NeuroImage 48 (2009) 73–83

Contents lists available at ScienceDirect

NeuroImage

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

Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3

Weili Zheng [a], Michael W.L. Chee [b], Vitali Zagorodnov [a,*]

[a] School of Computer Engineering, Nanyang Technological University, Singapore
[b] Cognitive Neuroscience Laboratory, Duke-NUS Graduate Medical School, Singapore

ARTICLE INFO

Article history:
Received 18 December 2008
Revised 2 June 2009
Accepted 17 June 2009
Available online 25 June 2009

Keywords:
Non-uniformity correction
Cortex thickness estimation

ABSTRACT

Smoothly varying and multiplicative intensity variations within MR images that are artifactual, can reduce the accuracy of automated brain segmentation. Fortunately, these can be corrected. Among existing correction approaches, the nonparametric non-uniformity intensity normalization method N3 (Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. Nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17, 87–97.) is one of the most frequently used. However, at least one recent study (Boyes, R.G., Gunter, J.L., Frost, C., Janke, A.L., Yeatman, T., Hill, D.L.G., Bernstein, M.A., Thompson, P.M., Weiner, M.W., Schuff, N., Alexander, G.E., Killiany, R.J., DeCarli, C., Jack, C.R., Fox, N.C., 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. NeuroImage 39, 1752–1762.) suggests that its performance on 3 T scanners with multichannel phased-array receiver coils can be improved by optimizing a parameter that controls the smoothness of the estimated bias field. The present study not only confirms this finding, but additionally demonstrates the benefit of reducing the relevant parameter values to 30–50 mm (default value is 200 mm), on white matter surface estimation as well as the measurement of cortical and subcortical structures using FreeSurfer (Martinos Imaging Centre, Boston, MA). This finding can help enhance precision in studies where estimation of cerebral cortex thickness is critical for making inferences.

Introduction

Intensity non-uniformity artifact in images obtained from high field MR systems refers to the presence of smoothly varying and multiplicative intensity variations within the tissues, also referred to as bias field. The presence of this artifact may adversely affect qualitative and quantitative image analysis. Common causes include static field inhomogeneity B_0, eddy currents driven by the switching of field gradients, non-uniform sensitivity in the surface coil and specific permeability and dielectric properties of the imaged object (Vovk et al., 2007). The latter effect is particularly prominent at higher magnetic fields (Belaroussi et al., 2006), and is often observed in the form of "central brightening artifact" in head imaging. Despite measures to prospectively correct such artifacts using phantom or shimming techniques, and/or in hardware, using a multichannel phased-array receiver coil (Bernstein et al., 2006), additional retrospective correction is often necessary. Detailed reviews of recent developments in retrospective non-uniformity correction can be found in (Belaroussi et al., 2006; Hou, 2006; Styner and Leemput, 2005; Vovk et al., 2007).

Among existing approaches, the nonparametric non-uniformity intensity normalization method N3 (Sled et al., 1998) is one of the most frequently used. This approach iterates between three main steps, histogram sharpening, bias field estimation (based on the obtained sharpened histogram) and B-spline smoothing. High performance and robustness (Arnold et al., 2001) have practically turned N3 into an industry standard (Arnold et al., 2001; Gispert et al., 2004; Hou et al., 2006; Likar et al., 2001; Luo et al., 2005; Shattuck et al., 2001; Sled et al., 1997; Vovk et al., 2006). For example, it has been incorporated into FreeSurfer,[1] a widely used tool for estimating cortical thickness.

N3 contains several adjustable parameters, such as smoothing distance (the distance between B-spline nodes), deconvolution kernel size, stopping criteria for iterations, maximum number of iterations and image down-sampling ratio (Sled et al., 1998). However, in most publications (Arnold et al., 2001; Gispert et al., 2004; Hou et al., 2006; Likar et al., 2001; Luo et al., 2005; Shattuck et al., 2001; Vovk et al., 2006) default values reported in the original publication (Sled et al., 1998) were used. We contend that these parameters, intended for use on images acquired on 1.5 T scanners built over a decade ago need to be revised for data acquired on modern higher field research magnets.

Modern research MR scanners typically have a field strength of 3 T or higher and employ multichannel phased array coils. The images

* Corresponding author. Fax: +65 67926559.
E-mail address: zvitali@ntu.edu.sg (V. Zagorodnov).

[1] http://surfer.nmr.mgh.harvard.edu/.

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.06.039

they generate are associated with somewhat different non-uniformity profiles. For example, "center brightening artifact" is more prominent at 3 T (~30%) than at 1.5 T (~5%) (Bernstein et al., 2006). As such, default N3 parameters developed for legacy systems may not be appropriate for modern scanners. A recent study (Boyes et al., 2008) has suggested that reducing one of the N3's parameters, the smoothing distance, from the default 200 mm to 50–100 mm can lead to substantial improvement in the correction performance.

The current study had two goals. First, we wanted to replicate Boyes et al's findings regarding the effect of smoothing distance on the quality of intensity correction. Secondly, we wanted to determine the effect that changing smoothing distance has at different points along FreeSurfer segmentation pipeline, see (Dale et al., 1999; Fischl et al., 2002, 2004b, 1999) for detailed descriptions. This decision was motivated by a wide use of this software; as of March 2009 there were 58 published papers using FreeSurfer in various neurological studies.[2]

Critically, our goal of optimizing intensity correction on the basis of segmentation performance may not necessarily result in more reliable bias field estimates. In particular, by making the tissue intensity more uniform in the effort to improve segmentation, interesting differences in WM intensity caused by biological factors may be inadvertently altered (van Walderveen et al., 2003).

Our findings suggest that using N3 with smaller smoothing distances engenders better intensity non-uniformity correction, which, at least in the context of FreeSurfer segmentation pipeline, translates into more accurate segmentation of white matter (WM) surface and improved reliability of within and between scanner measurements of cerebral cortex thickness and the volume of selective subcortical structures.

Materials and methods

Subjects and data acquisition

MRI was performed on 3 T Siemens Allegra and Tim Trio (Siemens, Erlangen, Germany) systems using a standardized imaging procedure that incorporated a number of quality control measures. The T1-weighted MP-RAGE sequence used was modeled after that used by the Alzheimer's Disease Neuroimaging Initiative (ADNI) consortium (TR = 2300 ms (Allegra), 2530 (Trio); TE = 2.91 ms (Allegra); 1.64, 3.5, 5.36 and 7.22 ms (Trio, multi-echo acquisition); TI = 900 ms (Allegra), 1200 ms (Trio); flip angle = 9° (Allegra), 7° (Trio); Bandwidth 240 Hz/pixel (Allegra), 651 Hz/pixel (Trio); FOV 256 × 240 mm, 256 × 256 matrix; resulting voxel dimensions: 1.0 × 1.0 × 1.1 mm^3 (Allegra), 1.0 × 1.0 × 1.0 mm^3 (Trio). Acquisition time was 9 min 14 s (Allegra); 6:03 (Trio). Prescan normalization was used on both scanners. Parallel acquisition with a 4 channel phase array coil was not accelerated on the Allegra, whereas GRAPPA with an acceleration factor of 2 was used on the 12 channel phase array coil of the Tim Trio scanner. Both coils were stock Siemens coils without user modification.

Data Set 1: 15 healthy subjects (age 56–71, 9 males), each scanned once on a Siemens Allegra 3 T scanner.

Data Set 2: 15 healthy subjects (age 56–71, 4 males), each scanned once on a Siemens Allegra 3 T scanner. The main difference between this data set and data set 1 was the image quality. Data set 1 required little or no editing, while the current data set gave rise to occasional pial surface overgrowth due to inclusion of dura within the brain mask and an underestimation of WM surface, and consequently necessitated substantial manual editing.

Data Set 3: 24 healthy subjects (age 54–68, 8 males), each scanned once on a Siemens Tim Trio 3 T scanner.

Data Set 4: 8 healthy subjects (4 young subjects, age 21–26, 3 males, and 4 elderly, age 62–73, 3 males), scanned within a short time interval on Siemens Allegra 3 T and Siemens Tim Trio 3 T scanners. The four young subjects were also scanned twice on each scanner (Table 1). Biological change in the serial scans was assumed to be negligible.

Data processing

Data sets 1–3 were first processed by FreeSurfer (FS) segmentation pipeline (version 3.0.4) using default parameters and the resultant GM/WM output, edited by an expert, was used as ground truth. The editing affected only those areas which were classified as wrongly segmented by an expert and was done on original images that were not intensity corrected. After editing, FreeSurfer pipeline was rerun to propagate the changes. For example, in the case of WM underestimation, only the WM surface was edited while the corresponding change in the position of the pial surface was derived after rerunning the pipeline.

All data sets were then processed using FreeSurfer several times, once for each smoothing distance tested. Similar to (Boyes et al., 2008), we increased the maximum number of iterations from default 50 to 1000 in order to cope with the smaller smoothing distances, which lead to longer convergence times and thus require more iterations. Parameter modifications were achieved by supplying two additional arguments to 'mri_nu_correct.mni' function, namely '–distance' and '–proto-iters', which control smoothing distance and maximum number of iterations respectively. All other parameters were kept at default values.

Performance evaluation measures

Non-uniformity correction measure

As a measure of non-uniformity correction performance we used a commonly accepted coefficient of variation of the white matter CV_{WM} (Vovk et al., 2007), defined as the ratio of the standard deviation σ_{WM} and the mean μ_{WM} of intensities within WM region:

$$CV_{WM} = \frac{\sigma_{WM}}{\mu_{WM}}. \quad (1)$$

The underlying assumption of CV_{WM} is that increasing intensity non-uniformity leads to monotonic increase in standard deviation of image intensity within WM. Hence a larger reduction in CV would correspond to better non-uniformity correction. Partial volume voxels may cause an increase in σ_{WM} not originating from intensity non-uniformity. To avoid this undesirable effect, we excluded the outer layer of voxels from the ground truth of WM mask (Boyes et al., 2008; Sled et al., 1998).

Surface distance measure (segmentation performance)

FreeSurfer segmentation results in the generation of WM/GM and GM/CSF surface meshes, suggesting a performance measure based on surface-to-surface geometric distance, such as Hausdorff distance

Table 1
Acquisition scheme of test–retest data; A1 and A2 represent the data obtained from two scans on Siemens Allegra 3 T scanner, T1 and T2 — on Siemens Tim Trio 3 T scanner.

Scan/Subject	Subj1	Subj2	Subj3	Subj4	Subj5	Subj6	Subj7	Subj8
A1	×	×	×	×	×	×	×	×
A2					×	×	×	×
T1					×	×	×	×
T2	×	×	×	×	×	×	×	×

[2] http://surfer.nmr.mgh.harvard.edu/fswiki.

(Huttenlocher et al., 1993). Given two finite point sets $A=\{a_1,\ldots a_p\}$ and $B=\{b_1,\ldots b_q\}$, the Hausdorff distance is defined as

$$H(A,B) = \max(h(A,B), h(B,A)) \tag{2}$$

where

$$h(A,B) = \max_{a\in A}\min_{b\in B}||a-b|| \tag{3}$$

and $||\cdot||$ designates arbitrary norm (e.g. L_∞ or Euclidean norm).

However, mesh representation poses several difficulties to the interpretation of distance results. First, FreeSurfer's surface meshes are arbitrarily initiated, leading to a possible misalignment between meshes even when the surfaces are identical; this invariably results in non-zero Hausdorff distance (Fig. 1, top). This residual distance can equal half the sampling distance in the worst cases, becoming quite substantial in some sparse mesh locations. For example, FreeSurfer's sampling distance can be as large as 3 mm, which may produce an overestimation of up to 1.5 mm in the distance between surfaces. Another limitation of Hausdorff distance is its inability to distinguish between underestimation and overestimation.

To overcome these problems we modified the definition (2) in the following manner. Given two finite vertex sets, $A=\{a_1,\ldots a_p\}$ from the ground truth surface and $B=\{b_1,\ldots b_q\}$ from a test surface, and sets of surface triangles $S_A=\{s_1,\ldots s_r\}$ and $S_B=\{s_1,\ldots s_t\}$ formed by neighboring elements of A and B, the surface underestimation measure is defined as:

$$H_u = \max(d_-(A,B), d_+(B,A)) \tag{4}$$

and surface overestimation measure is defined as:

$$H_o = \max(d_+(A,B), d_-(B,A)) \tag{5}$$

where

$$d_+(A,B) = \max_{a\in A}\left(\min_{s_b\in S_B,\ \mathrm{sign}(||a-s_b||)=+1}||a-s_b||, 0\right)$$

$$d_-(A,B) = \max_{a\in A}\left(\min_{s_b\in S_B,\ \mathrm{sign}(||a-s_b||)=-1}||a-s_b||, 0\right)$$

Here $||x-s||$ designates the distance between point x and the plane formed by the triangular patch s; $\mathrm{sign}(||x-s||)$ is equal to $+1$ if the angle between the projection vector from x onto the plane and the surface normal vector is less than 90°, where surface normal is defined as facing outside the surface. Fig. 1 (bottom) illustrates the newly introduced underestimation and overestimation surface measures.

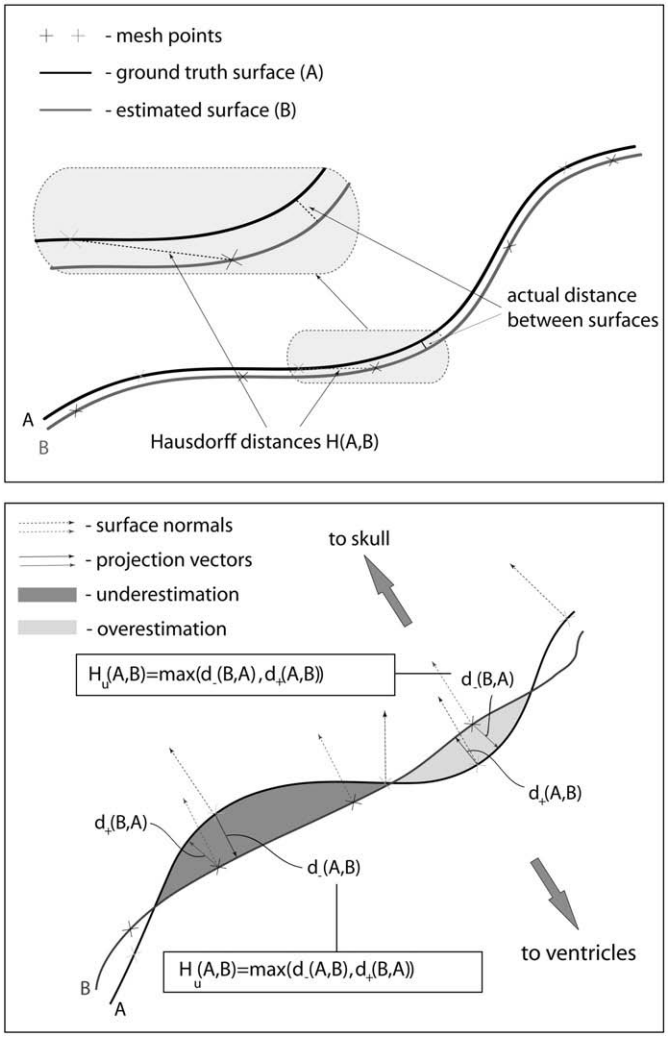

Fig. 1. Due to finite sampling rate, Hausdorff distance overestimates the distance between surfaces (top). Evaluation of surface underestimation H_u and overestimation H_o distances (bottom).

Within scanner and between scanner reliability (segmentation performance)

Within and between scanner reliability was adopted as a second measure of segmentation performance, based on measurements of cerebral cortex thickness and volumes of subcortical structures. Improving this reliability is essential for the detection of subtle brain structure changes over time. Among a variety of reliability measures used in structural MRI studies (Desikan et al., 2006; Feczko et al., 2009; Fischl and Dale, 2000; Fischl et al., 2002; Han et al., 2006; Scott and Thacker, 2005; Smith et al., 2002; van der Kouwe et al., 2008), we chose intraclass correlation (ICC), absolute thickness difference of cortical measurements and percentage volume change (PVC) of subcortical structures.

The ICC (McGraw and Wong, 1996; Shrout and Fleiss, 1979; Weir, 2005) is one of the most commonly used metrics in structural MRI (Desikan et al., 2006; Feczko et al., 2009). We have chosen ICC (2,1) in the nomenclature of Shrout and Fleiss (1979) due to its desirable sensitivity to systemic error. The ICC (2,1) is defined as an estimate of the ratio between the subject variance and the total variance, composed of between subject variance and variances of all other random factors (Shrout and Fleiss, 1979). In the case of within scanner (or test–retest Friedman et al., 2008) reliability ICC (2,1), these random factors include subjects, repeat scans and interaction between the scans and the subjects:

$$\rho_{\text{within\_scanner}} = \frac{\sigma^2_{\text{subj}}}{\sigma^2_{\text{subj}} + \sigma^2_{\text{w}} + \sigma^2_{\text{Iw}} + \sigma^2_{\text{e}}} \tag{6}$$

where σ^2_{subj} is between subject variance, σ^2_{w} is between scan variance, σ^2_{Iw} is the scan-subject interaction and σ^2_{e} is the variance of noise.

In case of between scanner reliability ICC (2,1), the random factors include subjects, scanners and interaction between the scanners and the subjects:

$$\rho_{\text{between\_scanner}} = \frac{\sigma^2_{\text{subj}}}{\sigma^2_{\text{subj}} + \sigma^2_{\text{b}} + \sigma^2_{\text{Ib}} + \sigma^2_{\text{e}}} \tag{7}$$

where σ^2_{subj} and σ^2_{e} are the same as in Eq. (6), σ^2_{b} is between scanner variance, σ^2_{Ib} is the scanner–subject interaction. The variance components in the above definitions were estimated through a

Table 2
Mean CV_{WM} for various N3 smoothing distances and three tested data sets.

Distance (mm)	Data Set 1 mean (std)	Data Set 2 mean (std)	Data Set 3 mean (std)
30	**0.044 (0.005)**	**0.043 (0.007)**	**0.041 (0.005)**
40	0.045 (0.005)	0.044 (0.008)	0.043 (0.004)
50	0.046 (0.005)	0.045 (0.007)	0.045 (0.005)
60	0.049 (0.007)	0.047 (0.008)	0.045 (0.005)
70	0.050 (0.007)	0.048 (0.009)	0.054 (0.006)
80	0.052 (0.008)	0.050 (0.009)	0.059 (0.007)
90	0.053 (0.009)	0.050 (0.008)	0.060 (0.007)
100	0.055 (0.010)	0.051 (0.009)	0.051 (0.008)
110	0.057 (0.011)	0.052 (0.009)	0.053 (0.008)
120	0.057 (0.011)	0.053 (0.009)	0.056 (0.009)
140	0.056 (0.010)	0.052 (0.009)	0.058 (0.009)
150	0.057 (0.011)	0.052 (0.009)	0.059 (0.009)
160	0.057 (0.011)	0.053 (0.009)	0.060 (0.010)
170	0.058 (0.011)	0.053 (0.009)	0.061 (0.010)
180	0.058 (0.011)	0.054 (0.009)	0.062 (0.010)
190	0.058 (0.010)	0.054 (0.010)	0.063 (0.010)
200	0.058 (0.011)	0.054 (0.010)	0.067 (0.009)

Bold values highlight the best achieved values of CV_{WM}.

two-way model ANOVA, see McGraw and Wong (1996), Shrout and Fleiss (1979) and Weir (2005) for more details.

The absolute thickness difference of a cortical region is defined as the absolute value of the difference between two measurements of its thickness (Han et al., 2006). The percentage of volume change (PVC) of a subcortical structure (Fischl et al., 2002) is a normalized version of its absolute volume difference and is defined as:

$$\mathrm{PVC} = \frac{2|V_1 - V_2|}{V_1 + V_2}. \quad (8)$$

Results

Reduced smoothing distances better suited for data from contemporary scanners

In this experiment we evaluated CV_{WM} for various N3 smoothing distance values, ranging in 10 mm increments from 30 to 200 mm. The results (Table 2) showed good agreement with (Boyes et al., 2008) on all three data sets (spanning two scanners), confirming that smaller N3 smoothing distance results in better non-uniformity correction. However, the best smoothing distance was revealed to be 30 mm rather than 50 mm, slightly lower than previously found (Boyes et al., 2008). This is possibly due to the more limited testing range (50–200 mm) used previously. Note that reduction in CV_{WM} at smaller smoothing distance does not imply overfitting to image structures. As shown in Fig. 2 on example of one of the subjects from data set 3, smoothing distance of 30 mm produces distinctly sharper estimate of bias field but the one that does not resemble the anatomy of the brain.

The advantage of using N3 with a smaller smoothing distance of 30 mm compared to the default setting of 200 mm is illustrated in Fig. 3. Assuming that the signal in parietal and temporal white matter should be identical, the use of the smaller smoothing distance narrowed the difference in signal intensity within these two regions by 21 points from 28 to 7, whereas the default smoothing distance achieved only a 12 point reduction.

Testing segmentation performance using a surface distance measure

To test whether improvements in intensity non-uniformity correction (Table 2) would result in better segmentation performance we ran FreeSurfer pipeline on data set 2 with five N3 smoothing distances (30, 50, 100, 150 and 200 mm) and compared the estimated WM surfaces with the ground truth. For more detailed analysis, surface distance measures were applied individually to 34 anatomical regions parcellated and labeled by FreeSurfer (Desikan et al., 2006; Fischl et al., 2004a). This resulted in two sets of 34(regions)×30(hemispheres) measurements for each smoothing distance used, one for assessing underestimation and the other for evaluating overestimation.

The results of underestimation measure H_u at the default smoothing distance of 200 mm are shown in Fig. 4. About 80% of underestimation measures were below 1.5 mm and were likely to result from measurement error rather than from segmentation (see earlier discussion of the surface distance measure limitations). The largest 20% of the underestimated measurements cannot be explained by measurement error and primarily originate from eight regions highlighted in Fig. 4.

For further analysis we divided the set of 34(regions)×30(hemispheres) into two groups. The first group contained data from the eight problematic regions where underestimation exceeded 1.5 mm. The rest of the data involved regions where underestimation was less than 1.5 mm. We hypothesized that smaller smoothing distances would reduce the underestimation in some of the regions in the first group and would have no effect on the second group.

Fig. 5 shows how underestimation was affected by smoothing distance. Using a smoothing distance of 30 mm reduced the average underestimation in the first group from 3.7 mm to 2.5 mm. The per-region *reduction* in underestimation ranged from 0.2 mm (middle temporal) to 1.87 mm (temporal pole). In four regions (entorhinal, fusiform, inferior temporal, and temporal pole) the reduction was statistically significant ($p<0.05$) with both 30 mm and 50 mm smoothing distances. Note that significance was estimated using a paired t-test and resulting p-values were Bonferroni corrected for multiple comparisons. Another three regions (lateral occipital, lateral orbito frontal, and superior temporal) showed large but not significant reductions in underestimation, e.g. 1.68 mm in lateral occipital region. The failure for these effects to reach statistical significance can be attributed to small sample size (only 7 hemispheres for lateral occipital region). In the last region (middle temporal) the underestimation

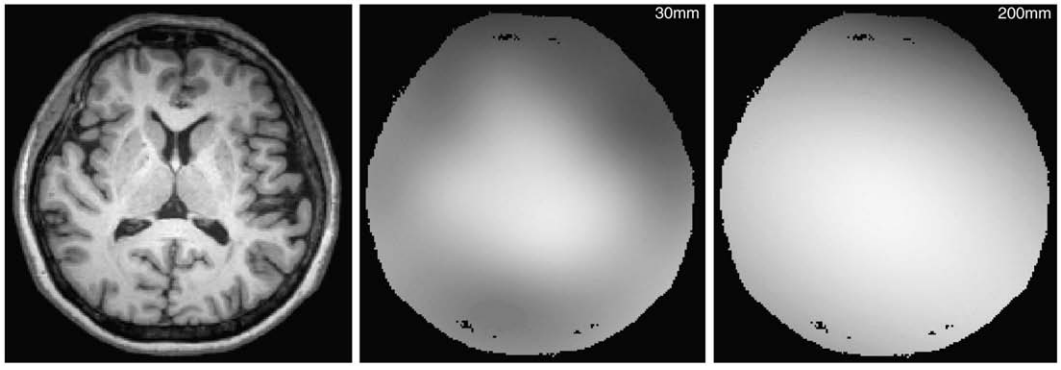

Fig. 2. The shape of estimated bias field for two values of N3 smoothing distance, 30 mm (middle) and 200 mm (right).

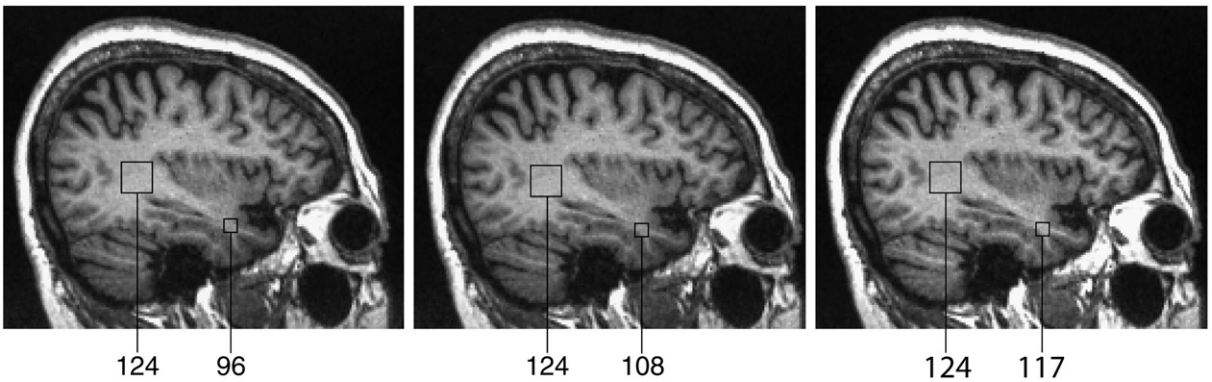

Fig. 3. Original image (left), image corrected using N3 with smoothing distance $d = 200$ mm (middle) and $d = 30$ mm (right). Average intensity values of selected ROIs are indicated below. White matter intensities were normalized between three images using the left ROI.

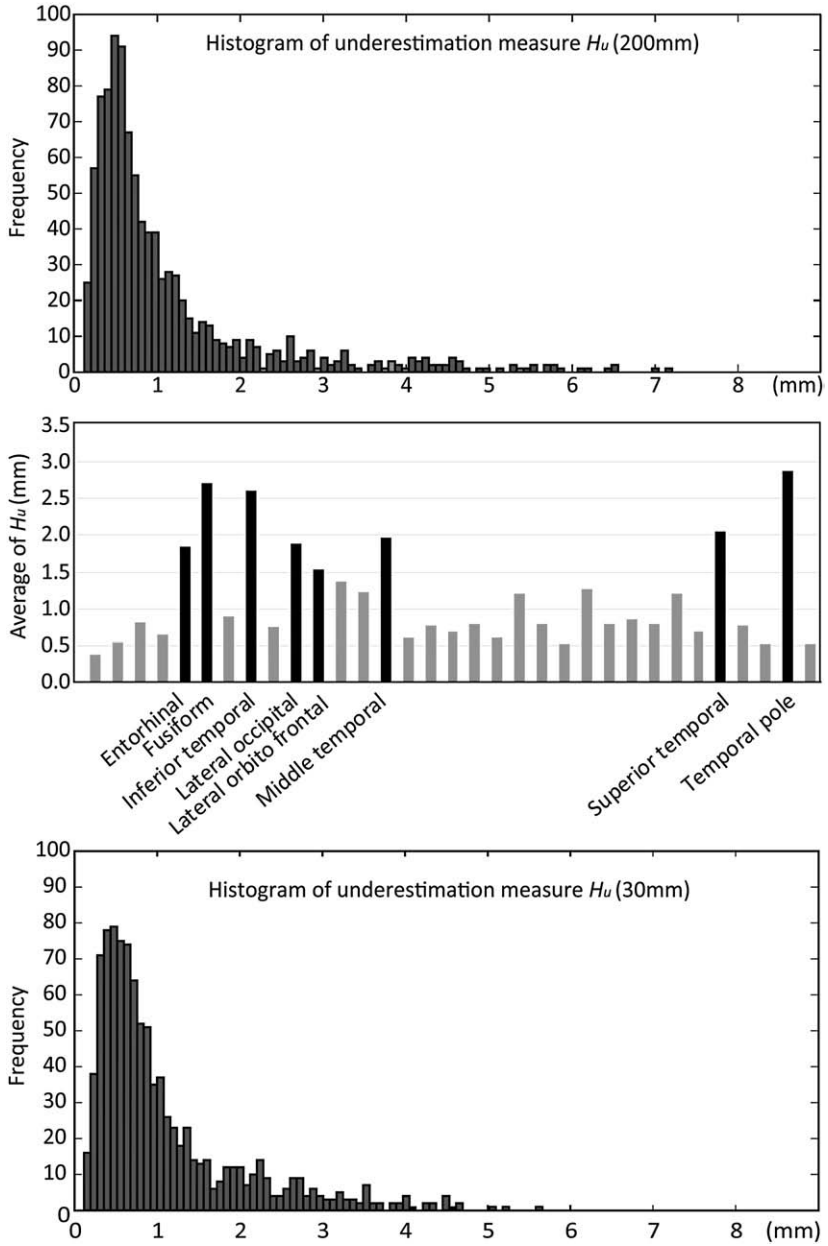

Fig. 4. Regional underestimation distances obtained with default N3 smoothing distance of 200 mm: histogram of all distances (top) and average distance per region (middle); highlighted bars represent values larger than 1.5 mm. Histogram of underestimation distances at 30 mm smoothing distance (bottom).

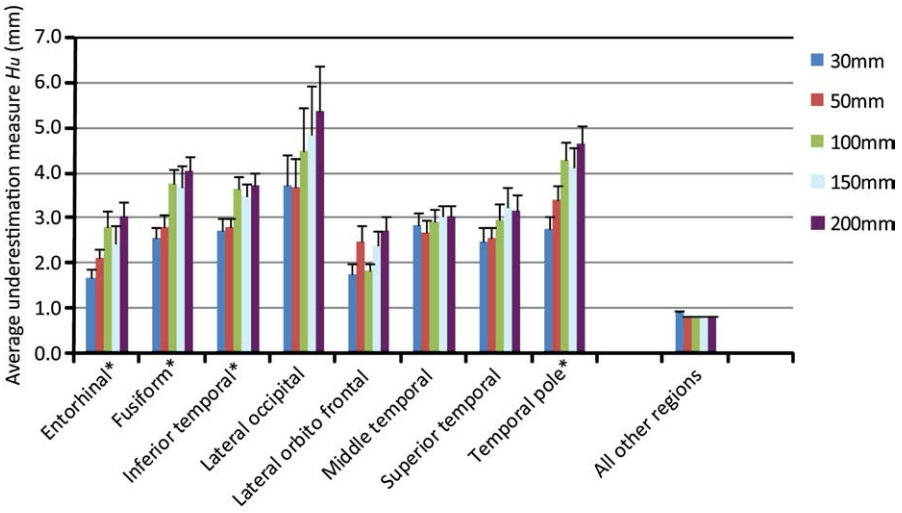

Fig. 5. Underestimation measures under varying N3 smoothing distances for eight selected regions and the average for the rest of the regions. *$p<0.05$ (Bonferroni corrected for multiple comparisons).

reduction was only 0.2 mm, suggesting that the underestimation problem was caused by factors other than intensity non-uniformity. Fig. 6 illustrates the extent to which reduction in underestimation could be achieved. Residual intensity inhomogeneity within the temporal lobe when using the default distance of 200 mm resulted in significant underestimation of the WM and pial surfaces (Fig. 6). Changing the smoothing distance to 30 mm led to better non-uniformity correction and improved segmentation. Improvement can also be observed in a tighter clustering of the underestimation measures' histogram at 30 mm (Fig. 4).

In the second group a small, non-significant increase in the average underestimation, from 0.78 mm at 200 mm to 0.89 mm at 30 mm, was observed (Fig. 5). This increase, however, was no longer present at the 50 mm smoothing distance. Regarding the overestimation measurements, none of the 34 regions displayed statistically significant change. This was expected as overestimation of the pial surface is primarily caused by inadequate skull stripping rather than intensity non-uniformity.

Reliability of segmentation performance

To test the effect of N3 on segmentation reliability we processed data set 4 in FreeSurfer pipeline, recording corresponding regional cortex thicknesses and volumes of subcortical structures. Five smoothing distances (30 mm, 50 mm, 100 mm, 150 mm and 200 mm) were tested, obtaining 5 sets of measurements, each containing 34(regions) × 24(hemispheres) (16 from the first scan + 8 from the repeated scan, see Table 1) of cortex thickness measurements and 37(volumes) × 24(hemispheres) of subcortical structure volume measurements (Fischl et al., 2002) from each scanner. It was hypothesized that smaller smoothing distances would produce better segmentation reliability in cortex thickness measurement. We did not

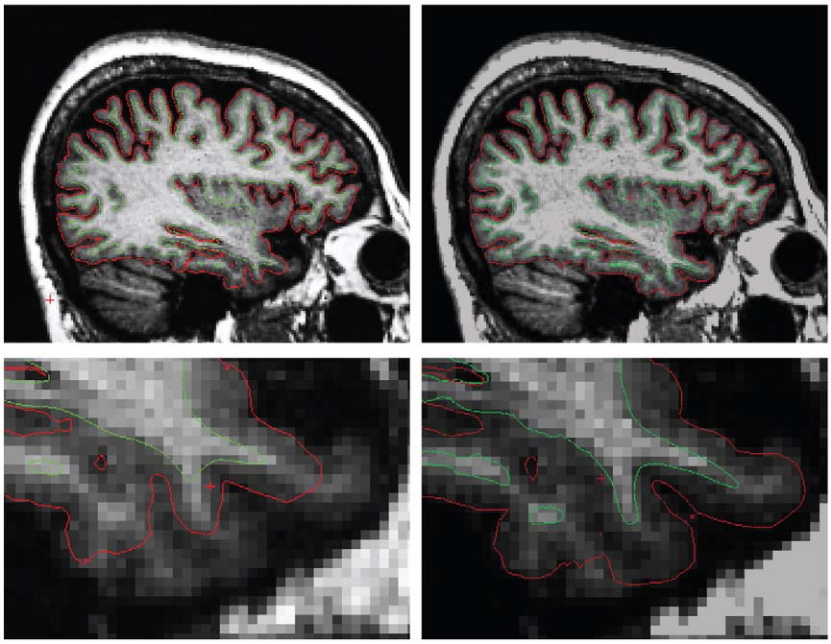

Fig. 6. Effect of intensity inhomogeneity correction on FreeSurfer segmentation performance: segmented white matter surface (green) and pial surface (red) overlaid on image corrected with N3 smoothing distance $d = 200$ mm (top left); segmented surfaces with N3 smoothing distance $d = 30$ mm (top right); enlargements of relevant areas (bottom).

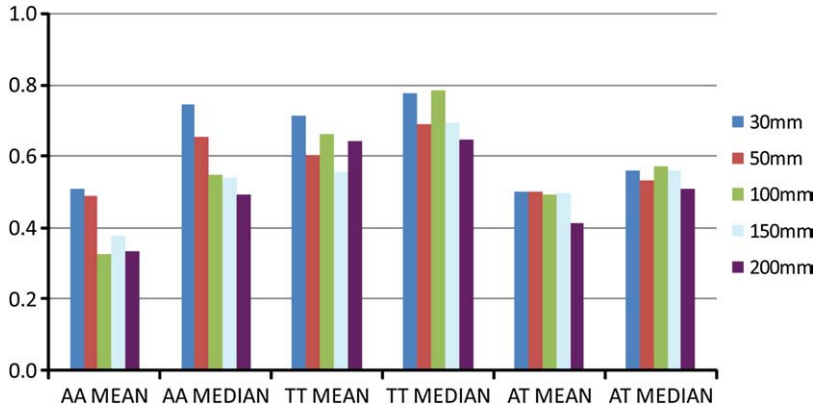

Fig. 7. Mean and median ICCs of cortex thickness measurements; AA: within Allegra scanner, TT: within Tim Trio scanner; AT: between Allegra and Trio scanners.

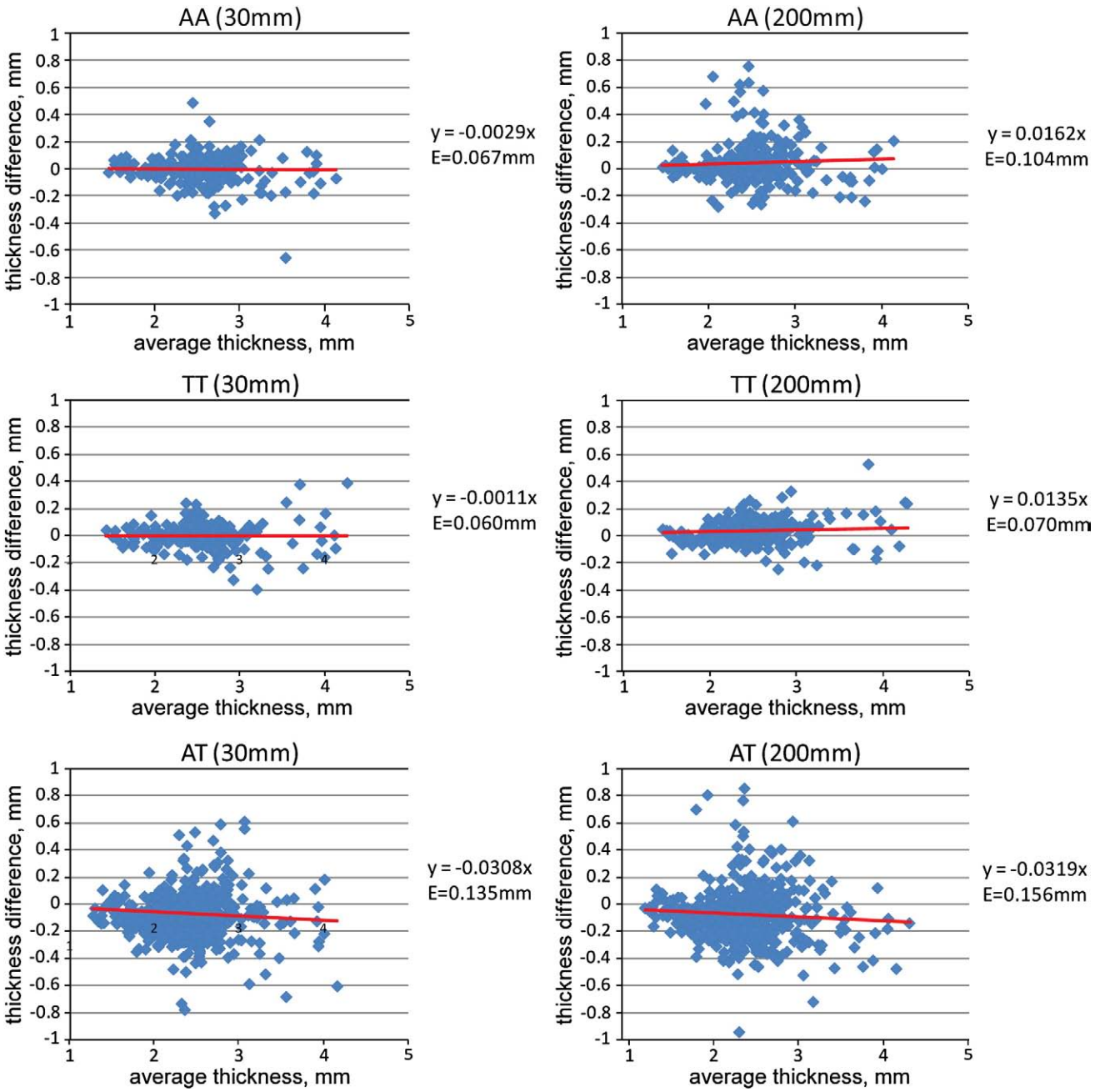

Fig. 8. Scatter plots of regional cortex thicknesses measured within Allegra scanner (AA_*), within Tim Trio scanner (TT_*) and between the two scanners (AT_*) with smoothing distances 30 mm and 200 mm; *E* represents the mean absolute thickness difference.

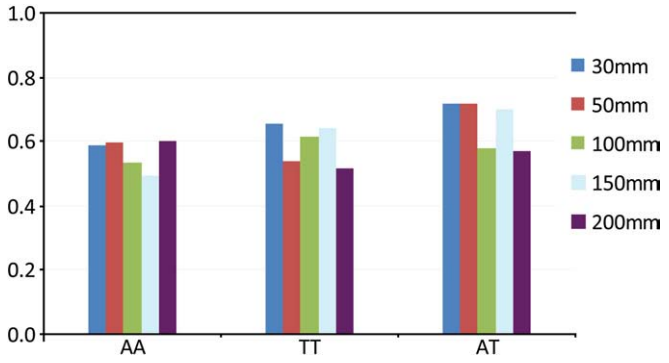

Fig. 9. Average ICCs of the volumes of subcortical structures, AA: within Allegra scanner; TT: within Tim Trio scanner; AT: between two scanners.

expect a change in reliability of subcortical structure measurement, as smooth intensity variations caused by non-uniformity can be neglected for structures with a small spatial extent.

The mean and median ICCs of cortex thickness measurements within and between two 3 T scanners (Siemens Allegra and Tim Trio) are shown in Fig. 7. We observed an increase in ICC for all measurements when using smaller smoothing distances of 30–50 mm. As the mean ICC values were affected by outliers, we computed median values that probably better reflect the actual performance. The largest improvement was observed in Allegra data where the median ICC was raised from 0.49 to 0.74. The change in ICC was smaller for within Tim Trio and between scanner measurements.

When reliability was visualized using Bland–Altman scatter plots (Fig. 8), the greatest benefit again arose from Allegra data. The dispersion of measurements was reduced and the slope was closer to zero using the 30 mm smoothing distance. For the Tim Trio data, there was no difference in cluster scatter, but an improvement in slope was observed. Finally, the between scanner measurements revealed no change in slope but a more compact clustering at 30 mm.

These changes in estimation of cortical thickness resulted in a reduction of the mean absolute thickness difference measure that was largest for Allegra data (from 0.104 mm to 0.067 mm), and smaller for between scanner (from 0.156 mm to 0.135 mm) and within Tim Trio measurements (from 0.070 mm to 0.060 mm).

The reliability of volume measurement within 14 subcortical structures (7 from each hemisphere) is shown in Fig. 9 (ICC) and Fig. 10 (PVC). These are the same structures that were estimated in prior publications (Fischl et al., 2002, 2004b; Han and Fischl, 2007); see the full list in Fig. 10. Smoothing distance had a small effect on average ICC, there being a modest improvement for the Tim Trio and between scanner measurements.

This increase was primarily due to an improved reliability in pallidum segmentation; its PVC was significantly reduced from 2.5–3.5% to 1–1.5% (Fig. 10). This could be explained by the fact that the pallidum is the largest subcortical structure and correspondingly, the most sensitive to intensity non-uniformity.

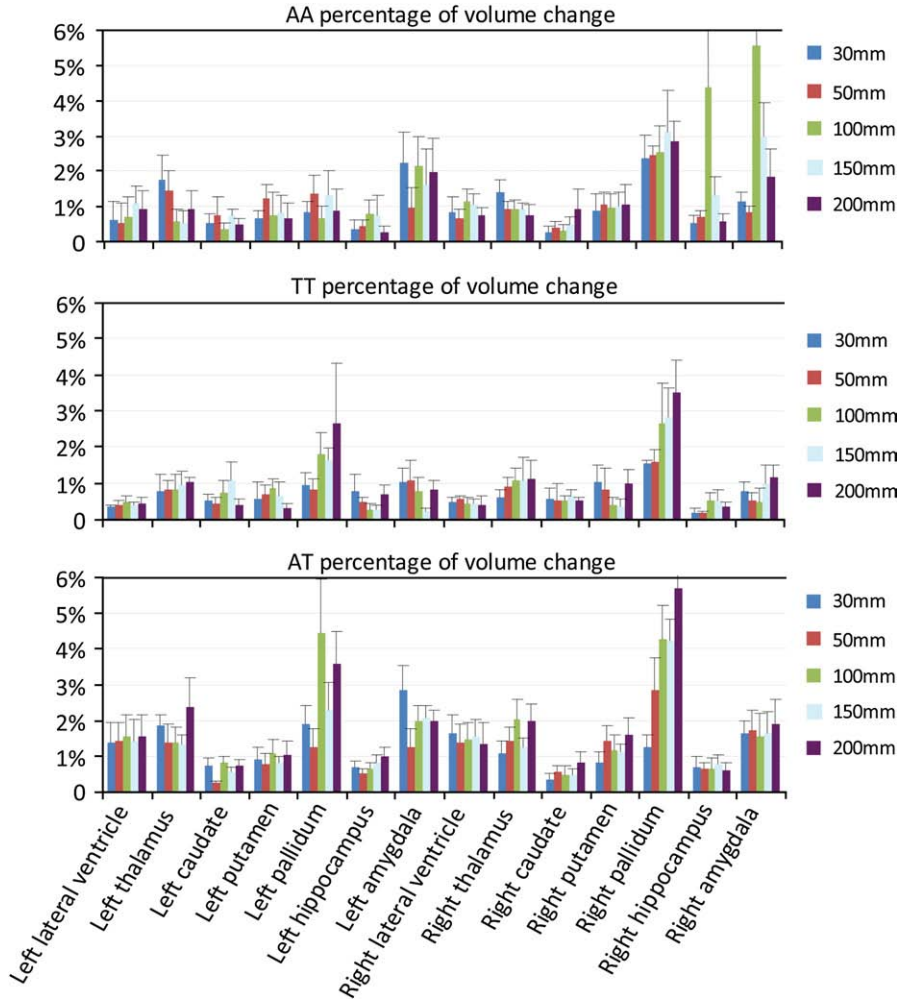

Fig. 10. Mean percentage of volume change for selected subcortical structures; AA: within Allegra scanner, TT: within Tim Trio scanner, and AT: between Allegra and Tim Trio scanners.

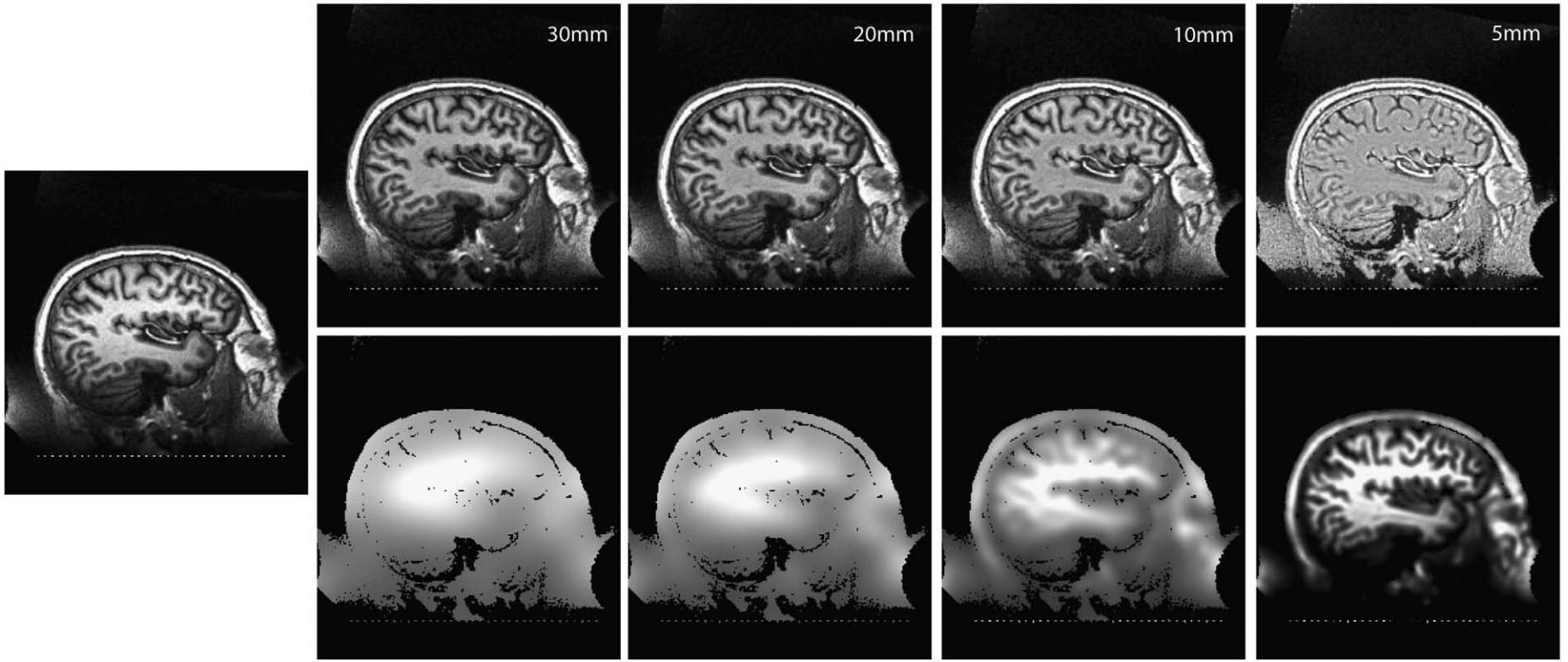

Fig. 11. Left — original image. Top row — images after N3-like correction with Gaussian smoothing. Bottom row — estimated bias fields. As FWHM of the Gaussian smoothing kernel gets smaller, the bias field overfits to image structures and reduces WM/GM contrast.

Discussion

Our experimental results provide substantial support for using smaller smoothing distances (30–50 mm rather than default value of 200 mm) with the widely used N3 non-uniformity correction. This seemingly trivial adjustment results in reduction in inhomogeneity of WM intensity that in turn improves the estimation of WM/GM surfaces, leading to improved reliability of cortical (and even some subcortical) segmentation.

Note that the lower bound on smoothing distance (30 mm) was chosen due to limitations of the current implementation of N3 algorithm, causing it to fail when the distance is set too low. N3's approach to smoothing is based on B-spline fitting, where the spline coefficients are obtained by solving a system of linear equations. The number of elements in the equation matrix grows inversely proportional to the sixth degree of the smoothing distance. Reducing the smoothing distance beyond a certain point under 30 mm generates an enormous matrix, which causes a memory error.

Additionally, smoothing distances below 30 mm are unlikely to be beneficial. To confirm this, we used our own implementation of N3 algorithm, substituting spline-based smoothing with a more conventional Gaussian smoothing. When the full width half maximum (FWHM) of the Gaussian kernel was reduced to below 30 mm, non-uniformity correction resulted in progressive overfitting to image structures, reducing contrast between WM and GM (Fig. 11).

Of 118 locations where WM was underestimated, complete resolution of the underestimation problem — to below 1.5 mm, was realized in 15 locations. The other locations displayed a partial resolution where underestimation distance was reduced, but was above 1.5 mm. This indicates that intensity non-uniformity is not the only factor causing WM underestimation. Other factors may include noise, poor subject-dependent WM/GM contrast (Fischl and Dale, 2000; Han et al., 2006) and biological factors that drive regional differences in WM intensity (van Walderveen et al., 2003).

As an optimal smoothing distance value, 50 mm is suggested, despite the fact that 30 mm facilitates a larger decrease in the underestimation error for problematic regions. The distance of 50 mm appears to be a 'safer' choice because, unlike 30 mm, it does not affect WM estimation in non-problematic locations.

The results also highlight that, using lower N3 smoothing distance values can improve sensitivity to small changes in brain structure volumes/thicknesses. For example, at a smoothing distance of 200 mm, one scanner had a reliability performance 1.5 times worse than another (mean absolute thickness difference of 0.10 mm vs. 0.07 mm), suggesting that the former scanner would be much less sensitive to small changes in regional cortex thickness. However, when applying smoothing distances of 30–50 mm, inter-scanner differences narrow (0.067 mm vs. 0.060 mm).

It remains to be seen whether the current findings can be extended to segmentation approaches other than FreeSurfer. It is well known that some segmentation algorithms, such as edge based (Mcinerney and Terzopoulos, 1996; Paragios, 2001; Ravinda and Rajapakse, 2003) and several variants of region growing approaches (Adams and Bischof, 1994; Hojjatoleslami and Kittler, 1998), are robust to slow variation in intensity. Intensity non-uniformity primarily affects the segmentation performance of methods relying on statistical classification of voxel intensities (Ahmed et al., 2002; Kovacevic et al., 2002; Marroquin et al., 2002; Pham and Prince, 1999; Van Leemput et al., 1999, 2003; Wells et al., 1996; Zhang et al., 2001). The sensitivity of the FreeSurfer segmentation approach to intensity non-uniformity follows from its reliance on a mixture of edge based and statistical classification segmentation algorithms (Fischl et al., 2002, 2004b; Han and Fischl, 2007). For example, while FreeSurfer's surface estimation is based on a deformable algorithm, the initialization of the surface deformation is provided by initial coarse segmentation, which is based on statistical classification and facilitated by a probabilistic atlas. An error in initialization can lead to the deformable surface algorithm being trapped at a wrong edge (Dale et al., 1999; Mcinerney and Terzopoulos, 1996). Moreover, the energy functional that guides the WM surface deformation process contains a term that penalizes the sum of intensity variances inside each tissue, which can also make the segmentation sensitive to intensity non-uniformity.

Conclusion

Adopting the FreeSurfer segmentation pipeline for illustrative purposes, we demonstrated that for N3, a non-uniformity correction technique in widespread use, a smoothing distance of 30–50 mm, significantly improves the accuracy and reliability of brain tissue segmentation. The finding could contribute to enhancing the accuracy of brain morphometry studies where cerebral cortex thicknesses, its change over time or following neurological disease are important endpoints.

Acknowledgments

This work is supported by grants SBIC C-012/2006 and BMRC 04/1/36/19/372 provided by A*STAR, Singapore (Agency for Science and Technology and Research).

References

Adams, R., Bischof, L., 1994. Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell. 16, 641–647.
Ahmed, M.N., Yamany, S.M., Mohamed, N., Farag, A.A., Moriarty, T., 2002. A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data. IEEE Trans. Med. Imag. 21, 193–199.
Arnold, J.B., Liow, J.-S., Schaper, K.A., Stern, J.J., Sled, J.G., Shattuck, D.W., Worth, A.J., Cohen, M.S., Leahy, R.M., Mazziotta, J.C., Rottenberg, D.A., 2001. Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. NeuroImage 13, 931–943.
Belaroussi, B., Milles, J., Carme, S., Zhu, Y.M., Benoit-Cattin, H., 2006. Intensity non-uniformity correction in MRI: existing methods and their validation. Med. Image Anal. 10, 234–246.
Bernstein, M.A., Huston, J., Ward, H.A., 2006. Imaging artifacts at 3.0 T. J. Magn. Reson. Imag. 24, 735–746.
Boyes, R.G., Gunter, J.L., Frost, C., Janke, A.L., Yeatman, T., Hill, D.L.G., Bernstein, M.A., Thompson, P.M., Weiner, M.W., Schuff, N., Alexander, G.E., Killiany, R.J., DeCarli, C., Jack, C.R., Fox, N.C., 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. NeuroImage 39, 1752–1762.
Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. NeuroImage 9, 179–194.
Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980.
Feczko, E., Augustinack, J.C., Fischl, B., Dickerson, B.C., 2009. An MRI-based method for measuring volume, thickness and surface area of entorhinal, perirhinal, and posterior parahippocampal cortex. Neurobiol. Aging 30 (3), 420–431.
Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. 97, 11050–11051.
Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. NeuroImage 9, 195–207.
Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.
Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D., Busa, E., Seidman, L., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M., 2004a. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14, 11–22.
Fischl, B., Salat, D.H., van der Kouwe, A.J.W., Makris, N., Segonne, F., Quinn, B.T., Dale, A.M., 2004b. Sequence-independent segmentation of magnetic resonance images. NeuroImage, Math. Brain Imag. 23, S69–S84.
Friedman, L., Stern, H., Brown, G.G., et al., 2008. Test–retest and between-site reliability in a multicenter fMRI study. Hum. Brain Mapp. 29, 958–972.
Gispert, J.D., Reig, S., Pascau, J., Vaquero, J.J., Garcia-Barreno, P., Desco, M., 2004. Method for bias field correction of brain T1-weighted magnetic resonance images minimizing segmentation error. Hum. Brain Mapp. 22, 133–144.
Han, X., Fischl, B., 2007. Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE Trans. Med. Imag. 26, 479–486.
Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B.,

Fischl, B., 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. NeuroImage 32, 180–194.
Hojjatoleslami, S.A., Kittler, J., 1998. Region growing: a new approach. IEEE Trans. Image Process. 7, 1079–1084.
Hou, Z., 2006. A review on MR image intensity inhomogeneity correction. Int. J. Biomed. Imag. 1–11.
Hou, Z., Huang, S., Hu, Q., Nowinski, W.L., 2006. A Fast and Automatic Method to Correct Intensity Inhomogeneity in MR Brain Images. Springer Verlag, Heidelberg, D-69121, Germany, Copenhagen, Denmark, pp. 324–331.
Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15, 850–863.
Kovacevic, N., Lobaugh, N.J., Bronskill, M.J., Levine, B., Feinstein, A., Black, S.E., 2002. A robust method for extraction and automatic segmentation of brain images. NeuroImage 17, 1087–1100.
Likar, B., Viergever, M.A., Pernus, F., 2001. Retrospective correction of MR intensity inhomogeneity by information minimization. IEEE Trans. Med. Imag. 20, 1398–1410.
Luo, J., Zhu, Y., Clarysse, P., Magnin, I., 2005. Correction of bias field in MR images using singularity function analysis. IEEE Trans. Med. Imag. 24, 1067–1085.
Marroquin, J.L., Vemuri, B.C., Botello, S., Calderon, F., Fernandez-Bouzas, A., 2002. An accurate and efficient Bayesian method for automatic segmentation of brain MRI. IEEE Trans. Med. Imag. 21, 934–945.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46.
Mcinerney, T., Terzopoulos, D., 1996. Deformable models in medical image analysis: a survey. Med. Image Anal. 1, 91–108.
Paragios, N., 2001. A variational approach for the segmentation of the left ventricle in MR cardiac images. Proceedings of the IEEE Workshop on Variational and Level Set Methods (VLSM'01), pp. 153–160.
Pham, D.L., Prince, J.L., 1999. Adaptive fuzzy segmentation of magnetic resonance images. IEEE Trans. Med. Imag. 18, 737–752.
Ravinda, G.N.M., Rajapakse, J.C., 2003. Nurbs snakes. Image Vis. Comput. 21, 551–562.
Scott, M.L.J., Thacker, N.A., 2005. Robust Tissue Boundary Detection for Cerebral Cortical Thickness Estimation. Springer Verlag, Heidelberg, D-69121, Germany, Palm Springs, CA, United States, pp. 878–885.
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.
Shrout, P.E., Fleiss, J., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1997. A comparison of retrospective intensity non-uniformity correction methods for MRI. Lect. Notes Comput. Sci. 1230, 459–464.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. Nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17, 87–97.
Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17, 479–489.
Styner, M., Leemput, K.V., 2005. Retrospective Evaluation and Correction of Intensity Inhomogeneity. Taylor&Francis Group, LLC.
van der Kouwe, A.J., Benner, T., Salat, D.H., Fischl, B., 2008. Brain morphometry with multiecho MPRAGE. NeuroImage 40, 559–569.
Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imag. 18, 897–908.
Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 2003. A unifying framework for partial volume segmentation of brain MR images. IEEE Trans. Med. Imag. 22, 105–119.
van Walderveen, M.A., van Schijndel, R.A., Pouwels, P.J., Polman, C.H., Barkhof, F., 2003. Multislice T1 relaxation time measurements in the brain using IR-EPI: reproducibility, normal values, and histogram analysis in patients with multiple sclerosis. J. Magn. Reson. Imaging 18, 656–664.
Vovk, U., Pernus, F., Likar, B., 2006. Intensity inhomogeneity correction of multispectral MR images. NeuroImage 32, 54–61.
Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans. Med. Imag. 26, 405–421.
Weir, J.P., 2005. Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Cond. Res. 19, 231–240.
Wells, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of MRI data. IEEE Trans. Med. Imag. 15, 429–442.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imag. 20, 45–57.