NeuroImage 49 (2010) 225–239

Contents lists available at ScienceDirect

NeuroImage

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

Skull stripping using graph cuts

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

[a] School of Computer Engineering, Nanyang Technological University, 50 Nanyang Ave, Singapore 639798, Singapore
[b] Cognitive Neuroscience Laboratory, Duke-NUS Graduate Medical School

ARTICLE INFO

Article history:
Received 18 November 2008
Revised 25 July 2009
Accepted 24 August 2009
Available online 2 September 2009

Keywords:
Skull stripping
Graph cuts
Brain morphometry
Segmentation
MRI

ABSTRACT

Removal of non-brain tissues, particularly dura, is an important step in enabling accurate measurement of brain structures. Many popular methods rely on iterative surface deformation to fit the brain boundary and tend to leave residual dura. Similar to other approaches, the method proposed here uses intensity thresholding followed by removal of narrow connections to obtain a brain mask. However, instead of using morphological operations to remove narrow connections, a graph theoretic image segmentation technique was used to position cuts that isolate and remove dura. This approach performed well on both the standardized IBSR test data sets and empirically derived data. Compared to the Hybrid Watershed Algorithm (HWA; (Segonne et al., 2004)) the novel approach achieved an additional 10–30% of dura removal without incurring further brain tissue erosion. The proposed method is best used in conjunction with HWA as the errors produced by the two approaches often occur at different locations and cancel out when their masks are combined. Our experiments indicate that this combination can substantially decrease and often fully avoid cortical surface overestimation in subsequent segmentation.

Introduction

Accurate "skull stripping" involves removal of the scalp, skull and dura, and is an important procedure in brain image analysis. Tissue classification, registration, volumetric analysis of the brain and cortical surface reconstruction all benefit from the accurate removal of these non-brain tissues. In particular, the removal of dura while leaving brain tissue untouched is especially important when estimating cortical thickness (Freesurfer (Dale et al., 1999), CLASP (Kim et al., 2005)) or gray matter volume (voxel based morphometry (VBM) (Ashburner and Friston, 2000), SIENAX (Smith et al., 2002), part of FSL (Smith et al., 2004)). Unintended removal of the cortical surface cannot be reversed downstream in the processing pipeline and will result in underestimation of cortical thickness. Inclusion of non-brain structures can result in reduced VBM sensitivity (Fein et al., 2006), while dural attachments can cause overestimation of cortical thickness (van der Kouwe et al., 2008). This has resulted in the proposal of a variety of skull stripping techniques, such as region-based (Hahn and Peitgen, 2000; Shattuck et al., 2001), boundary-based (Smith, 2002; Zhuang et al., 2006) and hybrid approaches (Huang et al., 2006; Segonne et al., 2004).

The majority of skull stripping algorithms treat the brain as a single connected region separated from non-brain tissues by a rim of cerebrospinal fluid (CSF). In reality, even with high-resolution T1W MR images, thin connections between the brain and other cranial structures exist in the form of dura and connective tissue lining venous sinuses (Lemieux et al., 1999), (Fig. 1A). These connections reduce the accuracy of skull stripping (Figs. 1B–E).

Each of the existing skull stripping techniques has strengths and weaknesses. Brain Surface Extractor (BSE) (Shattuck et al., 2001) uses a combination of anisotropic diffusion filtering, Marr–Hildreth edge detector and morphological operators to separate brain and non-brain tissues. Typical of edge-based approaches, it may lead to inappropriate removal of brain tissue, contributed in part by reduced brain signal intensity adjacent to false connections, (Fig. 1B). The Watershed Algorithm (WAT) (Hahn and Peitgen, 2000), an intensity-based approach, relies on a 3D algorithm with pre-flooding performed on the intensity inverted image, selecting the basin to represent the brain. It may fail to remove dura, skull and various non-brain structures in the neck/eye area (Fig. 1C). The irregular, anatomically implausible brain masks sometimes generated by BSE and WAT can be avoided by imposing additional smoothness constraints with a deformable surface model, which is then fitted onto the brain surface by a set of internal and external forces. This method is utilized by Brain Extraction Tool (BET) (Smith, 2002) and Hybrid Watershed Algorithm (HWA) (Segonne et al., 2004). Deformation of BET's mask is guided by constraints on surface smoothness and voxel intensities in the vicinity of the surface position. HWA is a hybrid approach combining watershed algorithm and deformable model, where the latter adds atlas based shape constraints in order to guarantee anatomically meaningful brain mask. However, even additional constraints

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

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.08.050

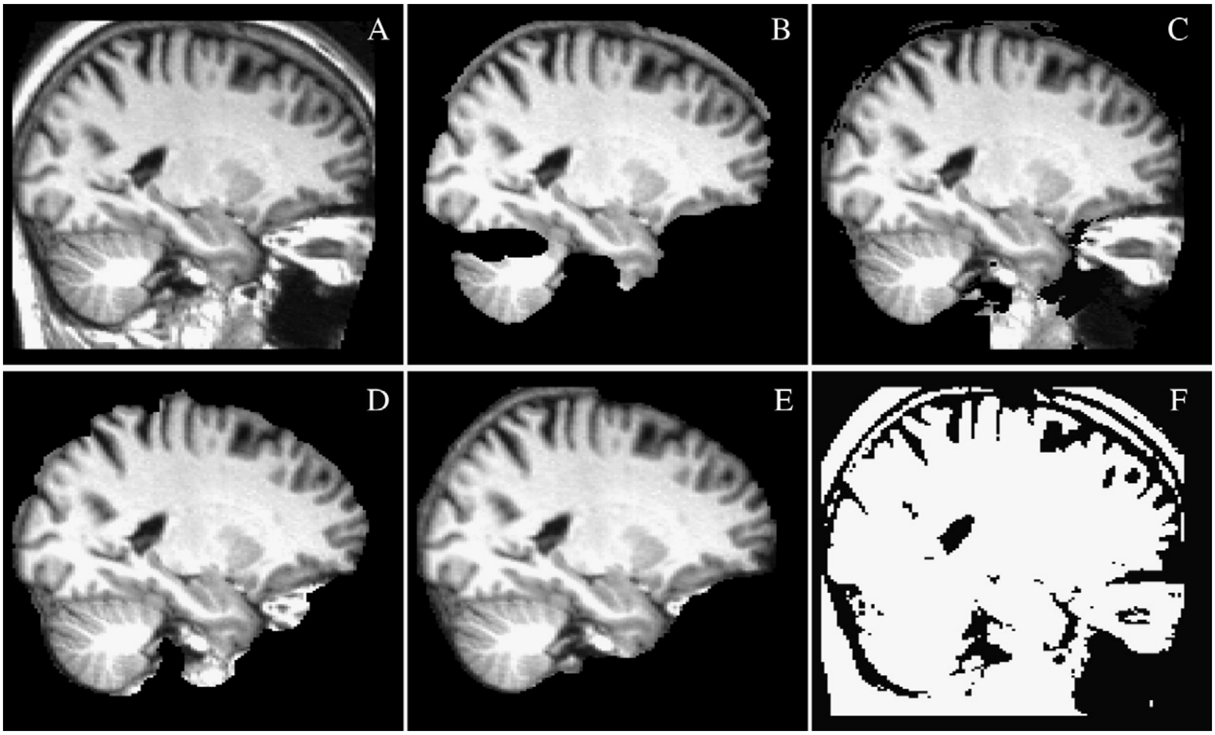

Fig. 1. Typical results of existing skull stripping techniques (A) Original image, (B) BSE, (C) WAT, (D) BET, (E) HWA, (F) Simple intensity thresholding.

do not always resolve the problem, because segmentation can be smooth and still lead to brain loss (Fig. 1D) or include non-brain tissues (Figs. 1D, E).

Despite the fact that among existing solutions HWA is the only approach that is very careful at preserving the brain, suiting it for subsequent cortical thickness estimation, it can greatly benefit from further stripping of the dura (Fennema-Notestine et al., 2006). The alternative approach proposed here sought to achieve further dura reduction while maintaining the same low rate of brain loss, in order to facilitate more accurate estimation of cortical thickness.

Our method was motivated by work (Atkins and Mackiewich, 1998; Kapur et al., 1996; Lemieux et al., 1999) that suggested segmenting the brain by using simple intensity thresholding followed by morphological opening operations to cut the narrow connections. Instead of morphological operations, which can only remove sufficiently narrow connections, the proposed approach relies on graph theoretic image segmentation techniques to position cuts that serve to isolate and remove dura. To establish the advantages of this approach, we evaluated its performance on legacy standardized test images as well as images collected using a more contemporary MR

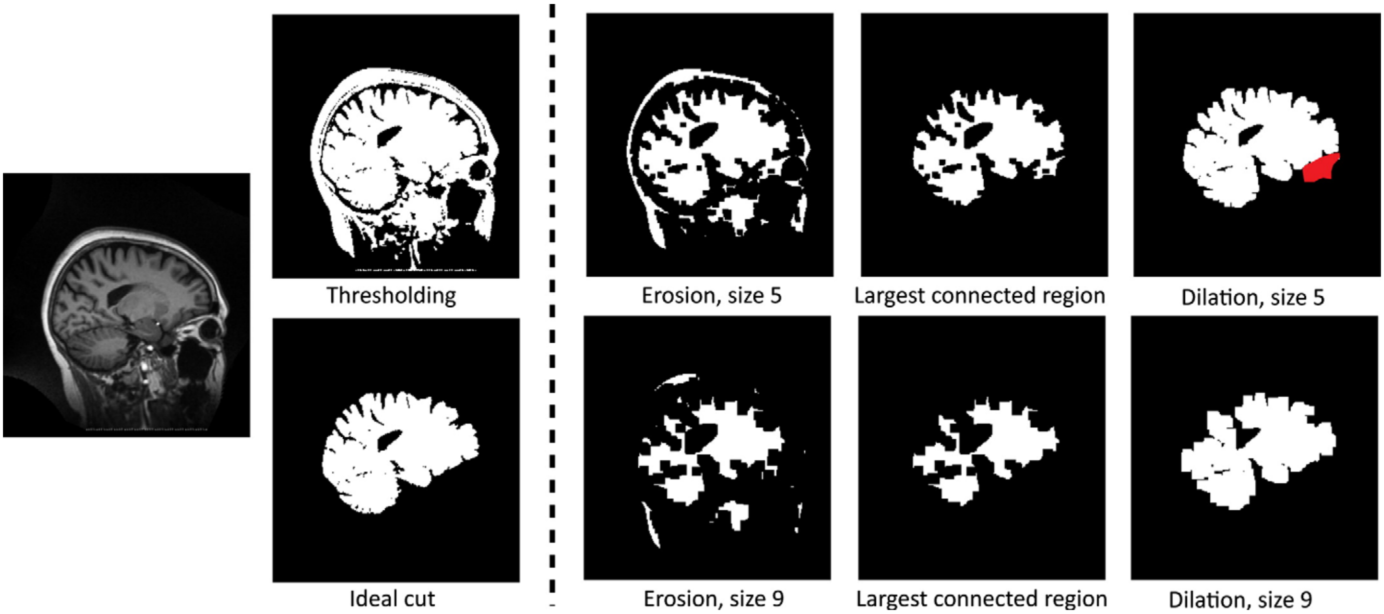

Fig. 2. Using morphological opening to cut narrow connections. Small erosion size (upper row) is useful for very narrow connections, leaving substantial non-brain tissue highlighted in red. Larger erosion size (lower row) cuts wider connections but leads to erosion of brain tissue.

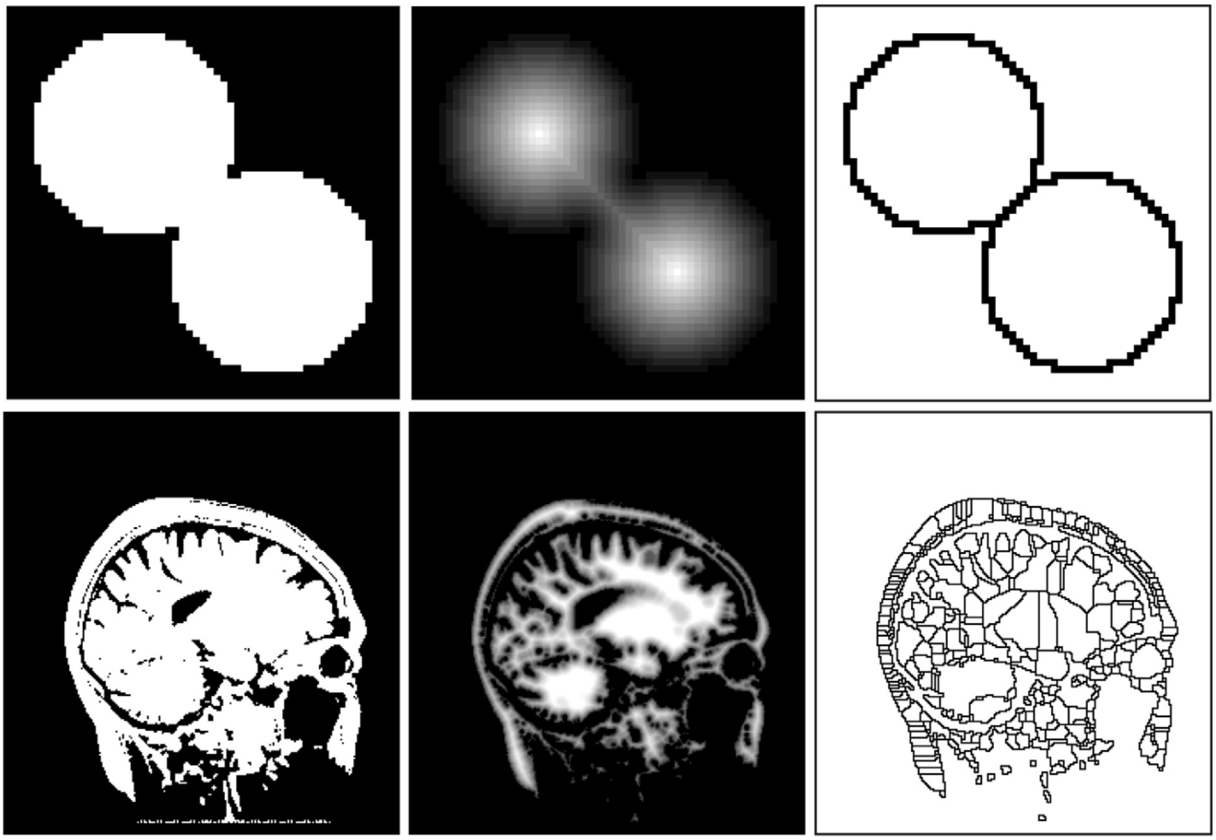

Fig. 3. Using distance transform followed by the watershed algorithm to cut narrow connections. Initial mask (left column) is processed by distance transform (middle column) followed by the watershed algorithm (right column). This approach worked well on simple shapes (upper row), but led to over-segmentation of the brain mask.

scanner. We compared our results with those obtained using current state-of-the-art techniques.

Materials and methods

Previous work on removal of narrow connections

The morphological framework for removal of narrow connections (MORPH), which forms the basis of (Atkins and Mackiewich, 1998; Kapur et al., 1996; Lemieux et al., 1999), is illustrated in Fig. 2 and consists of mask erosion, selection of the largest connected region and dilation of the same size as the erosion. Connections with widths less than the amount of erosion (controlled by the size of the structural element) are removed and are not restored following dilation. A drawback of this procedure is that only sufficiently narrow connections are removed (Fig. 2, upper row). Attempting to remove wider connections results in substantial elimination of brain tissue (Fig. 2, lower row).

Another frequently used technique involves distance transform followed by watershed algorithm (DWAT), see examples in (Grady, 2006; Park and Keller, 2001). The distance transform assigns each voxel in the binary mask a value equal to the shortest distance from this voxel to the boundary (Fig. 3). The watershed algorithm interprets lower values of the distance transform (occurring at voxels inside narrow connections) as valleys and higher values (occurring at voxels deeper inside the mask) as hills, as it segments the mask into a set of hills separated from each other by valleys. This works well for simple shape masks, (Fig. 3, upper row), but with brain data, valleys are created within the brain substance, negating the benefit of eliminating narrow connections (Fig. 3, lower row).

Our cutting algorithm was inspired by the isoperimetric graph partitioning (ISO) approach (Grady, 2006), which sets the problem in

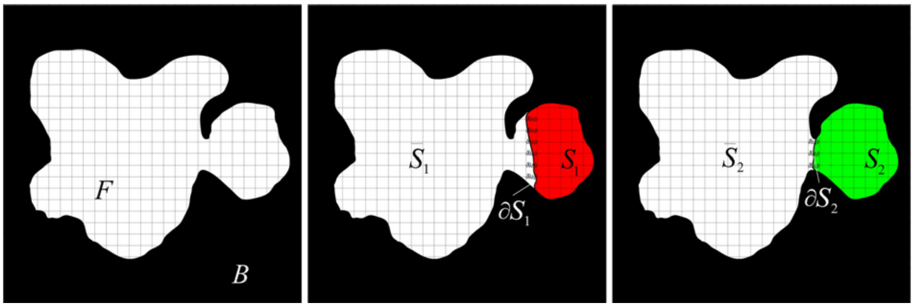

Fig. 4. Cutting narrow connections with ISO. Among all possible partitions of foreground region, the one with the smaller ratio of the cut value and the volume of smaller element was chosen.

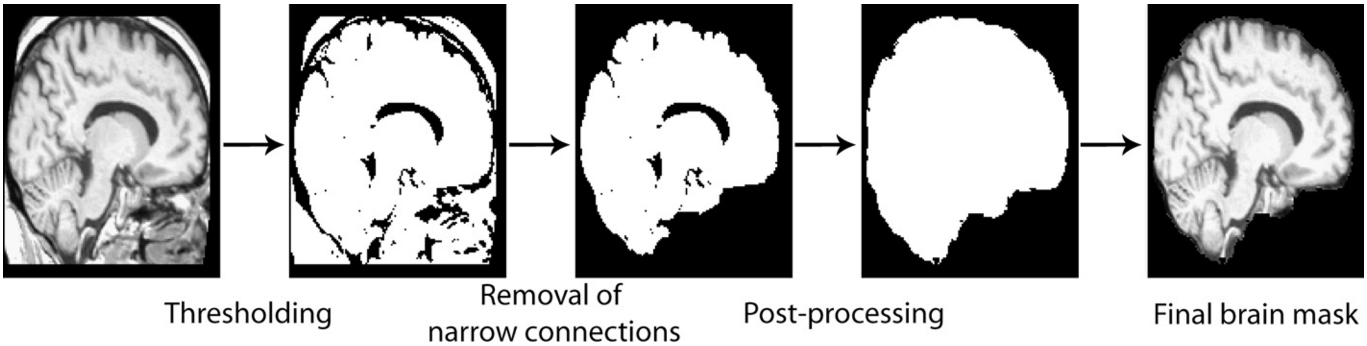

Fig. 5. The pipeline of proposed skull stripping approach.

a graph theoretic framework (Fig. 4). Here the image is treated as a weighted undirected graph $G = (V, E, W)$, where V, E and W are sets of vertices, edges and edge weights respectively. The edges refer to connections between any two vertices and the edge weights are arbitrary values assigned to each edge. The goal of ISO is to partition the vertices of the graph into two connected sets S and $\bar{S}$ so that isoperimetric ratio is minimized. The isoperimetric ratio is defined as a ratio of the cut value—the sum of edge weights along the boundary between the partitions, $|\delta S| = \sum_{i \in s} \sum_{j \in \bar{s}} w_{i,j}$, and the volume (the number of vertices) of the smaller partition element S, *Vol* (S):

$$c = \inf_{s} \frac{|\delta S|}{\mathrm{Vol}(S)}, \mathrm{Vol}(S) \leq \frac{1}{2} \mathrm{Vol}_{\mathrm{total}} \tag{1}$$

where $\mathrm{Vol}_{\mathrm{total}}$ designates the total number of vertices in the graph.

Starting with an initial mask, consisting of the foreground F and the background B regions (Fig. 4), ISO examines all possible ways to cut F into two sub-regions, choosing the cut with the smallest isoperimetric ratio. For example, between the two cuts shown in Fig. 4, $\mathrm{Vol}(S_1) < \mathrm{Vol}(S_2)$ and $|\delta S_1| = 6 > |\delta S_2| = 3$. Hence $c_1 = \frac{|\delta S_1|}{\mathrm{Vol}(S_1)} > c_2 = \frac{|\delta S_2|}{\mathrm{Vol}(S_2)}$ and the second cut will be chosen.

Proposed approach

The proposed method of skull stripping uses three groups of operations. These are: thresholding to obtain preliminary mask, removal of narrow connections using graph cuts and post-processing (Fig. 5). The goal of the last step is to reinstate partial volume gray matter (GM) voxels inadvertently removed following thresholding.

Obtaining preliminary mask

The initial thresholded mask must satisfy two conditions:

(1) The brain should be weakly connected to non-brain structures.
(2) The mask should preserve as much brain as possible, since the subsequent narrow connection removal can only further reduce the mask.

For T1W images, an appropriate intensity threshold lies somewhere between the mean intensities of GM and CSF. Threshold values that are too low may lead to the inclusion of CSF and dura, resulting in the appearance of strong connections between brain and the cranial vault (Fig. 6). Values that are too high may provide a clearer demarcation between brain and non-brain structures but at the expense of brain erosion. The desired threshold results in a mask with sufficiently narrow connections and acceptable brain loss that can be compensated for during post-processing (Fig. 6).

Existing methods of threshold selection utilize the image histogram—using the histogram's first valley (Atkins and Mackiewich, 1998) or fitting a function of Otsu's threshold (Lemieux et al., 1999). However, histogram features may not identify an appropriate threshold. For example, choosing the first valley of the histogram may result in a threshold that is too low (Fig. 6). In this example an appropriate threshold is located in between the first valley and peak of GM tissue distribution.

For T1W images, we found that a good threshold lies within 32–40% of white matter (WM) intensity, i.e. $0.32 I_{\mathrm{WM}} \leq T \leq 0.4 I_{\mathrm{WM}}$ and chose $T = 0.36 I_{\mathrm{WM}}$ for subsequent tests. “WM intensity” was estimated by averaging intensities within WM seed voxels, as explained in a later section on seed selection.

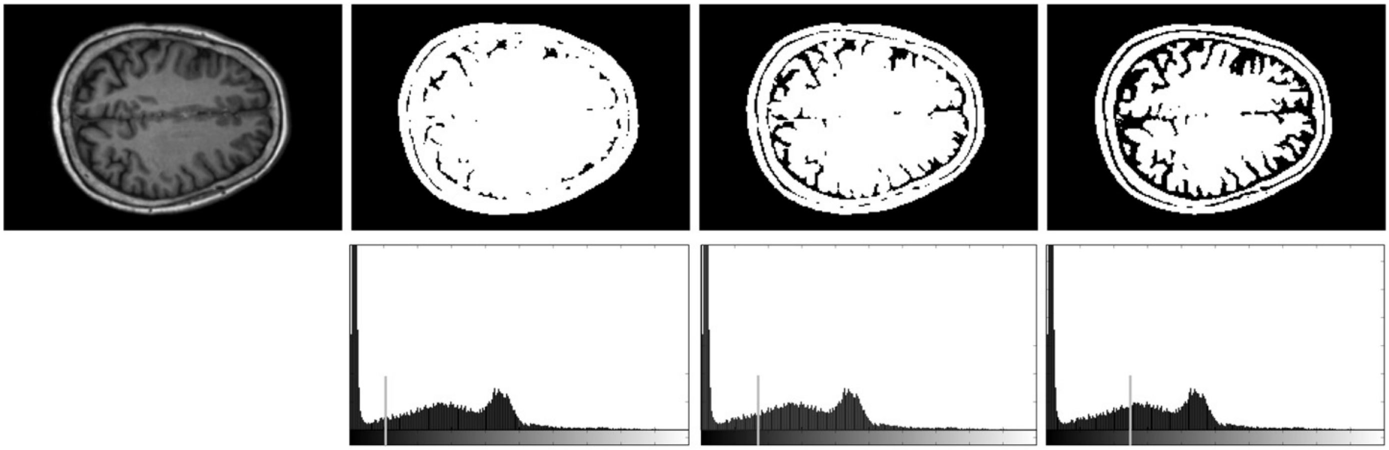

Fig. 6. Effect of different threshold values on the quality of initial mask. Too low threshold (second column) leads to insufficient separation between brain and non-brain structures, too high threshold (right column) results in brain loss.

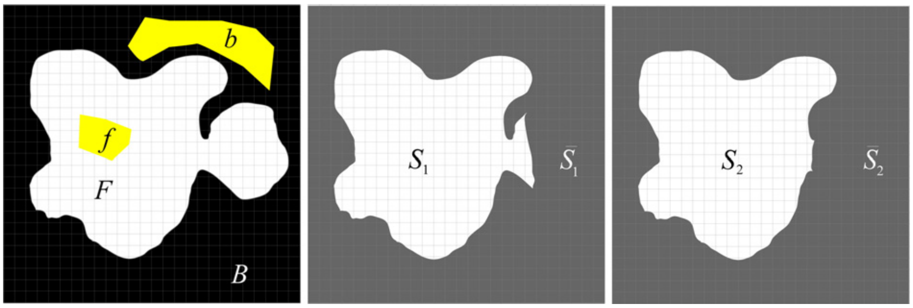

Fig. 7. Cutting narrow connections with graph cuts. After seeds are chosen, the algorithm searches for a partition with the smallest cut value that separates the seed regions.

Removing narrow connections with graph cuts

In our approach, instead of ISO we used graph cuts (Boykov and Jolly, 2001; Boykov et al., 2001), another graph theoretic segmentation approach, which uses a different minimization criterion and optimization procedure. The reasons for not using ISO were mostly technical- publicly available code, described in detail in Grady and Schwartz (2006), scales poorly with image size, limiting its application to 2D images. A more efficient solution based on tree representation of sparse matrices (Grady, 2006) has not been made publicly available yet. Another reason for using graph cuts approach is its optimization procedure, which allows reaching a globally optimal (vs. locally optimal in case of ISO) solution in a relatively short time (linear or weakly polynomial in the number of voxels), and its moderate memory requirements (Boykov and Kolmogorov, 2004).

Similar to ISO, the graph cuts approach treats the image as a weighted undirected graph $G=(V, E, W)$, where V, E and W are sets of vertices, edges and edge weights respectively. But instead of minimizing the ratio of the cut value $|\delta S|$ and the volume of the smaller partition element, it minimizes just the cut value, subject to constraint that the partition must correctly separate pre-defined sets of foreground and background seeds:

$$\hat{s} \;= \arg\inf_{S} |\delta S|, f \subseteq S, b \subseteq \bar{S} \tag{2}$$

The graph cut algorithm can be applied to the removal of narrow connection as follows. Given the initial mask (foreground F and background B), we define the graph on the whole image (not just on F as in ISO). The goal is then to choose appropriate seed regions and edge weights, so that the desired cut has the minimum value of all admissible cuts, i.e. cuts that separate the seed regions. For example, assume that (S_2,$\bar{S}_2$) is the desired cut (Fig. 7). First, we need the foreground and background seeds to be fully contained inside S_2 and B respectively, since the cut cannot be made through the seeds. To make the value of the cut (S_2,$\bar{S}_2$) smaller than that of (S_1,$\bar{S}_1$) or any other admissible cut, the weights of the edges surrounding S_2 should be made small while the weights of the edges surrounding other cuts large. While this is not an easy task in general, in the ensuing sections we will show how this selection can be achieved in the case of separation of brain from non-brain tissues.

Seed selection

Initial thresholding results in a binary segmentation (F, B), where the foreground region F is assumed to contain all brain as well as an unknown number of non-brain tissue voxels. Hence we can set the background seed to be equal to the background of the initial mask ($b=B$). Choosing the foreground seed is more challenging, as there is no a priori information concerning where actual brain tissue is located within the initial mask.

In our implementation we used the fact that in a T1W MR image, WM constitutes the largest region with uniformly high intensity. We began by partitioning the image into cubes of size 5 × 5 × 5 voxels. We then selected a cube that best fulfilled the combination of brightness (high mean intensity) and uniformity (low variance of intensity values). We found that this always resulted in a cube located inside the WM, (Fig. 8).

We then obtained an estimate of the region bounded by WM using region growing, initialized by the cube. Note that taking the cube alone as the foreground seed could lead to a trivial solution containing only this seed, as the number of edges surrounding the cube is small. The stopping criterion for region growing was conservative in that the grown region would still be separated from non-brain structures by a rim of GM and CSF, (Fig. 8).

Note that the described method of WM seed selection does not require preliminary brain masking and can be performed directly on the original image. In fact, WM seed selection was the first step in the actual implementation of the proposed skull stripping approach. The intensities of the voxels within WM seed were subsequently averaged to estimate the WM intensity, using the latter to determine initial threshold ($T=0.36I_{WM}$) and preliminary brain mask.

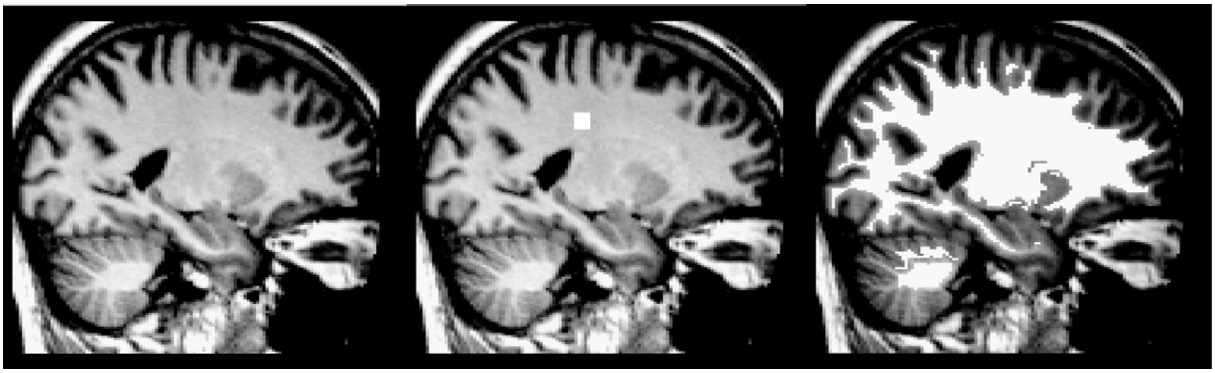

Fig. 8. Selection of foreground seed is performed by finding the brightest and most uniform cube (middle), followed by conservative region growing (right).

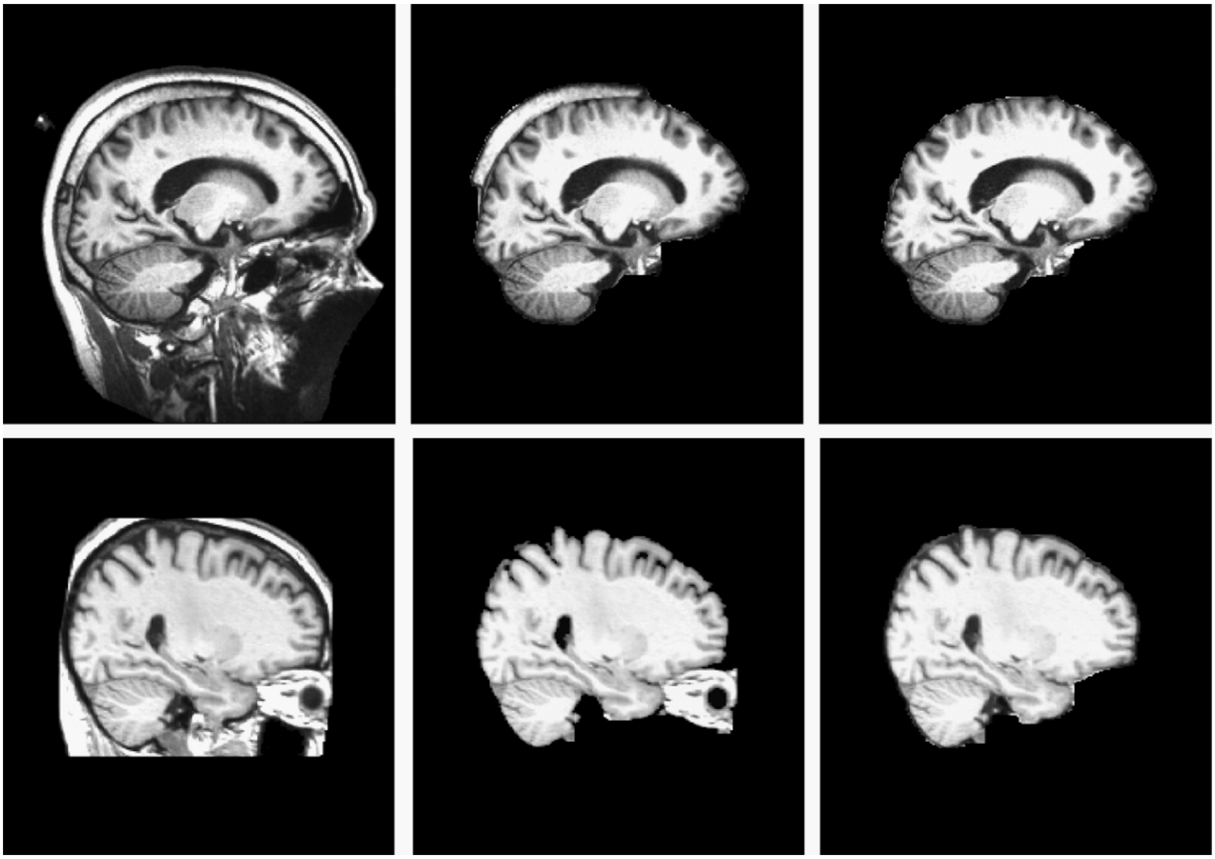

Fig. 9. Original image (left column), graph cuts with weight assignment based on distance transform (middle), graph cuts with weight assignment based on distance transform and intensity (right). First row is an image from data set 1, second row—from data set 2.

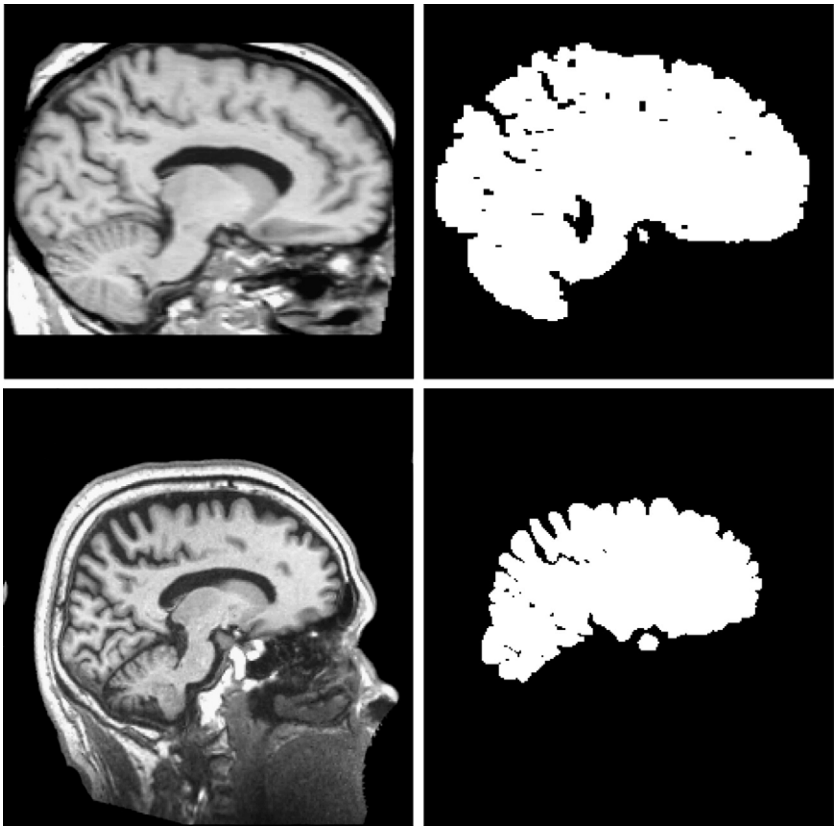

Fig. 10. Example of ground truth masks, top row—data sets 1 and 2 (IBSR), cerebellum included, bottom row—data sets 3 and 4 (Siemens Allegra 3T), cerebellum excluded.

Edge weight assignment

Recall that the weights of edges at the true boundary of the brain should be small relative to edge weights elsewhere. This was achieved by setting all the weights of the edges connecting foreground and background of the initial mask to 1. All other edge weights are defined as follows:

$$w_{i,j} = \max_{v_i \in F}\left(D(v_i), D\left(v_j\right)\right) \tag{3}$$

where $D(v_i)$ and $D(v_j)$ are distance transform values at voxels v_i and v_j. This assignment increases the weight of edges located deep within the foreground region, making cuts here less likely. On the other hand, weights inside the narrow connections remain small.

In the case of an occasional wide connection, this procedure fails because of high distance transform values inside the connection (Fig. 9, middle column). A simple solution to this is to use the fact that appropriate cuts between brain and non-brain structures usually go through voxels of relatively lower intensity, e.g. partial volume voxels between GM, CSF and dura. The weights of the edges that connect these pixels should thus be reduced, to favor cutting through them. We found the following assignment to work well (Fig. 9, right column):

$$w^{*}_{i,j} = w_{i,j} \cdot \left[\exp\left(k\,\frac{\min\limits_{v_i \in F}\left(I(v_i), I\left(v_j\right)\right) - T}{I_{WM} - T}\right) - 1\right] \tag{4}$$

where $w_{i,j}$ is given by (3), T is the threshold used to obtain the initial mask and k is a parameter that controls the contribution of voxel intensities. For T1W images, we obtained good results when $1 \leq k \leq 3$ and chose $k = 2.3$ for subsequent tests. The mean intensity of WM (I_{WM}) can be estimated by averaging intensities of the voxels inside the foreground seed. The graph cuts algorithm code was obtained from Boykov (Boykov and Kolmogorov, 2004).

Postprocessing

The initial thresholding procedure can inadvertently remove some darker partial volume voxels at the GM/CSF boundary. To improve results further, we applied a post-processing step to recover partial volume voxels and CSF. This was accomplished by performing morphological closing operation (10 mm voxel dilation and 10 mm voxel erosion, sizes rounded to the nearest integer) on the final mask and adding a layer of voxels at the cuts. This smoothes the mask and fills in the ventricles (Fig. 5).

Results

Data sets

We used the following four data sets for performance evaluation:

(1) *Data Set 1*: 18 T1W volumes from the Internet Brain Segmentation Repository (IBSR[1]), slice thickness 1.5 mm.

(2) *Data Set 2*: 20 T1W volumes of normal subjects from IBSR[1], slice thickness 3.1 mm.

(3) *Data Set 3*: 15 healthy subjects (age 56–71, 9 males), each scanned once on a Siemens Allegra (Erlangen, Germany) 3T scanner using the following parameters: TR = 2300.00 ms, TE = 2.91 ms, TI = 900 ms, FA = 9 degrees, resolution 1 × 1 × 1.1 mm.

(4) *Data Set 4*: 15 healthy subjects (age 56–71, 4 males), each scanned once on a Siemens Allegra 3T scanner using the following parameters: TR = 2300.00 ms, TE = 2.91 ms, TI = 900 ms, FA = 9 degrees, resolution 1 × 1 × 1.1 mm. In contrast with data set 3, segmentation of images from data set 4 gave rise to a variety of problems, e.g. inclusion of dura with GM, underestimation of WM surface, etc., which entailed substantial manual editing. The purpose of including this data set was to illustrate the potential benefit of the proposed approach for the subsequent segmentation of brain tissues, using FreeSurfer segmentation pipeline as an example.

Table 1
Estimated CNR between GM and dura/CSF, and coefficient of variation within WM for tested data sets.

Data set	CNR mean (SD) [range]	CV(WM) mean (SD) [range]
1	28.61 (11.70) [10.36 52.49]	0.10 (0.02) [0.06 0.13]
2	34.94 (12.15) [13.21 56.85]	0.13 (0.08) [0.08 0.33]
3	35.14 (6.0) [23.84 47.82]	0.09 (0.01) [0.08 0.11]
4	34.37 (6.98) [19.90 46.60]	0.09 (0.01) [0.07 0.10]

For all four data sets, the ground truth was defined as GM + WM (Hata et al., 2000; Lemieux et al., 1999; Stokking et al., 2000). Manual, expert segmentation containing GM, WM and subcortical structures (inclusive of cerebellum) was already included in the data sets 1 and 2 (Fig. 10, top row). To obtain ground truth for data sets 3 and 4, these were processed using FreeSurfer 3.04 (Dale et al., 1999). The resultant pial surfaces were edited by an expert and converted to volume masks. Note that the ground truth for data sets 3 and 4 excluded cerebellum (Fig. 10, bottom row).

To further highlight the differences between data sets, we evaluated two image quality metrics, namely contrast-to-noise ratio (CNR) between GM and dura/CSF and coefficient of variation (CV) of WM (Table 1). To estimate CNR, we first defined dura/CSF region by selecting 1-voxel thin layer external to GM. CNR was then defined as the difference between median intensities of GM and dura/CSF divided by the standard deviation of the noise, where the latter was estimated from a manually selected ROI in the air space outside the head. Images with dark CSF, good separation between GM and dura/CSF and low noise have higher CNRs and should be easier to skull strip. To estimate the CV of WM, we divided the standard deviation of image intensities within WM region (as defined by ground truth) by the mean intensity within the same region. This metric is often used in evaluation of nonuniformity correction algorithms as it is sensitive to slow variations in image intensities. However, it is also sensitive to noise and presence of imaging artifacts, sharp intensity variations, ghosting, all of which may be detrimental to skull stripping.

According to Table 1, data set 2 on average has better delineation between brain and non-brain structures and hence higher CNR, compared to data set 1. However, its images exhibited a large variety of imaging artifacts (strong intensity nonuniformity, ghosting), which is reflected in 30% increase in CV(WM). Images from data set 3 and 4 had similar CNR to that of images in data set 2, but better uniformity. Their quality was also much more consistent, which is reflected in halving of the metric variability compared to data sets 1 and 2. Data set 4 had slightly lower CNR compared to data set 3, possibly explaining why it was more problematic.

Evaluation metrics

(1) *Similarity coefficients*. We used Jaccard similarity, defined as $JS = \frac{M \cap N}{M \cup N}$, and Dice similarity, defined as $DS = \frac{2|M \cap N|}{|M| + |N|}$, where M and N refer to segmentation and ground truth respectively (Fennema-Notestine et al., 2006; Rehm et al., 2004; Segonne et al., 2004; Shattuck et al., 2001).

(2) *Segmentation error* (false positives (FP) and false negatives (FN)) (Hartley et al., 2006; Lee et al., 2003; Niessen et al., 1999; Segonne et al., 2004). We used $FP = \frac{|M \setminus N|}{|N|}$ and $FN = \frac{|N \setminus M|}{|N|}$.

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

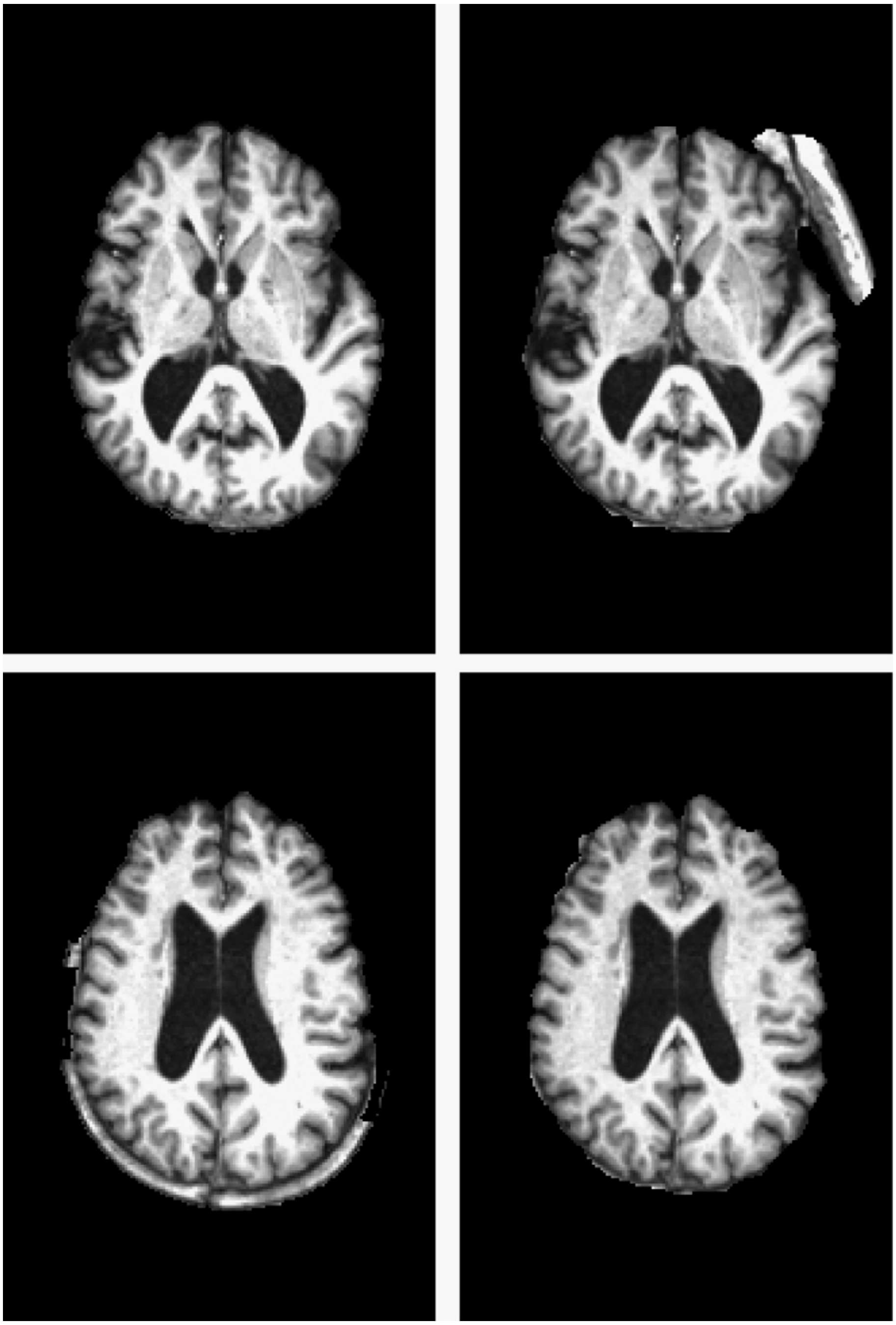

Fig. 11. Skull stripping result 1, FP = 27%, FP_adj = 8% (left), skull stripping result 2, FP = 31%, FP_adj = 6% (right).

One serious deficiency of these metrics in the setting of skull stripping is that they may lead to unfair comparisons between different approaches. This is because in addition to GM, WM and subcortical structures that must be preserved in a valid brain mask, there are also "optional" structures, e.g. brain stem and CSF, inclusion or exclusion of which does not materially affect the quality of the tissue segmentation that usually follows. If the ground truth contains only GM and WM, as was the case for all our data sets, these "optional" regions would contribute to artificial increase in FP rate and reduction in similarity coefficients. Hence skull stripping that preserves less CSF or cuts more of brain stem can be falsely declared superior to skull stripping which does the opposite, even if the two approaches are equivalent in terms of utility for subsequent processing or analysis.

Another problem that primarily affects existing formulation of FP rate is its insensitivity to the types of preserved non-brain structures, potentially making it a poor measure of subsequent segmentation performance. For example, consider two qualitatively different skull-stripping results in Fig. 11. The mask on the left contains large chunks of skull and dura; their proximity to a large area of the brain surface would likely to cause pial surface overgrowth. The mask on the right preserves a large portion of orbital contents/skull that are located further from the brain surface and are less likely to cause segmentation problems, despite causing a higher FP rate (31% vs. 27%).

To provide results that are less affected by CSF voxels we followed suggestions made by prior investigators (Boesen et al., 2004; Rex et al., 2004; Shattuck et al., 2001) and included *JS*, *DS* and *FP* metrics calculated without "dark" voxels, in addition to the traditional metrics. The dark voxels were intended to be a coarse estimate for CSF voxels and were classified as those voxels that had intensity below $0.36I_{\mathrm{WM}}$, the same threshold that was used to obtain

Table 2

Comparison of graph cuts skull stripping approach (GCUT) to cutting approaches based on morphological opening (MORPH) and distance transform followed by watershed algorithm (DWAT), data set 1 (18 1.5 mm scans, IBSR).

Method	DS (without dark pixels) mean (SD) [range]	JS (without dark voxels) Mean (SD) [range]	FN (%) mean (SD) [range]	FP (without dark voxels, %) mean (SD) [range]	FP_adj (%) mean (SD) [range]
Initial mask	0.74** (0.07) [0.60–0.83]	0.59** (0.09) [0.42–0.71]	0.003* (0.01) [0–0.04]	68.23** (21.93) [40.31–116.9]	9.40** (2.02) [6.67–14.29]
MORPH (size 5)	0.77** (0.09) [0.61–0.91]	0.64** (0.13) [0.44–0.84]	0.18 (0.54) [6.8e-05–2.28]	57.73** (27.17) [18.82–111.4]	7.76** (2.36) [3.32–13.61]
MORPH (size 11)	0.85** (0.09) [0.71–0.97]	0.75** (0.14) [0.55–0.95]	2.16* (2.97) [0.04–9.75]	31.22** (20.05) [5.31–62.32]	4.80** (2.19) [1.78–9.40]
DWAT	0.60** (0.11) [0.34–0.79]	0.44** (0.11) [0.21–0.65]	40.87** (15.03) [0.05–70.17]	29.64** (12.59) [15.50–55.3]	6.91** (2.43) [2.19–10.52]
GCUT	**0.95 (0.02)** [0.93–0.97]	**0.91 (0.03)** [0.86–0.95]	**0.03 (0.04)** [0.0001–0.15]	**7.09 (4.47)** [1.50–15.8]	**3.02 (2.05)** [0.49–7.01]

Bold emphasis designates the best value among all rows.

[*] $0.001<p<0.05$, where p designates the statistical significance of the difference between current value and GCUT's result.

[**] $p<0.001$.

preliminary mask, and did not belong to the ground truth. The new metrics can be calculated using the same formulas as above, through exclusion of dark voxels from segmented mask *M*. Note that *FN* rate did not need to be recalculated because ground truth mask *N* did not contain dark voxels by their definition. To reduce the influence of brain stem and cerebellum (for data sets 3 and 4 only), and to differentiate the skull/dura from other non-brain structures, we also provided an "adjusted FP rate" that excluded from consideration non-brain voxels located further than 5 mm external to the ground truth boundary. As shown in Fig. 11, this new metric is more sensitive to the amount of preserved dura and neighboring skull, resulting in a lower estimate for skull stripping result on the right.

The final experiment involved evaluating the effect of our skull stripping approach on subsequent brain segmentation performed using FreeSurfer 3.0.4 (Martinos Centre, Charlestown MA)[2].

Experiment 1. Comparing graph cuts to other cutting techniques

The purpose of this experiment was to highlight the advantage of using graph cuts over two alternatives, MORPH and DWAT. The same initial mask and post-processing were applied to each approach. For MORPH, we used our own implementation of dilation and erosion functions in 3D using two cubic structuring elements of size $5\times5\times5$ mm^3 and $11\times11\times11$ mm^3, rounding the size to nearest integer. For DWAT, we implemented 3D watershed algorithm as described in (Hahn and Peitgen, 2000). The results (averaged over all scans) pertaining to data set 1 are shown in Table 2.

Using the graph cuts based skull stripping pipeline (GCUT) as shown in Fig. 5 resulted in a 10-fold decrease in FP rate and 3-fold decrease in adjusted FP rate at the expense of an acceptable increase in FN rate, compared with the initial mask. MORPH traded FN performance for FP; selecting a small structuring element led to high FP rates (not much change from the initial mask), while choosing a large element resulted in substantial exclusion of brain tissue. Given that the connections between brain and non-brain structures vary in size, it is unlikely that an acceptable compromise can be found so as to realize desirable FN and FP rates. DWAT performed a little worse, and was associated with substantial brain loss, due to false valleys formed inside the brain.

Experiment 2. Comparison with existing skull stripping approaches

In the second experiment we compared GCUT with BET, BSE, WAT, and HWA on four chosen data sets, see results in Tables 3-6. A short description of each of these algorithms was provided previously. BET, WAT and HWA were used with default parameters. For BSE we changed the default parameter values as suggested in (Hartley et al., 2006) (diffusion constant = 35, diffusion iterations = 3, edge constant = 0.62, erosion size = 2); this resulted in better performance compared with default on four chosen data sets.

Substantial disparity between similarity coefficients and FP rates of data sets 1 and 2 compared to data sets 3 and 4 was due to ground truth definition; the ground truth for data sets 1 and 2 contained cerebellum whereas the ground truth for data sets 3 and 4 did not. Since all tested approaches preserved cerebellum as part of the brain, this led to higher FP rates and lower DS, JS for data sets 3 and 4. The adjusted FP rate was affected to a lesser degree, as it only counted the voxels within immediate vicinity of the brain surface.

Our performance evaluation of existing algorithms was consistent with previous findings. BSE's reported FN rate ranged from 2 to 12% (Boesen et al., 2004; Lee et al., 2003; Rex et al., 2004; Shattuck et al., 2001), depending on whether the parameters were fixed for the whole set or tuned for each individual volume (Boesen et al., 2004). The FN rates for BET and WAT were reported to be 2.7–4.3% (Lee et al., 2003; Rex et al., 2004) and 2% (Rex et al., 2004), respectively. Our findings fell in the same range, except that BSE and WAT performed poorly on data sets 2–4 in terms of FN rate. Our tests confirmed that HWA offers a favorable trade-off between FN and FP rates, leading to negligible (almost zero) brain loss at the expense of slightly higher FP rates (Fennema-Notestine et al., 2006). This trend was violated only for data set 2, where HWA resulted in substantial brain loss. Further investigation revealed that this poor performance was due to five subjects in data set 2 for which HWA either led to up to 50% brain loss or returned a running error. After exclusion of these subjects (see the bottom of Table 4), HWA's performance became consistent with that on other data sets. Note also that BET and BSE were similar to HWA in terms of adjusted FP rate but 2–4 times worse in terms of FP rates (data sets 3–4). This suggests that HWA tends to preserve smaller non-brain structures in the vicinity of the brain surface, e.g. skull and dura, while BET and BSE preserve larger non-brain structures in the eye and neck areas.

Overall, compared to HWA, our approach led to substantial decrease (10–30%) in adjusted FP rate, which was statistically significant for data sets 1 and 3 ($p<0.001$). The FN rates of two approaches were negligible; HWA did slightly better on data sets 1 and 3 while GCUT did better on data set 2 and 4. None of the differences was statistically significant. Note that HWA and GCUT performed similar to each other on data set 2, but only after excluding 5 subjects for HWA and one subject for GCUT. GCUT was superior when all subjects in data set 2 were used for performance evaluation.

The results discussed here relate to images that were not corrected for intensity nonuniformity. This said, intensity correction using N3 algorithm (Sled et al., 1998) with default parameters had little effect on the subsequent skull stripping performance of our algorithm, consistent with previous findings (Fennema-Notestine et al., 2006). For example, there was no change for JS, DS and FN rate, and a slight increase in adjusted FP rate from 2.95 (without N3 correction) to 3.2 (with N3 correction) for images in data set 3. For data set 4, correction helped improve JS from 0.79 to 0.8 and DS from 0.88 to 0.89. The FN rate was also slightly improved from 0.038 to 0.034, but at the expense of a slight increase in FP adjusted rate (from 3.92 to 4.09).

Since both GCUT and HWA had very low FN rates, an obvious way to further decrease the FP rate is by using the intersection of the two masks; the results are shown in Tables 3–6 under the name

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

Table 3
Comparison of graph cuts skull stripping approach (GCUT) with existing skull stripping approaches, Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), and Hybrid Watershed Algorithm (HWA), using data set 1 (18 1.5 mm scans, IBSR).

Method	DS mean (SD) [range]	JS mean (SD) [range]	FP (%) mean (SD) [range]	DS (without dark pixels) mean (SD) [range]	JS (without dark voxels) mean (SD) [range]	FP (without dark voxels, %) mean (SD) [range]	FN (%) mean (SD) [range]	FP_adj (%) mean (SD) [range]
BSE	0.91 (0.04) [0.84–0.97]	0.84 (0.07) [0.73–0.94]	**12.1** ** **(8.7)** [4.0–36.8]	0.92* (0.04) [0.84–0.98]	0.86* (0.08) [0.72–0.95]	7.9 (8.0) [0.6–30.9]	5.87* (7.82) [0.44–22.41]	**2.17** * **(1.98)** [0.10–7.40]
BET	**0.93 (0.04)** [0.79–0.96]	**0.86 (0.06)** [0.65–0.92]	14.0* (9.2) [7.7–48.6]	0.94 (0.04) [0.80–0.98]	0.89 (0.07) [0.67–0.95]	8.1 (7.5) [1.9–35.4]	1.93* (2.25) [0.11–6.73]	2.27* (1.28) [0.56–5.42]
WAT	0.91 (0.08) [0.60–0.96]	0.85 (0.11) [0.43–0.92]	18.8 (29.7) [5.3–134.3]	0.94 (0.03) [0.85–0.97]	0.89 (0.05) [0.73–0.94]	7.4 (8.2) [1.5–35.9]	2.45** (1.79) [0.08–7.08]	2.50* (2.08) [0.61–8.94]
HWA	0.88** (0.03) [0.82–0.91]	0.79** (0.04) [0.69–0.83]	27.1** (6.8) [20.0–44.4]	0.94** (0.02) [0.90–0.96]	0.89** (0.03) [0.82–0.93]	9.0** (5.4) [2.9–22.4]	**0.015 (0.02)** [0–0.07]	4.12** (2.27) [1.25–8.90]
GCUT	0.91 (0.02) [0.87–0.93]	0.84 (0.03) [0.78–0.87]	19.3 (4.0) [14.8–28.6]	**0.95 (0.02)** [0.93–0.97]	**0.91 (0.03)** [0.86–0.95]	7.1 (4.5) [1.5–15.8]	0.029 (0.04) [0–0.15]	3.02 (2.05) [0.49–7.01]
GCUT_HWA	0.92* (0.02) [0.88–0.97]	0.85** (0.03) [0.78–0.88]	18.0** (3.8) [13.8–27.9]	**0.95** * **(0.01)** [0.93–0.97]	**0.91** * **(0.03)** [0.87–0.95]	**6.4** * **(4.3)** [1.4–15.4]	0.044**(0.05) [0–0.17]	2.72* (1.97) [0.46–6.94]

Bold emphasis designates the best value among all rows.
GCUT_HWA stands for mask obtained by intersecting GCUT and HWA mask.
* $0.001<p<0.05$, where p designates the statistical significance of the difference between current value and GCUT's result.
** $p<0.001$.

GCUT_HWA. The intersection led to a small but tolerable increase in FN rate and further 5–20% decrease in the adjusted FP rate.

Experiment 3. Effect on FreeSurfer segmentation pipeline performance

To evaluate the practical usefulness of GCUT to the subsequent brain tissue segmentation, we applied it in the context of FreeSurfer segmentation pipeline. In FreeSurfer, the pial surface localization is limited to the brain mask obtained after the skull stripping procedure. An imperfect skull strip can affect the resultant pial surface in two ways. Inclusion of non-brain structures, such as dura matter, may result in overestimation of the pial surface. The opposite problem, brain loss inside the mask, will logically lead to underestimation of the pial surface. Given the ground truth, one way to assess the quality of segmentation is to determine the distance between corresponding points on the ground truth and test pial surfaces. However, we have found this difficult to implement in practice. First, this does not distinguish between underestimation and overestimation. Second, the surface mesh used in FreeSurfer has varying distances between vertices depending on the local surface complexity, making it difficult to find unique pairs of matching vertices. To circumvent these problems, we converted the pial surfaces into volume masks using FreeSurfer's built-in function “mri_surfmask” and evaluated performance in volume space. We also excluded from computation of FP and FN a layer of voxels one voxel thick closest to the ground truth volume boundary. This made computation more robust, as even small sub-voxel changes in the surface position can lead to inclusion or exclusion of an entire voxel. We also excluded a 10-mm-thick slab centered on the midsagittal plane because FreeSurfer segmentation is known to include dura at this location. The underestimation and overestimation of the pial surface were then expressed using standard FN and FP rates.

Our experiment was performed as follows. We first visually examined the HWA segmentation results of data set 4, processing the left and right hemispheres separately. We selected 15 hemispheres where the pial surface was overestimated (later referred to as problematic) and 15 where this did not occur (referred to as non-problematic). Finally, we

Table 4
Comparison of GCUT with existing skull stripping approaches using data set 2 (20 normal subjects, IBSR).

Method	DS mean (SD) [range]	JS mean (SD) [range]	FP (%) mean (SD) [range]	DS (without dark pixels) mean (SD) [range]	JS (without dark voxels) mean (SD) [range]	FP (without dark voxels, %) mean (SD) [range]	FN (%) mean (SD) [range]	FP_adj (%) mean (SD) [range]
BSE	0.79* (0.21) [0–0.95]	0.69 (0.22) [0–0.90]	**5.1** ** **(3.1)** [2.1–13.0]	0.80** (0.22) [0–0.95]	0.70** (0.22) [0–0.91]	**4.0** * **(3.1)** [1.5–12.7]	27.0** (24.1) [3.5–100]	**0.69** ** **(0.78)** [0–3.12]
BET	0.74* (0.14) [0.53–0.90]	0.61** (0.18) [0.36–0.81]	79.9* (59.3) [22.7–179.4]	0.80** (0.14) [0.59–0.95]	0.68** (0.18) [0.42–0.91]	57.8* (47.7) [9.5–139.3]	0.1** (0.1) [0.0–0.4]	6.42** (2.94) [2.52–13.02]
WAT	0.76* (0.14) [0.47–0.92]	0.64* (0.18) [0.31–0.86]	18.4* (14.1) [5.2–61.2]	0.80* (0.15) [0.52–0.96]	0.70* (0.19) [0.35–0.92]	7.5 (4.2) [3.4–17.0]	24.5** (22.7) [0.1–62.7]	2.23** (1.70) [0.68–6.31]
HWA	0.78* (0.21) [0.16–0.88]	0.68* (0.21) [0.09–0.78]	131.2 (308.2) [19.4–1060.2]	0.90 (0.13) [0.51–0.97]	0.83 (0.17) [0.35–0.94]	26.6 (49.9) [4.8–188.8]	1.9 (6.5) [0.0–28.9]	4.89 (3.82) [1.15–16.31]
GCUT	0.85 (0.09) [0.49–0.90]	0.75 (0.10) [0.33–0.81]	38.3 (40.1) [23.1–207.5]	**0.93 (0.09)** [0.56–0.97]	**0.87 (0.12)** [0.39–0.94]	18.5 (32.5) [5.0–155.3]	**0.01 (0.02)** [0.0- 0.06]	4.29 (2.84) [1.21–12.21]
GCUT_HWA	**0.86 (0.09)** [0.49–0.90]	**0.76 (0.11)** [0.33–0.82]	34.2* (40.9) [16.0–207.5]	0.92 (0.09) [0.56–0.97]	**0.87 (0.13)** [0.39–0.94]	16.7* (32.8) [4.5–155.3]	1.92 (6.53) [0–28.88]	3.47* (2.47) [0.95–12.21]
	Excluding 5 failed volumes for HWA and GCUT_HWA and 1 failed volume for GCUT and BSE							
BSE	0.84 (0.11) [0.62–0.95]	0.73 (0.15) [0.44–0.90]	4.7** (2.6) [2.1–11.5]	0.84** (0.11) [0.62–0.95]	0.74** (0.15) [0.45–0.91]	3.6** (2.4) [1.5–9.7]	23.1** (17.3) [3.51–54.53]	0.69** (0.78) [0–3.12]
HWA	0.86** (0.01) [0.83–0.88]	0.75** (0.02) [0.71–0.78]	32.6** (2.8) [28.1–40.4]	0.95* (0.01) [0.92–0.97]	0.90* (0.02) [0.85–0.94]	10.8* (3.4) [4.8–16.9]	0.012 (0.04) [0–0.17]	6.42** (2.94) [2.52–13.02]
GCUT	0.87 (0.02) [0.83–0.90]	0.77 (0.02) [0.71–0.81]	29.4 (4.3) [23.1–40.8]	0.94 (0.02) [0.89–0.97]	0.89 (0.03) [0.81–0.94]	11.3 (4.5) [5.0–23.2]	0.012 (0.02) [0–0.058]	2.23** (1.70) [0.68–6.31]
GCUT_HWA	0.88** (0.01) [0.86–0.90]	0.79** (0.02) [0.75–0.82]	26.0** (3.0) [21.2–32.6]	0.95* (0.01) [0.93–0.97]	0.91* (0.02) [0.87–0.94]	9.7* (3.3) [4.5–15.4]	0.02 (0.05) [0–0.20]	4.89 (3.82) [1.15–16.31]

Abbreviations are explained in Table 3′s caption.
Bold emphasis designates the best value among all rows.
* $0.001<p<0.05$, where p designates the statistical significance of the difference between current value and GCUT's result.
** $p<0.001$.

Table 5
Comparison of GCUT with existing skull stripping approaches using data set 3 (Siemens Allegra 3T scanner, good quality).

Method	DS mean (SD) [range]	JS mean (SD) [range]	FP (%) mean (SD) [range]	DS (without dark pixels) mean (SD) [range]	JS (without dark voxels) mean (SD) [range]	FP (without dark voxels, %) mean (SD) [range]	FN (%) mean (SD) [range]	FP_adj (%) mean (SD), [range]
BSE	0.63* (0.24) [0–0.89]	0.50** (0.24) [0–0.81]	111.3** (75.6) [20.4–200.5]	0.71* (0.22) [0–0.90]	0.58** (0.21) [0–0.82]	54.0** (31.2) [16.2–102.7]	9.54 (25.03) [2.34–100]	3.45 (1.44) [0.11–6.26]
BET	0.74** (0.10) [0.57–0.85]	0.60** (0.13) [0.40–0.74]	73.3* (44.0) [29.7–151.2]	0.77** (0.09) [0.62–0.85]	0.63** (0.11) [0.45–0.74]	51.2** (31.8) [19.4–112.2]	1.75** (1.47) [0.04–4.30]	5.50** (2.68) [2.93–12.15]
WAT	0.82 (0.10) [0.54–0.89]	0.71 (0.13) [0.37–0.80]	**21.8** (8.4)** [4.7–35.8]	0.83* (0.11) [0.53–0.89]	0.72* (0.13) [0.36–0.80]	**14.9* (5.5)** [2.4–22.6]	13.1* (18.2) [1.18–57.32]	**2.63 (0.89)** [1.02–4.52]
HWA	0.84** (0.01) [0.83–0.85]	0.72** (0.01) [0.70–0.74]	38.2** (2.1) [35.3–42.0]	0.88** (0.01) [0.86–0.90]	0.79** (0.02) [0.76–0.81]	18.2** (2.1) [14.9–21.5]	**0.013 (0.02)** [5e-4–0.05]	3.63** (0.54) [2.86–4.40]
GCUT	**0.86 (0.01)** [0.84–0.87]	0.75 (0.01) [0.73–0.77]	32.8 (2.0) [29.8–36.9]	**0.89 (0.01)** [0.87–0.90]	**0.80 (0.02)** [0.77–0.82]	17.2 (2.0) [14.2–20.6]	0.025 (0.03) [5e-4–0.14]	2.95 (0.42) [2.22–3.55]
GCUT_HWA	**0.86** (0.01)** [0.85–0.87]	**0.76** (0.01)** [0.74–0.77]	31.8** (1.9) [29.1–35.4]	**0.89** (0.01)** [0.87–0.90]	**0.80** (0.02)** [0.77–0.82]	16.9** (1.9) [14.0–20.3]	0.035* (0.04) [0.006–0.16]	2.81** (0.37) [2.16–3.33]

Abbreviations are explained in Table 3's caption.
Bold emphasis designates the best value among all rows.
* $0.001 < p < 0.05$, where p designates the statistical significance of the difference between current value and GCUT's result.
** $p < 0.001$.

compared FN and FP rates of three masks, HWA, GCUT and GCUT_HWA across the two sets of hemispheres, see results in Tables 7 and 8.

Contrary to our expectations, GCUT brain mask performed poorly; it led to a 2.5 fold increase in overestimation in problematic hemispheres. However, the intersection of HWA and GCUT masks, GCUT_HWA, performed very well, halving the FP rate of problematic hemispheres. For non-problematic hemispheres, the intersection mask did not change the FP rate (as expected) and also resulted in a small decrease in FN rate. The reason for such a decrease is unclear; it appears that the large FN rates reported in Tables 7 and 8 were not a result of inappropriate masking but a result of residual intensity nonuniformity that often causes pial surface underestimation on the medial and inferior surfaces of the temporal lobe. The negligible increase in FN rate after using GCUT_HWA mask suggests that the new mask does not lead to greater brain loss than HWA and that used alone, it is fine for non-problematic brains.

Visual examination of pial surfaces segmented with the help of GCUT_HWA showed that overestimation problem was completely resolved in 11 out of the 15 problematic hemispheres. Fig. 12 shows several examples of successful use of GCUT_HWA.

Robustness and sensitivity to algorithm's parameters

In this work, robustness refers to the ability of the algorithm to successfully process images whose appearance or quality substantially deviated from the norm. We applied our algorithm to examples of such images. For example, data set 1 contained a number of images with very poor contrast between gray matter and CSF. Images in data set 2 exhibited various imaging artifacts, such as strong intensity inhomogeneity and ghosting. Data set 4 was selected because it comprised brains that had strong connections between dura and GM as well as pronounced intensity inhomogeneity, characteristics that usually lead to segmentation errors.

To highlight robustness of HWA compared to GCUT we calculated the number of failures of each approach on the four data sets. “Failure” was defined qualitatively as either gross brain loss or preservation of majority of non-brain structures. Using this criterion, GCUT was more robust than HWA; the failures were observed only for data set 2, five for HWA (substantial brain loss or running code errors) and one for GCUT (preservation of almost all non-brain structures). GCUT's failure was on subject “7_8” and was caused by erroneous choice of seed position, inside the neck region rather than in WM. Usually, the neck region is highly heterogeneous, which avoids this problem, but in this particular subject it had a very homogeneous appearance. Interestingly, both HWA and BSE failed on the same subject, the latter due to wrong selection of largest connected component (the neck instead of the brain).

Motivated by the cortical thickness estimation problem, we have found it more informative to define failure quantitatively in terms of a brain mask with either moderately high FN rate (FN>0.1%) or moderately high adjusted FP rate (FP_adj>7%). Such masks are

Table 6
Comparison of GCUT with existing skull stripping approaches using data set 4 (Siemens Allegra 3T scanner, poor quality).

Method	DS mean (SD) [range]	JS mean (SD) [range]	FP (%) mean (SD) [range]	DS (without dark pixels) mean (SD) [range]	JS (without dark voxels) mean (SD) [range]	FP (without dark voxels, %) mean (SD) [range]	FN (%) mean (SD) [range]	FP_adj (%) mean (SD) [range]
BSE	0.66* (0.23) [0–0.87]	0.52** (0.22) [0–0.78]	98.9** (66.4) [26.0–191.2]	0.72* (0.22) [0–0.89]	0.60* (0.20) [0–0.80]	52.9** (30.1) [16.3–103.4]	8.59 (25.30) [1.12–100]	5.10* (2.14) [0.03–9.24]
BET	0.75** (0.10) [0.54–0.87]	0.60** (0.12) [0.37–0.77]	70.7* (42.3) [24.4–168.6]	0.78** (0.08) [0.6–0.88]	0.64** (0.10) [0.44–0.79]	49.8** (28.7) [16.3–116.2]	1.46* (1.79) [0.01–5.98]	6.21* (2.62) [2.79–11.87]
WAT	0.82 (0.09) [0.57-0.91]	0.70 (0.11) [0.40–0.83]	**26.1* (7.8)** [13.1–43.1]	0.83* (0.09) [0.57–0.91]	0.72* (0.12) [0.40–0.84]	**17.0 (4.7)** [8.40–23.55]	12.22* (14.9) [0.29–49.25]	3.76 (1.35) [1.55–6.15]
HWA	0.84** (0.02) [0.80-0.85]	0.72** (0.02) [0.67-0.75]	39.4** (4.6) [33.8–50.1]	0.88 (0.01) [0.85–0.90]	0.79 (0.02) [0.75–0.82]	19.6 (2.6) [13.3–23.2]	0.055 (0.07) [0.0015–0.26]	4.28 (0.79) [3.12–5.79]
GCUT	0.86 (0.02) [0.82–0.89]	0.75 (0.02) [0.70–0.81]	33.6 (4.3) [23.6–43.7]	0.88 (0.02) [0.83–0.90]	0.79 (0.03) [0.71–0.82]	18.7 (3.6) [10.4–27.1]	**0.038 (0.04)** [0.0010–0.13]	3.92 (1.40) [1.80–8.02]
GCUT_HWA	**0.87* (0.01)** [0.84–0.90]	**0.76* (0.02)** [0.73–0.81]	31.1* (3.2) [22.8–37.4]	**0.89* (0.01)** [0.86–0.90]	**0.80* (0.02)** [0.76–0.82]	17.4* (2.5) [10.2–20.2]	0.089* (0.09) [0.0025–0.38]	**3.19* (0.56)** [1.77–4.0]

Abbreviations are explained in Table 3's caption.
Bold emphasis designates the best value among all rows.
* $0.001 < p < 0.05$, where p designates the statistical significance of the difference between current value and GCUT's result.
** $p < 0.001$.

Table 7
Effect of brain masks on subsequent estimation of pial surface position in 15 hemispheres with prior overestimation problem.

	FP (%)				FN (%)		
	HWA	GCUT	GCUT_HWA	GCUT_HWA –HWA[a]	HWA	GCUT	GCUT_HWA
AVG	0.06	0.17	0.03	− 0.03	0.12	0.15	0.13
STD	0.02	0.34	0.03	0.03	0.17	0.16	0.14
p-value				<0.001			

[a] This column designates the difference between the results in GCUT_HWA and HWA columns.

often unsuitable for cortical thickness measurements. By this definition, GCUT was more robust than HWA on data sets 1 (two failures for GCUT vs. three for HWA) and 2 (two failures for GCUT vs. six for HWA) but less robust on data set 3 (one failure for GCUT vs. none for HWA). On data set 4 the two approaches were similar (two failures each).

GCUT was found to be robust to the choice of parameters: the intensity threshold T that is used to obtain preliminary mask and parameter k that controls the contribution of voxel intensities in deciding cut positions. This follows, first of all, from the fact that using the same values ($T=0.36I_{\mathrm{WM}}$ and $k=2.3$) resulted in excellent performance across all four data sets. Further, changes in the parameter values had limited influence on the overall performance (Tables 9 and 10). Changing T led to reciprocal changes in FN and FP rates—decreasing T led to reduction in FP rate and increase in FN rate as more voxels were lost in the preliminary mask (Table 9). Smaller k (higher influence of voxel intensities) led to lower FP rate but at the expense of larger FN rate, due to increased chance of cutting within the brain. Increasing k initially reduced but subsequently increased the FN rate, as a result of wrongly positioned cuts that were no longer guided by voxel intensities.

Discussion

Overall, in terms of similarity indices, FP and FN rates (both conventional and modified), our method (GCUT) performed better than all other approaches. GCUT was superior to BET according to all metrics on all data sets except the first one, where BET achieved somewhat better false positive rates and similarity indices at the expense of almost 2% false negatives compared to almost zero for our approach. Notably, the slight advantage of BET was observed with only one data set out of four.

Similar statements can be made about BSE and WAT. BSE was inferior to GCUT on data sets 3–4 according to all metrics, and superior to GCUT in terms of false positives on data sets 1–2, at the expense of unacceptable brain loss of 6–27%. WAT was consistently superior to GCUT in terms of false positives, but again due to very high FN rate of 2.5–24%. WAT's FN rate exceeded 12% on three data sets out of four.

Our approach was superior to HWA on all data sets and metrics, excluding FN rate, where the two approaches were roughly equivalent and achieved negligible brain loss much lower than that produced by BET, BSE and WAT. On FP adjusted rate, which is most sensitive to preserved dura, GCUT achieved 10–30% reduction compared to HWA. Taking into account other factors as well as effect on the subsequent surface estimation performance, the new skull stripping approach GCUT offers a different balance of advantages and disadvantages compared to HWA.

Advantages:

(1) Cleaner skull strip with less remaining dura (at least 10–30% reduction).
(2) Better robustness of output using legacy data.
(3) Freedom from adependence on shape priors, suggesting possible deployment of this method to developmental studies on humans and for studies involving animal brains.
(4) No need for alignment to standard space, which may fail for strongly misaligned brains.

Disadvantages:

(1) Trivial increase in brain loss for some data sets, which can still be considered negligible for practical applications.
(2) Despite less dura preservation, subsequent segmentation can be more problematic.

A caveat regarding the current study is that performance analyses are based on data obtained from relatively healthy adults. Further work is advisable to determine if the advantages outlined here will generalize to data from patient brains obtained in clinical settings where movement, tissue abnormalities and artifacts could be problematic.

However, the main value of our approach is not its standalone performance but rather its effect on subsequent segmentation when used in conjunction with HWA. As illustrated in the third experiment, the intersection of the two masks can completely solve the overestimation problem (11 out of 15 problematic cases in our study; Fig. 12). This effect can be attributed to GCUT's mask having complimentary properties to that of HWA, as illustrated in Fig. 13.

HWA's main problem appears to be the double boundary between scalp/skull/dura/GM structures, often resulting in inclusion of large chunks of skull/dura mater that run parallel to brain surface. This happens because image intensity information and smoothness (or shape) constraints are combined in a single energy function (Segonne et al., 2004). In the presence of double boundary, inclusion or exclusion of skull/dura mater results in similar mask shapes and equally dark boundary voxels, making it likely for the algorithm to choose the wrong mask, (top two rows of middle column in Fig. 13). In GCUT, the intensity information is used to create a preliminary mask after which the shape constraints are imposed by performing cuts. As the dura mater connections with GM are irregular in the preliminary mask, they are likely to be cut, reducing the likelihood of dural inclusion (top two rows of right column in Fig. 13).

On the other hand, HWA's ability to explicitly impose smoothness and shape constraints results in more regular brain mask shapes, which is particularly important when closely adherent skull/dura shares the same signal intensity as GM and becomes inseparable from GM (row 3 of Fig. 13). In such cases GCUT would make a wrong cut, resulting in somewhat irregular mask shape. Intersecting the two masks appears to resolve both problems and significantly decrease the overestimation of the pial surface (Fig. 12).

Table 8
Effect of brain masks on subsequent estimation of pial surface position in 15 hemispheres without prior overestimation problem.

	FP (%)			FN (%)		
	HWA	GCUT	GCUT_HWA	HWA	GCUT	GCUT_HWA
AVG	0.02	0.02	0.02	0.47	0.42	0.31
STD	0.01	0.01	0.01	0.38	0.30	0.26

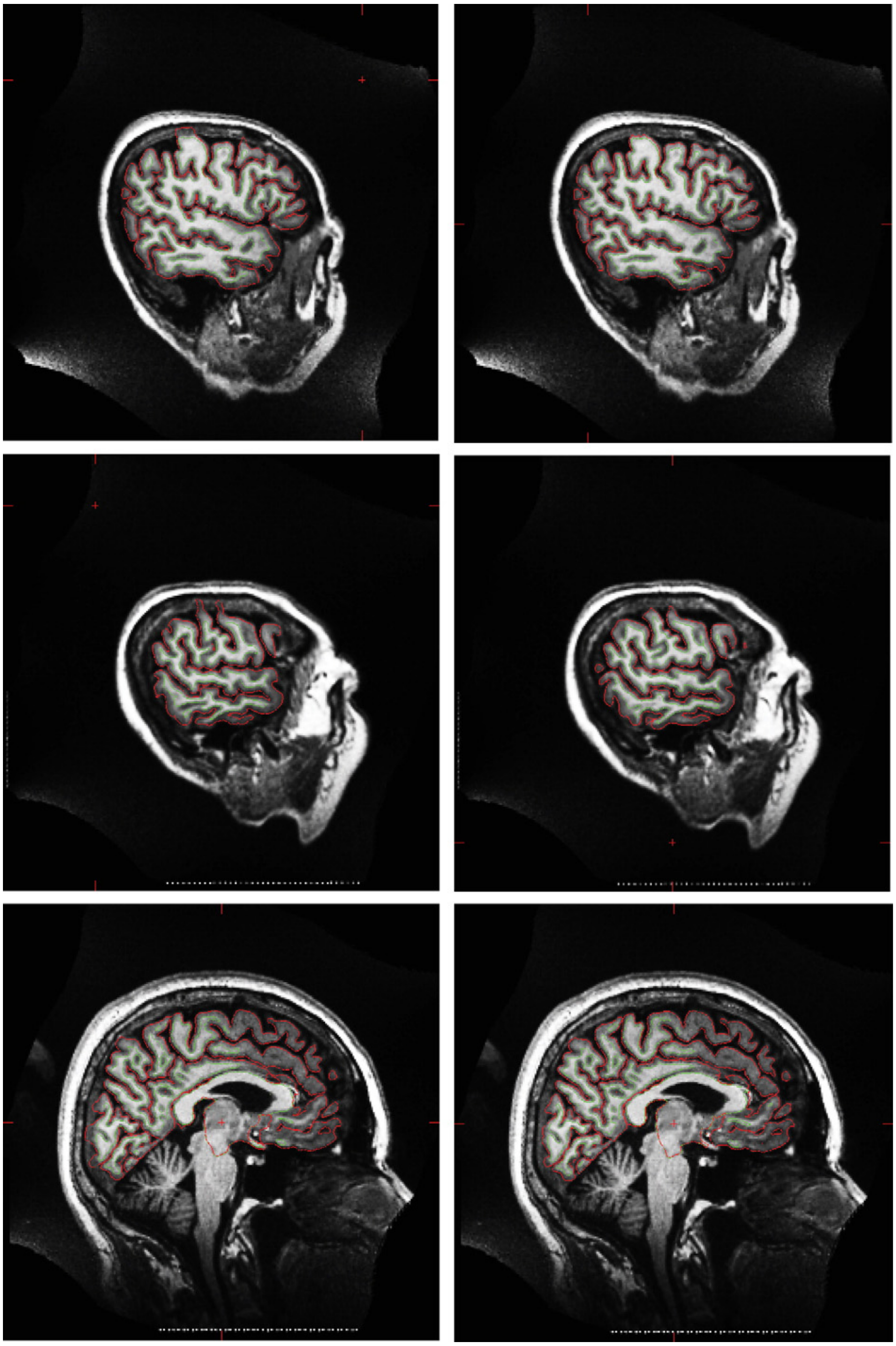

Fig. 12. Problematic FreeSurfer segmentation performance using HWA brain mask (left) is improved using GCUT_HWA (right).

It is likely that multispectral segmentation techniques made more feasible by availability of fast 3D T2W imaging (Hennig et al., 2003) will result in further improvement with respect to dura mater removal. An alternative strategy is to use multiple echoes to create intensity differences between dura adjacent to GM, particularly in the medial temporal region (van der Kouwe et al., 2008). While this is clearly promising, analyzing the vast amount of legacy data available will still benefit from the technique we describe.

Table 9

Sensitivity of GCUT performance to intensity threshold parameter for $k = 2.3$.

Data set	T (% of I_{WM})	30	32	34	**36**	38	40
1	FN (%)	0.015	0.021	0.018	**0.031**	0.0389	0.0726
	FP_adj (without dark voxels, %)	5.22	4.57	3.73	**2.94**	2.37	1.81
2	FN (%)	0.004	0.006	0.010	**0.012**	0.015	0.02
	FP_adj (without dark voxels, %)	16.30	14.43	12.81	**11.28**	9.73	8.76
3	FN (%)	0.023	0.023	0.023	**0.025**	0.029	0.0369
	FP_adj (without dark voxels, %)	4.28	3.74	3.41	**2.95**	2.68	2.45
4	FN (%)	0.030	0.030	0.033	**0.038**	0.058	0.086
	FP_adj (without dark voxels, %)	6.04	5.15	4.47	**3.92**	3.49	3.09

Bold emphasis designates the best value among all rows.

Conclusion

We proposed a new skull stripping approach that builds upon earlier work (Atkins and Mackiewich, 1998; Lemieux et al., 1999),

Table 10
Sensitivity of GCUT performance to parameter k controlling the influence of voxel intensity on cut positions for $T = 0.36\ I_{WM}$.

Data set	K	1	1.5	2	**2.3**	2.5	3
1	FN (%)	0.129	0.086	0.041	**0.029**	0.023	0.085
	FP_adj (without dark voxels, %)	2.80	2.86	2.93	**3.02**	3.14	3.30
2	FN (%)	0.016	0.013	0.011	**0.012**	0.013	0.029
	FP_adj (without dark voxels, %)	3.39	3.37	3.60	**3.87**	3.91	3.81
3	FN (%)	0.034	0.027	0.025	**0.025**	0.023	0.019
	FP_adj (without dark voxels, %)	2.91	2.92	2.96	**2.95**	2.98	3.14
4	FN (%)	0.092	0.052	0.049	**0.038**	0.035	0.069
	FP_adj (without dark voxels, %)	3.64	3.78	3.91	**3.92**	4.01	4.30

Bold emphasis designates the best value among all rows.

which used intensity thresholding followed by cutting of narrow connections to obtain a brain mask. Our approach consists of three steps: intensity thresholding, refining of the initial mask by cutting false connections between brain and non-brain structures, and post-processing to recover CSF and partial volume voxels.

Instead of using mathematical morphology for false connection removal (Atkins and Mackiewich, 1998; Lemieux et al., 1999), a method that only cuts sufficiently narrow connections, our algorithm uses a superior graph cuts approach that is capable of locating precise cut positions of varying widths.

By itself, our approach offers a good alternative to HWA—it is more robust on legacy data, can work on a larger variety of brain shapes and achieves at least 10–30% reduction in residual dura without significant increase in brain tissue removal. The greatest benefit of using our approach is realized when it is employed in conjunction with HWA, for example by using a simple intersection of the two masks. The errors produced by the two masks are complementary, resulting in significant improvement of subsequent segmentation performance when the masks are combined. In our experiments, the combined application of the two techniques resulted in the successful segmentation of 11 out of 15 volumes that were not adequately segmented using HWA alone.

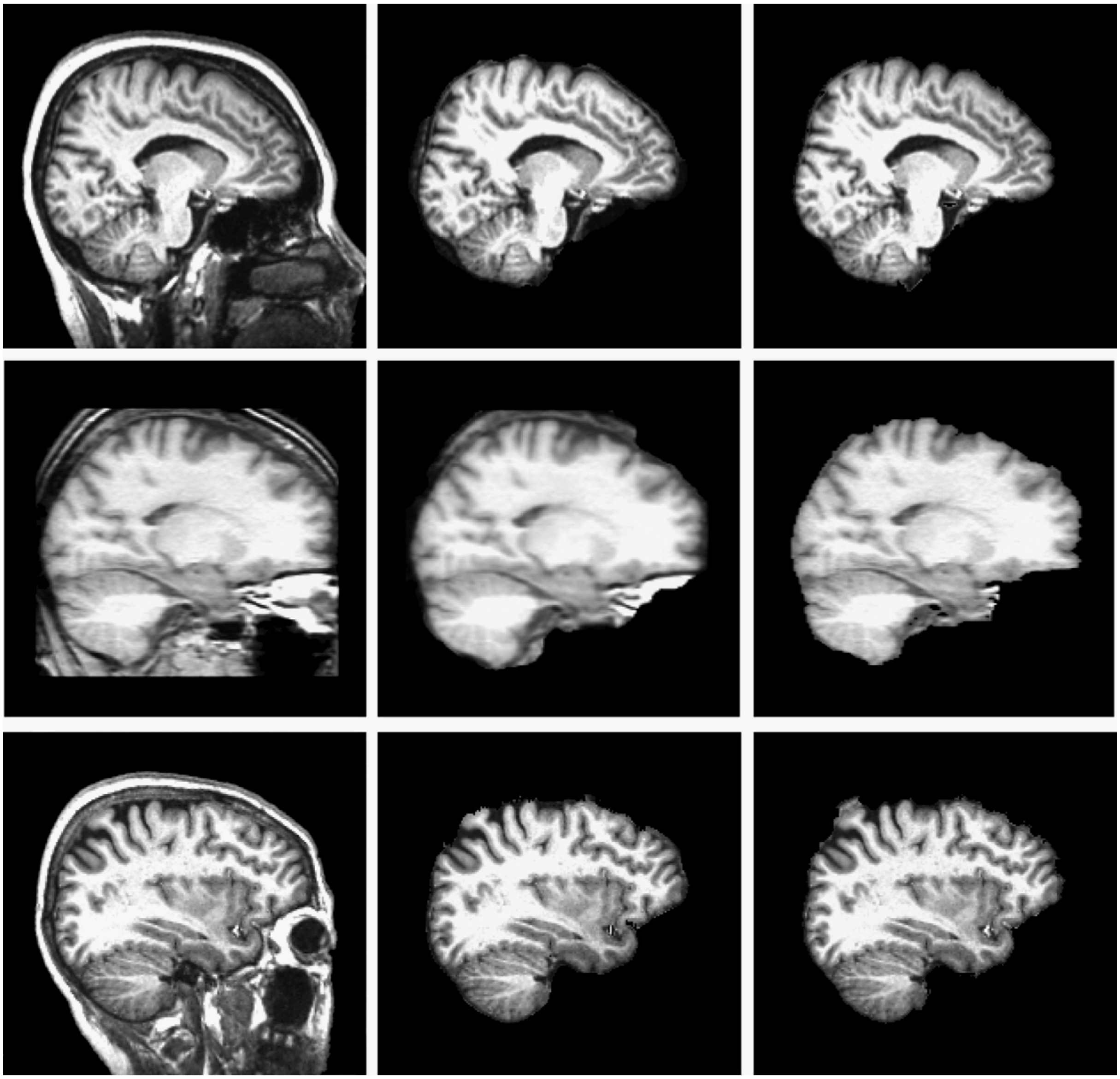

Fig. 13. Typical errors exhibited by HWA (middle) and GCUT (right). HWA is often confused by double boundary between scalp/dura/GM, resulting in inclusion of large chunks of skull/dura mater, rows 1–2. GCUT fails to cut connections where there is no noticeable intensity separation between dura and GM, row 3.

Acknowledgments

This work was 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

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11, 805–821.

Atkins, M.S., Mackiewich, B.T., 1998. Fully automatic segmentation of the brain in MRI. IEEE Trans. Med. Imag. 17, 98–107.

Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., Luders, E., Rottenberg, D., 2004. Quantitative comparison of four brain extraction algorithms. NeuroImage 22, 1255–1261.

Boykov, Y., Jolly, M.P., 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. pp. 105-112.

Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intelli. 26, 1124–1137.

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

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

Fein, G., Landman, B., Tran, H., Barakos, J., Moon, K., Di Sclafani, V., Shumway, R., 2006. Statistical parametric mapping of brain morphology: sensitivity is dramatically increased by using brain-extracted images as inputs. NeuroImage 30, 1187–1195.

Fennema-Notestine, C., Ozyurt, I.B., Clark, C.P., Morris, S., Bischoff-Grethe, A., Bondi, M.W., Jernigan, T.L., Fischl, B., Segonne, F., Shattuck, D.W., Leahy, R.M., Rex, D.E., Toga, A.W., Zou, K.H., Brown, G.G., 2006. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Hum. Brain Mapp. 27, 99–113.

Grady, L., 2006. Fast, quality, segmentation of large volumes—isoperimetric distance trees. ECCV 2006. Lect. Notes Comput. Sci. 3, 449–462.

Grady, L., Schwartz, E.L., 2006. Isoperimetric graph partitioning for image segmentation. IEEE Trans. Pattern Anal. Mach. Intelli. 28, 469–475.

Hahn, H.K., Peitgen, H.O., 2000. The skull stripping problem in MRI solved by a single 3D watershed transform. Medical Image Computing and Computer Assisted Intervention (MICCAI). Lect. Notes Comput. Sci. 134–143.

Hartley, S.W., Scher, A.I., Korf, E.S.C., White, L.R., Launer, L.J., 2006. Analysis and validation of automated skull stripping tools: a validation study based on 296 MR images from the Honolulu Asia aging study. NeuroImage 30, 1179–1186.

Hata, Y., Kobashi, S., Hirano, S., Kitagaki, H., Mori, E., 2000. Automated segmentation of human brain MR images aided by fuzzy information granulation and fuzzy inference. IEEE Trans. Syst. Man Cybern., Part C Appl. Rev. 30, 381–395.

Hennig, J., Weigel, M., Scheffler, K., 2003. Multiecho sequences with variable refocusing flip angles: optimization of signal behavior using smooth transitions between pseudo steady states (TRAPS). Magn. Reson. Med. 49, 527–535.

Huang, A., Abugharbieh, R., Ram, R., Traboulsee, A., 2006. MRI brain extraction with combined expectation maximization and geodesic active contours. IEEE International Symposium on Signal Processing and Information Technology.

Kapur, T., Grimson, W.E., Wells III, W.M., Kikinis, R., 1996. Segmentation of brain tissue from magnetic resonance images. Med. Image Anal. 1, 109–127.

Kim, J.S., Singh, V., Lee, J.K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J.M., Kim, S.I., Evans, A.C., 2005. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. NeuroImage 27, 210–221.

Lee, J.-M., Yoona, U., Namb, S.H., Kima, J.-H., Kima, I.-Y., Kim, S.I., 2003. Evaluation of automated and semi-automated skull-stripping algorithms using similarity index and segmentation error. Comput. Biol. Med. 33, 495–507 2003.

Lemieux, L., Hagemann, G., Krakow, K., Woermann, F.G., 1999. Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. Magn. Reson. Med. 42, 127–135.

Niessen, W.J., Vincken, K.L., Weickert, J., Romeny, B.M.T.H., Viergever, M.A., 1999. Multiscale segmentation of three-dimensional MR brain images. Int. J. Comput. Vis. 31, 185–202.

Park, J., Keller, J.M., 2001. Snakes on the watershed. IEEE Trans. Pattern Anal. Mach. Intelli. 23, 1201–1205.

Rehm, K., Schaper, K., Anderson, J., Woods, R., Stoltzner, S., Rottenberg, D., 2004. Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. NeuroImage 22, 1262–1270.

Rex, D.E., Shattuck, D.W., Woods, R.P., Narr, K.L., Luders, E., Rehm, K., Stolzner, S.E., Rottenberg, D.A., Toga, A.W., 2004. A meta-algorithm for brain extraction in MRI. NeuroImage 23, 625–637.

Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. NeuroImage 22, 1060–1075.

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.

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

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.

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.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage Math. Brain Imag. 23, S208–S219.

Stokking, R., Vincken, K.L., Viergever, M.A., 2000. Automatic morphology-based brain segmentation (MBRASE) from MRI-T1 data. NeuroImage 12, 726–738.

van der Kouwe, A.J.W., Benner, T., Salat, D.H., Fischl, B., 2008. Brain morphometry with multiecho MPRAGE. NeuroImage 40, 559–569.

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