NeuroImage: Clinical 11 (2016) 264–275

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Temporal filtering of longitudinal brain magnetic resonance images for consistent segmentation

Snehashis Roy[a,*], Aaron Carass[b,c], Jennifer Pacheco[d], Murat Bilgel[b,d], Susan M. Resnick[d], Jerry L. Prince[b], Dzung L. Pham[a]

[a] Center for Neuroscience and Regenerative Medicine, Henry M. Jackson Foundation for the Advancement of Military Medicine, United States
[b] Image Analysis and Communications Laboratory, Department of Electrical and Computer Engineering, Johns Hopkins University, United States
[c] Department of Computer Science, Johns Hopkins University, United States
[d] Laboratory of Behavioral Neuroscience, National Institute on Aging, United States

ARTICLE INFO

Article history:
Received 28 August 2015
Received in revised form 13 January 2016
Accepted 12 February 2016
Available online 16 February 2016

ABSTRACT

Longitudinal analysis of magnetic resonance images of the human brain provides knowledge of brain changes during both normal aging as well as the progression of many diseases. Previous longitudinal segmentation methods have either ignored temporal information or have incorporated temporal consistency constraints within the algorithm. In this work, we assume that some anatomical brain changes can be explained by temporal transitions in image intensities. Once the images are aligned in the same space, the intensities of each scan at the same voxel constitute a temporal (or 4D) intensity trend at that voxel. Temporal intensity variations due to noise or other artifacts are corrected by a 4D intensity-based filter that smooths the intensity values where appropriate, while preserving real anatomical changes such as atrophy. Here smoothing refers to removal of sudden changes or discontinuities in intensities. Images processed with the 4D filter can be used as a pre-processing step to any segmentation method. We show that such a longitudinal pre-processing step produces robust and consistent longitudinal segmentation results, even when applying 3D segmentation algorithms. We compare with state-of-the-art 4D segmentation algorithms. Specifically, we experimented on three longitudinal datasets containing 4–12 time-points, and showed that the 4D temporal filter is more robust and has more power in distinguishing between healthy subjects and those with dementia, mild cognitive impairment, as well as different phenotypes of multiple sclerosis.

1. Introduction

Segmentation of brain magnetic resonance (MR) images is an important step in analyzing brain structures. By providing quantitative measures of the shape and size of brain structures, a better understanding of normal aging (Resnick et al., 2003; Thambisetty et al., 2010), as well as diseases such as Alzheimer's (Querbes et al., 2009) and multiple sclerosis (MS) (Shiee et al., 2012) can be gained. Longitudinal segmentation in brain imaging provides additional insights into the dynamics of brain anatomy for monitoring atrophy and other structural changes that may be related to disease progression. For example, longitudinal imaging studies have revealed accelerated brain volume decline in mild cognitive impairment (Driscoll et al., 2009) and accelerated gray matter atrophy in MS progression (Fisher et al., 2008). The goal of this work was to develop an algorithm that improves the accuracy and stability of brain segmentations in longitudinal data.

The primary challenge of a 4D image segmentation method, in contrast to 3D analysis, is ensuring consistency or stability of the results while retaining sensitivity. Longitudinal processing is aimed toward quantifying time-varying changes of a subject. However, the presence of image artifacts and noise can reduce the sensitivity of a 4D segmentation method by overshadowing the time-varying effects. An example is shown in Fig. 1 where four longitudinal T_1-weighted SPGR (spoiled gradient recalled) scans of a healthy volunteer are processed with independent 3D Freesurfer (FS) (Dale et al., 1999) and a state-of-the-art 4D segmentation method, longitudinal Freesurfer (or 4D Freesurfer) (Reuter et al., 2012). Each scan is separated by approximately one year. The inconsistency is visually evident in the hard segmentations (orange arrow in Fig. 1 second row), where the cortical gray matter shrinks and grows over time. Although all images are scanned in the same scanner and with the same pulse sequence parameters, small differences in the noise level or intensities across time-points give rise to inconsistency in the 4D segmentations. In comparison, use of our 4D filter produces a more consistent segmentation, where the cortex shrinks gradually, as expected in normal aging (Fig. 1 bottom row). In this paper, we address this issue of segmentation stability in

* Corresponding author.
E-mail address: snehashis.roy@nih.gov (S. Roy).

http://dx.doi.org/10.1016/j.nicl.2016.02.005

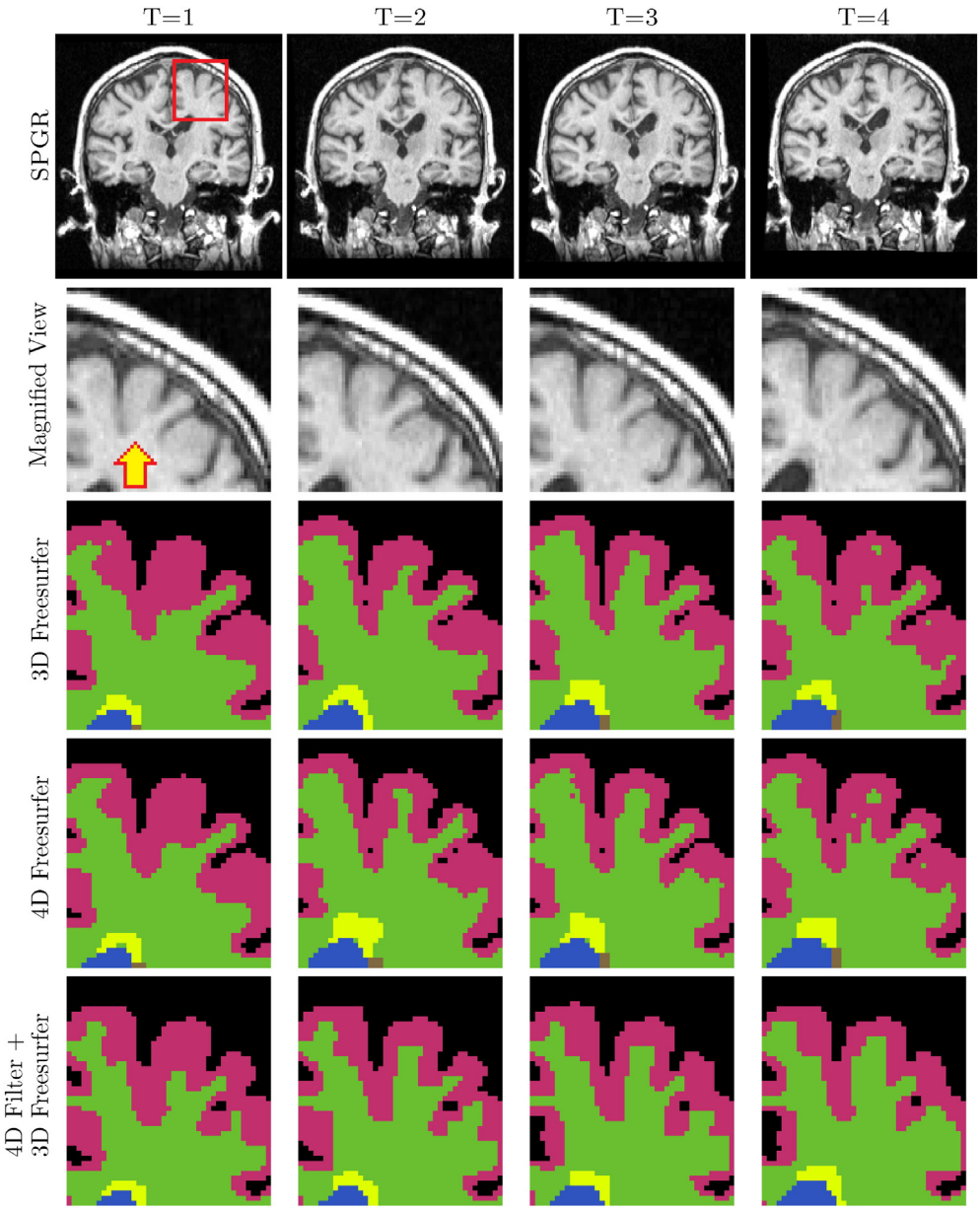

Fig. 1. Top row shows SPGR scans of four consecutive time-points of a healthy volunteer (age 68 at $T=1$), with each time point being one year apart. Next four rows show magnified views of the cortex on the original images, segmentations from 3D Freesurfer, 4D Freesurfer, and 3D Freesurfer following by our 4D filtering. An arrow shows where the cortex shrinks and grows periodically on both 3D and 4D Freesurfer segmentations, while it is more consistent with the 4D filter.

longitudinal imaging studies of aging or diseases where one might expect gradual changes.

Several methods have been previously proposed to analyze 4D images in a longitudinally consistent manner. CLASSIC (Xue et al., 2006) is a 4D segmentation algorithm that contains a 4D registration and 4D segmentation loop. First, all 3D time-points are co-registered using a 4D registration (Shen and Davatzikos, 2004; Prastawa et al., 2012; Kim et al., 2013) to account for the anatomical changes and then are segmented by a 4D version of a fuzzy C-means algorithm (Pham, 2001) to account for temporal smoothness in segmentation. Here, smoothness refers to lack of sudden changes or discontinuities in intensities across time. A 4D segmentation pipeline was also proposed in (Dai et al., 2012), where temporal constraints on cortical thickness are enforced to obtain longitudinally smooth cortical thickness measures. In 4D Freesurfer (Reuter et al., 2012), individual time-points are first segmented with 3D Freesurfer (Dale et al., 1999), a mean template is created, and then the time-points are segmented using the mean template as a target in an unbiased fashion. However, 4D FreeSurfer does not take advantage of the fact that much of the brain often remains unchanged over time, which could be used to improve robustness to noise and other artifacts.

In this paper, we propose a patch based 4D temporal filtering algorithm as a pre-processing step to any segmentation method so as to obtain temporally consistent longitudinal brain segmentations. A *patch* is a small 3D subimage (e.g., $3\times3\times3$ voxels) centered at a voxel of interest. For a longitudinal dataset where the scans are of similar contrast, time-points are first co-registered using a rigid transformation. Once the images are rigidly registered, we assume that some changes in anatomy can be modeled by smooth temporal changes in image intensities. Patches are used to model the temporal change in order to take advantage of the contextual information within the neighborhood around a voxel (Roy et al., 2013a; Roy et al., 2011; Roy et al., 2014). We then automatically distinguish between two types of patches, (1) patches that show gradual temporal atrophy, but can be corrupted by noise and minor intensity variations, and (2) patches that do not follow any gradual trend. The intensities of only the first type of patches

are filtered using an auto-regressive model of first order (AR(1)), under the assumption that the patch intensities of the t^{th} time-point can be predicted from the $(t\text{-}1)^{th}$ time-point. The intensities of the second type of patches are not changed. Once the images are temporally filtered, they can be segmented using any segmentation method. We show that by using temporal filtering as a pre-processing step, stable and robust 4D segmentations are obtained. We have presented preliminary elements of this work in conference form (Roy et al., 2013c, 2013b).

The paper is organized as follows. First, the motivation for intensity based regression of normal anatomical changes is proposed in Section 2.2. Then the auto-regressive model on the intensities is described in Section 2.3. Next, we evaluate the accuracy of the method on simulated longitudinal atrophies and show the improvement in robustness on test–retest scans of real data in Section 3.2. We also show that 4D filtering improves discrimination of dementia or mild cognitive impairment when compared against healthy groups using a variety of segmentation algorithms in Sections 3.3 and 3.4. In particular, we show that using 3D FreeSurfer with our proposed approach improves upon 4D Freesurfer especially when the number of time-points is large (Section 3.4). Finally, we explore in Section 3.5 the effect of 4D filtering in finding GM atrophy on subjects with multiple sclerosis (MS).

2. Materials and methods

2.1. Data sets

We experimented on 6 sets of data, listed in Table 1. First, we estimated the parameter f based on the image contrast and scanners, as noted in Section 2.3. We simulated atrophy on a dataset containing two subjects (denoted by Sim-2), one with both SPGR (GE 1.5T, $0.94\times0.94\times1.5$ mm^3, $T_R=6.9$ ms, $T_E=3.4$ ms, flip angle $=8^\circ$) and MPRAGE (GE 1.5T, $0.94\times0.94\times1.5$ mm^3, $T_R=35$ ms, $T_E=5$ ms, flip angle $=45^\circ$) and one with MPRAGE from a Philips 3T ($0.82\times0.82\times1.1$ mm^3, $T_R=10.21$ ms, $T_E=6$ ms, flip angle 8°) scanner. These two subjects were chosen from the BLSA and MS dataset, respectively (explained later in this section).

To estimate the test–retest stability of the method, we use two datasets. For the first set, MPRAGE scans ($0.82\times0.82\times1.17$ mm^3, $T_R=10.2$ ms, $T_E=6$ ms, flip angle 8°) from a 3T Siemens scanner (denoted by TR-10) were acquired every week for 10 consecutive weeks on a healthy male volunteer. Precautions were taken to have identical scanning and patient conditions every time, e.g., scans were done at the same time of the day, and the volunteer was allowed the same amount of sleep and fluid intake on the day of scanning. The second dataset consists of SPGR scans (GE 3T, 1.0 mm^3, $T_R=8.06$ ms, $T_I=450$ ms, flip angle 8°) of 29 healthy subjects, each subject having 10 scans over a month, separated by 3 days. This dataset (denoted by CoRR-29) is obtained from the Consortium for Reliability and Reproducibility (CoRR) database (Zuo et al., 2014), specifically the HNU dataset (Hangzhou Normal University) (http://dx.doi.org/10.15387/fcp_indi.corr.hnu1). Note that although the scans were available for 30 subjects, we discarded one due to an incomplete field of view.

To demonstrate the discriminative power in longitudinal analysis versus cross-sectional analysis, 49 subjects with 3-5 time-points (age range 60-92) were selected from the OASIS (Open Access Series of Imaging Studies) (Marcus et al., 2007) longitudinal database (dataset denoted by OA-49). Each visit was separated by approximately one year and there was an average of three visits per subject. For every subject and every visit, there were 3–4 repeat MPRAGE scans ($1\times1\times1.25$ mm^3, $T_R=9.7$ ms, $T_E=4$ ms) from a Siemens 1.5T scanner. The repeat scans are co-registered and averaged to improve signal-to-noise ratio. Among the 49 subjects, 15 of them were diagnosed with dementia, and the other 34 were characterized as non-demented throughout the span of the study. Note that there were also 14 patients in the database who were initially non-demented, but were diagnosed with dementia at later time-points. We did not include them in our subset.

To evaluate our method when applied to a greater number of time-points, we experimented on 39 subjects from the BLSA (Baltimore Longitudinal Study of Aging) database (Resnick et al., 2000), with 15 of them diagnosed with mild cognitive impairment (MCI) at all time-points (dataset denoted by BL-39). The other 24 subjects are healthy controls. Each subject contains 4–11 visits (average 9 visits per subject), each separated by approximately one year. SPGR images were acquired axially ($0.94\times0.94\times1.5$ mm^3, $T_R=35$ ms, $T_E=5$ ms, flip angle $=45^\circ$) for each visit on a GE 1.5T scanner.

The 4D filtering was also applied on a set of MS patients, where we explored the progression of atrophy on different phenotypes of MS. The data set includes 59 patients with MS (dataset denoted by MS-59), with each patient having 3–8 visits (average 4), each visit separated by a year. The average age of the participants was 44 years (range 22–67), with an average disease duration 9 years. Among the 59 patients, 22 were diagnosed with relapsing remitting MS (RRMS), 15 with primary progressive (PPMS), and 22 with secondary progressive MS (SPMS). All subjects had T_1-w MPRAGE ($0.82\times0.82\times1.1$ mm^3) and FLAIR ($0.82\times0.82\times2.2$ mm^3) scans acquired on a 3T Philips scanner (TR/TE $=10.21/6$ ms, flip angle 8°).

To demonstrate the flexibility of our approach, we apply it as a preprocessing step to several different segmentation algorithms. These algorithms each use different underlying methodologies. The Freesurfer (Dale et al., 1999) algorithm employs a Markov random field model in combination with statistical atlases and deformable surfaces. Because it includes both a 3D and 4D implementation, a natural comparison that we focus on is the combination of temporal filtering and 3D Freesurfer against 4D Freesurfer (Reuter et al., 2012). Other algorithms that we tested include Atropos (Avants et al., 2011), which is based on an expectation–maximization approach, FIRST (Patenaude et al., 2011), which is based on active appearance models, TOADS (Bazin and Pham, 2008), which uses fuzzy clustering with topological constraints, and MALP-EM (Ledig et al., 2015), which combines multi-atlas label fusion with expectation–maximization.

2.2. Motivation

Most existing 4D segmentation methods involve a nonlinear 4D registration step where either the individual images (Dale et al., 1999) or their segmentations (Xue et al., 2006) are deformably registered to a common template. Thus the temporal change in anatomy is modeled

Table 1
The table describes the details of the datasets used for all the experiments.

	Name	Contrast	#	Scanner	Resolution (mm^3)	T_R (ms)	T_E (ms)	Flip angle
Simulation	Sim-2	SPGR & MPRAGE	2	GE 1.5T & Philips 3T	$0.94\times0.94\times1.5$ & $0.82\times0.82\times1.1$	6.9 & 35	3.4 & 5.0	8° & 45°
Test–retest	TR-10	MPRAGE	1	Siemens 3T	$0.82\times0.82\times1.17$	10.2	6	8°
	CoRR-29	SPGR	29	GE 3T	$1.0\times1.0\times1.0$	8.06	N/A	8°
OASIS	OA-49	MPRAGE	49	Siemens 1.5T	$1.0\times1.0\times1.25$	9.7	4	10°
BLSA	BL-39	SPGR	49	GE 1.5T	$0.94\times0.94\times1.5$	35	5	45°
MS	MS-59	MPRAGE	59	Philips 3T	$0.82\times0.82\times1.1$	10.21	6	8°

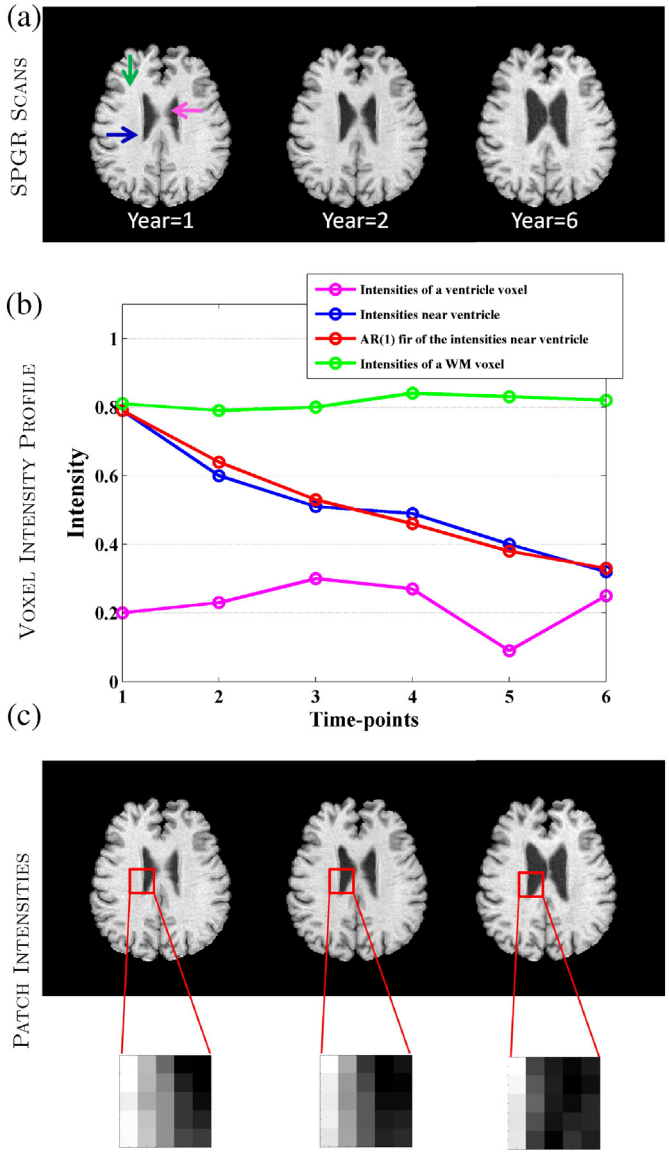

Fig. 2. The top row shows SPGR scans of three time-points of a normal subject having 6 longitudinal scans, separated by a year. The second row shows the intensity profiles of three voxels, one inside ventricle (magenta), one in deep WM (green), and one on the ventricle-WM boundary (blue). The AR(1) fit of the blue line is shown in red. The bottom row shows the intensities of 5×5 patches around the blue voxel over the 3 time-points. The leftmost and rightmost voxels in the patches represent WM and ventricle voxels that remain unchanged over the time course, while the voxels in the middle columns of the patches are generally decreasing in intensity, indicating enlargement of the ventricle. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by a geometric "morphing" of the images. However, it has been shown that the same change in anatomy can also be modeled by an intensity change (Miller and Younes, 2001).

An example of how longitudinal anatomical changes can be explained by intensity changes is shown in Fig. 2(a), where three time-points (years 1, 2, and 6) of a healthy volunteer are shown. Each of the time-points is rigidly registered (Jenkinson and Smith, 2001) to a reference space (e.g., an atlas or an average of all time points). Each of the images are scaled so that the mode of the white matter (WM) intensity is at unity. The mode is automatically found by fitting a smooth kernel estimator to the image histogram (Pham and Prince, 1999). Fig. 2(b) shows the temporal intensity profiles of three voxels, one each in deep WM, ventricles, and on the ventricle-WM boundary. As the subject is healthy, the anatomical change associated with normal aging, e.g., enlargement of the ventricles, is manifested in the monotonic decrease in intensity of the voxel on the WM-ventricle boundary (blue line), with perturbations due to image noise. Also, the voxel inside WM (or ventricle) remains WM (or ventricle) for all the time-points, thus its intensity remains in the realm of the corresponding tissue intensities. A similar trend can also be observed for a patch as well. Fig. 2(c) shows the same 5×5 patch (in 2D) over the three time-points, demonstrating the shifting boundary of the ventricles. We propose that after rigid registration, the trend in some patches can be attributed to various factors like noise (e.g., green and magenta lines in Fig. 2(b)), actual gradual anatomical changes (e.g., blue line in Fig. 2(b)), or small registration errors. To remove the noise and other artifacts from those patches, we propose to fit the longitudinal intensities to a smooth first order auto-regressive model.

2.3. Algorithm

Image patches are defined as $p\times q\times r$ 3D subimages associated with each voxel of the image. Patches are typically small and centered on the voxel of interest—e.g., $p=q=r=3$. It is convenient for writing out the mathematics to describe a patch as the 1D vector of size $d\times 1$, where $d=pqr$. The voxels within a patch are always ordered in the same way, using a consistent rasterization, to create this 1D representation.

We assume that there are T time-points available for a subject, $\{s_1,\ldots,s_T\}$. Each of the time-points $\{s_1,\ldots,s_T\}$ is rigidly registered to a reference template. Each scan is skull-stripped with the same mask (generated from either the baseline or a mean template) and corrected for any intensity inhomogeneities independently. They are also scaled so that the mode of the WM intensities of each image is unity (Pham and Prince, 1999; Shinohara et al., 2014; Roy et al., 2013d). The images are also assumed to be scanned with the same acquisition protocol. Each image is first decomposed into $d\times 1$ patches, $y_i^{(t)}, i=1,\ldots,N, t=1,\ldots,T$, where N is the total number of voxels in each image. Since the images are rigidly registered, the trend at the i^{th} voxel is obtained from the collection of patches $\{y_i^{(1)},\ldots,y_i^{(T)}\}$.

At every voxel, an observed image patch $y_i^{(t)}$ is assumed to consist of a "*true patch*" $x_i^{(t)}$ and additive noise, expressed as.

$$y_i^{(t)} = x_i^{(t)} + \boldsymbol{\varepsilon}_i^{(t)}, 1\le t\le T, \tag{1}$$

where the observed patch is a noisy perturbation of the true patch at each voxel. Here $\boldsymbol{\varepsilon}_i^{(t)}$ accounts for both image noise as well as small intensity variations. We assume that there are two types of patches: patches that follow a gradual trend and patches that do not. Our assumption is that aging or neurodegenerative processes typically cause gradual shrinking of brain structures. These constitute the first type of patch, while other patches fall into the second type. We discuss this in greater detail in the remainder of the section.

As shown in Fig. 2, there are some patches for which the atrophy results in smooth intensity change, such as ventricular enlargement. We assume that the gradual anatomical changes can be modeled by smooth changes in intensities; therefore, for those patches, $x_i^{(t)}$ are modeled as an AR(1) regression,

$$x_i^{(t)} = M_i\, x_i^{(t-1)},\ t>1. \tag{2}$$

Assuming a_i is the i^{th} true patch for the baseline image ($t=1$), then $x_i^{(t)}$ can be re-written as,

$$x_i^{(t)} = M_i^{t-1}\, a_i, t\ge 1. \tag{3}$$

The $d\times d$ matrix M_i is positive definite and contains the parameters of the AR model. The exponent on M_i signifies multiple products with itself ($(t-1)$ power). We note that the intensities are a nonlinear function of the time t; therefore the model has the capability of detecting nonlinear changes in the intensity profile. However, because M_i itself is not a

function of t, $x_i^{(t)}$ must follow a monotonic trend. The first order model is based on the assumption that the serial images of a subject are obtained at approximately uniform intervals, which is true for all subjects in our database. However, when a subject is scanned at highly nonuniform intervals, higher order models can be used in the same framework.

There are some patches that do not follow monotonic change in intensities. Such patches might include acute lesions in MS subjects which may appear and disappear in successive time-points, the presence of a new tissue-boundary in developing brains that is not possible to predict from the previous time-point, or tissue variations due to hydration level changes (Nakamura et al., 2015). We distinguish between these two types of tissue patches. The patches that do not follow smooth, monotonic intenity changes are where the fitting error of the AR regression (Eq. (3)) is sufficiently large. Therefore, for such patches, the original intensities must be preserved. Eq. (3) is modified to satisfy this criteria for all patches,

$$z_i^{(t)} = w_i^{(t)} M_i^{t-1}\, a_i + \left(1 - w_i^{(t)}\right) y_i^{(t)}, t>1. \tag{4}$$

Here, $z_i^{(t)}$ is an estimated patch accounting for both gradual and non-gradual atrophies. The scalar $w_i^{(t)}$ is a weight that determines if the trend at the i^{th} patch is due to a gradual anatomical change (such as growing ventricles in normal aging, $w_i^{(t)} \to 1$) or some atrophy that is not gradual. In the presence of such non-gradual atrophies, the intensities cannot be smoothed with an AR model, thus $w_i^{(t)}$ is set to 0.

The weights $w_i^{(t)}$ are computed in a data-dependent manner so as to distinguish between the two types of patches. If the model fitting error $\| y_i^{(t)} - x_i^{(t)} \|^2$ is too large, it indicates the presence of a non-monotonic atrophy that cannot be explained by smooth change in intensities. Following this notion, $w_i^{(t)}$ is defined as,

$$w_i^{(t)} = \frac{1}{\sqrt{1 + \dfrac{\| y_i^{(t)} - x_i^{(t)} \|_2^2}{f^2}}}, t \geq 1. \tag{5}$$

where f is a smoothing parameter that also acts as a soft noise threshold. Note that $w_i^{(t)}$ depends on the deviation of the observed patch $y_i^{(t)}$ from the true patch $x_i^{(t)}$, when the patch is assumed to follow a gradual atrophy. If this deviation $\|\delta_i^{(t)}\| = \| y_i^{(t)} - x_i^{(t)} \| \gg f$, $w_i^{(t)} \approx 0$, then the patch is largely unaffected by the smoothing filter, i.e., $x_i^{(t)} \approx y_i^{(t)}$. Estimation of f is described in Section 3.1.

Based on the data $y_i^{(t)}$, we first estimate the weights $w_i^{(t)}$ and AR(1) parameters M_i and a_i. Then the estimated patches $z_i^{(t)}$ are found from Eq. (4) using these estimates. For mathematical simplicity, we make the assumption that M_i is a positive definite diagonal matrix, $M_i = diag\{m_{i,1}, \ldots, m_{i,d}\}$, and thus $M_i^t = diag\{m_{i,1}^t, \ldots, m_{i,d}^t\}$. Then the filtered intensities $x_i^{(t)}$ are found for the i^{th} patch by minimizing the ℓ_2 norm of the error with respect to M_i and a_i. The total sum of squared error of the estimated patches is given by

$$\begin{aligned} E_{\text{Total}} &= \sum_{i=1}^{N} \sum_{t=1}^{T} \| y_i^{(t)} - z_i^{(t)} \|_2^2 \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} \left(w_i^{(t)}\right)^2 \| y_i^{(t)} - M_i^{t-1}\, a_i \|_2^2 \\ &= f^2 \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\| y_i^{(t)} - M_i^{t-1}\, a_i \|_2^2}{f^2 + \| y_i^{(t)} - M_i^{t-1}\, a_i \|_2^2}. \end{aligned} \tag{6}$$

Note that although we employ an autoregressive model of the first order, it is evident from Eq. (6) that all time-points are used to estimate the model parameters.

Estimates of M_i and a_i are obtained by differentiating E_{Total} with respect to these variables and setting the gradients to zero. Under the diagonal assumption of M_i, the update equations are as follows,

$$\sum_{t=1}^{T} (t-1)\left(y_{i,\ell}^{(t)} - m_{i,\ell}^{t-1} a_{i,\ell}\right) m_{i,\ell}^{t} \frac{1}{\left(f^2 + \|\delta_i^{(t)}\|_2^2\right)^2} = 0, \tag{7}$$

$$\sum_{t=1}^{T} \left(y_{i,\ell}^{(t)} - m_{i,\ell}^{t-1} a_{i,\ell}\right) m_{i,\ell}^{t-1} \frac{1}{\left(f^2 + \|\delta_i^{(t)}\|_2^2\right)^2} = 0, \tag{8}$$

$$\delta_i^{(t)} = y_i^{(t)} - x_i^{(t)}, \ell = 1, \ldots, d,$$

where $y_{i,\updownarrow}^{(t)}$ and $a_{i,\updownarrow}$ denote the $\updownarrow^{\text{th}}$ component of $y_i^{(t)}$ and a_i, respectively, $\updownarrow = 1, \ldots, d$. Initializing $m_{i,\updownarrow} = 1$ and $a_i = y_i^{(1)}$, Eqs. (7)–(8) are repeated until the values of $m_{i,\ell}$ and a_i converge. The converged values for M_i, a_i, and a chosen value for f (see Section 3.1) are used to provide the estimated patches $z_i^{(t)}$ based on Eq. (4). It is noted that the computation in Eqs. (7)–(8) is simplified by the diagonal assumption of M_i, and they are solved for each of the patch dimensions ℓ separately. Nevertheless, the patch-based error $\|\delta_i^{(t)}\|$ introduces contextual information in the computation of M_i. The diagonal M_i can be replaced by an arbitrary $d \times d$ positive definite matrix M_i. However, the number of time-points is usually much less than the patch dimension (i.e. $T \ll d$), introducing instability in the computation of an arbitrary positive definite matrix.

Fig. 3 shows four out of ten time-points of a healthy volunteer, each separated by two years, and the corresponding $m_{i,14}$ for the center voxel of each $3 \times 3 \times 3$ patch in the image ($d = 27$). Voxels that remain WM throughout the time span (deep WM denoted by a white arrow), have a $m_{i,14} \approx 1$ indicating their intensity remains stable over the time series. Enlargement of the ventricles (black arrow) is also evident, resulting in $m_{i,14} < 1$ shown in dark blue, which indicates a decrease in intensity over time, corresponding to the transition from WM voxels to ventricle voxels. Some cortical thinning is also observed with $m_{i,14} < 1$ near deep sulci (magenta and white arrows), where GM voxels turn into CSF voxels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The parameter f acts as a soft threshold for the AR model fitting to the temporal intensities of a patch. By design, it can also be a user input to the algorithm. Larger f implies less error ($\|\delta_i^{(t)}\|$) tolerance, indicating filtered intensities should follow the autoregressive model more strictly. Ideally, f should vary from patch to patch based on the tissue type and amount of expected atrophy at a patch. For example, with a ventricular CSF patch, the inherent signal-to-noise ratio is lower on T_1-weighted images (Sijbers et al., 1998) than a pure WM patch, indicating that the choice of f should be higher for CSF patches than WM. The choice of f should also depend on the acquisition protocol of the image, e.g., SPGR vs MPRAGE and the scanner. However, in this paper, we only address the variation of f with respect to the image contrast and scanner, and choose a global f based on a simulation study on an SPGR and MPRAGE images of different scanners in Section 3.1.

3. Results

The algorithm was implemented in MATLAB. We used $3 \times 3 \times 3$ patches, i.e., $d = 27$, for all our experiments. The processing time is about 2 hours on a 12-core 2.92GHz Intel processor for an 11 time-point data set with images of 1 mm^3 resolution, requiring about 2-3GB of memory. 3D and 4D Freesurfer require about 24 and 36 h, respectively on the same machine. Before temporal filtering, all images were pre-processed by skull-stripping (Carass et al., 2011), registration to MNI atlas (Mazziotta et al., 1995) (www.mristudio.org), resampling to 1 mm^3 isotropic resolution, and inhomogeneity correction using N4 (Tustison et al., 2010).

3.1. Simulation study

To estimate the parameter f in Eq. (5), we used an atrophy simulation algorithm (Karacali and Davatzikos, 2006) on the Sim-2 data (Section 2.1) to simulate anatomical changes in the brain. One subject with both SPGR and MPRAGE scans (GE 1.5T) from the BL-39 and one subject with MPRAGE from MS-59 dataset (Philips 3T) were used for this experiment. The simulation algorithm takes an MR image and its 3-class hard segmentation (CSF, GM, and WM) and generates an atrophied image around designated foci with a given radius of atrophy. We used decreasing radii of atrophy (20 to 15 mm) to simulate six time-points where the ventricles grow over time, simulating a common occurrence in late aging. A 3-class fuzzy c-means (Bezdek, 1980) (FCM) based segmentation was used as the input. The six time-points were filtered using varying $f \in (0,1)$ and the resultant images were again individually segmented using FCM. The same experiment was repeated for all of the three acquisitions. The misclassification rates in both cases, averaged over three classes, are obtained for each f. Misclassification rate is defined as the ratio of average number of misclassified voxels (over 6 time-points) and the average number of voxels involved in the simulated atrophy.

As mentioned earlier in Section 2.3, a very small f will not have any effect on the filtering, as smaller f imparts no change in intensities (as $f \to 0$, $w_i^{(t)} \to 0$ in Eq. (5)). With increasing f, the temporal intensities are smoothed more and more, thereby becoming more robust to noise, while at the same time smoothing some actual tissue changes. We obtained these optimal values: $f_{\text{SPGR1.5T}} = 0.31$, $f_{\text{MPRAGE1.5T}} = 0.21$, $f_{\text{MPRAGE3T}} = 0.19$, at less than 2% misclassification error. Since the f values for MPRAGE on both scanners are very close, we used $f_{\text{MPRAGE}} = 0.21$ for all the experiments with MPRAGE images described later. A larger value of f_{SPGR} compared to f_{MPRAGE} arises from the observation that the contrast to noise ratio in MPRAGE images is typically higher than in SPGR, indicating a smaller noise threshold is needed for detection of gradual atrophy.

3.2. Robustness on test–retest scans

In this section, we evaluate the robustness of the 4D filtering using test–retest scans of healthy volunteers, i.e. TR-10 and CoRR-29 data (Section 2.1). Ideally for a healthy subject, the tissue volumes are not expected to change much within a span of ten weeks or a month. For TR-10 dataset, the volumes of cortical GM, cerebral WM, and ventricles are plotted in Fig. 4 for individual 3D Freesurfer segmentation of the original images (blue triangles), 4D Freesurfer (magenta spheres), and 3D Freesurfer segmentation of temporally filtered images (green stars). Similar to a previously reported study on multi-center test–retest reliability of Freesurfer (Jovicich et al., 2013), 4D FS produces almost half the variability compared to 3D FS on GM and WM, while there is very little difference between the two on ventricles. We also observe that 4D Freesurfer has more variability than the 4D filtering method. The difference is primarily emphasized in ventricle volume, where the coefficient of variation is 1.67% for 4D Freesurfer, while it is 0.53% for our filtering method, shown in Table 2. As the ventricle volume is the most robust statistic for scans spanning over only ten weeks, a significantly smaller coefficient of variation for our temporal filtering indicates greater consistency of the segmentation. A similar decrease is shown for the other tissues. We note that 4D Freesurfer has a consistent bias from individual Freesurfer segmentations with larger ventricles or smaller WM. This is likely due to the difference in the transformation spaces of the methods.

Table 2
Coefficient of variation (in percent) of volume changes in ten consecutive weeks for a healthy subject (TR-10 dataset). Bold indicates lowest coefficient compared to other methods.

	Ventricle	Cortical GM	Subcortical GM	Cerebral WM
3D Freesurfer	1.529	1.756	1.523	1.332
4D Freesurfer	1.669	0.757	0.741	0.557
4D Filter + 3D Freesurfer	**0.532**	**0.568**	**0.589**	**0.380**

Table 3
Mean ± standard deviations of coefficient of variation (in percent) of volume changes in ten consecutive scans are shown for 29 healthy subjects (CoRR-29 dataset). Bold indicates significantly lowest ($p<0.01$) coefficient compared to other methods.

	Ventricle	Cortical GM	Subcortical GM	Cerebral WM
3D Freesurfer	3.48 ± 1.13	2.41 ± 1.21	2.07 ± 0.54	1.15 ± 0.33
4D Freesurfer	2.22 ± 0.60	1.40 ± 0.22	0.99 ± 0.33	1.00 ± 0.17
4D Filter ± 3D Freesurfer	**1.24 ± 0.49**	**1.25 ± 0.57**	0.98 ± 0.39	**0.57 ± 0.22**

A similar analysis was carried out for the CoRR-29 dataset, with 29 healthy volunteers having 10 scans spanned over a month. Table 3 shows the mean and standard deviations of coefficients of variations for the three methods. 4D Filter followed by 3D Freesurfer shows the lowest average coefficient of variation ($p<0.01$ using Wilcoxon signed rank test) among the three on ventricles, cortical GM and WM, indicating significantly improved segmentation stability.

To estimate the stability of the algorithm with respect to the number of time-points, we also created an augmented set of the TR-10 dataset by randomly sampling T images ($T = \{4, \ldots, 9\}$) from the 10 weekly scans. The sampling is done 10 times for each T. Therefore, this augmented dataset contains 60 subsets, each subset containing 4 to 9 scans from the weekly test–retest scans. Then the images were processed with the 4D filtering method. Then both the processed and original images are segmented with FAST (Zhang et al., 2001) to find CSF, GM, and WM segmentations. Since we do not expect the tissues to change much within a span of 10 weeks on a healthy control, Dice coefficients (Dice, 1945) for the three tissue classes were used as a similarity metric and were computed on each subset with respect to the baseline of that subset. The mean Dice coefficients for $T = 4, \ldots, 9$ on all three tissues were significantly larger ($p<10^{-5}$ with Wilcoxon rank-sum test) than the unfiltered images, indicating significant improvement in segmentation stability after 4D filtering. Also the Dice coefficients between T and $(T+1)$ time-points are not significantly different ($p>0.01$) for any T on any tissue, indicating that the 4D filtering is robust to the variation in number of time-points. Median WM Dice for $T = 4$ was 0.97, while it was 0.98 for $T = 9$, compared to 0.91 with unfiltered images.

We also tested our approach with a 3-class Expectation–Maximization (EM) based segmentation method Atropos, which also includes 3-D and 4-D implementations (Avants et al., 2011). Scans of the same healthy subject over 10 weeks are segmented using 4D Atropos with a temporal smoothness weight of 0.3. CSF, GM, and WM tissue volumes are plotted in Fig. 5 for 3D Atropos, 4D Atropos, and 3D Atropos on 4D filtered images. Use of filtering shows an improvement in segmentation stability of our temporal filter over a Markov random field temporal smoothness constraint on the segmentations. Clearly, 3D Atropos shows noisy volume trends, and 4D Atropos shows a slight decreasing trend in WM volume, indicating that the WM volume changes by 5%, while there is a 7% increase in CSF. Our 4D filtering with 3D segmentation produces the most stable segmentation without the spurious volume changes.

3.3. Atrophy detection on OASIS

As no longitudinal dataset with manual labels are freely available to test sensitivity of the proposed filtering method, we instead demonstrate that the 4D filter improves discrimination power between groups of healthy controls and patients with neurodegeneration on the

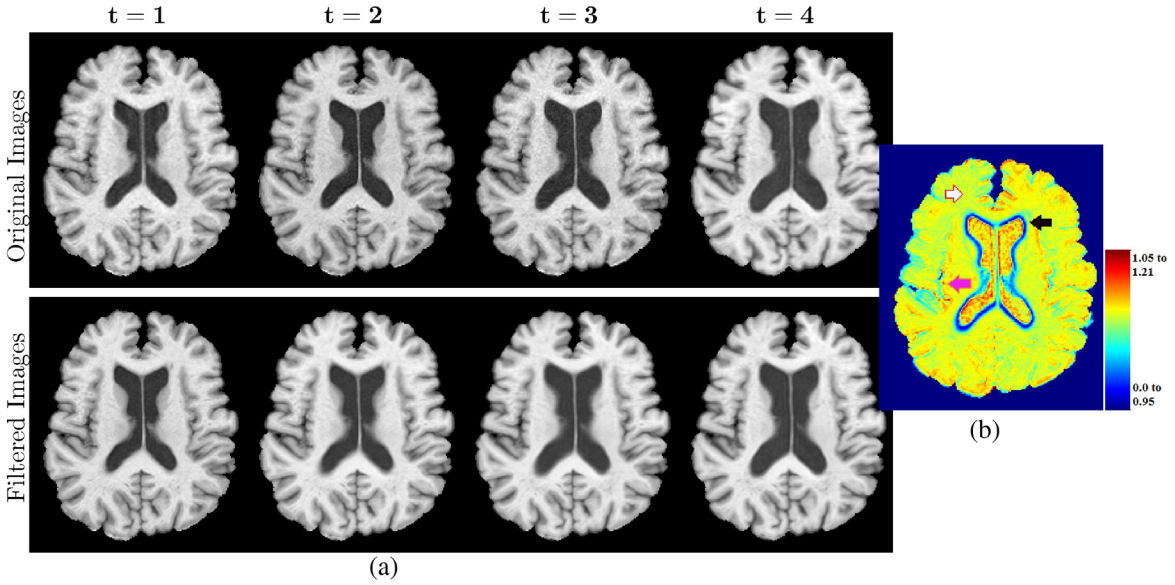

Fig. 3. (a) Four time-points of a subject before and after 4D processing, and (b) a map of $m_{i,14}$ using $3\times3\times3$ patches after temporal filtering ($d=27$). The white arrow points to a voxel inside deep WM with $m_{i,14}\approx1$, indicating no gradual change in intensities since the voxel remains as WM over the time-period. The magenta arrow indicates a voxel on the cortex with $m_{i,14}>1$, indicating a gradual increase in intensities (GM to WM), since the cortex shrinks over time. The black arrow indicates a voxel near the ventricle-WM boundary, where $m_{i,14}<1$. It indicates a gradual increase in ventricle size, where the voxel changes from WM to ventricle over time.

OA-49 data (Section 2.1). Since the filtering can be applied to any 3D segmentation method, we filtered and segmented both demented and non-demented subjects with TOADS (Bazin and Pham, 2008), FIRST (Patenaude et al., 2011), and 3D Freesurfer (Dale et al., 1999). 4D Freesurfer (Reuter et al., 2012) has already been shown to distinguish between the healthy and dementia groups on this dataset. Here, we replicated that result with 4D Freesurfer and showed that the combination of 4D filter and 3D Freesurfer performs similarly, if not better, in distinguishing the two groups than 4D Freesurfer. Similar improvement in distinguishing power after filtering was shown for TOADS and FIRST as well. TOADS provides a tissue segmentation of the whole brain into multiple labels, cerebellar and cerebral WM, GM, sulcal CSF, thalamus, caudate, putamen, and ventricles. FIRST generates subcortical GM segmentation into multiple labels, such as left and right thalamus, caudate, putamen, globus pallidus, hippocampus, and amygdala. Similar labels are also obtained from Freesurfer as well. To measure longitudinal atrophy in segmentation volumes, we propose *absolute percent volume change per year* (APV), defined as

$$\mathrm{APV} = \frac{|V_{t+1} - V_t|}{V_1 \Delta t}$$

where V_1 is a robust baseline volume for every subject, obtained from a linear fit of the longitudinal volumes (Reuter et al., 2012). V_t denotes the volume at t^{th} time-point, and Δt denotes the age difference between time-points $(t+1)$ and t. Note that in addition to the absolute volume changes, APV also normalizes with respect to the actual age difference between two scans. For subjects with more than two time-points, multiple APV values were obtained. However since the rates of atrophy may increase or decrease with age, we do not compute an average APV for a subject. Instead we use the APV values of all time-points of all subjects to compare the atrophy rates between demented and non-demented groups using non-parametric statistical testing.

Fig. 6 shows barplots of median APV when both original and filtered images are segmented with TOADS and FIRST. Subcortical GM volumes have been shown to be associated with dementia and early onset of Alzheimer's disease (Driscoll et al., 2009; Lehmann et al., 2010; Jovicich et al., 2009; den Heijer et al., 2010). Visually, the variation in APV decreases after 4D filtering (shown by lower interquartile range), indicating more confidence in discriminating demented vs non-demented groups. We hypothesized that APV for the demented group is larger than the non-demented group. A Mann–Whitney U test indicates that ventricle and putamen volumes for TOADS segmentations are indeed significantly ($p<0.01$) larger in the demented group after 4D filtering. The caudate and cerebral WM volumes do not show any significant difference after filtering, although the interquartile ranges decrease, indicating decrease in variances. Similarly, both the left and right hippocampus and amygdala volume changes obtained from FIRST are significantly larger ($p<0.001$) after 4D filtering. Left and right globus pallidus volume changes are significantly larger ($p<0.05$)

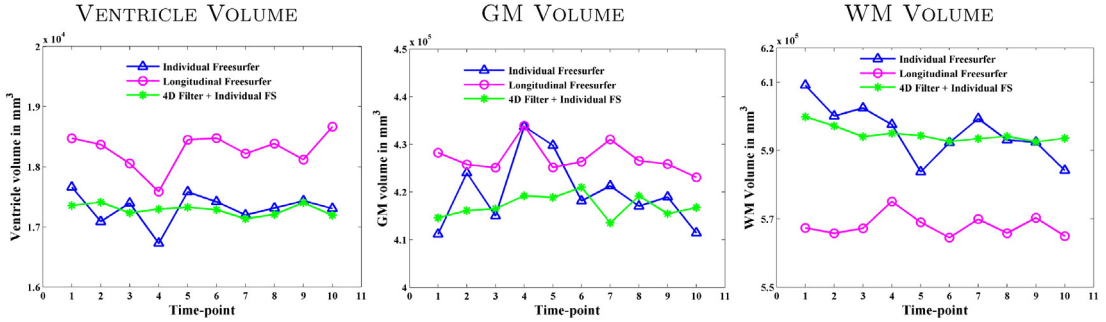

Fig. 4. Ventricle, WM, and cortical GM volumes (in mm^3) are plotted w.r.t. the time-points for a healthy subject, scanned weekly for ten consecutive weeks.

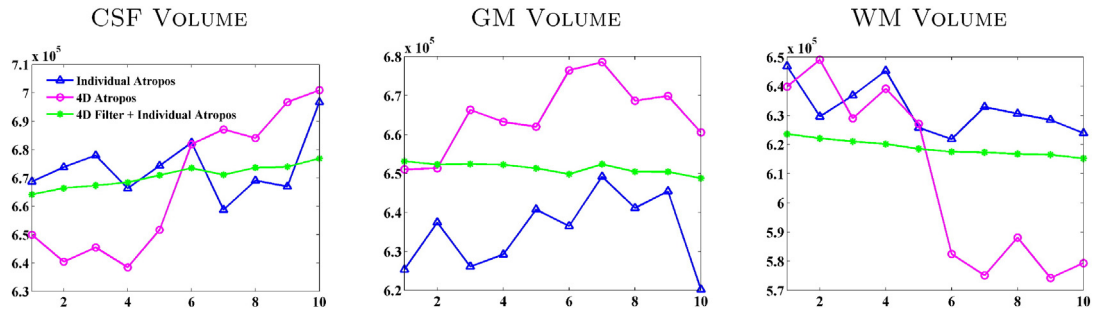

Fig. 5. MPRAGE scans of a healthy volunteer, scanned for 10 consecutive weeks, are segmented using both 3D and 4D EM based segmentation Atropos (Avants et al., 2011). We compared the 4D segmentation volumes in mm^3 (magenta lines) with individual 3D Atropos segmentations of original data (blue lines) and 4D filtered images (green lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

as well. Although we did not find any significant difference between demented and healthy subjects for thalamus volumes, the p-value decreases ($p = 0.067$ to 0.054) after filtering. Note that the statistic obtained from the Mann–Whitney U test is closely related to the area under an ROC (receiver operating characteristics) curve for a binary classification. Therefore, a lower p-value indicates a higher U statistic, which in turn indicates higher area under the curve, indicating improved sensitivity and specificity to distinguish between dementia and non-dementia based on atrophy rates.

Next we tested our approach when used with Freesurfer. Fig. 7 shows barplots of absolute percent volume change per year for 3D Freesurfer, 4D Freesurfer, and 4D filtering followed by 3D Freesurfer for both demented and non-demented groups. As shown earlier (Reuter et al., 2012), magnitude of volume changes, as well as their variances, are sometimes higher in 3D Freesurfer than 4D Freesurfer, e.g., left/right thalamus, putamen, and amygdala. Significant increase in atrophy was observed in the demented group on the left/right hippocampus ($p<0.05$) for all three methods, which is consistent with earlier results (Reuter et al., 2012). We observed significant increases in rates of volume changes in the demented group in the left/right ventricles and thalamus ($p<0.05$) on both 4D Freesurfer and 4D filtering followed by 3D Freesurfer, but not on 3D Freesurfer. We also observed significant increases in left putamen ($p<0.05$) volumes, though there is no significant increase on either 4D or 3D Freesurfer ($p>0.10$). Atrophy in the putamen has been shown to be associated with dementia (Moller et al., 2015). Therefore, our 4D filter followed by 3D Freesurfer has similar performance, if not better, as 4D Freesurfer in distinguishing neurodegeneration, with significant reduction in computation time (about 20–24 h of 3D Freesurfer processing).

3.4. Improvement in atrophy detection on BLSA

While the OASIS dataset has an average of three time-points per subject, we also experimented on the BL-39 dataset (Section 2.1), which has an average of nine time-points per subject separated by one year, to show the improvement in sensitivity and specificity in detecting atrophy when average number of time-points is larger. As before, we segmented the images with 3D and 4D Freesurfer, and 4D filtering followed by 3D Freesurfer. Fig. 8 shows barplots of median values and interquartile ranges of APV for these three segmentations. Visually,

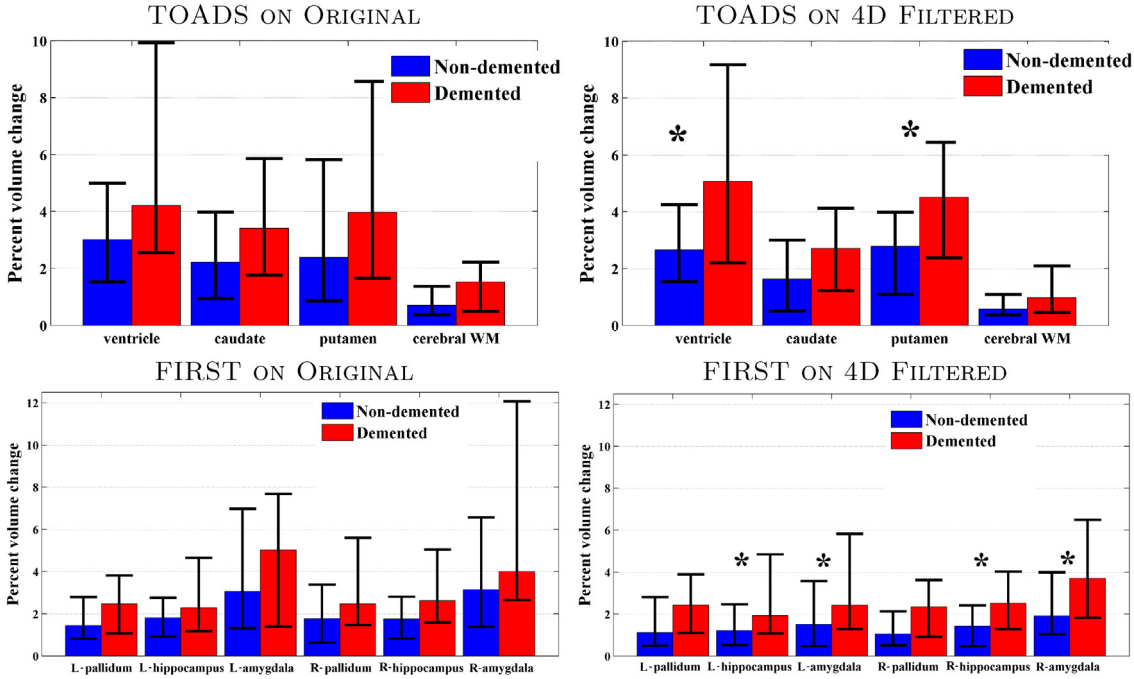

Fig. 6. Barplots show average percent volume change (APV) per year on both demented and non-demented groups from OASIS database, when segmented by TOADS and FIRST. The range corresponds to 25th and 75th percentiles. Asterisk indicates significantly ($p<0.01$) higher APV in demented group compared to the non-demented group. On TOADS segmentations, both ventricles and putamen show significant atrophy in dementia on 4D filtered images compared to un-filtered images. On FIRST segmentations, left and right hippocampus and amygdala show significant atrophy with 4D filtering, which was absent in the original images.

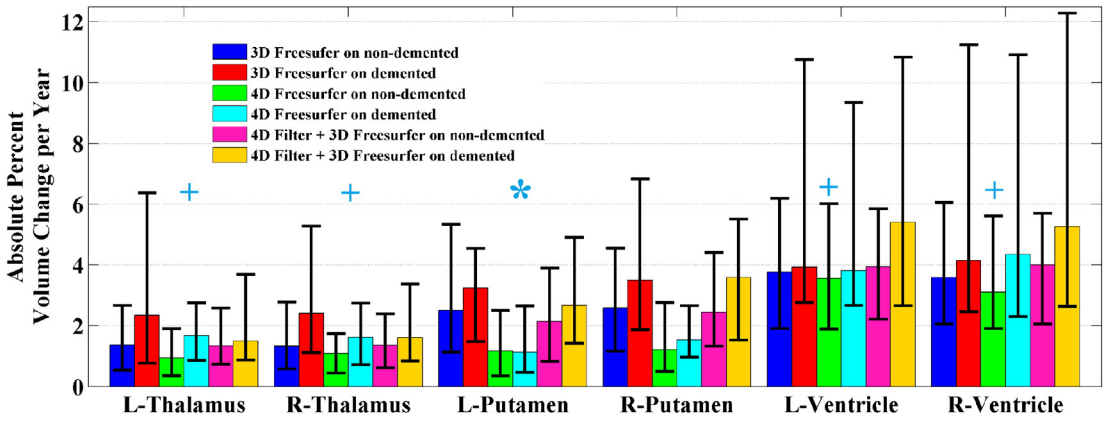

Fig. 7. Barplots show median of absolute percent volume change per year (APV) on left and right subcortical GM structures, when 49 subjects from OASIS dataset are segmented with 3D and 4D Freesurfer, as well as 4D filtering followed by 3D Freesurfer. The range corresponds to 25th and 75th percentiles. Plus indicates significantly ($p<0.05$) higher APV on left/right thalamus and ventricles from demented group observed in both 4D filter and 4D Freesurfer, but not on 3D Freesurfer. Asterisk indicates significant ($p<0.05$) atrophy on left putamen in demented group observed on 4D filtering, but not 3D or 4D Freesurfer.

magnitude and variations of atrophy rates of 3D Freesurfer are generally larger than 4D segmentations (e.g., left/right pallidum and amygdala), indicating longitudinally unstable segmentations. We observed significantly higher atrophy rates in the MCI group than the healthy volunteers for the left/right hippocampus and left/right ventricles ($p<0.05$) on all three methods. However, when the atrophy is more subtle and relatively small, both 3D and 4D Freesurfer failed to detect any significant change. Atrophies in thalamus and caudate are known to be associated with Alzheimer's Disease (Ryan et al., 2013). Significant atrophy in the MCI group was observed on the left/right thalamus ($p<0.01$), left caudate, right putamen, and right amygdala ($p<0.05$), which was not observed on both 4D and 3D Freesurfer segmentations (shown as asterisk in Fig. 8).

We also computed absolute percent change in cortical thickness per year. Accelerated cortical thinning has been observed in normal aging (Fjell et al., 2014), Alzheimer's Disease, as well as other neurodegenerative diseases (Mak et al., 2015). In the BLSA dataset, we observed in Fig. 9 that the MCI group showed a higher rate of cortical thinning than the controls on all three methods. However, the rate is not significant in 3D and 4D Freesurfer ($p>0.10$ in both cases), but it is significant ($p=0.015$) when the images are processed by our approach followed by 3D Freesurfer. Also, the variations in the rates of change of thicknesses are lower (smaller interquartile range) in the 4D filter, indicating a more stable segmentation. Therefore, our method is more sensitive to distinguishing diseased brains from normals than 4D Freesurfer when the number of time-points is large. The difference in performance of 4D Freesurfer can be explained by the absence of any explicit 4D smoothness model in segmentations. Therefore the effect of random temporal noise becomes more evident when the number of time-points becomes larger.

3.5. Atrophy in MS dataset

We have shown on two datasets, OA-49 and BL-39 in Sections 3.3–3.4, that our 4D filtering has superior sensitivity in distinguishing healthy subjects from patients with dementia and MCI, especially when the number of time-points is large. In this section, we explore the volumetric changes of GM in different phenotypes of MS on the MS-59 data (Section 2.1). In the previous experiments, we showed the advantage of the 4D filter with intensity based segmentation methods (FIRST, TOADS, and Freesurfer). Here we use a registration based label fusion algorithm MALP-EM (Ledig et al., 2015) after with and without the filtering as a pre-processing step.

Since the 4D intensity model does not account for lesions, the WM lesions in MPRAGE images are first segmented (Shiee et al., 2009), and then inpainted with WM intensities (Battaglini et al., 2012). The inpainted images are filtered with the 4D filtering, followed by segmentations with a recent multi-atlas label fusion method MALP-EM (multi-atlas label propagation using expectation maximization). MALP-EM produces whole brain labeling by first registering multiple atlases to a subject, then transferring the corresponding atlas labels to subject space, and combining the labelmaps in subject space via expectation maximization (EM). In this section, we show that the 4D filter in conjunction with the label fusion algorithm can detect changes in atrophy between different phenotypes of MS, compared to unfiltered images. The results are validated by corroborating with previous findings.

Increased decline in total brain and GM volume has been observed in MS patients (Battaglini et al., 2009) compared to controls, while atrophies in putamen, thalamus (Eshaghi et al., 2014), cerebellum, and ventricles (Ramasamy et al., 2009) are shown to be associated with the progression of the disease in cross-sectional studies. In different

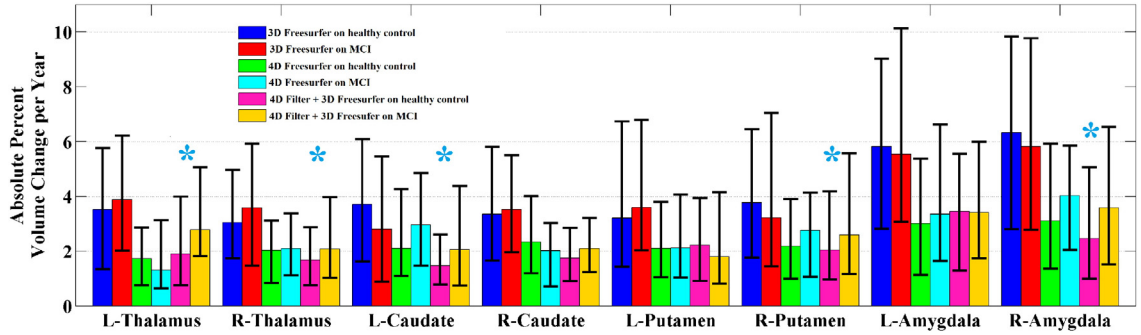

Fig. 8. Barplots show median of absolute percent volume change (APV) on left and right subcortical GM structures, when 39 subjects from BLSA dataset are segmented with 3D and 4D Freesurfer, as well as 3D Freesurfer after our 4D filtering. The range corresponds to 25th and 75th percentiles. Asterisk indicates significantly ($p<0.05$) higher APV on MCI group than controls observed in the combination of 4D filtering and 3D Freesurfer, when the APV is not significant in both 4D and 3D Freesurfer.

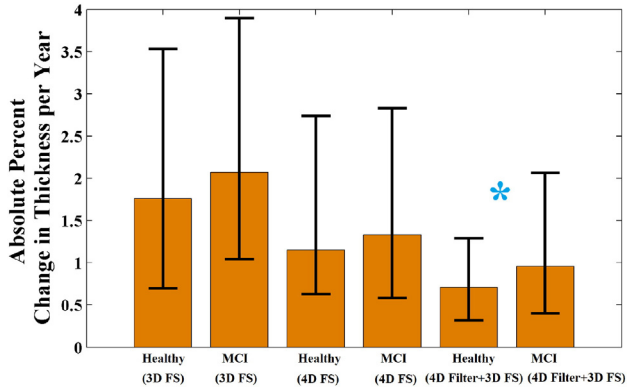

Fig. 9. Barplots show median and interquartile ranges of absolute percent change in cortical thickness, when 39 subjects from BLSA dataset are segmented with 3D and 4D Freesurfer, as well as 4D filtering followed by 3D Freesurfer. Asterisk indicates significantly ($p=0.015$) higher thinning in cortex in MCI group ($n=15$) compared to controls ($n=24$) observed in the combination of 4D filtering and 3D Freesurfer, while it is not significant in both 4D and 3D Freesurfer.

types of MS, rates of atrophy in different tissues are dominant in nature. E.g., ventricles, putamen, hippcampus show significant atrophy in RRMS, while caudate nucleus, along with regions of frontal, patietal, temporal and occipital cortex, show significant atrophies in SPMS and PPMS (Pagani et al., 2005). In a cross-sectional study, lower GM volume was observed in RRMS compared to SPMS (Roosendaal et al., 2011). It was also observed that total brain volume change is higher in SPMS compared to both RRMS and PPMS (Stefano et al., 2010). Specifically, brain volume change was shown to be a powerful statistic that can predict ($p<0.001$) relapses in RRMS (Horakova et al., 2009). Similarly, significantly higher volume loss was observed in the thalamus, left superior temporal gyrus, and left superior frontal sulcus on SPMS (Riccitelli et al., 2010) compared to PPMS.

Fig. 10(a)–(b) show barplots of median (and interquartile ranges) absolute percent volume changes per year (APV) on the 59 subjects with MS, where the segmentations are obtained from MALP-EM after 4D filtering. Significant atrophy is observed in left/right superior frontal ($p<0.001$), left superior occipital, left superior parietal, and right superior temporal ($p<0.05$) gyri in both SPMS and PPMS compared to RRMS (cf. Pagani et al., 2005) after filtering (Fig. 10(b)). Significant atrophy in SPMS was also observed in left/right thalamus ($p=0.01$) and left superior temporal gyrus ($p=0.05$) compared to PPMS (cf. (Riccitelli et al., 2010)). Higher rate of change was observed for total brain volume in SPMS compared to RRMS ($p=0.068$) (cf. Stefano et al., 2010), although it is similar to PPMS ($p=0.45$). Only right superior occipital gyrus and left thalamus shows significant atrophy in RRMS compared to SPMS ($p=0.009$ and 0.013 respectively) with raw images (Fig. 10(a)), indicating that 4D filter has more power in distinguishing between different types of MS. Note that although right superior occipital gyrus is significant in unfiltered images with 4D FS ($p=0.009$ RRMS vs SPMS), it becomes non-significant in filtered images ($p=0.285$).

We also segmented raw and filtered images with 4D Freesurfer and 3D Freesurfer, respectively. Significant atrophy was observed in the combination of 4D filter and 3D Freesurfer segmentations in left and right thalamus ($p<0.05$) in RRMS compared to both SPMS and PPMS, corroborating with the fact that the thalamus has been shown to have significant neurodegeneration in RRMS (Bergsland et al., 2012). Additionally, left ventricles in RRMS ($p=0.03$) have higher atrophy rates than PPMS (cf. Pagani et al., 2005), while left hippocampus has higher atrophy rates ($p=0.02$) in PPMS than RRMS. Total brain volumes from 3D Freesurfer does not show any significant difference between any groups ($p>0.10$). The only significant difference on 4D Freesurfer segmentations of raw images were observed on right thalamus, which was more atrophied in PPMS compared to both RRMS and SPMS.

4. Discussion and conclusions

Our method employs an auto-regressive (AR(1)) filter on the intensities of images collected longitudinally. Higher order autoregressive models can certainly be included in the framework, where intensity at the tth time-point can depend on all the previous time-points, or in a simpler case, $(t-1)$th through $(t-N)$th time-points. Eq. (6) can easily be modified to account for these higher order models. However, we use a simple

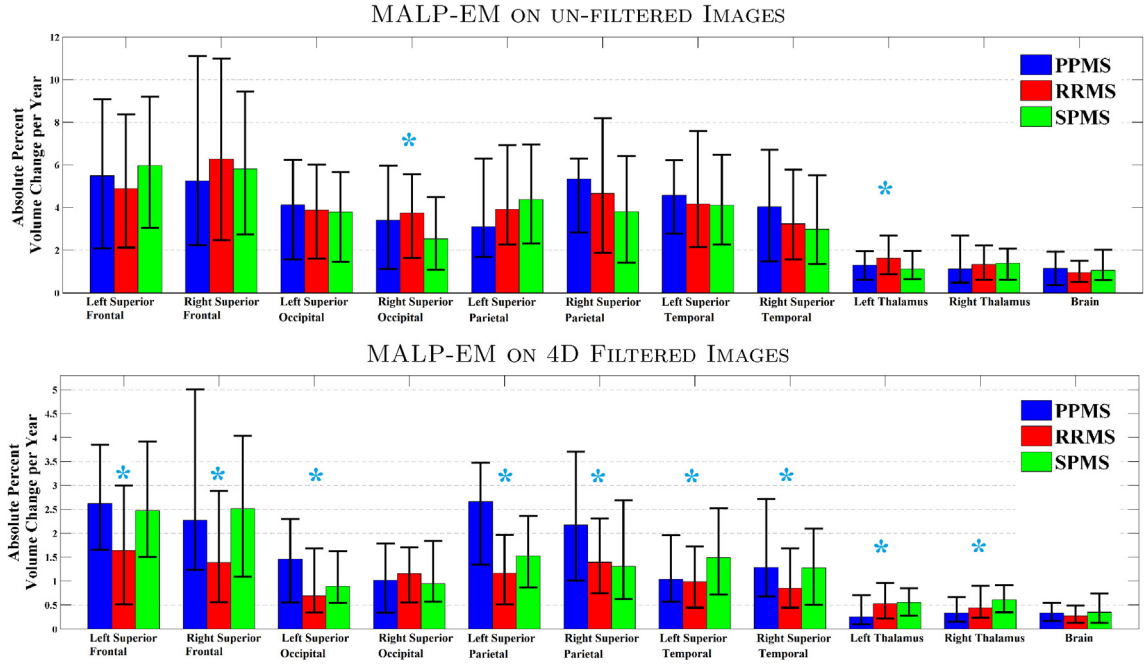

Fig. 10. Barplots show median (and interquartile range) of absolute percent volume change per year (APV) on 59 subjects from MS dataset, with label-fusion segmentations on (a) unfiltered images, and (b) 4D filtered images. Blue asterisk indicate structures where significant differences ($p<0.05$) are observed between at least one of the MS phenotypes. Brain indicates total brain volume. PPMS, RRMS, and SPMS stands for primary progressive, relapsing remitting, and secondary progressive MS. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

model to demonstrate that the 4D filtering can be used as a pre-processing step to any 3D segmentation algorithm, while being similar or superior to alternative approaches. We have shown that our 4D filter can detect tissue atrophy in dementia or MCI, as well as different types of MS.

The autoregressive model promotes the temporal intensity trend of a patch to be smooth. Other approaches have imposed a temporal smoothness penalty on segmentations via 4D Markov random field priors. In contrast, we enforce smoothness on the intensity values, so that the algorithm is not tied to any particular segmentation method. This can be a valuable advantage in many circumstances where the properties of a particular segmentation algorithm are desired. For example, when combined with the TOADS algorithm as shown in Section 3.3, our filtering can lead to stable topology-preserved longitudinal segmentations. However, like other 4D segmentations, a limitation of our method is that if a new scan is observed, the filtered images at all time-points must be re-computed.

The temporal auto-regressive filter in Eq. (4) has two components, one to enforce smooth temporal intensities (Eq. (3)), and a weight to penalize smoothness if the variation is greater than a noise threshold ($w_i^{(t)}$ in Eq. (5)). Other smoothing filters, e.g., a moving average filter, can be used in place of the AR model, but a key component of our work is the weight $w_i^{(t)}$ which ensures that the patches are smoothed only when the variation in the temporal dimension is small enough. The filtering depends on a suitable choice of the threshold f. In theory, f can vary spatially and also depend on the expected rate of change in the underlying anatomy. In future work, we will explore the effect of a temporally and spatially varying f.

In addition to the temporal stability of the brain segmentation, bias in the longitudinal analysis is a common problem, that can arise from the asymmetric interpolations when the baseline (first time point) is used as a target for 4D nonlinear registrations (Yushkevich et al., 2010) in deformation based morphometry. To account for the bias in registrations, two (or more) time-points of a subject can be transformed to a "halfway" space to remove the directional bias, as done in SIENA (Smith et al., 2001). 4D Freesurfer creates an unbiased "median" template of all the time-points and registers all the time-points to the template. In this work, we simply used the MNI atlas as the target for rigid registrations of all the time-points before applying the 4D filtering (as described in Section 2.3), since the focus of this paper is on the robustness of the segmentations. The rigid registration to the MNI atlas before filtering can be replaced by a rigid registration to an unbiased median template, generated by ANTS (Avants et al., 2008) or 4D Freesurfer.

4D FS has no explicit noise model; it only transforms all the time-points into a common unbiased space. Therefore, we would expect a very similar atrophy detection performance of 4D filter followed by 4D FS as the combination of 4D filter and 3D FS, since the noise has already been reduced by the filter. For the test–retest study (TR-10 and CoRR-29), we expect that combination of 4D FS and 4D filter to increase the robustness, i.e. decrease the coefficients of variations, similar to the comparison between 3D vs 4D FS.

In summary, we have described a temporal filtering algorithm to obtain stable longitudinal segmentations that can be used as a pre-processing step to any segmentation method. We used both intensity based (FIRST, TOADS, Freesurfer), as well as registration based (MALP-EM) segmentation algorithms to show that our method is significantly more stable than other approaches while remaining sensitive to actual longitudinal changes.

Acknowledgment

Support for this work included funding from the Department of Defense in the Center for Neuroscience and Regenerative Medicine. This work is supported in part by the Intramural Research Program of the NIA/NIH and the grants NIH/NINDS R01NS070906, NIH/NIBIB R21EB012765. We are grateful to the BLSA participants and neuroimaging staff for their dedication to these studies.

References

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.

Avants, B.B., Tustison, N.J., Wu, J., Cook, P.A., Gee, J.C., 2011. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9, 381–400.

Battaglini, M., Giorgio, A., Stromillo, M.L., Bartolozzi, M.L., Guidi, L., Federico, A., Stefano, N.D., 2009. Voxel-wise assessment of progression of regional brain atrophy in relapsing–remitting multiple sclerosis. J. Neurol. Sci. 282, 55–60.

Battaglini, M., Jenkinson, M., Stefano, N.D., 2012. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Hum. Brain Mapp. 33, 2062–2071.

Bazin, P.L., Pham, D.L., 2008. Homeomorphic brain image segmentation with topological and statistical atlases. Med. Image Anal. 12, 616–625.

Bergsland, N., Horakova, D., Dwyer, M.G., Dolezal, O., Seidl, Z.K., Vaneckova, M., Krasensky, J., Havrdova, E., Zivadinov, R., 2012. Subcortical and cortical gray matter atrophy in a large sample of patients with clinically isolated syndrome and early relapsing–remitting multiple sclerosis. AJNR Am. J. Neuroradiol. 33, 1573–1578.

Bezdek, J.C., 1980. A convergence theorem for the fuzzy ISO-DATA clustering algorithms. IEEE Trans. Pattern Anal. Machine Intell. 20, 1–8.

Carass, A., Cuzzocreo, J., Wheeler, M.B., Bazin, P.L., Resnick, S.M., Prince, J.L., 2011. Simple paradigm for extra-cerebral tissue removal: algorithm and analysis. NeuroImage 56, 1982–1992.

Dai, Y., Wang, Y., Wang, L., Wu, G., Shi, F., Shen, D., Alzheimer's Disease Neuroimaging Initiative, 2012. aBEAT: a toolbox for consistent analysis of longitudinal adult brain MRI. PLoS One 8, 60344.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis I: segmentation and surface reconstruction. NeuroImage 9, 179–194.

den Heijer, T., van der Lijn, F., Koudstaal, P.J., Hofman, A., van der Lugt, A., Krestin, G.P., Niessen, W.J., Breteler, M.M.B., 2010. A 10-year follow-up of hippocampal volume on magnetic resonance imaging in early dementia and cognitive decline. Brain 133, 1163–1172.

Dice, L.R., 1945. Measure of the amount of ecological association between species. Ecology 26, 297–302.

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, 1906–1913.

Eshaghi, A., Bodini, B., Ridgway, G.R., Garcia-Lorenzo, D., Tozer, D.J., Sahraian, M.A., Thompson, A.J., Ciccarelli, O., 2014. Temporal and spatial evolution of grey matter atrophy in primary progressive multiple sclerosis. NeuroImage 86, 257–264.

Fisher, E., Lee, J.C., Rudick, K.N.R.A., 2008. Gray matter atrophy in multiple sclerosis: a longitudinal study. Ann. Neurol. 64, 255–265.

Fjell, A.M., Westlye, L.T., Grydeland, H., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Dale, A.M., Walhovd, K.B., the Alzheimer Disease Neuroimaging Initiative, 2014. Accelerating cortical thinning: unique to dementia or universal in aging? Cereb. Cortex 24, 919–934.

Horakova, D., Dwyer, M.G., Havrdova, E., Cox, J.L., Dolezal, O., Bergsland, N., Rimes, B., Seidl, Z., Vaneckova, M., Zivadinov, R., 2009. Gray matter atrophy and disability progression in patients with early relapsing–remitting multiple sclerosis a 5-year longitudinal study. J. Neurol. Sci. 282, 112–119.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., Pacheco, J., Albert, M., Killiany, R., Blacker, D., Maguire, P., Rosas, D., Makris, N., Gollub, R., Dale, A., Dickerson, B.C., Fischld, B., 2009. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. NeuroImage 46, 177–192.

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartres-Faz, D., Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F., Hensch, T., Trankner, A., Schonknecht, P., Leroy, M., Lopes, R., Bordet, R., Chanoine, V., Ranjeva, J.P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargalloo, N., Blin, O., Frisoni, G.B., The PharmaCog Consortium, 2013. Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. NeuroImage 83, 472–484.

Karacali, B., Davatzikos, C., 2006. Simulation of tissue atrophy using a topology preserving transformation model. IEEE Trans. Med. Imaging 25, 649–652.

Kim, S.H., Fonov, V.S., Dietrich, C., Vachet, C., Hazlett, H.C., Smith, R.G., Graves, M.M., Piven, J., Gilmore, J.H., Dager, S.R., McKinstry, R.C., Paterson, S., Evans, A.C., Collins, D.L., Gerig, G., Styner, M.A., 2013. Adaptive prior probability and spatial temporal intensity change estimation for segmentation of the one-year-old human brain. J. Neurosci. Methods 212, 43–55.

Ledig, C., Heckemann, R.A., Hammers, A., Lopez, J.C., Newcombe, V.F.J., Makropoulos, A., Lotjonen, J., Menon, D.K., Rueckert, D., 2015. Robust whole-brain segmentation: application to traumatic brain injury. Med. Image Anal. 21, 40–58.

Lehmann, M., Douiri, A., Kim, L.G., Modat, M., Chan, D., Ourselin, S., Barnes, J., Fox, N.C., 2010. Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements. NeuroImage 49, 2264–2274.

Mak, E., Su, L., Williams, G.B., Watson, R., Firbank, M.J., Blamire, A.M., O'Brien, J.T., 2015. Progressive cortical thinning and subcortical atrophy in dementia with Lewy bodies and Alzheimer's disease. Neurobiol. Aging 36, 1743–1750.

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.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probablistic atlas of the human brain: theory and rationale for its development. NeuroImage 2, 89–101.
Miller, M., Younes, L., 2001. Group actions, homeomorphisms, and matching: a general framework. Int. J. Comput. Vis. 41, 61–84.
Moller, C., Dieleman, N., van der Flier, W.M., Versteeg, A., Pijnenburg, Y., Scheltens, P., Barkhof, F., Vrenken, H., 2015. More atrophy of deep gray matter structures in frontotemporal dementia compared to Alzheimer's disease. J. Alzheimers Dis. 44, 635–647.
Nakamura, K., Brown, R.A., Narayanan, S., Collins, D.L., Arnold, D.L., The Alzheimer's Disease Neuroimaging Initiative, 2015. Diurnal fluctuations in brain volume: statistical analyses of MRI from large populations. NeuroImage (0, 0–0).
Pagani, E., Rocca, M.A., Gallo, A., Rovaris, M., Martinelli, V., Comi, G., Filippi, M., 2005. Regional brain atrophy evolves differently in patients with multiple sclerosis according to clinical phenotype. AJNR Am. J. Neuroradiol. 26, 341–346.
Patenaude, B., Smith, S.M., Kennedy, D., Jenkinson, M., 2011. A bayesian model of shape and appearance for subcortical brain. NeuroImage 56, 907–922.
Pham, D.L., 2001. Spatial models for fuzzy clustering. Comp. Vision Image Underst. 84, 285–297.
Pham, D.L., Prince, J.L., 1999. Adaptive fuzzy segmentation of magnetic resonance images. IEEE Trans. Med. Imaging 18, 737–752.
Prastawa, M., Awate, S.P., Gerig, G., 2012. Building spatiotemporal anatomical models using joint 4-D segmentation, registration, and subject-specific atlas estimation. Mathematical Methods in Biomedical Image Analysis (MMBIA), 2012 IEEE Workshop on, pp. 49–56.
Querbes, O., Aubry, F., Pariente, J., Lotterie, J., Démonet, J., Duret, V., Puel, M., Berry, I., Fort, J.C., Celsis, P., The Alzheimer's Disease Neuroimaging Initiative, 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132, 2036–2047.
Ramasamy, D.P., Benedict, R.H.B., Cox, J.L., Fritz, D., Abdelrahman, N., Hussein, S., Minagar, A., Dwyer, M.G., Zivadinov, R., 2009. Extent of cerebellum, subcortical and cortical atrophy in patients with MS: a case–control study. J. Neurol. Neurosurg. Psychiatry 282, 47–54.
Resnick, S.M., Goldszal, A.F., Davatzikos, C., Golski, S., Kraut, M.A., Metter, E.J., Bryan, R.N., Zonderman, A.B., 2000. One-year age changes in MRI brain volumes in older adults. Cereb. Cortex 10, 464–472.
Resnick, S.M., Pham, D.L., Kraut, M.A., Zonderman, A.B., Davatzikos, C., 2003. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. J. Neurosci. 23, 3295–3301.
Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61, 1402–1418.
Riccitelli, G., Rocca, M.A., Pagani, E., Rodegher, M.E., Rossi, P., Falini, A., Comi, G., Filippi, M., 2010. Cognitive impairment in multiple sclerosis is associated to different patterns of gray matter atrophy according to clinical phenotype. Hum. Brain Mapp. 32, 1535–1543.
Roosendaal, S.D., Bendfeldt, K., Vrenken, H., Polman, C.H., Borgwardt, S., Radue, E.W., Kappos, L., Pelletier, D., Hauser, S.L., Matthews, P.M., Barkhof, F., Geurts, J.J.G., 2011. Grey matter volume in a large cohort of ms patients: relation to MRI parameters and disability. Mult. Scler. 17, 1098–1106.
Roy, S., Carass, A., Prince, J.L., 2011. A compressed sensing approach for MR tissue contrast synthesis. Inf. Proc. in Med. Imaging (IPMI), pp. 371–383.
Roy, S., Carass, A., Prince, J., 2013a. Magnetic resonance image example based contrast synthesis. IEEE Trans. Med. Imaging 32, 2348–2363.
Roy, S., Carass, A., Prince, J.L., 2013b. Longitudinal intensity normalization of magnetic resonance images using patches. Proceedings of SPIE Medical Imaging (SPIE), p. 86691J.
Roy, S., Carass, A., Shiee, N., Pham, D.L., Calabresi, P., Reich, D., Prince, J.L., 2013c. Longitudinal intensity normalization in the presence of multiple sclerosis lesions. Intl. Symp. on Biomed. Imag. (ISBI), pp. 1384–1387.
Roy, S., Jog, A., Carass, A., Prince, J.L., 2013d. Atlas based intensity transformation of brain MR images. Multimodal Brain Image Analysis, pp. 51–62.
Roy, S., He, Q., Carass, A., Jog, A., Cuzzocreo, J.L., Reich, D.S., Prince, J.L., Pham, D.L., 2014. Example based lesion segmentation. Proc. SPIE Med. Imaging (SPIE) 9034, 90341Y.
Ryan, N.S., Keihaninejad, S., Shakespeare, T.J., Lehmann, M., Crutch, S.J., Malone, I.B., Thornton, J.S., Mancini, L., Hyare, H., Yousry, T., Ridgway, G.R., Zhang, H., Modat, M., Alexander, D.C., Rossor, M.N., Ourselin, S., Fox, N.C., 2013. Magnetic resonance imaging evidence for presymptomatic change in thalamus and caudate in familial Alzheimer's disease. Brain 136, 1399–1414.
Shen, D., Davatzikos, C., 2004. Measuring temporal morphological changes robustly in brain MR images via 4-D template warping. NeuroImage 21, 1508–1517.
Shiee, N., Bazin, P.L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2009. A Topology-Preserving Approach to the Segmentation of Brain Images With Multiple Sclerosis Lesions NeuroImage.
Shiee, N., Bazin, P.L., Zackowski, K.M., Farrell, S.K., Harrison, D.M., Newsome, S.D., Ratchford, J.N., Caffo, B.S., Calabresi, P.A., Pham, D.L., Reich, D.S., 2012. Revisiting brain atrophy and its relationship to disability in multiple sclerosis. PLoS One 7, e37049.
Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Mateen, F.J., Calabresi, P.A., Jarso, S., Pham, D.L., Reich, D.S., Crainiceanu, C.M., the Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing, the Alzheimer's Disease Neuroimaging Initiative, 2014. Statistical normalization techniques for magnetic resonance imaging. NeuroImage 6, 9–19.
Sijbers, J., Dekker, A.J., Scheunders, P., Van Dyck, D., 1998. Maximum-likelihood estimation of rician distribution parameters. IEEE Trans. Med. Imaging 17, 357–361.
Smith, S.M., De-Stefano, N., Jenkinson, M., Matthews, P.M., 2001. Normalized accurate measurement of longitudinal brain change. J. Comput. Assit. Tomogr. 25, 466–475.
Stefano, N.D., Giorgio, A., Battaglini, M., Rovaris, M., Sormani, M.P., Barkhof, F., Korteweg, T., Enzinger, C., Fazekas, F., Calabrese, M., Dinacci, D., Tedeschi, G., Gass, A., Montalban, X., Rovira, A., Thompson, A., Comi, G., Miller, D., Filippi, M., 2010. Assessing brain atrophy rates in a large population of untreated multiple sclerosis subtypes. Neurology 74, 1868–1876.
Thambisetty, M., Wan, J., Carass, A., An, Y., Prince, J.L., Resnick, S.M., 2010. Longitudinal changes in cortical thickness associated with normal aging. NeuroImage 52, 1215–1223.
Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320.
Xue, Z., Shen, D., Davatzikos, C., 2006. CLASSIC: consistent longitudinal alignment and segmentation for serial image computing. NeuroImage 30, 388–399.
Yushkevich, P.A., Avants, B.B., Das, S.R., Pluta, J., Altinay, M., Craige, C., the Alzheimer's Disease Neuroimaging Initiative, 2010. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: An illustration in ADNI 3 tesla MRI Data. NeuroImage 50, 434–445.
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. Imaging 20, 45–57.
Zuo, X.N., Anderson, J.S., Bellec, P., Birn, R.M., Biswal, B.B., Blautzik, J., Breitner, J.C.S., Buckner, R.L., Calhoun, V.D., Castellanos, F.X., Chen, A., Chen, B., Chen, J., Chen, X., Colcombe, S.J., Courtney, W., Craddock, R.C., Martino, A.D., Dong, H.M., Fu, X., Gong, Q., Gorgolewski, K., Han, Y., He, Y., He, Y., Ho, E., Holmes, A., Hou, X.H., Huckins, J., Jiang, T., Jiang, Y., Kelley, W., Kelly, C., King, M., LaConte, S.M., Lainhart, J.E., Lei, X., Li, H.J., Li, K., Li, K., Lin, Q., Liu, D., Liu, J., Liu, X., Liu, Y., Lu, G., Lu, J., Luna, B., Luo, J., Lurie, D., Mao, Y., Margulies, D.S., Mayer, A.R., Meindl, T., Meyerand, M.E., Nan, W., Nielsen, J.A., OConnor, D., Paulsen, D., Prabhakaran, V., Qi, Z., Qiu, J., Shao, C., Shehzad, Z., Tang, W., Villringer, A., Wang, H., Wang, K., Wei, D., Wei, G.X., Weng, X.C., Wu, X., Xu, T., Yang, N., Yang, Z., Zang, Y.F., Zhang, L., Zhang, Q., Zhang, Z., Zhang, Z., Zhao, K., Zhen, Z., Zhou, Y., Zhu, X.T., Milham, M.P., 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data 9, 140049.