demonstrates that atlas selection is not a trivial task, and the methods should be further developed.

We also compared our atlas selection with the STAPLE algorithm. When performed for each structure separately, no difference was observed between the approaches when using IBSR data. When using ADNI data, however, atlas selection outperformed the STAPLE. The reason for the difference remained unclear and requires more studies.

Two methods were proposed for post-processing where multi-atlas segmentation is combined with statistical modelling of intensity distributions: a method based on (1) the graph cuts algorithm and (2) the EM algorithm. When compared with standard multi-atlas segmentation, the accuracy was increased by 0.01−0.02 by both algorithms. The improvement is comparable with the results presented in van der Lijn (2008) but our approach avoids the tedious and restricting training phase. Our results clearly show that an intensity-based refinement step improves the accuracy of multi-atlas segmentation. Both the graph cuts and EM algorithms produce approximately similar improvements. One practical difference can be noticed between the methods proposed. The EM algorithm can be applied directly to multi-object segmentation while the graph cuts algorithm must be applied separately to each object to be segmented.

The techniques proposed in this work increased the segmentation accuracy. However, the improvements were relatively small in terms of similarity index. We believe that there is not anymore much space for dramatic improvements in the accuracy because the segmentation error of subcortical structures, reported in many publications, start to approach the inter-observer error of manual segmentations; the similarity index is about 0.85 for hippocampus between manual segmentations.

The computation time for the multi-atlas segmentation took 3−4 min using a standard multi-core PC computer. The value was clearly lower than what has been reported in many articles previously. For example, in van der Lijn et al. (2008), the computation time was several hours. Our results are comparable to the ones recently reported in Chupin et al. (2009); they reported the similarity index of 0.85 for hippocampus (average for three cohorts) with a computation time of 15 min. The registration was done using SPM5.

When the segmentation accuracy of hippocampus is considered, a clear difference in the similarity index was observed between the ADNI data (0.88) and the IBSR data (0.82). There are several potential reasons for this. First, the image quality in ADNI data is better than in the IBSR data. Second, the protocol used in manual segmentation can be different making some protocols more favourable to automatic algorithms. Third, the clinical status and the demographic data of the subjects were different: IBSR contained data from children to aged

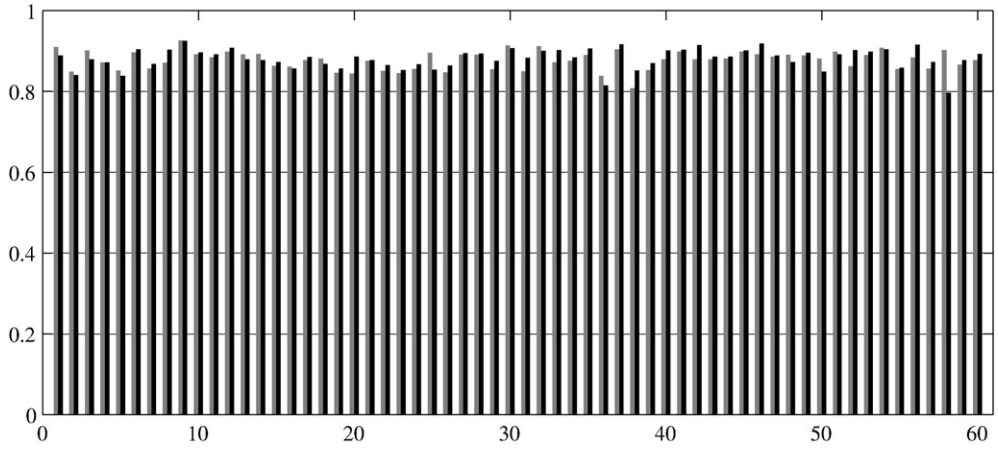

Fig. 6. Similarity indices for left and right hippocampus of 60 ADNI cases using 13 atlases selected in a template space and graph-cut approach.

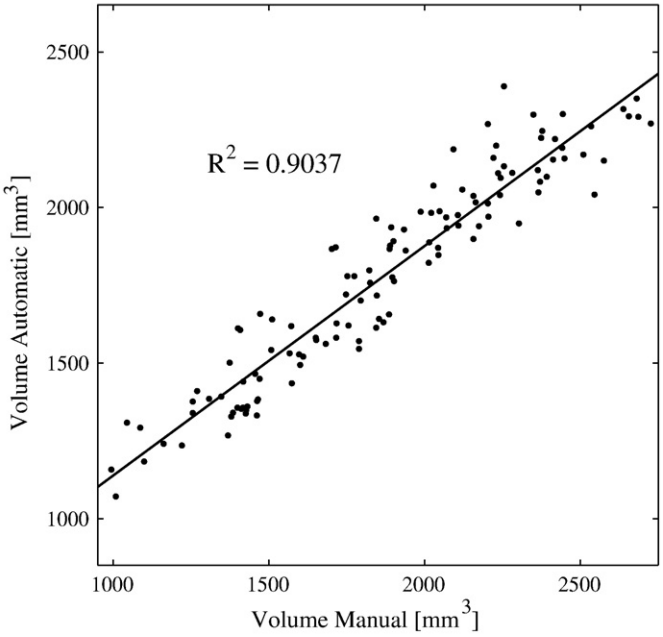

Fig. 7. Manually and automatically defined volumes for hippocampus using 13 atlases selected in a template space and graph cuts approach.

subjects while ADNI data consisted of aged controls and Alzheimer's disease patients. A careful analysis of these reasons and their effect to the robustness of segmentation is a highly relevant topic for future studies.

The methods proposed in this work are generic and can be incorporated also into many other tools available. Comparison of our results with the ones obtained with multi-atlas segmentation based on an established registration algorithm (Rueckert et al., 1999) showed identical results.

If the main results of our pipeline optimisation are summarised, the following three observations are made: (1) Intensity difference can be used instead of NMI in non-rigid registration without compromising the segmentation accuracy if intensities are normalised properly. This leads considerably shorter computation time. (2) Performing atlas selection in the template space after applying non-rigid registration provides a good compromise between the improved accuracy and the computation time needed. (3) Combining intensity modelling with the multi-atlas segmentation improves clearly the segmentation accuracy. Either the graph cuts or EM algorithm can be used.

Accurate and fast segmentation of images is a central component when the information in MRI images is exploited in the diagnostics. Despite promising results, several topics remain for future research. Our development work will focus on further reducing the computation time. For example, optimising the pre-processing steps, which have not been optimised yet at all, could lead to clear improvements in the computation time. Another important topic is guaranteeing the robustness when heterogeneous and sometimes non-optimal clinical data are used.

Acknowledgments

This work was partially funded under the 7th Framework Programme by the European Commission (http.//cordis.europa.eu/ist; EU-Grant-224328-PredictAD; Name: From Patient Data to Personalised Healthcare in Alzheimer's Disease) and Tekes–Finnish Funding Agency for Technology and Innovation (www.tekes.fi; Name: Extraction of diagnostic information from medical images).

The Foundation for the National Institutes of Health (www.fnih.org) coordinates the private sector participation of the $60 million ADNI public–private partnership that was begun by the National Institute on Aging (NIA) and supported by the National Institutes of Health. To date, more than $27 million has been provided to the Foundation for NIH by Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffmann-La Roche, Schering-Plough, Synarc Inc., and Wyeth, as well as non-profit partners the Alzheimer's Association and the Institute for the Study of Aging.

Appendix A. Graph cuts formulation

To assign a label $f_p \in L$ to each voxel $p \in I$, a MRF-based energy function is defined as

$$E(f) = \lambda \sum_{p \in I} D_p\left(f_p\right) + \sum_{\{p,q\} \in H} V_{p,q}\left(f_p, f_q\right) \quad (4)$$

where H is a neighbourhood of voxels and f is the labelling of I (Boykov et al., 2001). The data term D_P measures the disagreement between a prior probabilistic model and the observed data. $V_{p,q}(f_p, f_q)$ is a smoothness term penalising discontinuities in H. The parameter λ was in our experiments empirically set to $\lambda = 2$.

To optimise the previous equation with graph cuts, a graph $G = <V, E>$ with a node $v \in \boldsymbol{V}$ for each voxel p is defined on image I. Its edges $\boldsymbol{e} \in \mathbf{E}$ consist of connections between each node v and two terminal nodes s, t as well as connections between neighbouring voxels. The terminals s and t represent the two labels describing foreground and background. By determining an s–t cut on G the desired segmentation can be obtained (Boykov et al., 2001). The data term in the MRF model defines the weights of the edges connecting each node with both terminals and the smoothness term encodes the edge weights of neighbouring nodes.

Spatial prior

Following van der Lijn et al. (2008), our prior spatial probabilities are obtained from a subject-specific probabilistic atlas built from the labels obtained from multi-atlas segmentation (Heckemann et al., 2006). With multiple label maps f^j, the prior probability for a voxel p of its label being the foreground label f_{fore} is therefore:

$$P_A\left(f_p\right) = \frac{1}{N} \sum_{j=1,\ldots,N} \begin{cases} 1, f_p^j = f_{\text{fore}} \\ 0, f_p^j \neq f_{fore} \end{cases}. \quad (5)$$

P_A defines the spatial prior contribution to the data term in the graph cuts model.

Intensity model

The intensity prior for tissue classes or specific structures is usually modelled by a Gaussian probability distribution. To arrive at a generally applicable model, we directly estimate the parameters of the Gaussian distribution of the hippocampus from the unseen target image. It is estimated from all those voxels that at least 95% of the atlases assign to the hippocampus. The intensity component of the source link weight for a given voxel p with intensity I_p is denoted by P_S and is estimated from the intensity distribution model, i.e., $P_S\ (p,\ f_p) = P(I_p \mid f_{p,\text{fore}})$. Since the background of the hippocampus is not homogeneous, we use a spatially varying

mixture of Gaussians (MOG) model to describe it. The MOG model is defined by the Gaussian distributions of the three tissue classes (CSF, GM, WM) t_k, $k=\{1, 2, 3\}$ based on the method described in van Leemput et al. (1999) and non-rigidly aligned spatial priors for the three tissue classes. The probability of a voxel being in the background of the hippocampus is therefore estimated by:

$$P\left(I_p|f_{p,\text{back}}\right) = \left(1 - P_A\left(f_p\right)\right) \sum_{k=1,\ldots,3} \lambda_k P\left(I_p|t_k\right) \quad (6)$$

where λ_k is the tissue spatial prior.

This equation provides the intensity component of the edge weight in the graph cuts model. The intensity and spatial contributions are combined to give the data term in the graph cuts model.

Smoothness term

Combining intensity and local boundary information into the weights connecting neighbouring nodes has been applied successful for brain segmentation with graph cuts by Song et al., 2006. Following this approach, a smoothness term based on intensity I as well as the intervening contour probabilistic map B is used. With the gradient image G, B is defined for a voxel p as $B_p=1-\exp(-G_p/\sigma_G)$ with a normalisation factor σ_G. The weight of an edge connecting two neighbouring voxels p and q is then defined as:

$$V_{p,q}\left(f_p,f_q\right) = c\left(1 + ln\left(1 + \frac{1}{2}\left(\frac{|I_p - I_q|}{\sigma}\right)^2\right)\right)^{-1} + (1-c)\left(1 - max_{x\in M_{p,q}}(B_x)\right) \quad (7)$$

where $M_{p,q}$ is a line joining p and q, and σ is the robust scale of image I (Song et al., 2006). The parameter c controls the influence of the boundary- and intensity-based part and is empirically set to 0.5.

Appendix B. Expectation maximisation formulation

The labelling f of the image I minimising an energy functional was searched:

$$f = \underset{f}{argmin}\, \lambda E_{\text{intensity}}(f) + E_{\text{prior}}(f), \quad (8)$$

where $E_{\text{intensity}}$ measures the likelihood that observed intensities are from specific classes and E_{prior} describes the prior knowledge of class labels. Different values for the parameter λ were tested and the value producing the highest accuracy ($\lambda=0.3$) was chosen. However, the accuracy was not very sensitive to the λ value; the similarity index changed only a thousandth when λ was halved.

The intensity of each structure k was assumed to have a Gaussian density function, described by the mean μ and standard deviation σ:

$$E_{\text{intensity}} = -\sum_{p\in I} lnp\left(I_p|f_p=k\right), \quad (9)$$

where

$$p\left(I_p|f_p=k\right) = \frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{\left(I_p-\mu_k\right)^2}{2\sigma_k^2}\right). \quad (10)$$

The parameters μ_k and σ_k were estimated from the target volume by weighting each voxel with the probability that it belongs to the class k. The probability was estimated from labelled non-rigidly registered atlas volumes as described in Eq. (5).

The prior energy consisted of two components: spatial prior and regularity prior. The spatial prior was based on Eq. (5):

$$E_{\text{sprior}} = -\sum_{p\in I} lnp\left(f_p=k\right). \quad (11)$$

The regularity prior, based on Markov Random Fields, was defined for keeping the structures smooth. The prior is described in detail in van der Lijn et al. (2008).

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.10.026.

References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46, 726–738.

Andronache, A., von Siebenthal, M., Szekely, G., Catting, P.h, 2008. Non-rigid registration of multi-modal images using both mutual information and cross-correlation. Med. Image Anal. 12, 3–15.

Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de-Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imag. 28, 1266–1277.

Ashburner, J., Friston, K., 2005. Unified segmentation. NeuroImage 26, 839–851.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labelling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41.

Babalola, K.O., Petenaude, B., Aljabar, P., Schnabel, J., Kenneedy, D., Crum, W., Smith, S., Cootes, T.F., Jenkinson, M., 2008. Comparison and evaluation of segmentation techniques for subcortical structures in brain MRI. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (Eds.), Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008, Vol. 5241. Springer, pp. 409–416.

Bazin, P-L., Pham, D.L., 2007. Statistical and topological atlas based brain image segmentation. In: Ayache, N., Ourselin, S., Maeder, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention—MICCAI 2007, Vol. 4791. Springer, pp. 94–101.

Bhattacharjee, M., Pitiot, A., Roche, A., Dormont, D., Bardinet, E., 2008. Anatomy-preserving nonlinear registration of deep brain ROIs using confidence-based Block-Matching. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (Eds.), Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008, Vol. 5242. Springer, pp. 482–490.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Machine Intell. 23, 1222–1239.

Chupin, M., Hammers, A., Liu, R.S.N., Colliot, O., Burdett, J., Bardinet, E., Duncan, J.S., 2009. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. NeuroImage 46, 749–761.

Corso, J.J., Tu, Z., Yuille, A., Toga, A., 2007. Segmentation of sub-cortical structures by the graph-shifts algorithm. In: Karssemeijer, N., Lelievel, B (Eds.), Information Processing in Medical Imaging- IPMI 2007, Vol. 4584. Springer, pp. 183–197.

Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.t., Dickerson, B.C., Buckner, D., BalckeradnDale, Dale, A.M., Hyman, R.P., MaguireaKilliany, ., Killiany, M.S., Albertadn, R.J., 2006. An automated labeling system for subdiving the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980.

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 structures in the human brain. Neuron 33, 341–355.

Gholipour, A., Kehtarnavaz, N., Briggs, R., Devous, M., Gopinath, K., 2007. Brain functional localization: a survey of image registration techniques. IEEE Trans. Med. Imag. 26, 427–451.

Guimond, A., Roche, A., Ayache, N., Meunier, J., 2001. Three-dimensional multimodal brain warping using demons algorithm and adaptive intensity corrections. IEEE Trans. Med. Imag. 20, 58–69.

Han, X., Fischl, B., 2007. Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE. Trans. Med. Imag. 26, 479–486.

Hellier, P., 2003. Consistent intensity correction of MR images. Int. Conf. Image Process. (ICIP, 2003, pp. 1109–1112.

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

Jäger, F., Hornegger, J., 2009. Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging. IEEE Trans. Med. Imag. 28, 137–150.

Klein, A., Mensh, B., Ghosh, S., Tourville, J., Hirsch, J., 2005. Mindboggle: automated brain labeling with multiple atlases. BMC Medical Imaging 7.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage,

C., Rueckert, D., Thompson, P., Vercauteren, T., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 202–786.
Lötjönen, J., Koikkalainen, J., Thurfjell, L., Rueckert, D., 2009. Atlas-based registration parameters in segmenting sub-cortical regions from brain MRI-images. IEEE International Symposium on Biomedical Imaging- ISBI, 2009, pp. 21–24.
Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imag. 16, 187–198.
Morra, J., Tu, Z., Apostolova, L., Green, A., Avedissian, C., Madsen, S., Parikshak, N., Hua, X., Toga, A., Jack, C., Weiner, M., Thompson, P., The Alzhei, 2008. Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. NeuroImage 43, 59–68.
Nyúl, L., Udupa, J., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imag. 19, 143–150.
Pham, D.L., Prince, J.L., 1999. Adaptive fuzzy segmentation of magnetic resonance images. IEEE Trans. Med. Imaging. 18, 737–752.
Pohl, K.M., Fisher, J., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2006. A Bayesian model for joint segmentation and registration. NeuroImage 31, 228–239.
Rohlfing, T., Brandt, R., Menzel, R., Maurer, C., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brain. NeuroImage 21, 1428–1442.
Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Non-rigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imag. 18, 712–721.
Scherrer, B., Forbes, F., Garbay, C., Dojat, M., 2008. Fully Bayesian joint model for MR brain scan tissue and structure segmentation. In: Metaxas, D., Axel, L., Fichtinger, G., Szekely, G. (Eds.), Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008, Vol. 5242. Springer, pp. 1066–1074.
Schmidt, M., 2005. A method for standardizing MR intensities between slices and volumes. In: Univ. Alberta, Edmonton, AB, Tech. Rep. TR05-14.
Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. NeuroImage 22, 1060–1075.
Smith, S., 2002. Fast robust automated brain extraction. Hum. brain mapp. 17, 143–155.
Song, Z., Tustison, N., Avants, B., Gee, J.C., 2006. Integrated graph cuts for brain MRI segmentation. In: Larsen, R., Nielsen, M., Sporring, J. (Eds.), Medical Image Computing and Computer-Assisted Intervention—MICCAI 2006, Vol. 4191. Springer, pp. 831–838.
Studholme, C., Hill, D.L.G., Hawkes, D.J., 1997. Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures. Medical Physics 24, 71–86.
Studholme, C., Cardenas, V., Song, E., Ezekiel, F., Maudsley, A., Weiner, M., 2004. Accurate template-based correction of brain MRI intensity distortion with application to dementia and aging. IEEE Trans. Med. Imag. 23, 99–110.
van der Lijn, F., den Heijer, T., Breteler, M., Niessen, W., 2008. Hippocampus segmentation in MR images using atlas registration, voxel classification and graph cuts. NeuroImage 43, 708–720.
van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imag. 18, 897–908.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE. Trans. Med. Imag. 23, 903–921.
Wolz, R., Aljabar, P., Rueckert, D., Heckemann, R.A., Hammers, A., 2009. Segmentation of subcortical structures in brain MRI using graph-cuts and subject-specific a-priori information. IEEE International Symposium on Biomedical Imaging- ISBI, 2009, pp. 470–473.
Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C., Aizenstein, H., 2007. Optimum template selection for atlas-based segmentation. NeuroImage 34, 1612–1618.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. IEEE Trans. Med. Imag. 20, 45–57.