Published in final edited form as:

Acad Radiol. 2013 December ; 20(12): . doi:10.1016/j.acra.2013.09.010.

Multi-Atlas Skull-Stripping

Jimit Doshi, MSc, Guray Erus, PhD, Yangming Ou, PhD, Bilwaj Gaonkar, MSc, and Christos Davatzikos, PhD

Section of Biomedical Image Analysis, Department of Radiology, 3600 Market St. Suite 380, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Rationale and Objectives—We present a new method for automatic brain extraction on structural magnetic resonance images, based on a multi-atlas registration framework.

Materials and Methods—Our method addresses fundamental challenges of multi-atlas approaches. To overcome the difficulties arising from the variability of imaging characteristics between studies, we propose a study-specific template selection strategy, by which we select a set of templates that best represent the anatomical variations within the data set. Against the difficulties of registering brain images with skull, we use a particularly adapted registration algorithm that is more robust to large variations between images, as it adaptively aligns different regions of the two images based not only on their similarity but also on the reliability of the matching between images. Finally, a spatially adaptive weighted voting strategy, which uses the ranking of Jacobian determinant values to measure the local similarity between the template and the target images, is applied for combining coregistered template masks.

Results—The method is validated on three different public data sets and obtained a higher accuracy than recent state-of-the-art brain extraction methods. Also, the proposed method is successfully applied on several recent imaging studies, each containing thousands of magnetic resonance images, thus reducing the manual correction time significantly.

Conclusions—The new method, available as a stand-alone software package for public use, provides a robust and accurate brain extraction tool applicable for both clinical use and large population studies.

Keywords

Brain extraction; registration; multi-atlas; label fusion; Jacobian determinant

Brain extraction, or skull stripping is a very important preprocessing step preceding almost all automated brain magnetic resonance (MR) imaging (MRI) applications. It consists of the removal of the skull and the extracerebral tissues (e.g., scalp and dura) on brain MR images. An illustrative example of extraction on a T1-weighted image is shown in Figure 1. Brain extraction is known to be a difficult task, as the boundaries between brain and nonbrain tissues, especially those between the gray matter and the dura matter, might not be clear on MR images. Also, it is more prone to requiring manual intervention as the errors in this step propagate to most subsequent analysis steps, such as registration to a common space, tissue segmentation, cortical thickness and atrophy estimation, etc (1).

Particularly in large population studies, it is desirable to set up a fully automated processing pipeline with no or minimal manual intervention, to reduce processing time and to prevent any kind of human bias in the results. Thus, an automated brain extraction method should be accurate as well as robust, so that thousands of images within a study, potentially having different imaging characteristics and significant anatomical variations, could be successfully segmented.

Several brain extraction methods have been proposed since the early years of MRI research. Region- or boundary-based approaches (2–7) are simple, fast, and general. An example is the Brain Extraction Tool (BET) (5), a widely popular publicly available method, which is based on a deformable surface model that evolves to the boundaries of the brain. However, boundary-based methods might fail especially when the initial assumption of a clear separation between the brain and the nonbrain tissues is not completely satisfied.

Atlas-based (i.e., template-based) methods use an expert-defined segmentation on the atlas space as a prior for extracting the brain on the target image. In recent years, several multi-atlas-based methods have been proposed, producing very accurate, state-of-the-art segmentations (8–11). The main premise is that multiple atlases cover much wider anatomical variations and, when registered to the target image, they can correct errors among each other, thus providing an increased accuracy and robustness. The multi-atlas framework generally consists of three main modules: atlas selection based on similarity to the target scan, registration of the atlases to the target space, and fusion of the registered ground truth masks using various label fusion algorithms (11). In Carass et al (8), a probabilistic brain mask obtained using multi-atlas registration is combined with a fuzzy segmentation of the subject brain using topologically constrained morphological operators. In Iglesias et al (10), a hybrid (discriminative/generative) learning-based approach, named Robust Brain Extraction (ROBEX), is proposed: The discriminative component consists of a binary classifier that is trained on aligned and intensity-normalized atlases with brain masks. The initial classification is refined using a generative model of the brain shape that is learned from the atlases by applying an active shape model approach. The method is validated on three publicly available data sets against six popular, publicly available methods [BET (5), Brain Surface Extractor (BSE) (3), FreeSurfer (12), 3dSkullStrip in AFNI package (13), BridgeBurner (14), and GraphCuts (GCUT) (15)] and obtained a higher accuracy for almost every method/data set combination.

Our approach follows a multi-atlas-based brain extraction strategy, because of its demonstrated strength and promise in the aforementioned studies. In addition, our approach makes contributions in dealing with two fundamental challenges in multi-atlas approaches:

The first challenge is the variability of imaging characteristics between studies. We proposed a study-specific template selection strategy. This is different from the approach used in recent multi-atlas-based methods, such as in Eskildsen et al (9) and in Leung et al (11), where final templates were selected from a pre-defined template library with manual ground truth masks. In our approach, a set of representative templates are automatically selected within the study, corresponding ground truth masks are created semiautomatically, and these templates are used for brain extraction on all images in the study.

The second challenge is the quality of individual registrations, particularly because the registration is performed on raw images (i.e., unprocessed images that contain both brain and the skull). Extracerebral tissues and organs have an intersubject variation much higher than that of the brain, which may misguide the registration and cause significant errors. In this report, we propose to use a registration algorithm that is particularly adapted to this task, as it adaptively aligns different regions of the two images based on not only their similarity

but also the reliability of the matching between images, thus assigning lower importance to areas with missing correspondences and reducing the negative impact of outlier regions.

Several studies (16–18) have shown that a weighted voting (WV) strategy, which assigns higher weights to templates that are more similar to the target image, improves the segmentation quality. Following this approach, we applied a spatially adaptive WV strategy for combining coregistered template masks. However, instead of the commonly used intensity-based similarity metric, we used ranking of Jacobian determinant values to calculate local weights for each template. The underlying intuition is that a low volumetric change in registration between the template and the target images, as encoded by the Jacobian determinant value at each voxel, indicates higher local similarity between the two images.

Our method, which we call Multi-Atlas Skull-Stripping (MASS), is applied on three popular yet challenging public data sets, which have been widely used for validation in many recent brain extraction methods. The segmentations are compared to manual ground truth masks using Dice score and Hausdorff distance. The quantitative results are compared to the results reported by Iglesias et al (10).

METHODS

A general overview of the proposed method is given in Figure 2. Our framework consists of three components: template selection, registration, and label fusion. Each of these components is presented in the following subsections.

Template Selection

The quality of a registration is directly related to the similarity between the template and the target images. Due to either differences between populations (e.g., age, disease, etc) or changes in scanner type, technology, and protocol (e.g., 1.5 T to 3 T), images from two different studies might be significantly different. To increase the template–subject similarity, and hence to improve the registration accuracy, we select within-study templates, instead of the commonly used strategy of selecting templates from a predefined external template library. To limit the work required for the preparation of the ground truth brain masks for the selected templates, the selection should not be done for each subject individually, but the same set of templates should be used for processing all images in the study. With this aim, we use a clustering-based approach to select a set of templates that capture the population variability as much as possible.

Let $D = \{I_1, \dots, I_n\}$, $I_i \in \Omega$ be a set of rigidly aligned raw T1 images of n subjects from a new study. We want to select a set of images as templates $T = \{I_1^t, \dots, I_k^t\}$, where $k \ll n$ (practically $k \ll n$) and $T \subset D$. We apply a k-means clustering-based selection strategy. In k-means clustering a given set of observations is partitioned into k sets $S = \{S_1, \dots, S_k\}$ so as to minimize the within-cluster distances between observations:

$$\operatorname{argmin}_s \sum_{i=1}^k \sum_{j \in S_i} |x_j - \mu_i|^2 \quad (1)$$

where $x_j \in \mathbb{R}^d$ is a data vector obtained by concatenating the voxel values of I_j , and the cluster center μ_i is the mean of data vectors from images in cluster S_i . We used l_2 -norm as our distance metric. For each cluster, the image closest to the cluster center is selected as the template:

$$I_i^t = \underset{I_j \in S_i}{\operatorname{argmin}} |x_j - \mu_i|^2 \quad (2)$$

Template brain masks $M = \{M_1^t, \dots, M_k^t\}$ are created using a semiautomated approach: Brain extraction by MASS is performed on the selected templates using external templates, and the resultant brain masks are manually corrected.

Registration

We have chosen a recently developed publicly available registration method, DRAMMS (19), because of its ability to meet two major challenges specific to registering raw brain MR images. The first major challenge is the large amount of intensity inhomogeneity and background noise in raw brain MR images. DRAMMS finds voxelwise correspondences by looking at multiscale and multiorientation Gabor texture features around each voxel. Therefore, it is relatively robust to inhomogeneity and noise. The second major challenge in registering brain MR images with skull is the possible presence of outlier regions. Outlier regions, or missing correspondences, usually refer to regions that exist in one image but not in the other. For instance, the MR image of one subject may contain more neck regions or may have part of the superior skull missing due to a different field of view (FOV) during MRI acquisition. DRAMMS meets this challenge using the mutual salience weighting, as it adaptively finds and relies on voxels/regions that are more likely to establish reliable correspondences across images. This way, it reduces the negative impact of outlier regions compared to other registration methods that force matching for all voxels/regions.

Let the template image I^t (target) and the subject image I^s (source) be the two images to be registered. DRAMMS calculates a deformation $h(x) = x - u(x)$ in a three-dimensional (3D) space, indexed by i, j, k in three orientations. The deformation h transforms I^s into the shape of I^t , where x is the 3D coordinates of voxels and $u = (u_i, u_j, u_k)$ is the displacement field that defines the mapping from the coordinate system of I^s to I^t . The inverse transformation of h , denoted as h^{-1} , is applied on the template mask M^t to warp the mask to subject space, that is, $M_s^t = M^t(h^{-1}(x))$. In this way, binary brain masks in subject space from each template are obtained.

The Jacobian map J is the determinant of the Jacobian matrix of the deformation h . The Jacobian determinant value at each voxel encodes the local volumetric change between the source and target image on that voxel. The Jacobian matrix at a voxel x is defined as

$$\operatorname{Jac}(x) = \begin{pmatrix} \frac{\partial h^2(x)}{(\partial i)^2} & \frac{\partial h^2(x)}{\partial i \partial j} & \frac{\partial h^2(x)}{\partial i \partial k} \\ \frac{\partial h^2(x)}{\partial i \partial j} & \frac{\partial h^2(x)}{(\partial j)^2} & \frac{\partial h^2(x)}{\partial j \partial k} \\ \frac{\partial h^2(x)}{\partial i \partial k} & \frac{\partial h^2(x)}{\partial j \partial k} & \frac{\partial h^2(x)}{(\partial k)^2} \end{pmatrix} \quad (3)$$

and the determinant of this Jacobian matrix is

$$J(x) = \det(\operatorname{Jac}(x)). \quad (4)$$

The Jacobian determinant is greater than 1 for volume expansion, between 0 and 1 for volume shrinkage, 0 for volume vanish, and smaller than 0 if self-folding occurs. We compute the Jacobian map J_s^t for each template for the subsequent label fusion step.

Label Fusion

We adopt a spatially adaptive fusion strategy that takes into consideration the local similarities between the templates and the target image. At each voxel, a weight is assigned to each template such that a higher confidence is given to templates that are locally more similar (e.g., more easily mapped) to the target image. Our main premise here is that the Jacobian maps are good indicators of local similarities between source and target images. Large Jacobian values often correlate with large geometric differences between template and target images. It is preferable to assign high weights to labels from masks that are locally similar to the subject image, as we have more confidence on the registration when the source and target images are more similar. Such a weighting mechanism is also efficient for making the method more robust. If the registration of one (or a few in the extreme case) template completely fails, the corresponding Jacobian map will have extreme values in most voxels. Thus, the brain mask from this template will be ranked very low in general, and the template will not have any effect on the final extraction/segmentation.

Let k atlases, indexed by i , be each registered to the same subject image via a deformation h_i , from which the Jacobian determinant map J_i is calculated. A voxel $\mathbf{u}$ in the subject image space will tentatively have k segmentation labels propagated from all those k atlases,

denoted as $label(h_i^{-1}(\mathbf{u}))_{i=1,\dots,k}=\{0,1\}$, where the label 1 denotes the brain and 0 the background. To fuse them into a brain probability map, we use a WV strategy using Jacobian determinant ranking. Specifically, the probability of this voxel being brain is calculated as:

$$Pr(label(\mathbf{u})=1)=\frac{\sum_i Rank(J_i(\mathbf{u})) \cdot label(h_i^{-1}(\mathbf{u}))}{\sum_i Rank(J_i(\mathbf{u}))} \quad (5)$$

where the $Rank()$ operator represents ranking of atlases at each voxel based on the Jacobian determinant value on this voxel. Its value is k if an atlas ranks highest in this voxel, and is $(k + 1 - i)$ if an atlas ranks i -th among all atlases. The calculated probability values are thresholded for obtaining a binary brain mask. A threshold value of 0.5 has been used in all validation experiments. This default threshold value could be considered as a “majority voting on weighted votes” assigning the voxel to the label that has more than 50% of the votes after weighting.

Postprocessing

As the voting mechanism has been applied independently at each voxel, the label fusion may lead to some irregularities within the brain mask. To provide a smoother and a more regular brain mask, we apply a few postprocessing steps. A morphological opening operator with a 4-mm kernel size is applied to isolate the main, contiguous brain region from other smaller nonbrain clusters, and to smooth the masks’ boundaries. The kernel size is derived based on cross-validated average Dice scores obtained on a set of 20 ADNI images different from the ones used in validation of MASS. A hole-filling operator is then applied to guarantee that the brain structure is complete without any holes.

DATA SETS

We used three publicly available data sets for evaluating the accuracy and robustness of MASS. These data sets are the same as those used in Iglesias et al (10). The first data set is the Internet Brain Segmentation Repository (IBSR). It consists of 20 T1-weighted scans from healthy subjects (aged 29.0 ± 4.8 years) acquired at the Center for Morphometric Analysis at Massachusetts General Hospital, as well as their corresponding annotations. The brain was manually delineated by trained investigators in all scans. The second test data set

is the LPBA40 data set. It consists of 40 T1-weighted scans (20 men and 20 women, aged 29.20 ± 6.30 years) and their corresponding annotations. The third test data set consists of the first two discs (77 T1-weighted scans) of the cross-sectional MRI data set of the Open Access Series of Imaging Studies (OASIS) project. The population consists of 55 women and 22 men (aged 51.64 ± 24.67 years). The brain masks for this set were not manually delineated; instead, the brain was segmented with an automated method based on registration to an atlas. However, the output from the method was reviewed by human experts before releasing the data, so the quality of the masks is good enough at least to test the robustness of a method. Despite this lack of exactitude, this data set is very valuable because it includes scans from a very diverse population with a very wide age range as well as diseased brains. A more detailed description of these three data sets could be found in Iglesias et al (10).

We also used a fourth data set in the experiments concerning the internal validation of MASS. This data set consists of the baseline scans of 178 controls from the ADNI study (20). It includes standard T1-weighted images obtained using volumetric 3D magnetization prepared rapid acquisition gradient echo or equivalent protocols with slightly varying resolutions.

RESULTS

Single Registration Comparison

In a first set of experiments, we analyzed the contribution of the registration component in the accuracy of the final segmentation. We compared the performance of DRAMMS to two recent registration methods, ANTS (21) and DEMONS (22), and to a widely used public registration method, FNIRT (23), in a within-group single-atlas brain extraction task.

A set of 15 subjects were selected from three data sets, five each from ADNI, IBSR, and OASIS. Within each group, each subject was registered to four other subjects using the four registration methods. All registration methods were used with the optimized parameters as disclosed in Klein et al (24). For each template–target pair, the resulting deformation was applied to the template brain mask to obtain the final segmented brain.

The comparison between the segmented brain M and the manual ground truth mask N is performed using two metrics that have been commonly used in comparison of binary segmentations: Dice similarity coefficient (25), given by

$$D(M, N) = \frac{2|M \cap N|}{|M| + |N|} \quad (6)$$

is a general measure of the segmentation accuracy. It quantifies the amount of overlap between the two binary masks. The Hausdorff distance (26), the maximal surface-to-surface distance between the two masks, is given as

$$H(M, N) = \max\{h(M, N), h(N, M)\} \quad (7)$$

where

$$h(M, N) = \max_{m \in M} \min_{n \in N} \|m - n\| \quad (8)$$

and $\|m - n\|$ is the Euclidian distance between the voxels m and n . It is generally used to measure the spatial consistency of the overlap between the two masks. As this metric is very sensitive to noise or local outliers in both masks, we calculated the 95% percentile of the surface-to-surface distance (Hausdorff 95%) as a more robust metric.

Figure 3 shows the box plots of the Dice scores for each registration method. The average Dice scores and 95% of the surface-to-surface distances are given in Table 1.

While the average performance of each method was highly variable on different data set, ANTS, DEMONS, and DRAMMS generally obtained high average Dice scores. FNIRT obtained a very low overall accuracy on ADNI and IBSR data sets compared to the other three methods. For these two data sets, DRAMMS obtained the highest accuracy, with significantly higher Dice scores, lower Hausdorff distances, and fewer outliers (i.e., those masks having important registration errors). ANTS performed better on the OASIS data, with higher Dice scores. However, DRAMMS performed comparably well on OASIS, with a 95.7 average Dice score and 4.3 average Hausdorff distance. DRAMMS obtained the first rank in both metrics for the overall results.

Multi-Atlas Label Fusion

In a second set of experiments, we evaluated (1) the performance of the multi-atlas framework compared to multiple single registration-based extractions and (2) the performance of the Jacobian determinant weighted label fusion (JWF) compared to MV and STAPLE (27), a statistical label fusion algorithm considered as a benchmark method.

The evaluation is performed on 20 images from the ADNI data set and 20 images from the OASIS data set, which have been selected randomly. The template selection component of our method is applied to select seven templates for each data set from the remaining images. Final brain masks have been created by using (1) each individual warped template mask and (2) the fusion of warped template masks using either MV, STAPLE, or JWF.

Table 2 shows the average Dice scores obtained for the ADNI and OASIS subjects using each different method. The box plots of the Dice scores for all 40 subjects are shown in Figure 4. For the single-atlas extraction, for each single subject, instead of calculating the average of the Dice scores from the seven registered template masks, we reported the median and the maximum Dice score values (SA_{med} and SA_{max}). The median Dice score gives a robust estimate of the accuracy of the single-atlas approach, while the maximum Dice score gives an upper bound of its performance (i.e., how it would perform if the optimal template was selected).

The multi-atlas strategy had a significantly higher accuracy compared to single-atlas extraction. Furthermore, for most subjects, the multi-atlas extraction performed better than the best single-atlas extraction (i.e., the warped template that obtained the maximum Dice score).

The three label fusion methods obtained very similar scores with no significant differences between them. The MV, while being the most straightforward fusion algorithm of the three, obtained slightly higher Dice scores than both STAPLE and JWF.

Comparative Evaluation

The overall performance of MASS is compared to the results reported in Iglesias et al (10). That report presented a quantitative comparison of a new multi-atlas-based method, ROBEX, against six popular brain extraction methods. The evaluation was performed on IBSR, OASIS, and LPBA40 data sets, three publicly available data sets with highly variable

characteristics in terms of acquisition and study populations. Ground truth brain masks were provided for each data set.

MASS is applied using seven automatically selected templates from each data set. Images that were selected as templates have been processed using the remaining six templates.

We calculated the Dice score, the average surface-to-surface distance, the Hausdorff distance, and the surface-to-surface distance in the 95% confidence interval as metrics for measuring the accuracy of the final masks compared to the ground truth masks. We also reported

$$\text{sensitivity} = \frac{|TP|}{|N|} \quad (9)$$

and

$$\text{specificity} = \frac{|TN|}{|N|} \quad (10)$$

values between the automatic segmentation M and the ground truth mask N , where the true positive set is defined as $TP = M \cap N$, the true negative set is defined as $TN = M \cap \bar{N}$, and $|\bullet|$ denotes the number of elements in a set. Table 3 displays the means and standard deviations of the calculated metrics for MASS, BET, and ROBEX.

MASS outperformed the six benchmark methods and the ROBEX method for all three different data sets in terms of the average Dice score. Also, MASS obtained the lowest surface-to-surface distance, both the average distance over the whole brain and the maximum distance in the 95 % confidence interval. Figure 5 presents examples of final brain masks. The first four rows show highly accurate automated segmentations, each one for a subject from a different data set. In the last two rows, we presented two example cases that obtained low Dice scores.

To visualize the spatial distribution of the segmentation errors, we calculated mean error projections on axial, sagittal, and coronal axes. First, error maps have been created. On these maps, voxels for which the automatic segmentation was different from the ground truth mask were set to 1. Each subject's T1 image is registered to a standard template using DRAMMS. The corresponding error map is warped to the template space using the calculated deformation field with the nearest-neighbor interpolation. Average error volumes on two-dimensional planes have been calculated by averaging all warped error maps and projecting them to axial, sagittal, and coronal axes. Figure 6 shows the mean error projections.

DISCUSSION

Multi-atlas label fusion methods provide a valuable generic framework for obtaining accurate and robust segmentation of anatomical structures. One of the major factors affecting the quality of the final segmentation is the quality of the individual registrations. Here, we addressed this issue from two different perspectives: selecting the optimal templates, and applying a registration algorithm well adapted to our task.

Most recent multi-atlas-based brain extraction papers are validated on data sets for which ground truth brain masks are provided for all subjects. Consequently, either a large number

of atlases (eventually all images) are used in the fusion without need to select templates, or template selection can be done for each subject independently. However, practically, these approaches may not be optimal for processing images from a new study, as they would require using external templates that could be significantly different from the study images. Our study-specific template selection strategy aims to provide a more adaptive alternative for practical applications of the method on future large scale studies.

We assume that ideally all study images are available before template selection. However, if this assumption does not hold, templates selected from a representative subset of the study population could still be used for processing new images without the need to select new templates each time a new case is introduced in the data set, as long as imaging characteristics of the new data do not differ significantly from the old data (e.g., due to changes in acquisition protocol).

We have shown that both DRAMMS and the use of multiple atlases significantly contributed to improve the accuracy of the final brain masks. Compared to MV and STAPLE, the Jacobian determinant-based weighted MV did not improve the segmentation accuracy. However, all three approaches resulted in very similar segmentations. After a visual inspection of individual registered masks for various subjects, we observed that these masks had a very high agreement in general. The disagreement area was limited to a narrow ribbon in the brain boundary (as shown in Figure 7). Consequently, the high agreement between individual registrations might have decreased the impact of the label fusion algorithms.

Another challenge for the validation is that the quality of the manually created ground truth masks is not always perfect, as manual segmentation of the brain is a cumbersome and time-consuming task. The inaccuracies on the ground truth masks may influence the results in two ways: Those on the templates would be translated to the final mask in the target space, while those on the target ground-truth mask would reduce the calculated Dice score. In our qualitative evaluation, we visually inspected regions of incorrect segmentations on several images. In some of these regions, we observed inaccuracies on the manual mask, while the automated mask was better in delineating the brain. Figure 8 shows a few cases where the disagreement between the masks comes mostly from inaccuracies on the ground truth masks. We observed that the multi-atlas approach was robust against inaccuracies on the template masks, as those were mostly randomly distributed. However, in the practical application of our software for a new study, a careful visual inspection and correction of the automatically created template brain masks are strongly recommended, as those would improve the quality of the segmentations for all images in the data set.

We quantitatively compared our method's performance to a recent multi-atlas-based brain extraction method, which was evaluated against six benchmark methods on three data sets with significantly different characteristics and obtained the highest rankings. Most other recent methods are not publicly available and/or are validated on data sets with custom ground truth masks. For this reason, a quantitative comparison against these methods was not possible.

The method has few parameters to set (i.e., DRAMMS parameters, the number of templates, and the threshold value of the JWF mask). All results reported here are based on the default parameters of DRAMMS registration. The default parameters have been carefully tuned in the DRAMMS software to work with brain images (raw and skull-stripped), breast images, and cardiac images, as tested in a number of public data sets with ground truth for measuring registration accuracy. To measure the effect of the number of templates on the final segmentation, we performed a comparative evaluation on all data sets. The accuracy

increases with more templates but converges to a stable value after seven templates (Figure 9). The JWF mask was thresholded at 50% of its maximum value in all validation experiments. In a semiautomated setting, for cases with significant errors, a more appropriate threshold value than the default one could be easily selected manually.

The computational time is a limitation of the method, as it requires nonlinear registration of the templates to the subject space. However, that could certainly be addressed through more advanced computing mechanisms, such as parallel computing.

We successfully applied MASS on several recent imaging studies, each containing thousands of MR images. The output masks created by MASS have been verified by an expert rater, and they have been manually corrected only if they were not good enough to be used for further processing. The method was easy to apply on new studies (i.e., with minimal extra manual work), and it gave very accurate brain masks, which did not require any manual correction for further processing, for an important percentage of the images.

Directions for future work would address main shortcomings of the proposed approach. For example, we did not directly use image intensity information, specifically from the target image, for the final segmentation. Incorporation of fuzzy segmentation maps, intensity features extracted from local neighborhoods, and intensity-based image similarity metrics into the framework are alternative approaches that could be followed. Also, the final postprocessing steps using morphological operators could be updated using techniques that are well suited for shape recognition, such as level set methods (28).

The software is available for public use and could be downloaded from <http://www.rad.upenn.edu/sbia/software/mass>.

Acknowledgments

Grants supporting the research: 5R01AG014971-11, Computational Neuroanatomy of Aging and AD via Pattern Analysis, and 5R01EB009234-04, Computer Analysis of Brain Vascular Lesions in MRI: Evaluating Longitudinal Change.

References

1. Battaglini M, Smith SM, Brogi S, et al. Enhanced brain extraction improves the accuracy of brain atrophy estimation. *Neuroimage*. 2008; 40(2):583–589. [PubMed: 18255315]
2. Zeng X, Staib LH, Schultz RT, et al. Segmentation and measurement of the cortex from 3D MR images using coupled surfaces propagation. *IEEE Transactions on Medical Imaging*. 1999; 18:927–937. [PubMed: 10628952]
3. Shattuck D, Sandor-Leahy S, Schaper K, et al. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*. 2001; 13(13):856–857. [PubMed: 11304082]
4. Huh S, Ketter TA, Sohn KH, et al. Automated cerebrum segmentation from three-dimensional sagittal brain MR images. *Comput Biol Med*. 2002; 32(5):311–328. [PubMed: 12102751]
5. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapping*. 2002; 17(3):143–155.
6. Huang, A.; Abugharbieh, R.; Tam, R., et al. MRI brain extraction with combined expectation maximization and geodesic active contours. *IEEE International Symposium on Signal Processing and Information Technology*; 2006. p. 107–111.
7. Wang, Y.; Nie, J.; Yap, P-T.; Shi, F.; Guo, L.; Shen, D. Robust deformable-surface-based skull-stripping for large-scale studies. *Proceedings of the 14th International Conference on Medical Image Computing and Computer-Assisted Intervention: Volume Part III, MICCAI'11*; Heidelberg: Springer-Verlag Berlin; 2011. p. 635–642.
8. Carass A, Cuzzocreo J, Wheeler MB, et al. Simple paradigm for extra-cerebral tissue removal: algorithm and analysis. *NeuroImage*. 2011; 56(4):1982–1992. [PubMed: 21458576]

9. Eskildsen SF, Coupe P, Fonov V, et al. Alzheimer's Disease Neuroimaging Initiative, BEaST: brain extraction based on nonlocal segmentation technique. *NeuroImage*. 2012; 59(3):2362–2373. [PubMed: 21945694]
10. Iglesias J, Liu C-Y, Thompson P, et al. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging*. 2011; 30(9):1617–1634. [PubMed: 21880566]
11. Leung, K.; Barnes, J.; Modat, M., et al. Automated brain extraction using Multi-Atlas Propagation and Segmentation (MAPS). *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*; 2011. p. 2053-2056.
12. FreeSurfer. <http://surfer.nmr.mgh.harvard.edu> (sw)
13. AFNI. <http://afni.nimh.nih.gov> (sw)
14. Mikheev A, Nevsky G, Govindan S, et al. Fully automatic segmentation of the brain from T1-weighted MRI using Bridge Burner algorithm. *J Magn Reson Imaging*. 2008; 27(6):1235–1241. [PubMed: 18504741]
15. Sadananthan S, Zheng W, Chee M, et al. Skull stripping using graph cuts. *NeuroImage*. 2010; 49(1):225–239. [PubMed: 19732839]
16. Artaechevarria X, Munoz-Barrutia A, Ortiz de Solorzano C. Combination strategies in multi-atlas image segmentation: application to brain MR data. *IEEE Trans Med Imaging*. 2009; 28(8):1266–1277. <http://dx.doi.org/10.1109/TMI.2009.2014372>. [PubMed: 19228554]
17. Isgum I, Staring M, Rutten A, et al. Multi-atlas-based segmentation with local decision fusion: application to cardiac and aortic segmentation in CT scans. *IEEE Trans Med Imaging*. 2009; 28(7):1000–1010. [PubMed: 19131298]
18. Sabuncu MR, Yeo B, Thomas T, et al. A generative model for image segmentation based on label fusion. *IEEE Trans Medical Imaging*. 2010; 29(10):1714–1729. <http://dx.doi.org/10.1109/tmi.2010.2050897>.
19. Ou Y, Sotiras A, Paragios N, et al. DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. *Med Image Analysis*. 2011; 15(4):622–639.
20. Jack C, Bernstein M, Fox N, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J Magn Reson Imaging*. 2008; 27(4):685–691. [PubMed: 18302232]
21. Avants BB, Epstein CL, Grossman M, et al. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Analysis*. 2008; 12(1):26–41.
22. Vercauteren T, Pennec X, Perchant A, et al. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage*. 2009; 1(45):S61–S72. [PubMed: 19041946]
23. Andersson, J.; Smith, S.; Jenkinson, M. FNIRT—FMRIB non-linear image registration tool. Fourteenth Annual Meeting of the Organization for Human Brain Mapping—HBM; 2008.
24. Klein A, Andersson J, Ardekani B, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*. 2009; 3(46):786–802. [PubMed: 19195496]
25. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945; 26(3): 297–302.
26. Huttenlocher DP, Klanderman GA, Rucklidge WA. Comparing images using the hausdorff distance. *IEEE Trans Pattern Anal Mach Intell*. 1993; 15(9):850–863.
27. Warfield SK, Zou KH, Wells WM. Simultaneous Truth and Performance Level Estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004; 23:903–921. [PubMed: 15250643]
28. Sethian, JA. Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. 2. UK: Cambridge University Press; 1999.

Figure 1.
Brain extraction example.

Figure 2.

Outline of the method. In the registration component (*middle*), binary brain masks registered to target space are shown in *yellow*; in the label fusion component (*bottom*) the fused brain mask is shown using a *blue-red* color map, and the final binary brain mask is shown in *yellow*. (Color version of the figure is available online.)

Figure 3.

Box plot of Dice scores for within-group single-atlas-based brain extraction using four different registration methods. The plots show scores for each dataset independently (ADNI, IBSR, OASIS), and for all images combined (ALL).

Figure 4.
Dice scores for brain masks obtained by single-atlas and multi-atlas fusion strategies.

Figure 5.

Final brain masks generated by Multi-Atlas Skull-Stripping (MASS). Left: The ground truth brain masks are overlaid (in *yellow*) on the coronal and sagittal views of the brain. Right: Final MASS masks are overlaid (in *blue*) on the same views. From top to bottom: the first four rows, example segmentations from ADNI, OASIS, LONI, and IBSR, respectively; last two rows, two cases with segmentation errors (undersegmented or oversegmented brain). (Color version of figure is available online.)

Figure 6. Mean error projections along the axial, sagittal, and coronal axes for the Multi-Atlas Skull-Stripping (MASS) segmentation. (Please note that the values are normalized by the number of slices on each axis for better interpretability; as such, they represent the rate of error voxels along each axis). (Color version of figure is available online.)

Figure 7.

Agreement of individual registered masks. The probabilistic brain mask is calculated by the voxelwise sum of all binary template masks in the subject space, normalized by the number of templates. (Color version is available online.)

Figure 8. Examples of segmentation inaccuracies on the ground truth masks. (Color version is available online.)

Figure 9.
Dice scores for different number of templates.

TABLE 1

Average Dice Scores and 95% of the Surface-to-Surface Distances for within-Group Single-Atlas-Based Brain Extraction Using Four Different Registration Methods

	ANTS	DEMONS	DRAMMS	FNIRT
Dice score				
ADNI	93.1 ± 1.1	92.5 ± 4.6	98.0 ± 0.3	70.6 ± 12.3
IBSR	91.7 ± 3.7	94.6 ± 1.8	95.9 ± 2.5	74.3 ± 8.9
OASIS	96.8 ± 0.9	95.7 ± 0.6	95.7 ± 0.5	94.2 ± 1.6
ALL	93.8 ± 3.1	94.3 ± 3.1	96.5 ± 1.8	79.7 ± 13.6
Hausdorff distance (95%)				
ADNI	6.05 ± 7.97	11.66 ± 10.69	2.27 ± 2.82	23.79 ± 28.74
IBSR	7.14 ± 7.17	4.78 ± 5.84	4.68 ± 6.71	20.88 ± 36.03
OASIS	3.28 ± 0.98	4.50 ± 0.61	4.3 ± 0.49	5.82 ± 1.52
ALL	5.49 ± 2.89	6.98 ± 6.14	3.75 ± 2.25	16.83 ± 9.89

ALL: Images from the three datasets combined.

TABLE 2

Average Dice Scores for Brain Masks Obtained Using Single-Atlas Registration (SA_{med} , SA_{max}), Multi-Atlas Label Fusion Using Majority Voting (MV), STAPLE, and the Jacobian Determinant-Weighted Label Fusion (JWF)

	SA_{med}	SA_{max}	MV	STAPLE	JWF
ADNI	97.8 ± 0.2	98.1 ± 0.2	98.7 ± 0.1	98.7 ± 0.2	98.7 ± 0.1
OASIS	95.8 ± 0.4	96.5 ± 0.7	97.1 ± 0.2	97.1 ± 0.2	97.0 ± 0.2

TABLE 3
Comparison of MASS to BET and ROBEX on IBSR, OASIS, and LPBA40 Data Sets

Dataset	Method	Dice	Average Distance	Hausdorff	Hausdorff 95%	Sensitivity	Specificity
IBSR	BET	93.8 ± 2.9	2.20 ± 1.20	19.10 ± 9.20	6.20 ± 6.2	99.0 ± 3.5	89.1 ± 2.8
	ROBEX	95.6 ± 0.8	1.50 ± 0.30	13.30 ± 2.60	3.80 ± 0.70	99.2 ± 0.5	92.3 ± 1.9
	MASS	97.7 ± 0.8	0.71 ± 0.30	15.01 ± 10.72	2.81 ± 1.10	97.5 ± 0.9	99.8 ± 0.2
OASIS	BET	93.1 ± 3.7	2.70 ± 1.40	23.70 ± 8.30	8.20 ± 5.50	92.5 ± 5.4	94.2 ± 5.0
	ROBEX	95.5 ± 0.8	1.80 ± 0.30	9.80 ± 1.70	4.40 ± 0.60	93.8 ± 2.1	97.4 ± 12.0
	MASS	96.1 ± 1.0	1.60 ± 0.36	7.72 ± 1.51	3.81 ± 0.77	94.7 ± 2.6	99.2 ± 0.7
LPBA40	BET	97.3 ± 0.5	1.0 ± 0.20	14.20 ± 4.30	3.0 ± 1.0	97.0 ± 1.3	97.7 ± 0.8
	ROBEX	96.6 ± 0.3	1.20 ± 0.10	13.30 ± 2.50	3.10 ± 0.40	95.6 ± 9.0	97.7 ± 7.0
	MASS	98.2 ± 0.4	0.66 ± 0.13	9.01 ± 4.07	1.95 ± 0.34	98.2 ± 0.6	99.7 ± 0.2