HAL
open science

NICE: Non-local Intracranial Cavity Extraction

José V. Manjón, Simon F. Eskildsen, Pierrick Coupé, José E. Romero, Louis D. Collins, Montserrat Robles

To cite this version:

José V. Manjón, Simon F. Eskildsen, Pierrick Coupé, José E. Romero, Louis D. Collins, et al.. NICE: Non-local Intracranial Cavity Extraction. International Journal of Biomedical Imaging, 2014. hal-01060348

HAL Id: hal-01060348

https://hal.science/hal-01060348v1

Submitted on 3 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NICE: Non-local Intracranial Cavity Extraction

José V. Manjón[1], Simon F. Eskildsen[2], Pierrick Coupé[3], José E. Romero[1], D. Louis Collins[4], Montserrat Robles[1]

[1] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.

[2] Center of functionally integrative neuroscience, Department of Clinical Medicine, Aarhus University, Nørrebrogade 44, Denmark

[3] Laboratoire Bordelais de Recherche en Informatique, Unité Mixte de Recherche CNRS (UMR 5800), PICTURA Research Group, 351, cours de la Libération F-33405 Talence cedex, France.

[4] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, 3801 University Street, Montreal, Canada

Abstract. Automatic and accurate methods to estimate normalized regional brain volumes from MRI data are valuable tools which may help to obtain an objective diagnosis and follow-up of many neurological diseases. To estimate such regional brain volumes, the Intracranial Cavity Volume (ICV) is often used for normalization. However, the high variability of brain shape and size due to normal inter-subject variability, normal changes occurring over the lifespan, and abnormal changes due to disease makes the ICV estimation problem challenging. In this paper, we present a new approach to perform ICV extraction based on the use of a library of pre-labeled brain images to capture the large variability of brain shapes. To this end, an improved non-local label fusion scheme based on BEaST technique is proposed to increase the accuracy of the ICV estimation. The proposed method is compared with recent *state-of-the-art* methods and the results demonstrate an improved performance both in terms of accuracy and reproducibility while maintaining a reduced computational burden.

*** Corresponding author:** José V. Manjón. Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain.

Tel.: (+34) 96 387 70 00 Ext. 75275 Fax: (+34) 96 387 90 09.

E-mail address: jmanjon@fis.upv.es (José V. Manjón)

Keywords: Intracranial cavity extraction, non-local patch-based segmentation, brain volume analysis, BEAST, MRI.

1. Introduction

Automated brain image analysis has a huge potential to objectively help in the diagnosis and follow-up of many neurological diseases. To perform such analysis tasks, one of the first image processing operations is the delimitation of the area of interest. For brain image analysis, this operation has received many different names such as brain extraction, skull-stripping or intracranial cavity masking. In each case, the aim is to isolate the brain or intracranial tissues (depending on area definition) from the raw image. The accurate estimation of the intracranial volume plays crucial role to obtain robust and reliable normalized measurements of brain structures [1].

The importance of this operation is reflected by the large number of methods proposed over the past decade [2-13]. Many of these methods are based on the modeling of brain intensities (normally using T1 weighted images due to their excellent contrast for brain tissues) combined with a set of morphological operations [3,5,12] or atlas priors [9].

The most widely used automated methods correspond to those that are publically available. For example, the BET (Brain Extraction Tool) software from the FSL image processing library [2] is one of the most used techniques probably due to its accuracy, ease of use and low computational load. Other techniques like 3dIntracranial [6], Hybrid Watershed algorithm (HWA) [5] or Brain Surface Extractor (BSE) [13] have been also widely used.

Intracranial cavity extraction can also be obtained indirectly as part of a full modeling of brain intensities using a parametric model such as done in Statistical Parametric Mapping (SPM) [14] or VBM8 (http:/dbm.neuro.uni-jena.de/vbm) software packages.

Over the last decade, methods have been proposed to automatically measure the Intracranial Cavity Volume (ICV) by using non-linear registration atlas-based approaches [15,16].

More recent works of special interest for the brain extraction problem are methods like MAPS [10] and BEaST [11]. Both methods rely on the application of a multi-atlas label fusion strategy. MAPS uses multiple non-linear registrations followed by a voxel-wise label fusion while BEaST uses a single linear registration in combination with non-local patch-based label fusion. Both techniques scored well on the LONI segmentation validation engine (SVE) [17] comparison for brain extraction (see http://sve.loni.ucla.edu) although MAPS has a much larger computational load compared to BEaST.

In this paper we present an extension of the BEaST methodology where we aim to improve the accuracy while reducing the computational load. The main contributions of the proposed method are threefold: First, the use of a new pipeline for the multi-atlas library construction for

improved normalization between template library subjects. Second, the use of a new bilateral patch similarity measure to better estimate pattern similarities. Finally, a block-wise labeling approach that enables significant savings in computational cost and imposing at the same time a regularization constraint that increases the method's accuracy.

2. Materials and methods

Since the method proposed in this paper is based on the use of a library of pre-labeled cases to perform the segmentation process, we will first describe the template library construction and then present the proposed method.

2.1. Template library construction

2.1.1. Library dataset description

A library of manually labeled templates was constructed using subjects from different publically available datasets. To include as large age range as possible, different datasets nearly covering the entire human life-span were considered. MRI data from the following databases were used:

- **Normal adults dataset**: Thirty normal subjects (age range: 24-75 years) were randomly selected from the open access IXI dataset (http://www.brain-development.org/). This dataset contains images from nearly 600 healthy subjects from several hospitals in London (UK). Both 1.5 T (7 cases) and 3 T (23 cases) images were included in this dataset. 3T images were acquired on a Philips Intera 3T scanner (TR = 9.6 ms, TE = 4.6 ms, flip angle=8°, slice thickness=1.2 mm, volume size=256×256x150, voxel dimensions = 0.94×0.94×1.2 mm^3). 1.5 T images were acquired on a Philips Gyroscan 1.5T scanner (TR = 9.8 ms, TE = 4.6 ms, flip angle=8°, slice thickness = 1.2 mm, volume size=256×256x150, voxel dimensions = 0.94×0.94×1.2 mm^3).
- **Alzheimer Disease (AD) dataset**: Nine patients with Alzheimer's disease (age range= 75-80 years, MMSE= 23.7±3.5, CDR = 1.1±0.4) scanned using a 1.5 T General Electric Signa HDx MRI scanner (General Electric, Milwaukee, WI) were randomly selected. This dataset consisted of high resolution T1-weighted sagittal 3D MP-RAGE images (TR=8.6 ms, TE=3.8 ms, TI=1000 ms, flip angle=8°, slice thickness=1.2 mm, matrix size=256×256, voxel dimensions=0.938×0.938×1.2 mm^3). These images were downloaded from the brain segmentation testing protocol [18] website (https://sites.google.com/site/brainseg/) while they belong originally to the open access OASIS dataset (http://www.oasis-brains.org/).

- **Pediatric dataset**: Ten infant datasets were also downloaded from the brain segmentation testing protocol [18] website (https://sites.google.com/site/brainseg/).

These data were originally collected by Gousias et al. [19] and are also available at http://www.brain-development.org (this dataset is property of the Imperial College of Science Technology & Medicine and has been used after accepting the license agreement). The selected 10 cases are from the full sample of 32 two-year old infants born prematurely (age = 24.8 ± 2.4 months). Sagittal T1 weighted volumes were acquired from each subject (1.0 T Phillips HPQ scanner, TR=23 ms, TE=6 ms, slice thickness=1.6 mm, matrix size= 256×256, voxel dimensions= 1.04×1.04× 1.6 mm^3 resliced to isotropic 1.04 mm^3).

Downloaded images from the different websites consisted of raw images with no preprocessing and no intracranial cavity masks were supplied with these data. To generate the template library, all 49 selected T1-weighted images were preprocessed as follows:

2.1.2. Denoising and inhomogeneity correction

All images in the database were denoised using the Spatially Adaptive Non-Local Means (SANLM) Filter [20] to enhance the image quality. The SANLM filter can deal with spatially varying noise levels across the image without the need of explicitly estimate the local noise level which makes it ideal to process data with either stationary or spatially varying noise (as in the case of parallel imaging) in a fully automatic manner. To further improve the image quality, an inhomogeneity correction step was applied using the N4 method [21]. The N4 method is an incremental improvement of the N3 method [22] that has been implemented in the ITK toolbox [23] and has proven to be more efficient and robust.

2.1.3. MNI space registration

In order to perform the segmentation process, templates and the subject to be segmented have to be placed in the same stereotactic space. Therefore, a spatial normalization based on a linear registration to the Montreal Neurological Institute (MNI152) space was performed using ANTS software [24]. The resulting images in the MNI space have a size of 181 x 217 x 181 voxels with 1 mm^3 voxel resolution.

2.1.4. Intensity normalization

As the proposed method is based on the estimation of image similarities using intensity-derived measures, every image in the library must be intensity normalized in order to make the intensity distributions comparable among them. We use a tissue-derived approach to force mean intensities of white matter (WM), grey matter (GM) and cerebro-spinal fluid (CSF) to be as similar as possible across subjects of the library in a similar manner as done by Lotjonen et al. [25]. For this purpose, mean values of CSF, GM and WM tissues were estimated using the

Trimmed Mean Segmentation (TMS) method [26] which robustly estimates the mean values of the different tissues by excluding partial volume voxels from the estimation jointly with the use of an unbiased robust mean estimator. Such estimation was performed using only voxels within the standard brain mask area of MNI152 template to minimize the inclusion of external tissues. Finally, a piecewise linear intensity mapping [25,27] was applied ensuring that WM had an average intensity of 250, GM 150 and CSF of 50 (see Fig. 1).

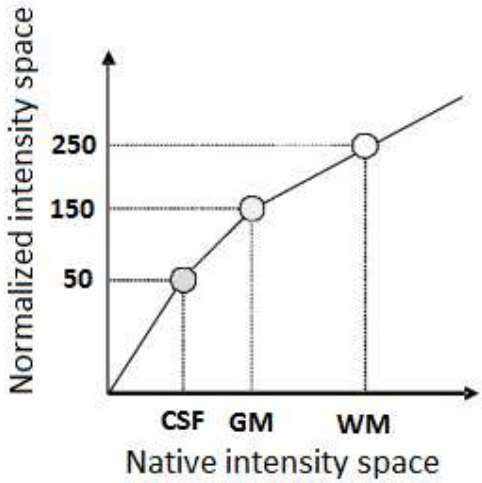

Figure 1. *Proposed intensity normalization via a piecewise linear mapping. CSF, GM and WM mean values are automatically estimated using TMS method and mapped to their corresponding normalized values (50,150 and 250).*

2.1.5. Manual labeling

As commented previously, there is no standard definition of what should be included in brain or intracranial masks (it all depends in what you are looking for). In BEaST, the mask definition included the following tissues:

- All cerebral and cerebellar white matter
- All cerebral and cerebellar gray matter
- CSF in ventricles (lateral, 3rd and 4th) and the cerebellar cistern
- CSF in deep sulci and along the surface of the brain and brain stem
- The brainstem (pons, medulla)
- Internal brain blood vessels

In the present work we extended that definition by including all external CSF (thus covering total CSF of IC) and therefore selecting most of the intracranial cavity volume. We have not included other intracranial tissues in our mask definition such as dura, exterior blood vessels or veins because they are normally of no interest for brain analysis. This mask definition has been traditionally used to estimate the Total Intracranial Volume (TIV) in many methods such RBM

[28], SPM8 or VBM8 methods to normalize brain tissue volumes [29,30] as it is expected to be nearly constant in each subject during the adult life-span.

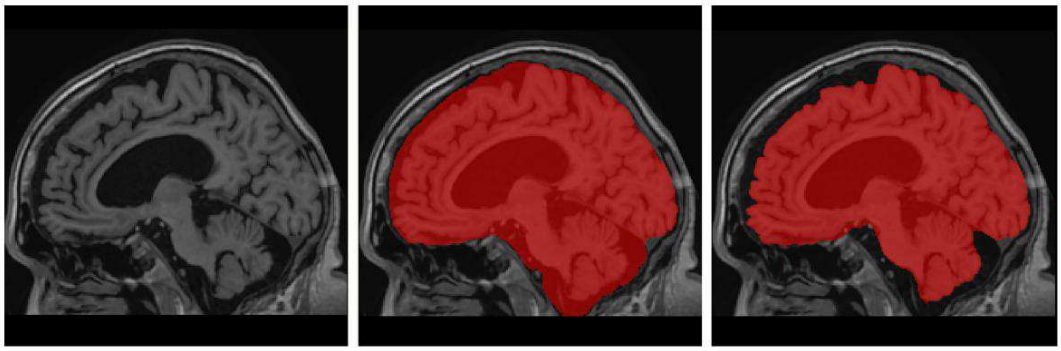

Figure 2. *Example of mask differences between our mask definition (center) and BEaST mask (right) for an Alzheimer case (Left). As can be noticed, all external CSF is included in NICE mask while this is not case at the corresponding BEaST mask (example case from Oasis dataset).*

To generate the template masks we followed a similar approach as described in BEaST paper since full manual labeling was too time consuming and error prone as discussed in Eskildsen et al [11]. All template images in the library were automatically segmented using BEaST software to have an initial mask. Conditional mask dilation (only over CSF voxels) was applied to include external CSF not already included in the BEaST mask. Finally, all the images were manually corrected by an expert on brain anatomy using the ITK-SNAP software [31] to remove segmentation errors. In Fig. 2 we show an example of our mask definition compared to BEaST definition for a patient with Alzheimer's disease.

To further increase the number of available priors on the library all the cases were flipped along the mid-sagittal plane using the symmetric properties of the human brain yielding a total number of 98 labeled templates (original and flipped) as done in BEaST paper [11].

Compared to BEaST template library creation, main differences are the use of a denoising method to improve data quality, the use of a different registration method (ANTS instead of ANIMAL) and the application of different intensity normalization method. The scheme of the template library construction pipeline is summarized in Fig. 3.

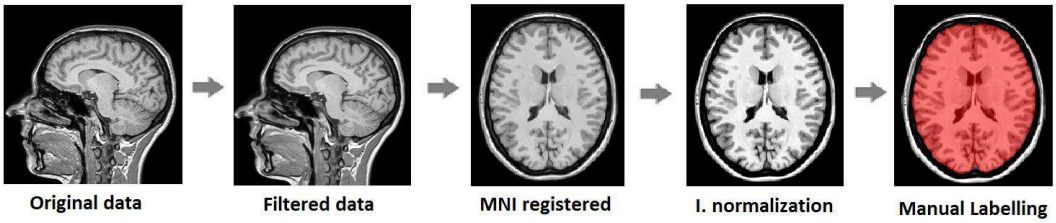

Figure 3*. Template library construction pipeline.*

2.2. Proposed Method

While the BEaST technique was designed to improve downstream analysis such as the assessment of cortical thickness, our proposed method has extended the mask definition to include extra-cerebral spinal fluid as it can be interesting to obtain normalized brain and tissue specific volumes in many neurological diseases such as Alzheimer or Parkinson. We will refer our proposed method as NICE (**N**on-local **I**ntracranial **C**avity **E**xtraction). Since the method proposed in this paper is an evolution of the BEaST brain masking method [11] we refer the reader to the original paper for the detailed method overview. Here, we summarize the NICE method and present the main improvements introduced to increase the method performance.

2.2.1. Preprocessing

To segment a new case, it must be first preprocessed using the proposed normalization pipeline (see section 2.1 and Fig. 2) so that the new case is spatially aligned with the template library and to ensure that it has the same intensity characteristics.

2.2.2. Improved Nonlocal means label fusion

In the classical nonlocal means label fusion technique proposed by Coupe et al. [32], for each voxel x_i from the new image to be segmented the method estimates the final label by performing a weighted label fusion $v(x_i)$ of all surrounding samples inside the search area V_i from N subjects selected from the library:

$$v(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j}).l_{s,j}}{\sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j})} \tag{1}$$

where $l_{s,j}$ is the label from the voxel x_{sj} at the position $_j$ in the template library subject s and $w(x_i, x_{s,j})$ is the weight calculated by patch comparison which is computed depending on the similarity of the surrounding patch for x_i and for $x_{s,j}$ This weight is estimated as follows:

$$w(x_i, x_{s,j}) = \begin{cases} \exp^{\frac{-\left\| P(x_i) - P(x_{s,j}) \right\|_2^2}{h}} & if \quad ss > th \\ 0 & otherwise \end{cases} \tag{2}$$

where $P(x_i)$ is the patch around the voxel x_i , $P(x_{s,j})$ is the patch around the voxel x_j in the templates and $\|.\|_2$ is the normalized L2-norm (normalized by the number of elements) calculated by the distance between each pair of voxels from both patches $P(x_i)$ and $P(x_{s,j})$ and

modulated by *h* parameter. If *ss* (structural similarity index [33] between patches) is less than a threshold *th* then *w* is not computed thus avoiding unneeded computations.

The structural similarity index *ss* is calculated as follows:

$$ss = \frac{2\mu_i\mu_{s,j}}{\mu_i^2+\mu_{s,j}^2} \cdot \frac{2\sigma_i\sigma_{s,j}}{\sigma_i^2+\sigma_{s,j}^2} \tag{3}$$

where μ and σ are the mean and standard deviation of the patches surrounding x_i and $x_{s,j}$ at location *j* of the template *s*.

Finally, the final label $L(x_i)$ is computed as:

$$L(x_i) = \begin{cases} 1 & v(x_i) \geq 0.5 \\ 0 & v(x_i) < 0.5 \end{cases} \tag{4}$$

In this paper, we introduce two modifications to this strategy. First, we make use of the fact that all the images are registered to a common space and therefore a *locality principle* can be used, assuming that samples that are spatially closer are likely to be more similar in their labels. However, this locality principle is limited by residual anatomical variability and registration errors in the template library space. Therefore, we redefined the similarity weight to take into account not only intensity similarity but also spatial patch proximity:

$$w(x_i, x_{s,j}) = \begin{cases} \exp^{-\left[\frac{\|x_i - x_j\|^2}{\sigma_d} + \frac{\|P(x_i)-P(x_{s,j})\|_2^2}{h}\right]} & if \quad ss > th \\ 0 & otherwise \end{cases} \tag{5}$$

where x_i and x_j are the coordinates of patch centers and σ_d is normalization constant. We set σ_d= 8 mm experimentally which curiously coincides with the typical Gaussian blurring kernel size normally used on Voxel Based Morphometry (VBM) to deal with registration error and subject anatomical variability. This approach shares some similarities to the bilateral filter proposed by Tomasi and Manduchi [34] for image denoising. We experimentally set the threshold *th* to 0.97 instead of 0.95 as used in BEaST (this difference can be explained due to the use of filtered data and a different intensity normalization method). Also a comment about *h* parameter of equation (5) is required since it plays a major role in the weight computation process. In [32] this value was set to:

$$h(x_i) = \lambda \underset{x_{s,j}}{\operatorname{argmin}} \quad \|P(x_i) - P(x_{j,s})\|_2 + \varepsilon, \tag{6}$$

where ε is a small constant to ensure numerical stability. In [32] λ was set to 1 but we found experimentally that a value of 0.1 produced better results in the proposed method possibly due to the improved intensity normalization.

The second modification concerns the voting scheme. Classical non-local label fusion works in a voxel-wise manner which sometimes results in a lack of regularization on the final labels. Given that we wish to segment a continuous anatomical structure, some level of regularization can be used as a constraint achieved by a block-wise vote scheme, similar to the one used by Rousseau [35] for label fusion and derived from MRI denoising [36]. This bilateral block-wise vote is computed as follows:

$$v(B(x_i)) = \frac{\sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j}) l(B(y_{s,j}))}{\sum_{k=0}^{M} \sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j})} \tag{7}$$

where $B(x_i)$ is a 3D region which is labeled at the same time. Finally, the vote for the voxel x_i is obtained in an overcomplete manner by averaging over all blocks containing x_i and the label $L(x_i)$ is decided as in Eq. (4).

With overlapping blocks, it is worth noting that the distance between adjacent block centers can be increased to be equal or higher than 2 voxels. Therefore, we can obtain important accelerating factors compared to the voxel-wise version of the algorithm (for example for a distance equal to 2 voxels in all three directions a speedup factor of $2^3=8$ can be obtained). The described approach is used within the multiresolution framework as described in the BEaST paper [11].

3. Experiments

3.1. Experimental datasets

To validate the proposed method different datasets were used. These datasets can be classified two categories: a) those that were used to measure the accuracy of the different methods compared and b) those used to measure their reproducibility.

3.1.1. Accuracy datasets

LOO dataset: To measure the accuracy of the proposed method we used the template library dataset by using a leave-one-out (LOO) cross validation. The characteristics of this dataset have already been described in section 2.1. Each of the 49 (non-flipped) library images was processed with the remaining images as priors (after removing the current case and its flipped version). The resulting segmentation was compared to the corresponding manual labels in the library.

Independent validation dataset: To avoid any factor associated to our IC mask definition that could bias the comparison of the compared methods we decided to use an independent dataset with its corresponding manual segmentations. Therefore, we performed a validation using an independent dataset available in the online Segmentation Validation Engine (SVE) [17]. The SVE IC segmentation followed rules similar to those used here. This dataset consists of 40 T1w MRI scans and its associated manual labels (20 males and 20 females; age range 19–40). This high-resolution 3D Spoiled Gradient Echo (SPGR) MRI volumes were acquired on a GE 1.5T system as 124 contiguous 1.5 mm coronal brain slices (TR range 10.0 ms–12.5 ms; TE range 4.22 ms – 4.5 ms; FOV 220 mm or 200 mm; flip angle 20°) with in-plane voxel resolution of 0.86 mm (38 subjects) or 0.78 mm (2 subjects).

3.1.2. Reproducibility dataset

Although the accuracy of a method is very important another important feature is its reproducibility. Indeed, the capability to detect changes induced by the pathology in a consistent manner is a key aspect. To measure the reproducibility of the different compared methods we used the reproducibility dataset of the brain segmentation testing protocol website (https://sites.google.com/site/brainseg/). This dataset consists of a test-retest set of 20 subjects scanned twice in the Same Scanner and Sequence (SSS) and another set of 36 subjects scanned twice on Different Scanner and Different magnetic Field strength (DSDF) (1.5 and 3 Tesla).

SSS dataset: To measure the reproducibility of the different methods compared on the same subjects and using the same MRI scanner, we used a subset of the OASIS (www.oasis-brains.org) dataset consisting in 20 subjects (age=23.4±4.0 years, 8 females) who were scanned using the same pulse sequence two times (1.5 T Siemens Vision scanner, TR=9.7 ms, TE=4 ms, TI=20 ms, flip angle=10°, slice thickness=1.25 mm, matrix size=256×256, voxel dimensions=1×1×1.25 mm3 resliced to 1×1×1 mm3, averages=1) [37].

DSDF dataset: To determine the consistency of the segmentations when different MRI scanners and different magnetic field strength were used, 36 adult subjects were scanned using two MRI scanners (1.5 T and 3.0 T General Electric Signa HDx scanner),mean inter-scan interval between 1.5 T and 3 T scanner=6.7±4.2 days) [18].

3.2. Method parameter settings

To study the impact of the method parameters an exhaustive search of the optimum values was performed using the LOO dataset using the library segmentations as gold standard references. Each one of the 49 subjects in the library was processed using the remaining cases of the library as priors and the resulting segmentation was compared to the manual labeling. To measure segmentation accuracy, the Dice coefficient [38] was used. Method parameters such as patch size and search area were set as in BEaST method while an exhaustive search for the optimal number of templates N used for the segmentation process was carried out (see Fig. 4). This search demonstrated that the segmentation accuracy stabilizes around N=[20-30] range which is in good agreement with previous results from BEaST. We decided to use N=30 as default value given the reduced computational cost of the proposed method.

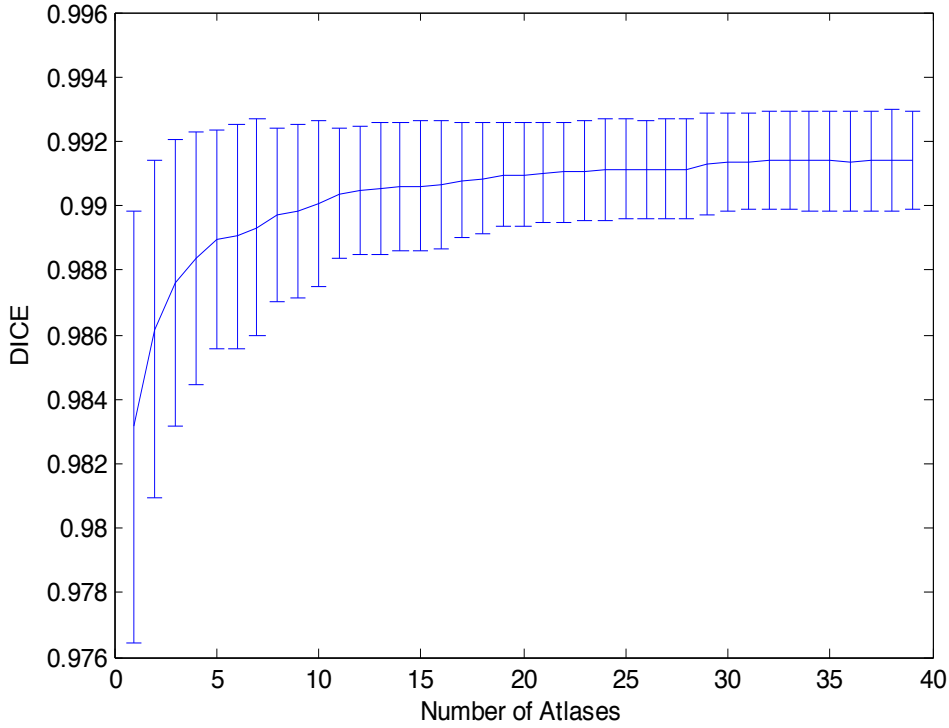

Figure 4. *Evolution of segmentation accuracy in function of the number of training subject templates used in the segmentation process.*

Another parameter of our proposed block-based approach is the spacing between adjacent blocks which jointly with patch size defines the degree of overlap between blocks. We observed experimentally that the optimal value for that parameter was 2 voxels in all 3 dimensions since the resulting accuracy was virtually the same from full overlap (1 voxel spacing) while computation time was greatly reduced. This is in good agreement with previous results on block-wise MRI denoising [36]. Higher block spacing resulted in worse segmentation results.

3.3. Compared methods

The proposed method was compared with BEaST and VBM8 methods. Both BEaST and VBM8 methods were selected because their public accessibility and because they are among the highest ranking methods on the online Segmentation Validation Engine website [17] (http://sve.loni.ucla.edu/archive/).

To ensure a fair comparison all three methods used the same preprocessing pipeline with the exception of the intensity normalization step (that is using ANTS registration to ensure the same image space and the same homogenization and filtering to ensure the same image quality). In this way only the labeling process was evaluated eliminating other sources of variability.

Both NICE and BEaST were run with the same number of preselected templates (N=30) to ensure a fair comparison. We used release 435 of VBM8, which was the latest version at the time of writing. To compare the segmentation results of the different methods several quantitative metrics were used: DICE coefficient [38], sensitivity and specificity.

4. Results

4.1. Accuracy results

In Table 1, the average DICE coefficient, sensitivity and specificity for all 49 cases of LOO dataset using the different methods compared are provided. Results for all the cases together and separated by dataset subtype are provided (Alzheimer Disease (AD), normal infants (Infant) and normal adult subjects (Adult)). As can be noticed, NICE method obtained the best results in all the situations. Table 2 shows the statistical significance of these differences (paired t-test).

Intracranial cavity volume is normally used to normalize brain tissue volumes to provide a tissue measure independent of head size. Therefore, the ability of the compared methods to provide an accurate ICV estimation has to be assessed. To this end, volume estimations using the different compared methods were obtained and compared to gold standard manual volumes. Figure 5 shows the automatic vs. manual volume correlation for all the compared methods and dataset used. As can be noticed, the NICE method had highest overall correlation (0.976) while BEaST and VBM8 had 0.923 and 0.778 respectively. In Fig. 6, a visual comparison of the segmentation results of three examples belonging to the three different subject populations can be performed.

Table 1. *Average DICE coefficient for the different methods compared on the different used datasets. The best results from each column are in bold.*

Method	Data	All (N=49)	Adults (N=30)	AD (N=9)	Infants (N=10)
NICE	DICE	**0.9911±0.0020**	**0.9921±0.0015**	**0.9892±0.0016**	**0.9899±0.0019**
	SEN	**0.9907±0.0036**	**0.9916±0.0035**	**0.9887±0.0029**	0.9898±0.0038
	SPE	**0.9971±0.0012**	**0.9975±0.0010**	**0.9964±0.0015**	**0.9965±0.0009**
BEAST	DICE	0.9880±0.0032	0.9891±0.0030	0.9857±0.0018	0.9866±0.0034
	SEN	0.9889±0.0062	0.9902±0.0060	0.9830±0.0049	**0.9900±0.0050**
	SPE	0.9955±0.0019	0.9958±0.0017	0.9960±0.0016	0.9940±0.0019
VBM8	DICE	0.9762±0.0052	0.9788±0.0026	0.9690±0.0064	0.9750±0.0033
	SEN	0.9740±0.0121	0.9796±0.0051	0.9587±0.0132	0.9710±0.0138
	SPE	0.9926±0.0027	0.9924±0.0019	0.9931±0.0033	0.9926±0.0041

Table 2. *NICE compared to the other two methods (p-values). Significant differences (p<0.05) are in bold.*

Method	Data	All (N=49)	Adults (N=30)	AD (N=9)	Infants (N=10)
BEAST	DICE	**6.30×10^{-8}**	**5.50×10^{-6}**	**5.16×10^{-4}**	**0.014**
	SEN	0.074	0.283	**0.012**	0.913
	SPE	**2.54×10^{-6}**	**2.97×10^{-5}**	0.498	**0.002**
VBM8	DICE	**1.26×10^{-33}**	**3.97×10^{-32}**	**3.35×10^{-7}**	**3.29×10^{-10}**
	SEN	**4.99×10^{-15}**	**2.60×10^{-15}**	**1.61×10^{-5}**	**6.18×10^{-4}**
	SPE	**2.06×10^{-18}**	**2.73×10^{-19}**	**0.010**	**0.009**

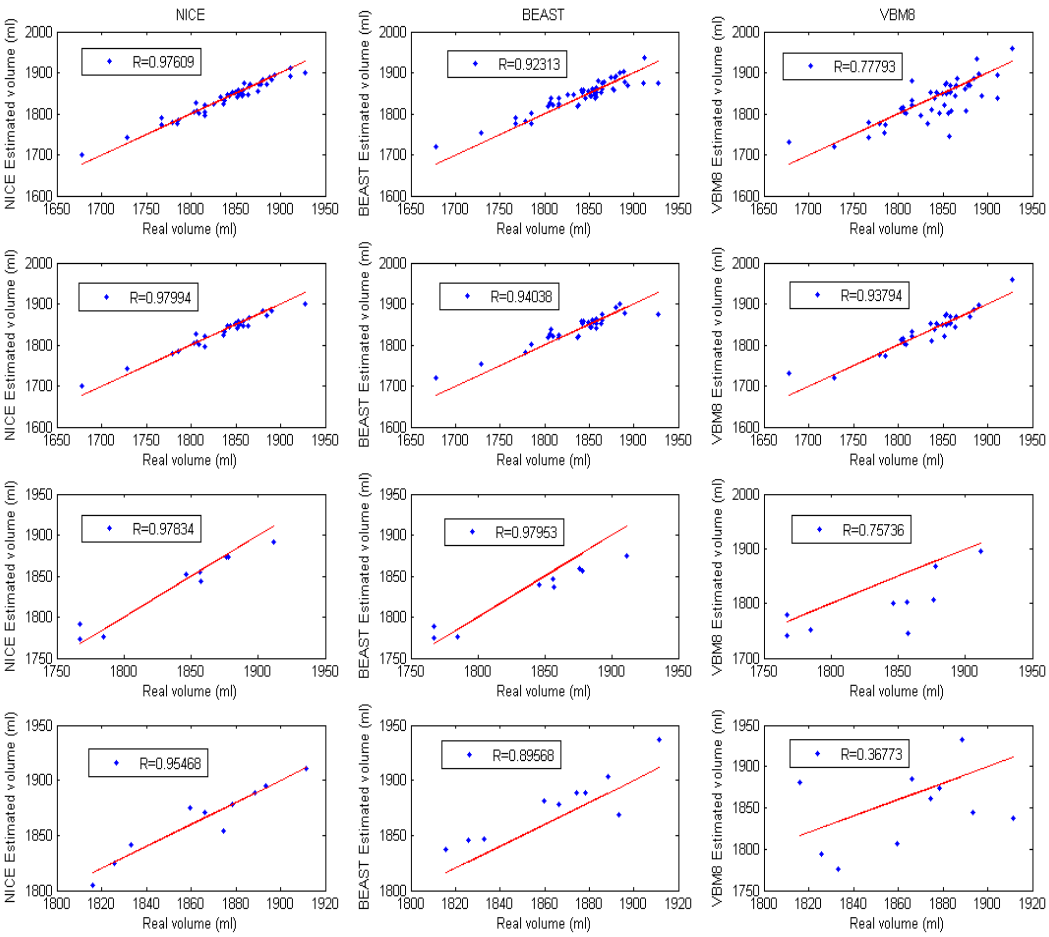

Figure 5. *Comparison of intracranial cavity volume estimation results. Automatic vs. manual volume correlation for all the compared methods and datasets used. The first row shows results for the whole library (N=49), the second only for normal adults (N=30), the third only for AD subjects (N=9) and the fourth only for infant cases (N=10). Red line represents ideal mapping between estimated and real volumes to highlight eventual over or under volume estimations (it does not represents the fitting line).*

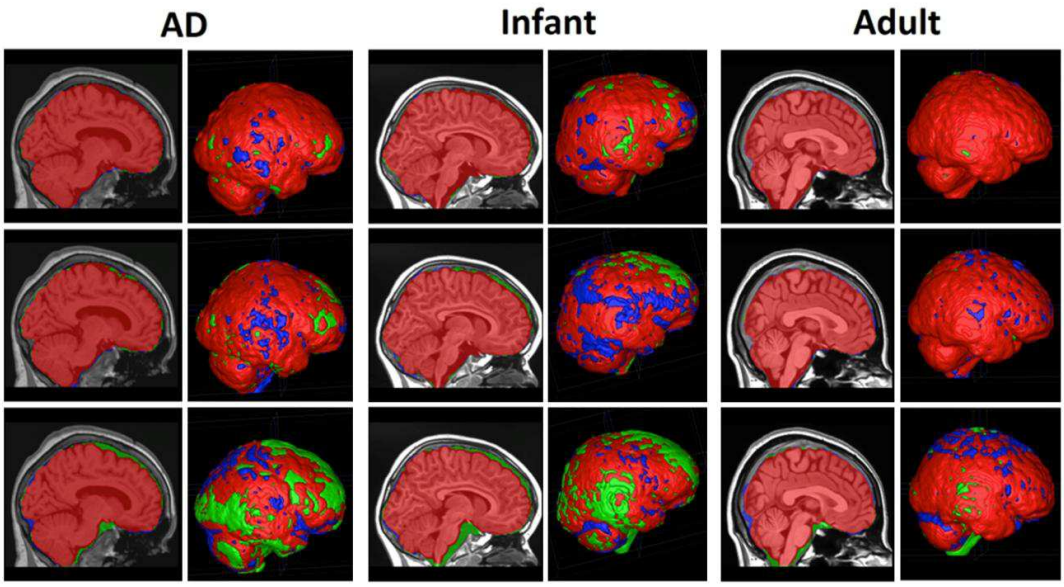

Figure 6. *Example segmentation results using NICE (first row), BEaST (second row) and VBM8 (third row) methods on the three different population samples. Sagittal slices and 3D renderings of the segmentations are shown. Red voxels correspond to correct voxels in the segmentation compared to the gold standard. Blue voxels are false positives and green voxels are false negatives (AD case belongs to Oasis dataset and the adult and infant cases to the IXI dataset).*

To perform an independent validation of the compared methods the SVE dataset was used. The SVE web service allows the comparison of results with hand-corrected brain masks. As done in BEaST and MAPS papers we used the brain masks provided by the SVE website which included all the internal ventricular CSF and some external sulcal CSF. Although this mask definition slightly differs from our mask definition (not all CSF was included) this does not represents a problem for the method´s comparison since all the methods shared the same references.

Validation of NICE using the SVE test dataset resulted in a mean DICE of 0.9819±0.0024 (see http://sve.loni.ucla.edu/archive). At the time of writing, this result was the best ($p<0.01$) of all the methods published on the website. BEaST had a DSC of 0.9781±0.0047 and VBM8 obtained a DSC of 0.9760±0.0025. Sensitivity and specificity results are also included in Table 3. A visual representation of false positive and false negative as supplied by the website is presented at figure 7.

Table 3. *Segmentation results of SVE dataset using different quality measures. NICE was compared to the other two methods (p-values). Best results are in bold (note that in this dataset, BEaST method obtained a significant higher sensitivity than NICE at the expense of a lower specificity).*

Method	DICE	Sensitivity	Specificity
NICE	**0.9819 ± 0.0024**	0.9857 ± 0.0044	**0.9960 ± 0.0015**
BEAST	0.9781 ± 0.0047 ($p = 9.82 \times 10^{-7}$)	**0.9887 ± 0.0035** ($p = 0.0001$)	0.9940 ± 0.0025 ($p = 2.11 \times 10^{-5}$)
VBM8	0.9760 ± 0.0025 ($p = 2.56 \times 10^{-15}$)	0.9840 ± 0.0046 ($p = 0.06$)	0.9942 ± 0.0014 ($p = 4.31 \times 10^{-8}$)

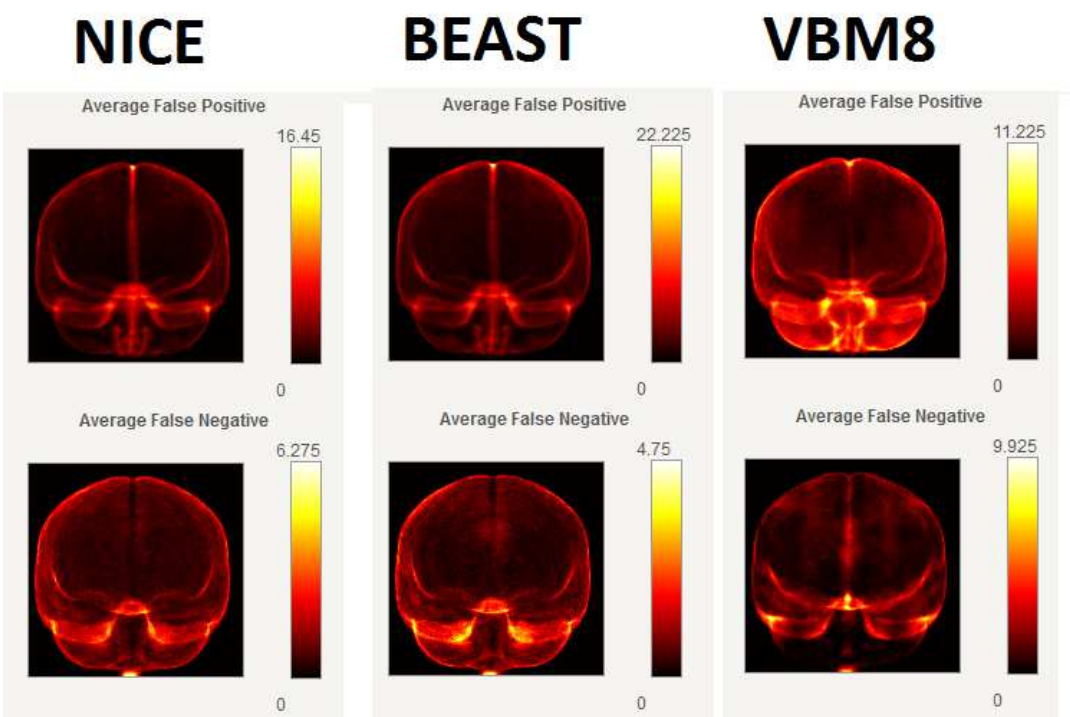

Figure 7. *False positive and false negative maps for NICE, BEaST and VBM8 on SVE dataset. VBM8 tended to produce a systematic over-segmentation compared to the used manual gold standard. The errors obtained by NICE and BEaST were more uniformly distributed indicating non-systematic segmentation errors. Note that in the images provided by the SVE website the vertical scale measuring error is not the same over all images.*

4.2. Reproducibility results

In Table 4, the average ICV differences for the different methods and datasets is provided. As can be noticed, NICE method obtained the most reproducible results in all situations. For the SSS dataset experiment (test-retest) NICE significantly improved BEaST method while these differences were not significant for VBM8 method. For the DSDF dataset experiment volume

differences were higher than in the previous experiment. In this case, NICE was found yield significantly improved estimates ($p<0.05$) compared to the two other methods.

Table 4. *Percent mean IC volume differences for the SSS dataset. NICE was compared to the other two methods (p-values). Best results are in bold.*

Method	SSS Dataset	DSDF Dataset
NICE	**1.4046± 0.2447**	**3.1856± 1.0280**
BEAST	2.1463± 0.6622 ($p= 3.21x10^{-4}$)	5.4696± 1.9097 ($p= 1.55x10^{-10}$)
VBM8	1.4268± 0.3843 (p= 0.8073)	3.7741± 1.1627 (p= 0.0242)

Finally, execution times of the different methods were compared. NICE method took around 4 minutes (NICE was implemented as a multithreaded MEX C file), BEaST method took around 25 minutes (we have to note that no multithreading optimizations were used here) and VBM8 took around 8 minutes on average (in this time it was also included the different tissue segmentations). All the experiments were performed using MATLAB 2009b 64 bits (Mathworks Inc.) on a desktop PC with an Intel core i7 with 16 GB RAM running windows 7.

However, it is worth noting that if we reduce the number of selected templates to 10 cases we can reduce the processing time to less than 1 minute with only a small reduction of the segmentation accuracy (0.9911 to 0.9901 in the LOO accuracy experiment).

5. Discussion

We have presented a new method for intracranial cavity extraction that outperforms related *state-of-the-art* methods and a previously proposed method (BEaST) by our group both in terms of accuracy and reduced computational load. In addition, we demonstrated that the new proposed method is more robust in terms of measurement reproducibility.

This last point is of special interest since in many cases we are not only interested in the specific brain volume at one time point but in its evolution in a longitudinal study. NICE method was demonstrated to be significantly more reproducible and accurate than BEaST method. In addition, VBM8 was found to be almost as reproducible as NICE but at the expense of introducing larger systematic errors on the segmentations. The high level of reproducibility of VBM8 may be explained by the fact that it uses a single template and thus a more deterministic pipeline is applied. Also the fact that it operates at 1.5 mm^3 resolution introduces a blurring effect which increases the method reproducibility at the expense of the accuracy. The increased reproducibility/accuracy of our proposed method may have a significant impact on the brain image analysis methods by increasing their sensitivity to detect subtle changes produced by the disease. While the advantage of NICE in segmentation accuracy of 0.9911 vs. 0.9762 for VBM8 may appear small when compared over the three datasets evaluated, it is statistically significant and corresponds to more than a 2 fold reduction in error, from 2.38% to 0.89%. In a large volume such as the intracranial cavity (1500cc), this reduction in error can represent a volume of approximately 20cc, a non-negligible amount. The improvement over BEaST is smaller (35%) but still statistically significant. When evaluated on the SVE dataset, the NICE yields a Dice overall of 0.9819, while BEaST and VBM8 yield 0.9781 and 0.9760, corresponding to 20% and 32% less error on average, respectively.

The improved results of NICE over BEaST can be understood thanks to improvements on two parts of the proposed method. First, improvements on template library construction such as the improved intensity normalization yield more coherent and better defined priors. This fact positively impacts the intensity driven image similarities of the label fusion part. One limitation of the first part of our validation is in the use of manually corrected masks that may induce a favourable bias toward BEaST and NICE. However, after the conditional dilation and manual correction steps, almost all edge voxels were modified, thus minimizing any bias. Second, the block-wise and new bilateral label fusion scheme results in more regular and accurate segmentations. The advantages of using a 3D block-wise approach in comparison to the previously used voxel-wise are twofold: first, the fact that we label together the whole block imposes an intrinsic regularization which forces connected voxels to have similar labels and second if a space between block centers is used a significant speed-up factor can be obtained in comparison with the voxel-wise version. Finally, the new similarity measure using spatial distance weighting takes into account a locality principle that favors the contribution of closer

patches by assuming that after linear registration similar structures are close in a similar manner as done for the well-known bilateral filter for image denoising [34].

It is also worth noting that the segmentation accuracy depends on the preexistence of similar local patterns within the library. In our method, we do not need to have totally similar templates to the case to be segmented within the library since it is able to find locally similar patterns from different templates in the library. However, it is also true that if some specific pattern is not present in the library it will not be correctly identified and therefore the resulting segmentation will be incorrect. This risk is normally reduced when using non-linear registrations at the expense of a much higher computational load and the introduction of interpolation artifacts in both images and associated labels. However, this issue can be solved more efficiently (mainly in terms of computational cost) by increasing library size with uncommon cases and their associated corrected manual labels making it unnecessary to perform costly non-linear registrations (but making necessary the manual label correction of new library cases). We experimentally found that increasing the size of the library just using the symmetric versions of the original library improved the segmentation results as previous reported in the BEAST paper. Finally, it is also possible to construct disease specific libraries (as done for templates in SPM) maximizing the likelihood to find suitable matches for the segmentation process or to improve templates preselection by adding extra information such as age or sex which could help to find optimal matches (specially useful when the library size will grow).

6. Conclusion

We have presented an improved method to perform intracranial cavity extraction that has been shown to be fast, robust and accurate. The improvements proposed have been shown to increase segmentation quality and reduce the computational load at the same time (the proposed method is able to work in a reasonable time of approximately 4 minutes). We plan to make the NICE pipeline publicly accessible through a web interface in the near future so everybody can benefit from its use independently of their location and local computational resources. Finally, the usefulness of the proposed approach to provide accurate ICV based normalization brain tissue measurements has to be addressed on future clinical studies.

Acknowledgments

We want to thank the IXI (EPSRC GR/S21533/02) and OASIS (P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584) dataset promoters for making available this valuable resource to the scientific community which surely will boost the research in brain imaging. This work has been supported by the Spanish grant TIN2011-26727 from Ministerio de Ciencia e Innovación. This work was funded in part by operating grants from the Canadian Institutes of Health Research, les Fonds de la recherche santé du Quebec, MINDLab

UNIK initiative at Aarhus University, funded by the Danish Ministry of Science, Technology and Innovation, grant agreement number 09-065250.

References

1. Nordenskjöld R., Malmberg F, Larsson E, Simmons A., Brooks S.J., Lind L., Ahlström H, Johansson L., Kullberg J.. (2013). Intracranial volume estimated with commonly used methods could introduce bias in studies including brain volume measurements. NeuroImage, 83:355-60.

2. Smith S.M. (2002). Robust Automated Brain Extraction. Human Brain Mapping, 17(3): 143-155.

3. Lemieux L, Hagemann G, Krakow K, Woermann F.G. (1999). Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data, Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine, 42 (1999), pp. 127–135.

4. Segonne, 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.

5. Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., (2001). Magnetic resonance image tissue classification using a partial volume model. Neuroimage 13, 856–876.

6. Ward, B.D., (1999). 3dIntracranial: Automatic segmentation of intracranial region.

7. Zhuang, A.H., Valentino, D.J., Toga, A.W. (2006). Skull-stripping magnetic resonance brain images using a model-based level set. Neuroimage 32, 79–92.

8. Sadananthan, S.A., Zheng, W., Chee, M.W., Zagorodnov, V. (2010). Skull stripping using graph cuts. NeuroImage 49, 225–239.

9. Ashburner J, Friston KJ. (2000). Voxel-based morphometry — the methods. Neuroimage 11:805-21.

10. Leung, K.K., Barnes, J., Modat, M., Ridgway, G.R., Bartlett, J.W., Fox, N.C., Ourselin, S., (2011). Brain MAPS: an automated, accurate and robust brain extraction technique using a template library. Neuroimage 55, 1091–1108.

11. Eskildsen S.F, Coupé P., Leung K.K., Fonov V., Manjón J. V., Guizard N., Wassef S., Østergaard L.R., Collins D. L. and ADNI. (2012). BEaST: Brain Extraction based on nonlocal Segmentation Technique. Neuroimage, 59(3):2362-73.

12. Park J.G. and Lee C. (2009). Skull stripping based on region growing for magnetic resonance brain images. Neuroimage, 47 (2009), pp. 1394–1407

13. Sandor, S., Leahy, R., (1997). Surface-based labeling of cortical anatomy using a deformable atlas. IEEE Trans. Med. Imaging 16, 41–54.

14. Ashburner J, Friston KJ. (2005). Unified segmentation. Neuroimage 26(3):839-51.

15. Buckner, R.L., Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., Snyder, A.Z., (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. NeuroImage 23 (2), 724–738.

16. Driscoll, I., Davatzikos, C., An, Y., Wu, X., Shen, D., Kraut, M., Resnick, S.M. (2009). Longitudinal pattern of regional brain volume change differentiates normal aging from MCI. Neurology 72 (22), 1906–1913.

17. Shattuck, D.W., Prasad, G., Mirza, M., Narr, K.L., Toga, A.W., (2009). Online resource for validation of brain segmentation methods. NeuroImage 45 (2), 431–439.

18. Kempton M, Underwood T, Brunton S, Stylios F, Schmechtig A, Ettinger U, Smith M , Lovestone S , Crum W, Frangou S, Williams S, Simmons A. (2011). A comprehensive testing protocol for MRI neuroanatomical segmentation techniques: Evaluation of a novel lateral ventricle segmentation method. Neuroimage, 58-1051:1059.

19. Gousias I, Rueckert D, Heckemann R. A, Dyet L. E., Boardman J. P., Edwards A. D. and Hammers A. (2007). Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. NeuroImage, 40(2): 672-684.

20. Manjón J.V, Coupé P, Martí-Bonmatí L, Robles M, Collins L. (2010). Adaptive Non-Local Means Denoising of MR Images with Spatially Varying Noise Levels. Journal of Magnetic Resonance Imaging, 31,192-203.

21. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. (2010). N4ITK: Improved N3 Bias Correction, IEEE Trans Med Imag, 29(6): 1310-1320.

22. Sled, J.G., Zijdenbos, A.P., Evans, A.C., (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17,87–97.

23. Ibáñez L., Schroeder W., Ng L., Cates J., Consortium T.I.S., Hamming R. (2003). The ITK Software Guide, Kitware, Inc.

24. Avants, B. B.; Epstein, C. L.; Grossman, M. & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal, 12, 26-41

25. Lotjonen J, Wolz R, Koikkalainen J, Thurfjell L, Waldermar G, Soininen H, Rueckert S. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage, 49(10): 2352-2365.

26. Manjón J.V, Tohka J , García-Martí G, Carbonell-Caballero J, Lull J.J., Martí-Bonmatí L. and Robles M. (2008). Robust MRI Brain Tissue Parameter Estimation by Multistage Outlier Rejection. Magnetic Resonance in Medicine, 59:866-873.

27. Nyul, L. G. and Udupa, J. K. (2000). Standardizing the MR image intensity scales: making MR intensities have tissue specific meaning. Medical Imaging 2000: Image Display and Visualization, 1, 496-504.

28. Keihaninejad S., Heckemann R, Fagiolo G, Symms M, Hajnal J, Hammers A. (2010). A robust method to estimate the intracranial volume across MRI field strengths (1.5T and 3T). NeuroImage, 50: 1427–1437.

29. Peelle J, Cusack R, Henson R. (2012). Adjusting for global effects in voxel-based morphometry: Gray matter decline in normal aging. NeuroImage, 60: 1503–1516.

30. Pengas G, Pereira JM, Williams GB, Nestor PJ. (2009). Comparative Reliability of Total Intracranial Volume Estimation Methods and the Influence of Atrophy in a Longitudinal Semantic Dementia Cohort. J Neuroimaging. 2009 Jan;19(1):37-46

31. Yushkevich P.A., Piven J., Hazlett H.C., Smith R.G., Ho S., Gee J.C., and Gerig G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage, 31(3):1116-28.

32. Coupé, P., Manjón, J., Fonov, V., Pruessner, J., Robles, M., Collins, L., (2011). Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54, 940–954.

33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., (2004). Image quality assessment: from error visibility to structural similarity: IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 13, pp. 600–612.

34. Tomasi C and Manduchi R. (1998). Bilateral Filtering for Gray and Color Images. Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India.

35. Rousseau F., Habas P.A., Studholme C. (2011). A Supervised Patch-Based Approach for Human Brain Labeling. IEEE Trans. Med. Imaging 30(10): 1852-1862.

36. Coupé P, Yger P, Prima S., Hellier P., Kervrann C., Barillot C. (2008). An Optimized Blockwise NonLocal Means Denoising Filter for 3-D Magnetic Resonance Images. IEEE Transactions on Medical Imaging, 27(4):425–441.

37. Marcus, D.S.,Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19, 1498–1507.

38. Zijdenbos A.P., Dawant B.M., Margolin R.A., Palmer A.C., (1994). Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imaging, 13, pp. 716-724.