Quantitative Comparison of Automatic ICV Segmentation Methods Using 3T MR Images

Suja S
GE Global Research, Bangalore
Suja.S@geind.ge.com

Rakesh Mullick
GE Global Research, Bangalore
Rakesh.Mullick@ge.com

Abstract

Skull-stripping for the MR brain images and computation of the Intra-Cranial Volume (ICV) appears as a key requirement for almost all neuro-image analysis. Segmentation of brain/non-brain tissue is one of the most time consuming pre-processing steps performed in neuro-imaging laboratories. Numerous Brain Extraction Algorithms (BEAs) have been developed to perform this step automatically. While BEAs speed up overall image processing, their output quality varies greatly and can affect the results of subsequent image analysis. We therefore quantitatively compared the performance of six brain extraction methods against manual brain extraction using a set of high-resolution T1- weighted 3T MRI brain volumes. The optimized algorithms were benchmarked based on the Similarity Index. The overall evaluation confirms the variability observed visually and a computationally intensive template-based registration approach yields results comparable to human performance.

Index Terms— Quantification, MR Image segmentation, Similarity Index, Skull Stripping.

1. Introduction

Brain extraction algorithms (BEA) "remove the skull and scalp and maintains the brain" which usually includes white and grey matter as well as CSF. Intracranial volume is often used as a correcting factor, for example as an age and/or gender normalizing factor to compare the volume of cerebral substructures between subjects.

Many applications, such as presurgical planning, cortical surface reconstruction and brain morphometry, either require, or benefit from, the ability to accurately segment brain from non-brain tissue. However, despite the clear definition of this essential step, no universal solution has been developed that is robust to neuroanatomical variability and the types of noise present in standard MRI sequences.

The skull-stripping problem is a difficult task, which aims at extracting the brain from the skull, eliminating all non-brain tissue such as bones, eyes, skin, fat…This important stage requires a semi-global understanding of the image, as the brain is constituted of different structures with different geometric and intensity properties (cerebellum, cortex etc), as well as a local understanding (accurate localization of the pial surface). This is often more difficult than situations where purely local or purely global solutions are appropriate. In addition, the presence of imaging artifacts, anatomical variability, varying contrast properties and poor registration considerably increase the difficulty of this essential step in computational analysis of neuroimaging data. The image processing techniques found in literature can be divided into three groups:

➢ *Region-based*. The most common approaches sequentially apply morphological operations and manual editing. First, the white matter gray values are located using thresholding or seeded region growing, followed by a morphological opening that detaches the brain tissue from the surrounding tissue. Morphological dilation and closing are required for the segmentation to cover the whole brain without holes [1, 2, 3, 4, 5, 6].

➢ *Hybrid methods*. In order to account for the shortcomings of morphological multistep approaches, they have been combined with edge-based methods [2, 7, 8, 9]. For example, contours (snakes) are applied to the morphological segmentation as a final step [8].

➢ *Template-based*. More recently, some investigators succeeded in fitting a balloonlike surface to the intensity-normalized MR data in order to separate the brain from surrounding structures [10]. Another three-dimensional template based approach using volume data is described in [11].

Manual brain/non-brain segmentation methods are, as a result of the complex information understanding involved, probably more *accurate* than fully automated methods are ever likely to achieve. At the lowest, most localized, level (for example, noise reduction or tissue-type segmentation), humans often cannot improve on the numerical accuracy and objectivity of a computational approach. The same also often holds at the highest, most global, level; for example, in image registration, humans cannot in general take in enough of the whole-image information to improve on the overall fit that a good registration program can achieve. However, with brain segmentation, the appropriate size of the image "neighborhood" which is considered when outlining the brain surface is ideally suited to manual processing. For example, when following the external contours of gyri, differentiating between cerebellum and neighboring veins, cutting out optic nerves, or taking into account unusual pathology, semi-global contextual information is crucial in helping the human optimally identify the correct brain surface.

Of course, there are serious enough problems with manual segmentation to prevent it from being a viable solution in most applications. The first is time cost – manual brain/non-brain segmentation typically takes 2 hours per 3D volume (256x256x100(2mm) slices). The second is the requirement for sufficient training, and care during segmentation, that subjectivity is reduced to an acceptable level. For example, even a clinical researcher who has not been explicitly trained will be likely to make a mistake in the differentiation between lower cerebellum and neighboring veins. Accurate measurement of brain volume is important for longitudinal or comparative studies. These typically involve a large number of image volumes to be analyzed, thus calling for automatic image processing methods.

Due to the complexity of the brain surface, there is at present no segmentation method that proves to work automatically and consistently on any 3D MR images of the head. There is a definite lack of validation studies related to automatic brain extraction. In this paper we evaluated the validity of several well-known brain stripping algorithms (Brain Extraction Tool (BET), Brain Surface Extractor (BSE), ANALYZE 5.0, AFNI, McStrip and ITK) by comparing it with a manually labeled set of images as a gold standard. To compare the result quantitatively, we used the Similarity Index (SI).

In theory it might be possible to "hand-tune" a method once for a given MR pulse sequence, and the resulting parameters then work well for all images of all subjects acquired using this sequence. Following this assumption, we have done 'hand-tuning' only once for each algorithm for the set of images.

Finally, note that results from a brain extraction algorithm may improve if the image is pre-processed in certain ways, such as with an intensity inhomogeneity reduction algorithm. But, it is also true that the best intensity inhomogeneity reduction methods require brain extraction to have already been carried out.

Our study is based on T1 weighted 3T images. The great appeal of 3T MRI is improvement in image quality and resolution. Because the signal-to-noise ratio (SNR) correlates in approximately linear fashion with field strength, it is roughly twice as great at 3T as at 1.5T. Also, greater contrast is available at higher field strength, a fact already well known from comparisons of images obtained at 0.5T, 1T, and 1.5T.

2. Materials and Methods

2.1 Data

We used 18 T1 weighted MR images, taken from 3T GE scanner. The volumes have slice dimensions of 256 x 256, with resolution of .781 and 0.859mm. Interslice distance is 2 mm with the number of slices for each volume 54 or 96.

We reviewed the existing algorithms and surveyed all available algorithms. These include BET, BSE, ANALYZE 5.0, AFNI, McStrip and ITK. We give a brief high-level description of the algorithm below.

2.2 Brain Extraction Algorithms

2.2.1.Brain Extraction Tool (BET) (Smith [12]):

BET is developed by FMRIB (Oxford Centre for Functional Magnetic Resonance Imaging of the Brain) and is available at http://www.fmrib.ox.ac.uk/fsl/ for research purposes. In BET, the intensity histogram is processed to find "robust" lower and upper intensity values for the image, and a rough brain/non-brain threshold is determined. The center-of-gravity of the head image is found, along with the rough size of the head in the image. Next a triangular tessellation of a sphere's surface is initialized inside the brain, and allowed to slowly deform, one vertex at a time, following forces that keep the surface well-spaced and smooth, whilst attempting to move towards the brain's edge. If a suitably clean solution is not arrived at then this process is re-run with a higher smoothness constraint. Fig. 1 shows a flowchart of BET.

Optimization: In BET, we used an intensity threshold of 0.2 and a threshold gradient 0.5. These parameters were selected after hand-tuning the algorithm for one of the volumes to get the best DSC with manually segmented volume.

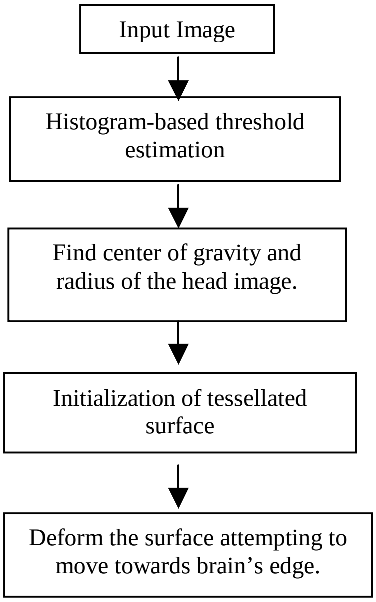

Figure 1: BET processing Flow chart.

2.2.2. Brain Surface Extractor (BSE) [13]

BSE is developed by NeuroImaging Research Group, University of Southern California and the executable is available from http://neuroimage.usc.edu/BSE/.

BSE is an edge- based method that employs anisotropic diffusion filtering. Edge detection is implemented using a 2D Marr-Hildreth operator, employing low-pass filtering with a Gaussian kernel and localization of zero crossings in the Laplacian of the filtered image. The final step is morphological processing of the edge map. Fig.2 shows a processing flow diagram of BSE.

Optimization: We used a diffusion constant of 5.0 and edge detector kernel as 0.98 for processing of all 18 volumes.

2.2.3. Analyze (Mayo Clinic) [14]

ANALYZE 5.0 is a comprehensive and interactive package for multidimensional image visualization, processing and analysis developed by The Biomedical Imaging Resource at Mayo Clinic, Rochester, MN (http://www.mayo.edu/bir/Software/Analyze/Analyze.html). The following steps describe the semiautomated, combined iterative segmentation procedure.

The data, after subjecting to grayscale inversion, is thresholded at the prefill value (For skull stripping of T1 weighted volumes, the pre-fill value is calculated as (0.11 * VolumeMax + 3.5 * n), where 'n' is the standard deviation of the high frequency noise in the background). Thresholding is followed by labeling of all connected components (basins) with the minimum intensity value within the basin. The pre-fill level is incremented. Voxels at the new prefill intensity are merged with the most neighboring basin. If a neighboring basin does not exist, these voxels create their own basins. If there are two neighboring basins, the basins are merged into the basin of lower intensity if the difference relative to the current voxel intensity is less than the original prefill value.

Optimization: Altering the default pre-fill value didn't have positive effect on the output image.

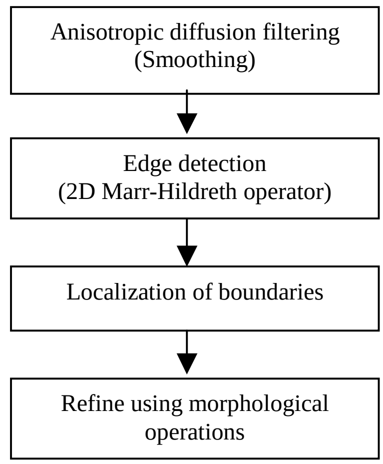

Figure 2: Flow diagram of BSE

2.2.4. AFNI [15]

AFNI (Analysis of Functional NeuroImages) is developed by NIH (National Institutes of Health) and can be downloaded from http://afni.nimh.nih.gov/afni/. The algorithm begin with the middle axial slice, voxels are preliminarily classified as belonging to the brain if their intensities are within the given bounds. The result is the set of (in_brain) voxels R. Note that R is usually a disconnected set. Since the brain needs to be connected, the set S is formed that is the largest subset of R, starting from the center, which is connected. Now, although S is connected, it may contain holes. To eliminate the holes, the set T is formed which is the largest connected set, starting from outside the brain, which excludes S. Therefore, the complement of T, $U = T^C$ contains S.

Optimization: We used a minimum value 25 and maximum value 125 as threshold. The minimum and maximum connectivity to enter and leave taken are the same as default values.

2.2.5. McStrip (Minneapolis Consensus Stripping): [16], [17]

McStrip is developed by International Neuroimaging Consortium (INC) and is available for download at http://www.neurovia.umn.edu/incweb/McStrip_download.html. McStrip is initialized with a warp mask using AIR [18], and dilates the AIR mask to form a Coarse Mask. It then estimates a brain/ non-brain threshold based on the intensity histogram, and automatically adjusts this threshold to produce a Threshold Mask. The volume of tissue within the Threshold Mask determines the choice of the BSE Mask from among a suite of 15 masks computed using parameter combinations spanning both smoothing and edge parameters. The final, McStrip Mask is a union of the Threshold and BSE masks after void filling and smoothing.

Optimization: Parameters used were as follows: anisotropic smoothing kernels: 5, 10, 15; iterations, 3; edge detection σ's: 0.60, 0.64, 0.70, 0.80, 0.90.

2.2.6. Insight Segmentation and Registration Tool kit (ITK) – Brain Strip Application [19]

ITK is an open-source software system developed by six principal organizations and is available at http://www.itk.org/.

In this atlas-based segmentation, delineation (manually or otherwise) is done only once on an "atlas" image. To segment other images, the atlas is non-rigidly registered ("Demons" Deformable Registration Algorithm based on histogram matching) to each image. If the registration is done correctly then each of the subject images can be segmented automatically. Any structure delineated in the atlas can be projected onto the subject image by applying the deformation field (obtained from the registration) to the atlas segmentation masks of each structure. Fig. 3 shows the flow chart of Atlas based segmentation of ITK.

Optimization: Number of match points used in the histogram matching: 256; Number of histogram bins in histogram matching: 14; Number of resolution levels in the multi-resolution demons registration algorithm: 4; Number of iterations of the demons algorithm at each resolution level: 256, 64, 32, 8.

2.3 Quantitative Comparison [20]

Performance of our brain volume segmentation was assessed using a kappa statistic based similarity index. The metric measures the overlap between two sets *A* and *B* as follows:

$$S = \frac{2|A \cap B|}{(|A| + |B|)}$$

Where |·| represents the size of the set, and ∩ represents the intersection of the two sets.

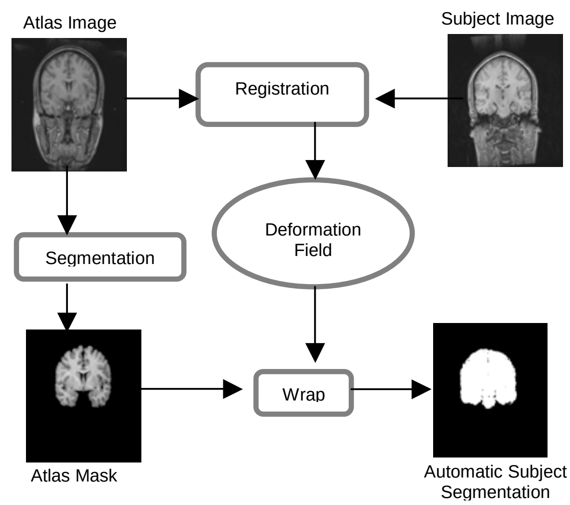

Figure 3: Atlas Based segmentation of ITK.

This index ranges from 0 (no overlap) to 1 (perfect alignment). The numerator of *S* is a measure of the overlap between the two sets and the denominator is the mean volume of the two sets. This index takes into account both the size and location of the overlap.

In our case, the comparison of the semi-automatic methods was done with manually segmented masks of Intracranial Volume (ICV). The confidence on the manually segmented masks was derived based on a 3-observer repeatability study conducted at two sites for the same set of 18 data sets. The similarity coefficient for the 3-way study for the ICV was 97.96 with a SD of 0.513. The Manual segmentations were done at Manipal Hospital, Bangalore, India and Albany Medical Center, Albany, NY.

3. Results

On these 18 data sets, ITK outperformed other methods, resulting in segmentations with better similarity coefficients. The performance of McStrip was unsatisfactory and it failed on 3 datasets. The reason for the poor performance of McStrip may be because the subject images (3T) varied a lot from the in-built template images. Analyze stripped away a large region of image and hence resulted in a low SI. Table 1 shows the result of each algorithm in terms of speed and SI when compared with manually segmented volume. Fig. 4 shows a plot of the similarity index of each method to the ground truth for the 18 volumes. Fig. 5 shows the output of each of the algorithms for the best case (Case-4), and fig. 6 shows the result for the worst case (Case-8) for 3 coronal slices of the volume.

Case	BSE	BET	AFNI	ITK	McStrip	Analyze
1	87.833	90.969	89.873	95.825	84.254	79.171
2	83.433	89.067	83.097	91.628	80.038	55.838
3	87.045	92.140	69.956	95.178	83.665	57.088
4	88.721	82.140	82.300	96.373	86.551	53.375
5	87.861	84.526	82.866	96.041	failed	84.446
6	88.726	91.520	88.204	93.171	90.282	36.967
7	83.642	89.940	84.545	93.693	90.789	57.957
8	83.986	92.521	82.263	90.294	77.162	45.313
9	88.187	90.622	87.825	92.237	89.847	48.535
10	88.303	90.626	83.961	94.103	90.253	46.983
11	87.025	91.086	84.679	93.357	failed	52.028
12	86.734	90.634	86.920	94.367	78.872	45.232
13	87.419	90.404	88.868	92.021	89.373	83.741
14	82.197	90.607	79.492	94.377	45.917	79.337
15	84.185	88.675	83.669	91.397	68.966	44.653
16	84.110	89.162	92.716	91.562	80.017	64.372
17	88.419	56.736	77.442	93.297	failed	78.809
18	88.667	61.467	81.046	94.040	76.052	79.203
Mean	86.472	86.269	83.873	93.498	80.803	60.725
SD	2.209	10.239	5.161	1.733	11.604	15.854
Time	20secs	15secs	5secs	36mins	~1 hr	< 5secs

Table 1: Similarity Index of 6 algorithms on 5 different volumes.

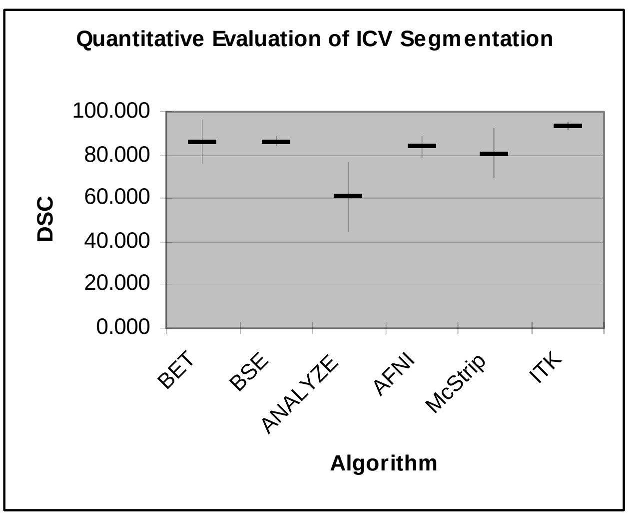

Figure 4: The similarity index (mean +- standard deviation) of each method to the ground truth.

4. Discussion and Conclusion

There are no exact locations along the brain boundary where any particular automatic segmentation technique consistently fails. In general, the image model segmentation method fails in some cases by including the orbital fat above the eye, or by excluding portions of the cerebellum.

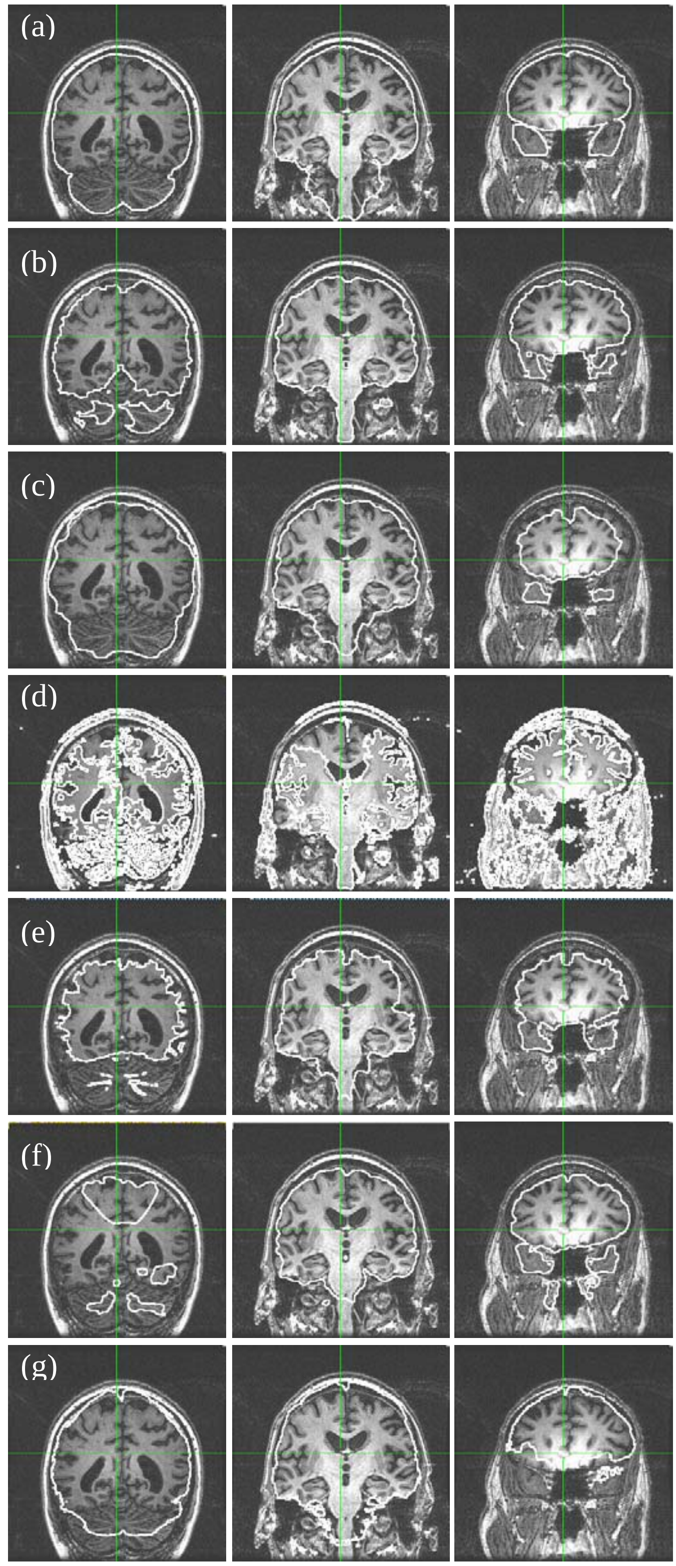

Figure 5: Results of skull-stripping of manual & 6 algorithms (a) Manual segmentation, (b) BET, (c) BSE, (d) ANALYZE (e) AFNI, (f) McStrip and (g) ITK on the best case (Case – 4).

The atlas model segmentation method fails due to misregistration error caused by a large amount of fat. This may cause missing a small portion of the cerebellum border, or including the dura and the scalp.

The result of ITK was more or less the same with the input parameters, where as the result of all other algorithms varied a lot depending upon the initial parameters. Even though not as good as the inter-observer variability measure of manual segmentation, the similarity index of ITK (Literature says SI > 70% is a good quality segmentation) met expectation for inclusion into the workflow.

It may be necessary, on occasion to manually touch up the segmented volumes when the subject volumes vary much from the atlas volume. The overall goal to enhance productivity of manual segmentation is achieved by a significant time reduction from almost 2 hours to a fraction of an hour (<15mins). It is observed that this gain in time, yielding a significant productivity gain for down stream brain image analysis.

5. Summary

The skull-stripping is concerned with the predominant tissues of the brain and can be classified into automated or semi-automated method according to the degree of user intervention. Despite the increasingly widespread use of skull-stripping methods as pre-processing stage for the various fields in neuroimage analysis, only a few attempts have been made to assess the performance of this procedure. Due to the lack of the ground truth in skull-stripping, the assessment of the actual skull-stripping accuracy can be quite subjective.

In this paper, we have presented an objective (quantitative) evaluation of several well-known skull-stripping algorithms-BET, BSE, ANALYZE 4.0, AFNI, McStrip and ITK Brain Strip Application. All methods had been tested on routine/volunteer T1-weighted MRI. In the analysis, ITK outperformed other algorithms in terms of SI measure when compared with manual segmentation (ground truth). Though our analysis is based on 3T MR, a similar study for 1.5T MR is being considered.

6. Acknowledgment

We are grateful to Dr. Uday Patil and Dr. Dawn Thomas, Manipal Hospital, Bangalore, India, who spent an incredible time tracing brain contours in this research. The authors also thank Ajay Narayanan for his VTK based tool for computing SI.

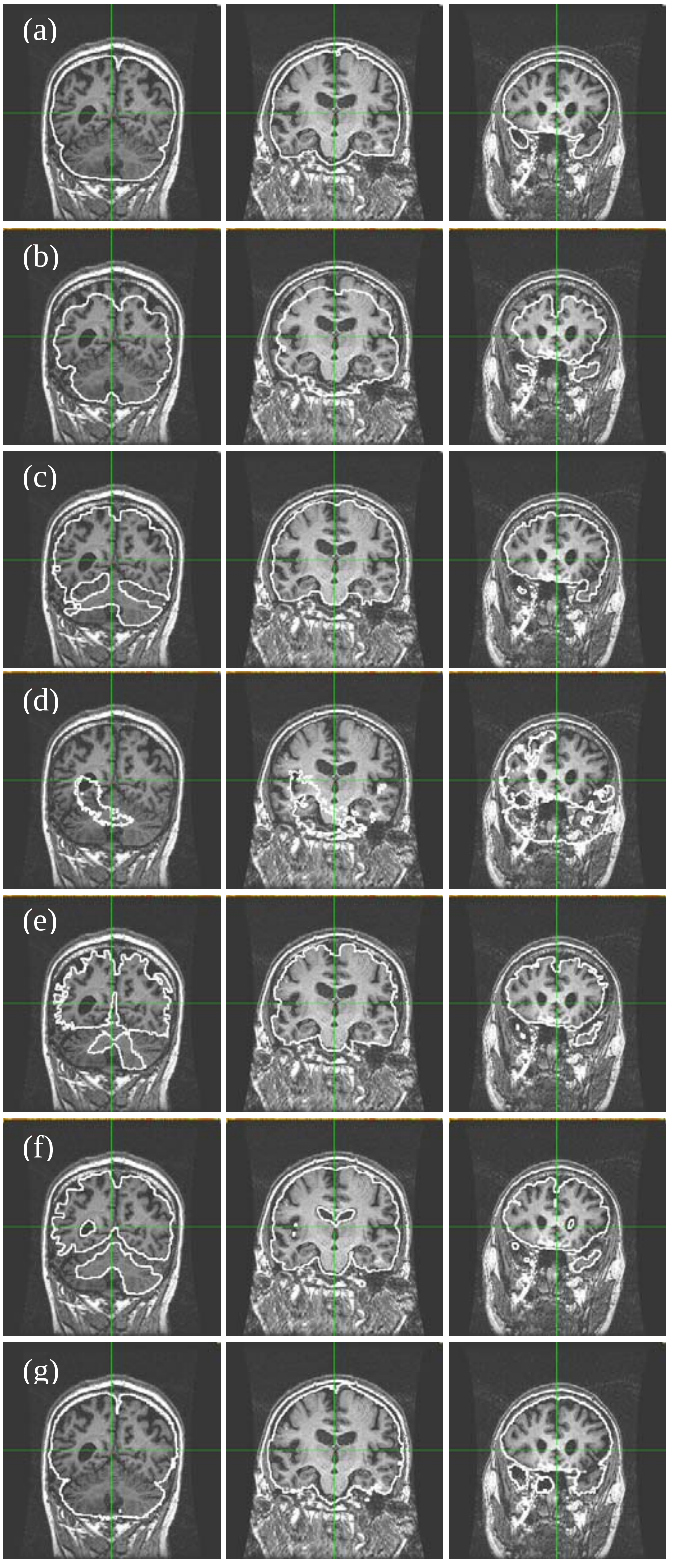

Figure 6: Results of skull-stripping of manual & 6 algorithms (a) Manual segmentation, (b) BET, (c) BSE, (d) ANALYZE (e) AFNI, (f) McStrip and (g) ITK on the worst case (Case - 8).

References

[1] R.Kikinis, M.E.Shenton, G.Gerig, J.Martin, M.Anderson, D.Metcalf, C.R.G.Guttmann, R.W.McCarley, W.Lorensen, H.Cline, and F.A.Jolesz. “Routine Quantitative Analysis of Brain and Cerebrospinal Fluid Spaces with MR Imaging,” J. Magnetic Resonance Imaging, 2: 619-629, 1992.

[2] T.Kapur, W.E.L.Grimson, W.M.Wells, and R.Kikinis. “Segmentation of Brain Tissue from Magnetic Resonance Images,” Medical Image Analysis 1(2): 109-127, 1996.

[3] S.Sandor and R.Leahy. “Surface-Based Labeling of Cortical Anatomy using a Deformable Atlas” IEEE Trans. Med. Imaging. 16(1): 41-54, Feb. 1997.

[4] P.A.Freeborough, N.C.Fox, and R.I.Kitney. “Interactive Algorithms for the Segmentation and Quantification of 3D MRI Brain Scans” Comput. Metho. Progr. Biomed. 53(1): 15-25, May 1997

[5] A.F.Goldszal, C.Davatzikos, D.L.Pham, M.X.H.Yan, R.N.Bryan, and S.M.Resnick. “An Image Processing System for Qualitative and Quantitative Volumetric Analysis of Brain Images” J. Comput. Assist. Tomogr. 22(5): 827-837, 1998.

[6] S.A.Hojjatoleslami, F.Kruggel, and D.Y.von Cramon. “Segmentation of White Matter Lesions from Volumetric MR Images” Proc. Medical Image Computing and Compter- Assisted Intervention MICCAI '99, Cambridge. Springer LNCS, vol. 1679: 52-61, 1999.

[7] G.B.Aboutanos and B.M.Dawant. “Automatic Brain Segmentation and Validation: Image- Based versus Atlas-Based Deformable Models” Proc. SPIE Med. Imaging '97, 299-310,Feb. 1997.

[8] M.S.Atkins and B.T.Mackiewich. “Fully Automatic Segmentation of the Brain in MRI” IEEE Trans. Med. Imaging. 17(1): 98-107, Feb. 1998.

[9] K.Rehm, D.Shattuck, R.Leahy, K.Schaper, and D.Rottenberg. “SemiAutomated Stripping of T1 MRI Volumes: I. Consensus of Intensity- and Edge-Based Methods” Proc. 5th Int. Conf. on Functional Mapping of the Human Brain HBM '99, Düsseldorf, poster no. 86, 1999, abstract pub. in NeuroImage.

[10] A.M.Dale, B.Fischl, and M.I.Sereno. “Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction” NeuroImage 9: 179-194, 1999.

[11] R.S.J.Frackowiak, K.J.Friston, C.D.Frith, R.J.Dolan, and J.C.Mazziotta. Human Brain Function, Academic Press, San Diego, 1997. Statistical Parametric Mapping SPM, http://www.fil.ion.ucl.ac.uk/

[12] S.M. Smith, Fast robust automated brain extraction, Hum. Brain. Mapp. 17 (2002) 143–155.

[13] D.W. Shattuck, S.R. Sandor-Leahy, K.A. Schaper, D.A. Rottenberg, R.M. Leahy, Magnetic resonance image tissue classification using a partial volume model, NeuroImage 13 (2001) 856–876.

[14] http://www.mayo.edu/bir/Software/Analyze/Analyze.html

[15] R. Cox. AFNI software package. afni.nimh.nih.gov.

[16] Rehm K, Shattuck D, Leahy R, Schaper K, Rottenberg DA. Semi-automated stripping of T1 MRI volumes: I. Consensus of intensity- and edge-based methods. NeuroImage. 9(6) S86, 1999.

[17] Rehm K, Schaper K, Anderson J, Woods R, Stoltzner S, Rottenberg D. Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. NeuroImage, November, 2003.

[18] http://bishopw.loni.ucla.edu/AIR5/

[19] http://www.itk.org/cgi-in/cvsweb.cgi/InsightDocuments/Validation/AtlasSegmentation/AtlasSegmentationStudy.doc?cvsroot=Insight&rev=1.2

[20] Zou et al. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. Acad Radiol 2004; 11:178-189.