Published in final edited form as:

Neuroimage. 2011 June 15; 56(4): 1982–1992. doi:10.1016/j.neuroimage.2011.03.045.

SIMPLE PARADIGM FOR EXTRA-CEREBRAL TISSUE REMOVAL: ALGORITHM AND ANALYSIS

Aaron Carass^{*,a}, Jennifer Cuzzocreo^b, M. Bryan Wheeler^a, Pierre-Louis Bazin^b, Susan M. Resnick^c, and Jerry L. Prince^a

^a Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD

^b Department of Radiology & Radiological Science, The Johns Hopkins University, Baltimore, MD

^c Laboratory of Personality & Cognition, National Institute on Aging, Baltimore, MD

Abstract

Extraction of the brain — *i.e.* cerebrum, cerebellum, and brain stem — from T1-weighted structural magnetic resonance images is an important initial step in neuroimage analysis. Although automatic algorithms are available, their inconsistent handling of the cortical mantle often requires manual interaction, thereby reducing their effectiveness. This paper presents a fully automated brain extraction algorithm that incorporates elastic registration, tissue segmentation, and morphological techniques which are combined by a watershed principle, while paying special attention to the preservation of the boundary between the gray matter and the cerebrospinal fluid. The approach was evaluated by comparison to a manual rater, and compared to several other leading algorithms on a publically available data set of brain images using the Dice coefficient and containment index as performance metrics. The qualitative and quantitative impact of this initial step on subsequent cortical surface generation is also presented. Our experiments demonstrate that our approach is quantitatively better than six other leading algorithms (with statistical significance on modern T1-weighted MR data). We also validated the robustness of the algorithm on a very large data set of over one thousand subjects, and showed that it can replace an experienced manual rater as preprocessing for a cortical surface extraction algorithm with statistically insignificant differences in cortical surface position.

Keywords

Brain extraction; skull stripping; watershed principle; segmentation; medical image processing

Introduction

Quantitative neurological image processing based on structural magnetic resonance images (MRI) typically requires the preliminary step of isolating the brain from non-brain tissues. This *skull stripping* or *brain extraction* step is equivalent to a whole brain segmentation that

© 2011 Elsevier Inc. All rights reserved.

*Please address correspondence to: Aaron Carass, Dept. of Electrical and Computer Engineering, The Johns Hopkins University, 105 Barton Hall, 3400 N. Charles St., Baltimore, MD 21218., aaron_carass@jhu.edu, Phone: 410-516-6820 Fax: 410-516-5566.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

correctly classifies the gray matter (GM) and white matter (WM) of the cerebrum, cerebellum, and brain stem from other tissues such as cerebrospinal fluid (CSF), skull, meninges, etc. In spite of the wide variety of automatic approaches that have been proposed (Ward, 1999; Goldszal et al., 1998; Kapur et al., 1996; Ashburner and Friston, 2000; Dale et al., 1999; Smith, 2002; Ségonne et al., 2004; Shattuck et al., 2001; Hahn and Peitgen, 2000; Rex et al., 2004; Rehm et al., 2004; Shan et al., 2002; Lemeiux et al., 1999; Yoon et al., 2001; Lee et al., 2003; Boesen et al., 2004; Fennema-Notestine et al., 2006; Sadananthan et al., 2010), the gold standard for skull stripping remains that of the human rater. This is not a satisfactory situation for two reasons: human raters invariably introduce unintended biases into their work and can be prone to errors from overwork and they also require considerable time to perform the require task, anywhere from two to four hours depending on the quality and resolution of the image and the experience of the human rater.

There are several key difficulties to overcome in the development of an automated robust approach to the problem of extra-cranial tissue removal. The primary difficulty is the lack of discernible contrast between many of the tissue types that compose the extra-cerebral tissue and the brain. Because of this, automated algorithms will often include non-brain tissues in the resulting brain mask or cut off brain tissues—particularly cortical gray matter—by accident. This difficulty is present in the example shown in Fig. 1. Here, the original image in panel (a) is shown stripped by an experienced human rater in (b) and then by the Brain Extraction Tool (BET version 2.1) (Smith, 2002) and our approach in (c) and (d), respectively. Companion zoomed images are shown for comparison in (e), (f), (g), and (h). Fig. 1(i), (j), and (k) provide additional representation of the different masks generated by our human expert (i), BET (j), and our algorithm (k), with the masks used as a color overlay on the original MR image. Another problem arises from the handling of different MR contrasts (e.g., T1-weighted, T2-weighted, etc.) and even the different intensity contrast or dynamic range within a single type of acquisition. Other difficulties include patient orientation differences, wrap-around artifacts, pixel resolution differences, and the scan field of view.

There have been several studies concerned with the comparison and consistency of raters as well as the more widely distributed automated algorithms (Lee et al., 2003; Boesen et al., 2004; Fennema-Notestine et al., 2006; Hartley et al., 2006). The general conclusions are that each automated algorithm may have utility but a thorough exploration of the biases and suitability of a given method should be explored before adopting any approach. Also, the choice of skull stripping technique should be motivated by the subsequent processing of the data. Consider the case of a PIB/PET imaging study which uses gray matter to help normalize the PIB/PET image data (Davis et al., 2003; Lowe et al., 2009), would require a reasonable representation of the gray matter, that would be further segmented. Whereas in segmentation of the internal capsule preservation of the peak intensity of white matter is a key objective so that it can be used for histogram equalization (Maldjian et al., 2001). Unfortunately, to date, there have been few studies detailing how human raters or automatic algorithms influence the performance of a neuroimage processing pipeline. One recent paper explored the effects of skull stripping on intensity inhomogeneity correction on GM segmentation and voxel based morphometry analysis (Acosta-Cabronero et al., 2008). They drew the conclusion that voxel based morphometry benefits from preprocessing with BET (Smith, 2002), over alternative methods (Shattuck et al., 2001; Ségonne et al., 2004), however the differences were not of a significant nature. Of course, as imaging studies and cohorts increase in size and complexity it becomes evermore difficult for minor processing errors to be noticed, thus increased robustness is of increased importance. Ultimately, it is the analysis of this kind of endpoint that determines whether a skull stripping approach is useful in a given application.

In this paper, we present *Simple Paradigm for Extra-Cerebral Tissue REmoval* (SPECTRE), a brain extraction algorithm that combines elastic registration, tissue segmentation, and morphological techniques, all guided by a novel watershed principle. SPECTRE is specifically designed to retain cortical gray matter so that subsequent processing designed to find the cortex will not be forced into making errors due to skull stripping mistakes. SPECTRE was designed for use with T1-weighted images, and this paper is exclusively concerned with an exposition of SPECTRE on T1-weighted data. However after a simple modification, described below, it can be applied to both T2 and PD weighted data.

The performance of SPECTRE is evaluated in three ways: firstly by comparison against six other skull stripping algorithms on the 38 subjects available from the Internet Brain Segmentation Repository at the Center for Morphometric Analysis (Center for Morphometric Analysis (CMA), 1995; Tsang et al., 2008); secondly by comparing SPECTRE to a human rater on a data set of 1046 scans; and thirdly by assessing the effect of skull stripping—manual versus SPECTRE—on the accuracy of finding the inner and outer cortical surfaces using the CRUISE neuroimage processing pipeline (Han et al., 2004) applied to the skull-stripped data.

We note that a preliminary version of SPECTRE was previously reported in a conference paper (Carass et al., 2007). The present paper provides a complete description of the algorithm, and presents a more complete set of validation experiments.

Background

Existing brain extraction algorithms fall into four broad classes. Firstly, there are the *morphology based* techniques, which are based on successive morphological operations on the MRI volume. Classically these methods have used user input to determine certain thresholds, the region of interest, or a seed for a region growing scheme. Generally the morphological operations of dilation and erosion are repeated, until a stopping criterion is satisfied or a user-specified end point is reached (Ward, 1999; Goldszal et al., 1998; Sandor and Leahy, 1997; Park and Lee, 2009).

The second class of algorithms is the *atlas based methods*; they register existing brain segmentations—or templates—into the subject in order to determine the brain region to extract. Existing methods use either affine registration (Ashburner and Friston, 2000) or higher dimensional registration methods (Kapur et al., 1996). Of course, all such methods incur the risk of bias towards the atlas images and there remains the question of “How many atlases are required to be representative of a population?”

Deformable surfaces best describe the third class of methods. These approaches, such as BET, evolve a surface using various forces to find the boundary of the brain from an initialization based on user input or other criteria such as the center of mass of the whole head. The brain extraction is thus mostly data-driven, but the initialization of the deformable surface can be critical (Dale et al., 1999; Smith, 2002).

The fourth class of brain extraction methods can be identified as *hybrid methods* because they comprise some combination of the first three classes. Many algorithms that could be previously classified distinctly in one of the first three groups are now being refactored to incorporate some benefit of the other methods (cf. (Shattuck et al., 2001; Ségonne et al., 2004)). The basic idea is to combine the skull stripping results from different approaches to cancel out the bias or inaccuracies inherent to one of the other methods. The BEMA (Rex et al., 2004) approach, for example, does this in a very explicit manner. SPECTRE is best described as a hybrid method as it is a combination of segmentation, registration, and morphological operations.

Objectives

The original goal of our algorithm was to provide an accurate segmentation of the brain as input to CRUISE (Han et al., 2004), a neurological image processing pipeline that reconstructs a precise surface representation of the cerebral cortex (Tosun et al., 2006). Because the goal of CRUISE is to reconstruct a precise cortical boundary representation—both GM/WM and GM/CSF boundaries—it is essential that the brain extraction preprocessing step removes skull, meninges, and other tissues without excising any cortical gray matter. In addition, because the CSF surrounding the cortical surface is used to define the GM boundary, we devised SPECTRE to consistently include a thin layer of CSF around the brain whenever feasible. In this context, feasible means we retain CSF around the brain where there is available CSF to retain, meaning that if there is at least one voxel of CSF between brain and non brain tissues we include it our mask.

With this objective in mind, how should we report the efficacy of our method? Firstly, in keeping with the literature, we report the *Dice coefficient* (Dice, 1945), which measures the degree of overlap between two sets of voxels, one corresponding to the brain extracted automatically (either SPECTRE or another automated method) and the other being the result from a human rater. In keeping with our objective to retain all brain including a small layer of CSF, we also report the *containment index* (CI), which is a measure of the fraction of the human rater's result that is included in the result of an automatic algorithm. Finally, since the ultimate objective is to find the cortex of the brain, we also compute the distances between manually selected cortical landmarks and the cortical surfaces that are found by CRUISE when initialized using brains extracted by either our automatic method or by hand.

Method

Assuming our method is to be applied to T1-weighted MR images (which covers the vast majority of neuroimaging scenarios), the following *watershed principle* applies:

There is a path connecting any two GM/WM voxels, such that from the point of highest intensity along the path to either end point the intensity is never increasing.

This principle can be understood by taking a slice of a T1-weighted MRI data set and viewing its intensities as heights (see Fig. 2). A similar connectivity principle can be stated for PD and T2-weighted images, with due care paid to the nature of T2-weighted images. As in the *fall set* watershed principle of Rutovitz (Rutovitz, 1978), voxels are considered to be connected if they are on the same *hill* and disconnected if they are separated by a *valley*. This implies that we can extract the brain by first identifying the peaks within the white matter and then descend the hill until encountering a valley, which must be the subarachnoid space containing cerebrospinal fluid (CSF). Although the two-dimensional visualization in Fig. 2 aids our understanding, the actual process is carried out in three dimensions. Further, although our overall computational process is guided by this principle, there are several additional steps that need to be taken in order to provide robust and accurate results.

Algorithm

Fig. 3 provides a flowchart overview of SPECTRE. The first objective is to identify the *crest of the hill* corresponding to the connected component comprising both the gray matter (GM) and white matter (WM) together. To achieve this, we first use the *adaptive bases algorithm* (ABA) (Rohde et al., 2003) to deformably register N atlas brains to the subject brain (see **B** in Fig. 3) using the mutual information similarity criterion. Each atlas comprises both an image and a manually delineated mask, which in this paper is the full brain comprising

cerebrum, cerebellum, and brain stem together. In order to save computation time, each registration task is run after downsampling both atlas and subject image by a factor of four.

Each mask is transformed into subject space based on its corresponding registration result and the masks are then combined into a *probability image*, where the value of the i th voxel is

$$P_i = \begin{cases} \frac{\# \text{Masks that include } i}{N}, & \text{when } \frac{\# \text{Masks that include } i}{N} \geq \frac{1}{2}, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

as illustrated in Fig. 4(b).

The second step used in computing an initial mask is to generate a tissue segmentation of the whole head (see **C** in Fig. 3). FANTASM, a robust tissue classification method based on the fuzzy c-means methodology (Pham, 2001), is applied using four classes. The resulting classification result represents an approximation to the tissue classes: Γ_1 comprising CSF, bone, and background, Γ_2 comprising mostly gray matter, Γ_3 comprising mostly white matter, and Γ_4 comprising mostly skin and adipose tissue. Fig. 4(c) shows a result of this classification process.

We combine the registration and segmentation results to compute our initial mask as follows:

$$\mathcal{M}_1 = \{i | i \in \Gamma_2, P_i = 1\} \cup \left\{ i | i \in \Gamma_3, P_i \geq \frac{N-1}{N} \right\}. \quad (2)$$

While providing a respectable representation of the GM and WM, this mask (shown in Fig. 4(d)) may include extraneous tissues such as dura or meninges in the sub-arachnoid space or portions of the brain stem. To address this, we perform a morphological erosion operation four times using a structuring element of size 1 mm and then retain the largest 6-connected component within that result. Fig. 4(e) shows the mask that remains after this process; it represents a “core” from which the final mask will be grown using our watershed principle.

The next major step in our approach is *hill descent*. We do this in a robust manner combining the registration and segmentation results together with the underlying data. Starting from the eroded initial mask, we add a voxel i on the boundary of the mask $\mathcal{M}$ if it meets either of the following two criteria:

1. $i \in \Gamma_2 \cup \Gamma_3, I_j \geq I_i, P_i \geq \frac{1}{N},$
2. $i \in \Gamma_1 \cup \Gamma_2 \cup \Gamma_3, I_j > I_i,$

with $j \in \mathcal{M}, j \in \Gamma_2 \cup \Gamma_3$ and $i \in \mathcal{N}_{18}(j)$, where I_i denotes the intensity of voxel i and $\mathcal{N}_{18}(j)$ is the 18-connected neighborhood of voxel j . Fig. 4(f) shows the mask after the *hill descent*.

The first condition allows us to grow the mask in regions of non-increasing intensity assuming the appropriate tissue classes and the probability mask is encouraging us to grow. The second condition allows us to expand the mask up to the inclusion of a CSF voxel only if the neighbor which we are descending from, j , is in the GM/WM tissue classes and we are strictly descending in image intensity. Thus, if CSF is evident outside the pial surface, then

the mask will grow to incorporate no more than one voxel of sulcal CSF. We allow the mask to dilate in this fashion until there are no more voxels which satisfy either growth condition.

There may still be some voxels that have not been included in the mask which we desire to be added. There are two such types of voxels which we are concerned with, those that are perturbed by noise within the pial surface and cannot be added during the *hill descent* and those WM voxels which have been misclassified during the tissue segmentation by FANTASM to be in the class meant to contain adipose, dura, and skin (Γ_4). Both of these cases comprise a small number of voxels and we handle them by performing a *topologically constrained morphological closing*. A traditional morphological closing operation consists of a dilation followed by an erosion using the same structuring element. With our topologically constrained morphological closing we first carry out a dilation, then perform a hole-filling, and then carry out an erosion. To think about this topologically, the hole filling will make the mask topologically equivalent to a sphere, which is a desirable property for a brain mask to have. Both the erosion and the dilation use a cubic 1mm structuring element. The hole filling finds the largest 6-connected component that is background and sets all other background components to be part of the mask. The rationale behind doing this as opposed to just a hole filling is simply to avoid those situations where the misclassified voxels are 6-connected to the background.

The next step is to do a “sanity check” of the mask to ensure that there have been no obvious failures. We calculate the volume of the current mask ($\mathcal{M}_S$) and compare it to the volume of the mask given by the ABA registrations ($\mathcal{M}_{ABA}$). We do not expect the mask to be too much smaller than the ABA registered mask, nor do we expect it to be significantly larger than this mask. If the following inequality holds:

$$0.9 \times \# \mathcal{M}_{ABA} < \# \mathcal{M}_S < 1.2 \times \# \mathcal{M}_{ABA}, \quad (3)$$

then we assume the mask is reasonable.

In the event that the inequalities are not satisfied, then we repeat the algorithm in the following manner:

1. Create the initial mask from the ABA registration and FANTASM segmentation in the same way as stated previously. Essentially proceeding from the mask denoted as $\mathcal{M}_1$.
2. Erode the initial mask but decrement by one the number of erosions used.
3. Perform the *hill descent* in an identical fashion as stated above.

We repeat this hierarchical approach until either the above inequality (Eqn. 3) is satisfied or the number of erosions applied to the initial mask is zero. As we start with the number of erosions set to three, we will at most be going through this cycle four times. It might be argued that one could selectively increase or decrease the repetition of the erosion of the initial mask based on which inequality was not satisfied. It has, however, been our experience that those problematic data sets are a result of too much erosion, usually to the point of leaving only one hemisphere of the cerebral cortex and thus make it difficult to properly recover a correctly skull stripped data set. The final result of our approach is shown in Fig. 5(d), with the original input being Fig. 5(a). For comparison the results of two human raters on the same subject are shown in Fig. 5(b) and (c).

Experiments

In this section we present three comprehensive experiments to demonstrate the robustness and accuracy of our approach. The first experiment compares seven algorithms including SPECTRE on the 38 subjects available from the Internet Brain Segmentation Repository (IBSR) at the Center for Morphometric Analysis (Center for Morphometric Analysis (CMA), 1995; Tsang et al., 2008), with the objective of establishing a benchmark for accuracy of our algorithm and a comparison to the state of the art. The second experiment compares SPECTRE to a manual rater on a large data set comprising 1046 scans. The goal is to establish whether there is a statistical difference in the algorithm and manual rater performances as measured by Dice coefficient and containment index and to demonstrate the robustness of our algorithm. The third experiment studies the impact of brain extraction on the performance of a cortical surface segmentation algorithm that accepts the skull stripped brain as input. The goal of this experiment is to determine whether the automatic algorithm can replace the human rater when a complex postprocessing pipeline is involved and a precise end measurement is required.

The Dice coefficient is the most commonly used volumetric measure for comparing the quality of automatic brain extraction. The Dice coefficient between the SPECTRE output mask $\mathcal{M}_S$ and that of a “gold standard” segmentation $\mathcal{M}_G$ is defined by

$$\text{Dice}(\mathcal{M}_S, \mathcal{M}_G) = \frac{2|\mathcal{M}_S \cap \mathcal{M}_G|}{|\mathcal{M}_S| + |\mathcal{M}_G|}. \quad (4)$$

Since we are particularly interested in retaining all cortical gray matter, we will also report the Containment Index (CI), which is a measure of how much of the “gold standard” mask is contained within the SPECTRE mask. It is defined as

$$\text{CI}(\mathcal{M}_S, \mathcal{M}_G) = \frac{|\mathcal{M}_S \cap \mathcal{M}_G|}{|\mathcal{M}_G|}. \quad (5)$$

The Dice coefficient, which is a measure of set agreement, has a range of [0, 1]. A Dice coefficient of 1.0 corresponds to perfect agreement between the algorithm and the rater, while a score of 0.0 represents complete discord between the two. An example of a Dice score of 0.88481 is shown in Fig. 6, the score is based on a comparison between Fig. 6(b) and Fig. 6(c). CI also has a range of [0, 1]. A CI of 1.0 means the algorithm result completely contains the rater result, and a value of 0.0 implies that the algorithm failed to contain any portion of the rater’s result; in essence it failed to identify any portion of the image correctly. The CI for the example shown in Fig. 6 is 0.99779. The proximity of this value to 1.0 indicates that the automated algorithm almost completely contains the human rater’s mask, which can be clearly seen in the figure.

We also use the paired t -test to see if we can distinguish results on the populations. The paired t -test is used to demonstrate that there is a statistically meaningful difference between different sets of results, either between human raters and our algorithm, or between our approach and other skull stripping methods. Unless otherwise stated, the significance level of all used t -tests is 0.001. Additionally, we use the Wilcoxon Rank Sum (Wilcoxon, 1945) test to test the null hypothesis that two populations have the same continuous distribution. We have also computed the co-efficient of variation, where appropriate, which is the standard deviation divided by the mean and is usually reported as a percentage — *i.e.* scaled

by 100. Coefficient of variation, values do not have a global scale, the results depend on the nature of the distributions, though smaller values are considered to show less variation and are therefore more stable.

Experiment 1

In this experiment, we compared SPECTRE with six leading skull stripping algorithms, described below, on the Internet Brain Segmentation Repository (IBSR) data sets. Portions of these results were previously presented in the work of Sadananthan et al. (2010). The IBSR data is comprised of two cohorts:

1. IBSR Set 1: 18 T1-weighted images with, slice thickness 1.5 mm.
2. IBSR Set 2: 20 T1-weighted images with, slice thickness 3.1 mm.

GCUT (Sadananthan et al., 2010) is a graph based approach to skull stripping. The method, at its core, uses a cutting algorithm based on isoperimetric graph partitioning (Grady, 2006), in which the image is treated as a undirected graph and the edge connections are cut so as to minimize the isoperimetric ratio. The steps of GCUT are to threshold the image to obtain an initial mask, removal of narrow connections (between brain and non-brain tissues) and a post-processing step to catch any partial volume GM voxels removed by the thresholding step. BET, the brain extraction algorithm (Smith, 2002), uses a deformable surface model that evolves to fit the brain's surface. BSE, brain surface extractor (Shattuck et al., 2001), is an edge based approach that uses anisotropic diffusion filtering derived from a 2D Marr-Hildreth operator. WAT, the watershed approach (Hahn and Peitgen, 2000), is an intensity-based approach which performs a watershed to an intensity inverted image to all slices and all orientations of an image, then selecting a basin to represent the brain. HWA, the hybrid watershed algorithm (Ségonne et al., 2004), combines the watershed algorithm with a deformable model which adds shape based constraints to guarantee brain like structure. GCUT-HWA denotes the mask generated by intersecting the masks of GCUT and HWA, it is argued in (Sadananthan et al., 2010) that the two algorithms differ in their errors in different regions thus the intersection of the two methods should give a more robust result.

The results for this experiment are shown in Tables 1–3 and in Figs. 7 and 8. Table 1 show the results of the comparison, on IBSR Set 1, between SPECTRE and the six named methods, the details of how each of the methods were run, including any non-standard parameter choices, are given in (Sadananthan et al., 2010). With respect to the Dice Coefficient, SPECTRE, followed by BET, are the best performers in this comparison. We can also see from Fig. 7, that SPECTRE ranks highest on 11 out of the 18 data sets. Paired *t*-tests between SPECTRE and each of the other algorithms (see Table 1) demonstrate that there is a statistically significant difference between three of the methods and SPECTRE with respect to the Dice score. The methods that do not reach statistical significance are BSE, BET, and WAT. All had mean Dice scores below that of the SPECTRE. BSE was better than SPECTRE on four of the 18 subjects, while also being the worst method on four of the 18 subjects. BET was better than SPECTRE on three of the 18 subjects, while also performing quite poorly on one of the subjects (see Subject 10 in Fig. 7). WAT was better than SPECTRE on four of the 18 subjects, but it performed the worst on two of the subjects, including one failure (see Subject 15 in Fig. 7). We can, additionally, look at the paired rank sum comparison between SPECTRE and the other methods, shown in the left column of Table 3. Here we see that SPECTRE is statistically significantly different from three of the methods, again BSE, BET, and WAT. SPECTRE also has the lowest coefficient of variation of all the methods on IBSR Set 1. In particular, it is lower than BSE, BET, and WAT.

The results for IBSR Set 2, are shown in Tables 2 and 3, and in Fig. 8. Again, with respect to the Dice score, SPECTRE is the best performer, this time followed by the Graph Cutting

(GCUT) approach. However, the comparison results on IBSR Set 2 are much more complicated to interpret than those of IBSR Set 1. Fig. 8 shows that SPECTRE ranks highest in eight of the 20 subjects in the data set, with the next best performer being BSE which ranked highest on five of the subjects. BSE, however, performs poorly on several of the data sets; it is the worst performer in nine of the 20 subjects. By considering either the *t*-test scores, shown in Table 2, or the rank sum scores, shown in Table 3, we can say that SPECTRE is statistically significantly better than BET and WAT. There is not enough statistical power to state conclusively that SPECTRE is better than the other algorithms, however. Of particular note in Fig. 8, is how poorly (Dice Coefficient ≤ 0.85) all of the methods did on this data set.

Experiment 2

In this experiment, we compared SPECTRE against 160 subjects from the Baltimore Longitudinal Study of Aging (BLSA) (Shock et al., 1984; Resnick et al., 2003). The human rater is a certified radiological technologist, with two decades experience in neuroradiography and over 15 years in cerebrum extraction. The skull stripping protocol used, by the human rater, for the experiment is a semi-automated approach (Goldszal et al., 1998), explained in greater detail in Bazin et al. (2007). All subjects were scanned on a GE Signa 1.5 T scanner (GE Healthcare, Waukesha, WI) using a T_1 -weighted SPGR imaging protocol (TR = 35ms, TE = 5ms, flip angle = 45° , NEX = 1). The subjects range in age from 48 to 93 (mean = 73.44 years, standard deviation = 7.98 years). There were 92 male subjects and 68 female subjects. In total there are 1046 scans of these one hundred and sixty subjects, with a mean of 6.5 scans per subject and a standard deviation of 2.8 scans per subject. The cerebrum extraction was done independently on each of the 1046 scans. Four atlases, used by SPECTRE, were selected from the BLSA data pool but not from within the sample of 1046 scans. Each atlas comprises an original MR image and an accompanying image that had been manually segmented. No preselection of atlases was done to influence or enhance results. Data sets in the experiment population did not exclude participants that had suffered some form of brain trauma, such as stroke.

Table 4 shows the results of SPECTRE's cerebrum extraction as compared to our human rater for these 1046 BLSA scans. Fig. 6 shows the worst results generated by SPECTRE on this sample. In this case, SPECTRE included the lesion — which we know to be the result of trauma — while the human rater excluded it from the brain mask. In this case it is not clear which is the better result as it is best defined by the nature of the subsequent analysis tasks.

The size of this cohort made it possible to explore the question of the effects of age and gender on our skull stripping algorithm. A linear regression model was used for this analysis. We used a backward elimination from a full model of *age* * *sex*, to establish the significance of either term, or the cross product term. Age was statistically significant (p-value of 2×10^{-16}), while gender never reached significance. Fig. 9 shows a plot of Age *versus* Dice scores, revealing a prominent trend of a decreasing Dice score with age. We believe that the observed dependence of Dice score with age can be explained by considering brain atrophy (Rettmann et al., 2006). In particular, SPECTRE is able to include a thin layer of cortical CSF more easily in the older brain while the rater continues to exclude everything that is not brain (and choosing masks that exclude sulcal CSF, for example). The plot of Age *versus* CI, shown in Fig. 10 supports this explanation. In particular, the linear fit of CI *versus* age (shown in Fig. 10) though not reaching significance, does show an increasing trend suggesting that the masks are containing more of the human rater result. To further substantiate this idea, we computed the CSF volumes in the SPECTRE brain masks using the topology-preserving, anatomy-driven segmentation (TOADS) (Bazin and Pham, 2007) method. The CSF volume represents only sulcal and subarachnoid CSF, and does not include CSF volume in the ventricles. The linear fit of these

CSF volumes with respect to age was also found to have a significant positive slope (p-value $< 2 \times 10^{-16}$). A plot of sulcal and subarachnoid CSF volume against age for these subjects is shown in Fig. 11, along with the corresponding linear fit.

Experiment 3

This experiment focuses on the impact of skull stripping on the estimation of the cortical surface using CRUISE (Han et al., 2004). CRUISE reconstructs the GM/WM interface (inner surface), the central surface and the GM/CSF interface (pial surface or outer surface) of the cerebral cortex from a T1-weighted MR brain image. CRUISE is currently designed to take a data set that has already had extra-cranial tissue excised as its initial input. To demonstrate the usefulness of SPECTRE as a replacement for human raters in this process, we have conducted two experiments in which a human rater picked landmarks on an MR brain image which were meant to represent one of the three surfaces generated by CRUISE. In the first experiment the human rater picked 10 central surface landmarks on each of three data sets. Five landmarks were chosen on each hemisphere. We then had two human raters, different from the rater who chose the landmarks, manually skull strip all three data sets. A comparison between the output of CRUISE given either the manually or SPECTRE skull stripped data as input was generated. The distance of each landmark from the central surfaces given by our two human raters and SPECTRE are shown in Table 5. Performing a paired *t*-test between each set of results for the central surface landmarks shows that the distributions cannot be distinguished from each other.

A more comprehensive set of inner and outer surface landmarks were selected by the same rater who selected our central surface landmarks on an additional data set. This data set was stripped by our two human raters and the result was given as input to CRUISE. The landmark rater picked a total of 420 landmarks to represent the inner surface, corresponding to ten landmarks within each of 14 fundi, 14 banks and 14 gyral crowns near major sulci on both data sets. The task was repeated for what the rater perceived to be the outer surface, yielding a total of 840 landmark points. All landmarks were picked on the original MR image. To avoid confusion with the central surface landmark experiment we will refer to this subject as “Subject 4”. The results of measuring the distance from the resultant CRUISE surface to the human selected landmarks are shown in Table 6. An example image, Fig. 12, shows the output CRUISE surfaces for one of the human raters and SPECTRE, as well as showing some of the landmarks used.

We perform a paired *t*-test between each rater on each surface with the results given in Table 7, showing statistically significant differences in the surfaces, generated based on SPECTRE input and those of our human raters. For Subject 4, on the inner surface, we can see that SPECTRE has a statistically significant difference with one of the human raters, while the two raters have given rise to very different distributions in the distances. Our second human rater ($\mathcal{M}_{H2}$) having the lower mean on this surface would appear to have done a better job than either SPECTRE or the human rater. There are also significant differences for the outer surfaces between SPECTRE and either rater. Looking at the mean distance between landmarks for each of $\mathcal{M}_{H1}$, $\mathcal{M}_{H2}$, and $\mathcal{M}_S$, we can see that SPECTRE has outperformed both humans on the outer surface. This result suggests that SPECTRE could be used to replace a human rater in the preprocessing steps of CRUISE.

Discussion

In Experiment 1, we demonstrated that SPECTRE is superior to several existing methods on modern T1-weighted data, IBSR Set 1, with through plane resolutions of 1.5mm. Whereas on legacy data, IBSR Set 2, where the through plane resolution is 3.1mm, SPECTRE is

comparable to existing technologies. Visual inspection of Fig. 8 suggests that SPECTRE is superior – at least less prone to dramatic failures as evidenced by the other methods – but our statistical tests do not reveal significance. It is important to note that all the methods performed quite poorly on this legacy data, which shows that there is still significant room for improvement in this area.

Experiment 2 showed that SPECTRE is robust, at least with respect to the task of cerebrum extraction on SPGR data sets for a large sample, one hundred and sixty subjects with over one thousand scans being processed, for the given study (Baltimore Longitudinal Study of Aging). It is interesting to observe that a *t*-test between the Dice scores for the SPECTRE results from Experiment 1 on IBSR Set 1 and those achieved on the BLSA subjects in Experiment 2 have a *p*-value of 0.05638. This does not reach the level of significance, but does suggest that SPECTRE has a similar performance across these two populations, or more concisely: SPECTRE has a similar performance across similar resolution data.

Though volume metrics such as the Dice coefficient and the CI are valuable in determining how well a given skull stripping algorithm works, it is the authors' contention that surface based markers representing important tissue boundaries are also of significance. The landmarks used in Experiment 3 demonstrated that when SPECTRE was worse than either of our human raters the difference was measured in terms of hundredths of a millimeter. Additionally, we noted that with regards to cortical reconstruction from CRUISE, SPECTRE contributed to a more accurate (*c.f.* Table 6) outer surface than either of the human raters. This is most likely a side effect of our decision to include at least one voxel of cerebrospinal fluid at the boundary on the brain, thus allowing CRUISE to more accurately articulate the boundary rather than being constrained by the boundary imposed by the skull stripping. Experiment 3 demonstrated achievement of our primary goal; development of a new automated skull stripping algorithm that would provide an accurate segmentation of the brain as input to CRUISE (Han et al., 2004).

Unlike other methods, which may fail by removing too much cortical GM, as shown in Fig. 6, when SPECTRE performs poorly it is because it includes more dura than other methods, which can also be seen in Fig. 1. In our application, cortical reconstruction, this is an appropriate and acceptable error. There are of course other post processing tasks for which the inclusion of excess extra cranial tissue would cause undue harm, and alternative skull stripping methods may be more useful. In specific comparison to other methods, Experiment 1 demonstrated that when BSE fails it can do so dramatically (one subject had a Dice Coefficient of 0.0), this was a result of the wrong selection of the largest connected component (the neck instead of the brain) (Sadanathan et al., 2010). BET was also capable of producing some poor results, as measured by the Dice metric. WAT, though capable of producing good results, had large standard deviations of Dice and CI, showing that the results are inconsistent. This can easily be seen in Fig. 8. The other methods (HWA, GCUT, GCUT-HWA) do quite well in containing the human rater's mask (CI scores at or above 0.990). However, because the Dice scores for these methods are low, we can conclude that they produce masks that are generally much too big.

We can observe that the SPECTRE mask is also slightly larger than that of human raters, as shown in Fig. 5. This larger mask comes about in two ways. Firstly, SPECTRE is recovering a better mask, as evidenced by considering the zoomed region shown in Fig. 1(e), (f), and (h). Secondly, SPECTRE always tried to include a single voxel of CSF on the boundary of the cortical GM, again clearly shown in Fig. 1(h) and also visible in Fig. 5(d). The human rater masks shown in Fig. 5 also demonstrate their failure to include all cortical GM.

Conclusions

Our goal in developing this new skull-stripping software was to create a tool that would be an automated replacement for manual skull stripping in large neurological studies and would have no adverse affects on subsequent processing of the data. Our experiments demonstrate that SPECTRE can accurately perform skull stripping and can be applied to large data sets. The experimental results suggest that SPECTRE is quite robust in comparison to other skull stripping methods, see Experiment 1, and to a human rater on a large cohort, see Experiment 2.

Acknowledgments

Funding support for this work was provided by the National Institute of Neurological Disorders and Stroke (NINDS), (R01-NS37747, R01-AG016324 and R01-NS054255), and by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) (R21-EB009900), both of which are part of the National Institute of Health (NIH). This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. We are grateful to the BLSA participants and neuroimaging staff for their dedication to these studies. The authors gratefully acknowledge the help of Dr. Vitali Zagorodnov, of Nanyang Technological University, for providing the numerical results of his experiments, original published in Sadananthan et al. (2010). The authors wish to thank Navid Shiee of Johns Hopkins University for his help in preparing this manuscript for publication. We thank the anonymous reviewers for their careful analysis of our paper, which helped to greatly improve this manuscript. The software is to be made publicly available through integration into the 3D Slicer (<http://www.slicer.org/>) software package distributed through NA-MIC (<http://www.na-mic.org/>).

References

- Acosta-Cabronero J, Williams GB, Pereira JMS, Pengas G, Nestor PJ. The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel based morphometry. *NeuroImage*. 2008; 39 (4):1654–1665. [PubMed: 18065243]
- Ashburner J, Friston KJ. Voxel-based morphometry: the methods. *NeuroImage*. 2000; 11 (6):805–821. [PubMed: 10860804]
- Bazin PL, Cuzzocreo JL, Yassa MA, Gandler W, McAuliffe MJ, Bassett SS, Pham DL. Volumetric neuroimage analysis extensions for the MIPAV software package. *Journal of Neuroscience Methods*. 2007; 165:111–121. [PubMed: 17604116]
- Bazin PL, Pham DL. Topology-Preserving Tissue Classification of Magnetic Resonance Brain Images. *IEEE Trans Med Imag*. 2007; 26 (4):487–496.
- Boesen K, Rehm K, Schaper K, Stoltzner S, Woods R, Luders E, Rottenberg D. Quantitative comparison of four brain extraction algorithms. *NeuroImage*. 2004; 22 (3):1255–1261. [PubMed: 15219597]
- Carass, A.; Wheeler, MB.; Cuzzocreo, J.; Bazin, PL.; Bassett, SS.; Prince, JL. A joint registration and segmentation approach to skull stripping. *Fourth IEEE International Symposium on Biomedical Imaging (ISBI 2007)*; 2007. p. 656-659.
- Center for Morphometric Analysis (CMA). Internet Brain Segmentation Repository: 18 T1-weighted MR scans with expert segmentations of 43 individual structures. 1995. URL <http://www.cma.mgh.harvard.edu/ibsr/>
- Dale AM, Fischl B, Sereno MI. Cortical Surface-Based Analysis I: Segmentation and Surface Reconstruction. *NeuroImage*. 1999; 9 (2):179–194. [PubMed: 9931268]
- Davis MR, Votaw JR, Bremner JD, Byas-Smith MG, Faber TL, Hoffman RJ, Grafton ST, Kilts CD, Goodman MM. Initial Human PET Imaging Studies with the Dopamine Transporter Ligand ¹⁸F-FECNT. *J Nucl Med*. 2003; 44 (6):855–861. [PubMed: 12791810]
- Dice L. Measure of the amount of ecological association between species. *Ecology*. 1945; 26 (3):297–302.
- Fennema-Notestine C, Ozyurt IB, Clark CP, Morris S, Bischoff-Grethe A, Bondi MW, Jernigan TL, Fischl B, Ségonne F, Shattuck DW, Leahy RM, Rex DE, Toga AW, Zou KH, BIRN, Brown GG. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy

- images: Effects of diagnosis, bias correction, and slice location. *Human Brain Mapping*. 2006; 27 (2):99–113. [PubMed: 15986433]
- Goldszal AF, Davatzikos C, Pham DL, Yan MXH, Bryan RN, Resnick SM. An image processing system for the qualitative and quantitative volumetric analysis of brain images. *J Computer Assisted Tomography*. 1998; 22:827–837.
- Grady, L. Fast, quality, segmentation of large volumes—isoperimetric distance trees; *Proceedings of ECCV 2006*; 2006. p. 449–462.
- Hahn, H.; Peitgen, HO. The skull stripping problem in MRI solved by a single 3D watershed transform. *Proc. 3rd Int'l Conf. Med. Imag. Comp. Comp. Assist. Inter. (MICCAI)*; Springer-Verlag; 2000. p. 134–143.
- Han X, Pham DL, Tosun D, Rettmann ME, Xu C, Prince JL. CRUISE: Cortical reconstruction using implicit surface evolution. *NeuroImage*. 2004; 23 (3):997–1012. [PubMed: 15528100]
- Hartley SW, Scher AI, Korf ESC, White LR, Launer LJ. Analysis and validation of automated skull stripping tools: A validation study based on 296 MR images from the Honolulu Asia aging study. *NeuroImage*. 2006; 30 (4):1179–1186. [PubMed: 16376107]
- Kapur T, Grimson WEL, Wells WM, Kikinis R. Segmentation of brain tissue from magnetic resonance images. *Medical Image Analysis*. 1996; 1 (2):109–127. [PubMed: 9873924]
- Lee JM, Yoon U, Nam SH, Kim JH, Kim IY, Kim SI. Evaluation of automated and semi-automated skull stripping algorithms using similarity index and segmentation error. *Comp Biol Med*. 2003; 33 (6):495–507.
- Lemeux L, Hagemann G, Krakow K, Woermann FG. Fast, accurate and reproducible automatic segmentation of the brain in T1-weighted volume MRI data. *Magn Reson Med*. 1999; 42 (1):127–135. [PubMed: 10398958]
- Lowe VJ, Kemp BJ Jr, CRJ, Senjem M, Weigand S, Shiung M, Knopman GSD, Boeve B, Mullan B, Petersen RC. Comparison of ¹⁸F-FDG and PiB PET in Cognitive Impairment. *J Nucl Med*. 2009; 50 (6):878–886. [PubMed: 19443597]
- Maldjian JA, Chalela J, Kasner SE, Liebeskind D, Detre JA. Automated CT Segmentation and Analysis for Acute Middle Cerebral Artery Stroke. *Am Jnl Neuroradiology*. 2001; 22 (6):1050–1055.
- Park JG, Lee C. Skull stripping based on region growing for magnetic resonance brain images. *NeuroImage*. 2009; 47 (4):1394–1407. [PubMed: 19389477]
- Pham, DL. Robust fuzzy segmentation of magnetic resonance images. *Proc. 14th IEEE Symp. on Computer-based Medical Systems (CBMS 2001)*; Bethesda, MD: 2001. p. 127–131.
- Rehm K, Schaper K, Anderson J, Woods R, Stoltzner R, Rottenberg D. Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. *NeuroImage*. 2004; 22 (3):1262–1270. [PubMed: 15219598]
- Resnick SM, Pham DL, Kraut MA, Zonderman A, Davatzikos C. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *J Neurosci*. 2003; 23:3295–3301. [PubMed: 12716936]
- Rettmann ME, Kraut MA, Prince JL, Resnick SM. Cross-sectional and longitudinal analyses of anatomical sulcal changes associated with aging. *Cerebral Cortex*. 2006; 16:1584–1594. [PubMed: 16400155]
- Rex DE, Shattuck DW, Woods RP, Narr KL, Luders E, Rehm K, Stoltzner SE, Rottenberg DA, Toga AW. A meta-algorithm for automated brain extraction in MRI. *NeuroImage*. 2004; 23 (2):625–637. [PubMed: 15488412]
- Rohde GK, Aldroubi A, Dawant BM. The adaptive bases algorithm for intensity based nonrigid image registration. *IEEE Trans Med Imag*. 2003; 22 (11):1470–1479.
- Rutovitz, D. Expanding picture components to natural density boundaries by propagation methods. The notions of fall-set and fall-distance. *Proc. 4th Int. J. Conf. Patt. Recog*; Kyoto, Japan. 1978. p. 657–664.
- Sadanathan SA, Zhenga W, Chee MWL, Zagorodnov V. Skull stripping using graph cuts. *NeuroImage*. 2010; 49 (1):225–239. [PubMed: 19732839]
- Sandor S, Leahy R. Surface-based labeling of cortical anatomy using a deformable atlas. *IEEE Trans Med Imag*. 1997; 16 (1):41–54.

- Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*. 2004; 22 (3):1060–1075. [PubMed: 15219578]
- Shan ZY, Yue GH, Liu JZ. Automated histogram-based brain segmentation in T1-weighted three-dimensional magnetic resonance head images. *NeuroImage*. 2002; 17 (3):1587–1598. [PubMed: 12414297]
- Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*. 2001; 13 (5):856–876. [PubMed: 11304082]
- Shock, NW.; Greulich, RC.; Andres, R.; Arenberg, D.; Costa, PT., Jr; Lakatta, E.; Tobin, JD. Normal human aging: The Baltimore longitudinal study of aging. U.S. Government Printing Office; Washington, D.C: 1984.
- Smith SM. Fast robust automated brain extraction. *Human Brain Mapping*. 2002; 17 (3):143–155. [PubMed: 12391568]
- Tosun D, Rettmann ME, Naiman DQ, Resnick SM, Kraut MA, Prince JL. Cortical Reconstruction Using Implicit Surface Evolution: Accuracy and Precision Analysis. *NeuroImage*. 2006; 29 (3): 838–852. [PubMed: 16269250]
- Tsang, O.; Gholipour, A.; Kehtarnavaz, N.; Panahi, I.; Gopinath, K.; Briggs, R. Comparison of Tissue Segmentation Algorithms in Neuroimage Analysis Software Tools. *Proceedings of the 30th IEEE EMBS Annual International Conference*. IEEE; 2008.
- Ward, BD. Intracranial segmentation. Milwaukee: Biophysics Research Institute, Medical College of Wisconsin. 1999. <http://afni.nimh.nih.gov/afni/>
- Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin*. 1945; 1 (6):80–83.
- Yoon UC, Kim JS, Kim JS, Kim IY, Kim SI. Adaptable fuzzy C-means for improved classification as a preprocessing procedure of brain parcellation. *J Digit Imaging*. 2001; 14 (2 Suppl 1):238–240. [PubMed: 11442112]

Figure 1.

(a) A cross section through an MR brain volume. The brain extracted by (b) a manual rater, (c) Brain Extraction Tool (BET version 2.1), and (d) our approach. (e) is the zoomed region represented by the red box for the MR brain volume, while (f) is the corresponding region for the manual rater. (g) and (h) are the zoomed regions for BET and our approach, respectively. The brain mask extracted by a manual rater is shown in (i) as a red overlay on the original, (j) is the Brain Extraction Tool (BET version 2.1) result shown as a green overlay, while (k) is a blue overlay of the mask generated by our algorithm, SPECTRE.

Figure 2.

A topographical representation of a slice of a T1 weighted MR image. The heights are the intensities of the image, shown inset, the colors also correspond to the intensities and are used for display purposes only.

Figure 3.

Flow Chart describing the basic components of SPECTRE. **A** is the input image. **B** denotes the flow from the input image registered against 4 atlas images to the creation of the probability mask, $\mathcal{M}_{ABA}$. **C** shows the hard tissue segmentation. **D** is the morphological operations phase of the algorithm. Image **4**, in **D**, must pass a *sanity check* before it is approved as the mask.

Figure 4. (a) Original image, (b) Probability mask, (c) Tissue classification, (d) initial mask $\mathcal{M}_1$, (e) mask after erosion and retaining largest connected component, (f) mask after *hill descent* but prior to the topologically constrained morphological closing.

Figure 5. (a) Original image, (b) ($\mathcal{M}_{H1}$) human rater, (c) ($\mathcal{M}_{H2}$) an alternative human rater and (d) ($\mathcal{M}_S$) output from SPECTRE.

Figure 6.

The worst result from the study of 1046 data sets, with the subject having a Dice coefficient of 0.88481. We show (a) the original image, (b) the human rater ($\mathcal{M}_G$), and (c) the result from SPECTRE ($\mathcal{M}_S$). The corresponding Containment Index score for this subject is 0.99779, see the text for an explanation of these numbers.

Figure 7.

The annotated line graph shows the Dice scores for each of the seven algorithms (SPECTRE, Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), Hybrid Watershed Algorithm (HWA), Graph Cuts algorithm (GCUT), and an approach based on the intersection of the masks of GCUT and HWA (GCUT-HWA)) on each of IBSR Set 1 (18 1.5mm scans). The y-axis is the corresponding Dice score (closer to 1.0 is better), while the x-axis is an index over the 18 IBSR Set 1 subjects. SPECTRE appears to perform better than the other approaches, it ranks highest in 11 of the 18 subjects.

Figure 8.

The annotated line graph shows the Dice scores for each of the seven algorithms (SPECTRE, Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), Hybrid Watershed Algorithm (HWA), Graph Cuts algorithm (GCUT), and an approach based on the intersection of the masks of GCUT and HWA (GCUT-HWA)) on each of IBSR Set 2 (20 3.1mm scans). The y-axis is the corresponding Dice score (closer to 1.0 is better), while the x-axis is an index over the 20 IBSR Set 2 subjects. SPECTRE appears to perform better than the other approaches, it ranks highest in eight of the 20 subjects.

Figure 9.

Shown is the linear fit for the Dice Coefficient against Age for the one thousand and forty six subjects used in Experiment 2. The Dice Coefficient is based on a comparison between our approach and an expert human rater. The p-value for the significance of Age in a linear model with the Dice Coefficient is less than 2×10^{-16} .

Figure 10.

Shown is the linear fit for the Containment Coefficient against Age for the one thousand and forty six subjects used in Experiment 2. The Containment Index is based on a comparison between our approach and an expert human rater. The linear model is increasing but is not statistically significant.

Figure 11. Shown is the linear fit for the sulcal and subarachnoid CSF volume against Age for the one thousand and forty six subjects used in Experiment 2. The p-value for the significance of Age in a linear model with the CSF volumes is less than 2×10^{-16} .

Figure 12.

(a) Shows the CRUISE outer surface derived from the skull stripping of a human rater, (b) is the CRUISE outer surface based on the output of our algorithm. Both (a) and (b) show the landmarks used on this slice, as red crosses in the posterior portion on the brain, these are some of the landmarks used in Experiment 3. In this case they are landmarks for the banks of the parieto-occipital sulcus on the outer surface. (c) & (d) show a zoomed in image centered on the landmarks (red crosses) with (c) being the outer surface derived from the human rater and (d) the corresponding result for our algorithm.

Comparison of SPECTRE with six other skull stripping approaches: Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), Hybrid Watershed Algorithm (HWA), Graph Cuts algorithm (GCUT), and an approach based on the intersection of the masks of GCUT and HWA (GCUT-HWA), on IBSR Set 1 (18 1.5mm scans). The Dice Coefficients of the BSE, BET, WAT, GCUT, and GCUT-HWA, methods were originally presented in Sadanathan et al. (2010).

Table 1

Method	Dice			CI			P-Value	CoV	[Range]	P-Value
	Mean (SD)	CoV	[Range]	Mean (SD)	CoV	[Range]				
SPECTRE	0.9440 (0.0115)	1.2216	[0.9257 – 0.9616]	–	0.9981 (0.0017)	0.1732	[0.9943 – 0.9999]	–		
BSE	0.9126 (0.0424)	4.6456	[0.8425 – 0.9679]	0.0032	0.9413 (0.0782)	8.3118	[0.7759 – 0.9956]	0.0070		
BET	0.9260 (0.0400)	4.3166	[0.7853 – 0.9583]	0.0392	0.9807 (0.0225)	2.2912	[0.9327 – 0.9989]	0.0050		
WAT	0.9128 (0.0832)	9.1129	[0.5980 – 0.9593]	0.1427	0.9755 (0.0179)	1.8383	[0.9292 – 0.9992]	0.0000		
HWA	0.8813 (0.0254)	2.8772	[0.8184 – 0.9091]	0.0000	0.9998 (0.0002)	0.0179	[0.9993 – 1.0000]	0.0006		
GCUT	0.9122 (0.0161)	1.7646	[0.8750 – 0.9311]	0.0000	0.9997 (0.0004)	0.0367	[0.9985 – 1.0000]	0.0017		
GCUT-HWA	0.9173 (0.0154)	1.6811	[0.8776 – 0.9350]	0.0000	0.9996 (0.0005)	0.0442	[0.9983 – 1.0000]	0.0038		

The P-Value is from a paired t -Test between the result of SPECTRE and the other algorithms for either the Dice or CI. SD denotes standard deviation, and CoV represents coefficient of variation.

Comparison of SPECTRE with six other skull stripping approaches: Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), Hybrid Watershed Algorithm (HWA), Graph Cuts algorithm (GCUT), and an approach based on the intersection of the masks of GCUT and HWA (GCUT-HWA), on IBSR Set 2 (20 3.1mm scans). The Dice Coefficients of the BSE, BET, WAT, GCUT, and GCUT-HWA, methods were originally presented in Sadanathan et al. (2010).

Table 2

Method	Dice			CI			P-Value	CoV	[Range]	P-Value
	Mean (SD)	CoV	[Range]	Mean (SD)	CoV	[Range]				
SPECTRE	0.8699 (0.0710)	8.1586	[0.7160 – 0.9426]	–	0.9890 (0.0113)	1.1411	[0.9662 – 0.9996]	–		
BSE	0.7941 (0.2139)	26.931	[0.0000 – 0.9482]	0.1050	0.7299 (0.2405)	32.952	[0.0000 – 0.9649]	0.0001		
BET	0.7429 (0.1441)	19.392	[0.5267 – 0.8976]	0.0015	0.9990 (0.0011)	0.1107	[0.9961 – 0.9999]	0.0010		
WAT	0.7635 (0.1437)	18.827	[0.4709 – 0.9237]	0.0019	0.7547 (0.2274)	30.132	[0.3727 – 0.9988]	0.0002		
HWA	0.7864 (0.2145)	27.333	[0.1587 – 0.8759]	0.0902	0.9809 (0.0653)	6.6522	[0.7113 – 1.0000]	0.6089		
GCUT	0.8532 (0.0867)	10.162	[0.4908 – 0.8962]	0.4514	0.9999 (0.0002)	0.0179	[0.9994 – 1.0000]	0.0004		
GCUT-HWA	0.8581 (0.0913)	10.645	[0.4908 – 0.9031]	0.6205	0.9808 (0.0653)	6.6531	[0.7112 – 1.0000]	0.6052		

The P-Value is from a paired *t*-Test between the result of SPECTRE and the other algorithms for either the Dice or CI. SD denotes standard deviation, and CoV represents coefficient of variation.

Table 3

Comparison of SPECTRE with six other skull stripping approaches: Brain Surface Extractor (BSE), Brain Extraction Tool (BET), Watershed Algorithm (WAT), Hybrid Watershed Algorithm (HWA), Graph Cuts algorithm (GCUT), and an approach based on the intersection of the masks of GCUT and HWA (GCUT-HWA). The table shows rank sum p-value comparisons between SPECTRE and the other six methods for the Dice measure on both the IBSR Set 1 (18 1.5mm scans) and IBSR Set 2 (20 3.1mm scans).

	IBSR 1 Dice P-Value	IBSR 2 Dice P-Value
SPECTRE	–	–
BSE	0.0200	0.3547
BET	0.1182	0.0035
WAT	0.0847	0.0068
HWA	0.0000	0.0515
GCUT	0.0000	0.2315
GCUT-HWA	0.0000	0.2766

Table 4

The mean, median, standard deviation (SD), maximum, and minimum Dice coefficient and containment index for both SPECTRE and a human rater over 1046 data sets from the BLA.

	Mean	Median	SD	Max	Min
Dice	0.93836	0.93911	0.01157	0.96611	0.88481
CI	0.99217	0.99392	0.00744	0.99894	0.91453

Table 5

Mean and standard deviation (SD) in millimeters for the absolute distance from the central surface landmarks to the corresponding central surface as output by CRUISE based on the corresponding skull stripped data set.

	Subject 1	Subject 2	Subject 3
	Mean (SD)	Mean (SD)	Mean (SD)
Human Rater 1 ($\mathcal{M}_{H1}$)	0.7559 (0.9273)	0.5972 (0.3512)	0.4096 (0.3138)
Human Rater 2 ($\mathcal{M}_{H2}$)	0.7226 (0.8305)	0.5724 (0.4280)	0.4339 (0.3477)
SPECTRE ($\mathcal{M}_S$)	0.7288 (0.7473)	0.5304 (0.3894)	0.4947 (0.3931)

Table 6

Mean and Standard Deviation in millimeters for the absolute distance from the set of 420 landmarks, per surface, to the corresponding CRUISE surface based on the input of either of the human raters ($\mathcal{M}_{H1}$, $\mathcal{M}_{H2}$) or SPECTRE ($\mathcal{M}_S$).

	Subject 4	
	Inner Surface	Outer Surface
	Mean (SD)	Mean (SD)
Human Rater 1 $\mathcal{M}_{H1}$	0.9433 (0.6819)	0.7909 (0.6477)
Human Rater 2 $\mathcal{M}_{H2}$	0.9360 (0.6816)	0.8466 (0.7187)
SPECTRE $\mathcal{M}_S$	0.9502 (0.6937)	0.7621 (0.6162)

Table 7

The upper triangular portion of the table is the p -value from a simple paired t -test between the distribution of distances from the CRUISE generated surface to the set of landmarks, the CRUISE surface was generated from either of the human raters ($\mathcal{M}_{H1}$, $\mathcal{M}_{H2}$) or SPECTRE ($\mathcal{M}_S$). The lower triangular portion of each table is the absolute difference in the means of the distances from the landmarks, in millimeters.

Subject 4	Inner Surface			Outer Surface		
	$\mathcal{M}_{H1}$	$\mathcal{M}_{H2}$	$\mathcal{M}_S$	$\mathcal{M}_{H1}$	$\mathcal{M}_{H2}$	$\mathcal{M}_S$
$\mathcal{M}_{H1}$	-	0.0363	0.0306	-	3.051×10^{-7}	7.367×10^{-4}
$\mathcal{M}_{H2}$	0.0073	-	1.107×10^{-4}	0.0557	-	2.302×10^{-8}
$\mathcal{M}_S$	0.0070	0.0142	-	0.0289	0.0846	-