Robust Brain Extraction Across Datasets and Comparison with Publicly Available Methods

Juan Eugenio Iglesias, Cheng-Yi Liu, Paul Thompson and Zhuowen Tu

***Abstract*—Automatic whole-brain extraction from magnetic resonance images (MRI), also known as skull stripping, is a key component in most neuroimage pipelines. As the first element in the chain, its robustness is critical for the overall performance of the system. Many skull stripping methods have been proposed, but the problem is not considered to be completely solved yet. Many systems in the literature have good performance on certain datasets (mostly the datasets they were trained/tuned on), but fail to produce satisfactory results when the acquisition conditions or study populations are different.**

In this paper we introduce a robust, learning-based brain extraction system (ROBEX). The method combines a discriminative and a generative model to achieve the final result. The discriminative model is a Random Forest classifier trained to detect the brain boundary; the generative model is a point distribution model that ensures that the result is plausible. When a new image is presented to the system, the generative model is explored to find the contour with highest likelihood according to the discriminative model. Because the target shape is in general not perfectly represented by the generative model, the contour is refined using graph cuts to obtain the final segmentation. Both models were trained using 92 scans from a proprietary dataset but they achieve a high degree of robustness on a variety of other datasets.

ROBEX was compared with six other popular, publicly available methods (BET [1], BSE [2], FreeSurfer [3], AFNI [4], BridgeBurner [5] and GCUT [6]) on three publicly available datasets (IBSR [7], LPBA40 [8] and OASIS [9], 137 scans in total) that include a wide range of acquisition hardware and a highly variable population (different age groups, healthy/diseased). The results show that ROBEX provides significantly improved performance measures for almost every method / dataset combination.

***Index Terms*—Skull stripping, Random Forests, point distribution models, minimum s-t cut, comparison.**

I. Introduction and Background

WHOLE brain segmentation, also known as skull stripping, is the problem of extracting the brain from a volumetric dataset, typically a T1-weighted MRI scan. This process of removing non-brain tissue is the first module of most brain MRI studies. Applications such as brain morphometry, brain volumetry, and cortical surface reconstructions require stripped MRI scans. Even early preprocessing steps such as bias field correction can benefit from skull stripping.

Automatic skull stripping is a practical alternative to manual delineation of the brain, which is extremely time consuming. Segmentation in MRI is in general a difficult problem due to the complex nature of the images (ill-defined boundaries, low contrast) and the lack of image intensity standardization. Over the last decade, the research community has produced a number of methods. However, as shown below, these systems fail to consistently provide highly accurate segmentations across datasets acquired with different protocols.

Some aspects of whole brain segmentation are not very well defined. There seems to be consensus that skull stripping is expected to follow the major folds on the surface; if the deeper sulci are to be extracted for brain surface analysis, subsequent post-processing can be performed. However, protocols differ on which parts of the brain to extract. Most include the cerebellum and brainstem in the segmentation (see Figures 1a and 1b), but others only extract the cerebrum, leaving the cerebellum and brainstem out (Figure 1c).

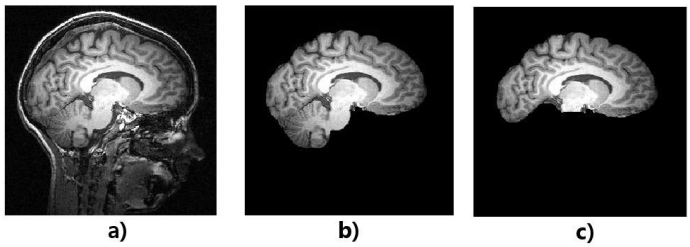

Fig. 1. a) Saggital slice of a T1-weighted MRI. b) Corresponding slice of the manually skull-stripped volume, including the cerebellum and brainstem. c) Skull stripping without the cerebellum or brainstem. It is the purpose of this study to automatically generate segmentations like the one in b).

The majority of skull stripping methods are designed to work with T1-weighted MRI for two reasons: 1. T1 it is the most frequent MRI modality in neuroimaging; and 2. even if another modality (such as T2 or FLAIR) is to be segmented, it is very likely that the data were acquired next to a T1 scan, which provides superior contrast. In that case, skull stripping is usually performed on the T1 volume and the resulting mask propagated to the other channels.

Most popular skull stripping methods have publicly available implementations. We consider six in this study. All of them include the cerebellum and brainstem in the segmentation, and all but one are designed to work only with T1-weighted data:

- The widely used Brain Extraction Tool (BET) [1], which is part of the FSL package, utilizes a deformable model which evolves to fit the brain surface. The model is initialized as a spherical mesh around the center of gravity of the brain as if it was a balloon, and locally adaptive forces “inflate” it towards the brain boundary. BET is very fast and relatively insensitive to parameter settings. It provides good results considering its simplicity, but often

produces undesired blobs of false positives around the brainstem. This can be solved using a two-pass scheme: after running BET once, the preliminary mask is used to guide the registration of the brain to an atlas. The mask of the atlas is then used to guide a second pass of BET. The only disadvantage is that the registration makes the method much slower. BET is the only method that also works with T2-weighted MRI.

- Another popular method is Brain Surface Extraction (BSE) [2]. BSE relies on a series of processes to detect the brain: anisotropic diffusion filtering, edge detection, and a chain of morphological operations. BSE can provide highly specific whole-brain segmentations, but it usually requires fine parameter tuning to work on specific images. Another noteworthy feature of BSE is that, as opposed to most other methods, it preserves the spinal chord.
- 3dSkullStrip, part of the AFNI package [4], is a modified version of BET that also uses the spherical surface expansion paradigm. It includes modifications for avoiding the eyes and ventricles, reducing leakage into the skull and using also data outside the surface (and not only inside) to guide the evolution of the surface, among other adjustments.
- BridgeBurner [5], part of the application FireVoxel, first finds a small cubic region in the brain white matter, and then uses its mean intensity to compute a window that can be used to create a coarse segmentation of the brain, as in AFNI. The surface of the preliminary mask is combined with the output of an edge detector to create a boundary set. Then, layers are "peeled" with morphological operations that eventually "burn" all the bridges between brain and non-brain tissue. One of the disadvantages of this algorithm is that, if a single bridge survives the process, the output can include large chunks of non-brain tissue. Also, BridgeBurner is not a skull stripping algorithm but a brain tissue segmentation method, meaning that cerebrospinal fluid within the brain (including the ventricles) is often left out of the segmentation. However, the algorithm can be modified to produce an output similar to the other methods by morphologically closing the output and then filling the holes in the mask.
- GCUT [6] is a recently proposed method based on graph cuts [10]. First, it finds a threshold between the intensities of the gray matter and the cerebrospinal fluid and uses it to generate a preliminary binary mask which ideally includes the brain, the skull and some thin connections between them. Then, graph cuts can be used to find a connected submask that minimizes the ratio between the cost of its boundary (a data term) and its volume. This can be seen as a simple shape prior. This submask is post-processed to obtain the final segmentation. GCUT is usually quite accurate but sometimes makes large mistakes by following a wrong edge (e.g., leaving the whole cerebellum out or an eye in).
- A very popular public method is the hybrid approach from [11], available as part of the software package FreeSurfer [3]. This method is more robust than the methods discussed above. It combines a watershed algorithm, a deformable surface, and a probabilistic atlas. The watershed algorithm creates an initial estimate of the mask assuming connectivity of the white matter. Then, a smooth surface is allowed to evolve to refine the mask, using the statistical atlas to disregard unreasonable outputs. The latest version of FreeSurfer uses GCUT to refine the output. Since both are very specific methods, the intersection (AND) of the masks from the two algorithms eliminates many false positives without affecting the sensitivity, improving the segmentation [6].

There are also other noteworthy methods in the literature. *Lemieux et al.* use automated thresholding and morphological operations in [12]. *Hahn et al.* proposed an algorithm based on a watershed transform in [13]. Two cortex extraction algorithms that provide the whole brain segmentation as by-product are presented in [14] and [15]. Level sets are used to approach the problem in [16] and [17]. A histogram-based approach is described by *Shan et al.* in [18]. Unfortunately, these algorithms are not publicly available. We believe that the six aforementioned methods (BET, BSE, AFNI, BridgeBurner, FreeSurfer and GCUT) are a very representative set of skull-stripping systems that are commonly used in the neuroscience research community.

Meanwhile, hybrid approaches combining generative and discriminative models have been widely used in medical imaging. Discriminative models are easy to train and capture the local properties of the data effectively, but cannot model the global shape information easily. On the other hand, generative approaches model shape with high specificity, but cannot be easily adapted to capture the local texture. Due to their complementary nature, it is natural to combine both types of model into robust hybrid systems. For example, Gaussian shape models are used in conjunction with probabilistic boosting trees in [19] and with a *k* nearest neighbor classifier in [20] and [21].

In this study we present a new hybrid approach to skull stripping of T1-weighted brain MRI data and compare it with the six aforementioned methods. The proposed system, henceforth denoted as ROBEX, is designed to work out of the box with no parameter tuning. ROBEX is programmed to preserve the cerebellum and brainstem in order to enable comparison with the previously described methods. It is also designed to be robust against intensity variations and to work well across datasets. ROBEX is a hybrid approach that fits a generative model (a point distribution model, PDM [22]) to a target brain using a cost function provided by a discriminative model (a Random Forest [23]). To the best of our knowledge, ROBEX is the first skull stripping method that is based on a hybrid generative / discriminative model. In the context of skull stripping, "hybrid" usually refers to combining region-based (such as [12]) and boundary-based methods (such as BET), as FreeSurfer [11] does.

The generative model assumes that the brain surface is a triangular mesh in which the *(x,y,z)* coordinates of the landmarks follow a Gaussian distribution. The discriminative model attempts to extract the interface between the skull and the rest of the data (dura matter, cerebellum, eyes, etc.) by

assigning to each voxel the probability that it is on the brain surface. Modeling the boundary suits the generative model very well because the cost of a certain brain shape can be efficiently computed by multiplying the likelihood that each mesh point is on the interface, making the fitting process fast. We propose exploring the shape model with coordinate descent after rotating the model basis with the varimax criterion, which reduces the interference between the optimization of the different coordinates.

The proposed combination of models also has the advantage of easily accommodating an efficient method of refining the fit of the PDM, which cannot match the target shape exactly in general. Each point in the PDM is allowed to shift along the normal to the object surface to reduce the value of the cost function under the constraint that the shifts of neighboring points must be within a certain margin (i.e. the resulting surface must be smooth). Graph cuts can be used to find the optimal solution of this problem [24].

The rest of this paper is organized as follows. Section II describes the datasets used in this study. Section III describes the methods: how the generative and discriminative models were created and how they are fitted to a test scan to produce a brain segmentation. The experiments and results are described in section IV. Finally, section V includes the discussion.

II. Datasets

Four different datasets were used in this study: one exclusively for training and three for evaluation. The evaluation datasets are all publicly available.

The training dataset consists of 92 T1-weighted scans from healthy subjects acquired with an inversion recovery rapid gradient echo sequence on a Bruker 4T system. The size of the volumes is $256 \times 256 \times 256$ voxels, and the voxel size is $0.9375 \times 0.9 \times 0.9375$ mm. Manual delineations of the brain by an expert physiologist are available for all of them. The first volume of the dataset was chosen as "reference volume". All the other 91 scans in the dataset were then registered to the reference. The software package Elastix [25] was used to optimize an affine transform using a mutual information metric. Because the scans are resampled to the resolution of the reference volume after registration, the reference was first downsampled to $1.5 \times 1.5 \times 1.5$ mm resolution to lighten the computational load of the algorithms. Rather than using an arbitrary scan as the reference, it is also possible to learn an unbiased mean volume from all the training images [26]. However, we found through pilot experiments that our algorithm was not sensitive to the choice of the reference.

The first test dataset is the Internet Brain Segmentation Repository (IBSR). It consists of 20 T1-weighted scans from healthy subjects (age 29.0$\pm$4.8 years) acquired at the Center for Morphometric Analysis at Massachusetts General Hospital, as well as their corresponding annotations. This dataset is available for download at http://www.cma.mgh.harvard.edu/ibsr/. The scans were acquired with a 3D spoiled gradient echo sequence on two different scanners. Ten scans on four males and six females were performed on a 1.5 Tesla Siemens Magnetom MR System with a FLASH pulse sequence and the following parameters: TR/TE = 40/8 ms, flip angle 50 degrees, slice thickness 3.1 mm, in-plane resolution 1×1 mm. Ten scans on six males and four females were performed on a 1.5 Tesla General Electric Signa MR System with a 3D-CAPRY pulse sequence and the following parameters: TR/TE = 50/9 ms, flip angle 50 degrees, slice thickness 3.0mm, in-plane resolution 1×1 mm. The brain was manually delineated by trained investigators in all the scans. Some of the scans have severe striation artifacts which, next to the large slice thickness, makes this dataset challenging to segment.

The second test dataset is the LPBA40 dataset [8], which can be downloaded from http://sve.loni.ucla.edu/. It consists of 40 T1-weighted scans (20 males, 20 females, age 29.20$\pm$6.30 years) and their corresponding annotations. The scans were acquired with a 3D spoiled gradient echo sequence on a GE 1.5T system. The acquisition parameters were: TR: 10.0ms - 12.5ms; TE range 4.22ms - 4.5 ms; flip angle 20 degrees. Coronal slices were acquired 1.5mm apart with in-plane resolution of 0.86 mm (38 subjects) or 0.78 mm (2 subjects).

The third test dataset consists of the first two discs (77 T1-weighted scans) of the cross-sectional MRI dataset of the OASIS project: http://www.oasis-brains.org/. The population consists of 55 females and 22 males, age 51.64 $\pm$24.67 years. Twenty subjects were evaluated as "demented and probable Alzheimer's disease". The scans were acquired on a 1.5T Siemens scanner with a MP-RAGE sequence, TR/TE/TI/TD=9.7ms/4.0ms/20ms/200ms, flip angle 10 degrees. Sagittal slices were acquired 1.5mm apart with in-plane resolution of 1 mm. The brain masks for this set were not manually delineated; instead, the brain was segmented with an in-house 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 dataset is very valuable because it includes scans from a very diverse population with a very wide age range as well as diseased brains.

III. Methods

The proposed segmentation method combines a discriminative model and a generative model to obtain the brain mask. The discriminative model (Section III-B) is a Random Forest classifier. The generative model (Section III-B) is a Gaussian distribution over a set of landmarks that defines the brain surface though a triangular mesh. Given a new volume, the segmentation is found as the instance of the generative model that maximizes the likelihood of the surface according to the discriminative model. The proposed optimization algorithm (described in Section III-C) consists of two steps: 1. optimizing the generative model with coordinate descent (Section III-C1); and 2. refining the output from the previous step using graph cuts (Section III-C2).

A. Discriminative model

1) Random Forests: The discriminative model in this study is a voxel-based Random Forest [23] classifier, which has been proven successful in a variety of domains and compares

favorably with other state-of-the-art algorithms [27]. Random Forests have only recently been adopted in medical imaging segmentation [28], [29]. Therefore, we provide a short description of how they work for the sake of completeness. A Random Forest is an ensemble of decision trees. Each tree is trained with a different subset of the training volumes ("bagging"), which improves the generalization ability of the classifier [30]. Voxels are pushed down each tree from the root by performing a simple binary test at each internal node until a leaf has been reached. The tests consist of comparing a certain feature with a threshold.

Training a forest implies finding the set of tests that best separate the data into the different classes. At each internal node, the feature space is searched for a test that maximizes the reduction of class impurity, typically measured with the class entropy or, as in this study, with the Gini index $G = 1 - \sum_i f_i^2$ (where $\{f_i\}$ are the fractions of the different classes). Rather than inspecting the full space of features at each node, a random subset of them is probed, and the best one selected. Even if this makes the individual trees weaker, it decreases the correlation between their outputs, increasing the performance of the forest as a whole. Each training voxel is sent to the corresponding child depending on the result of the test, and the process is recursively repeated until the number of samples in a node falls below a threshold, until a predefined maximum tree depth is reached or until all the samples belong to the same class. In that case, the node becomes a leaf, and the most frequent class of the training data at the node is stored for testing. Because the minimum number of samples can be reached at any depth, the tree is in general not perfectly balanced i.e. some leaves can be at deeper levels than others.

In testing, a previously unseen voxel is pushed down the different trees by running the tests corresponding to the nodes it travels along. When a leaf node is reached, the tree casts a vote corresponding to the class assigned to the node in the training stage. The final decision for a voxel is obtained by selecting the most voted class. Moreover, the probability that a voxel belongs to a class can be estimated as the number of votes for that class over the total number of trees.

2) MRI signal standardization: In this study, we train a Random Forest classifier to discriminate brain boundary voxels from any other voxels using features based on Gaussian derivatives and gradients. These features rely heavily on pixel intensities and therefore an intensity standardization process is required as a preprocessing step. The MRI signal levels are standardized as follows. First, the scan is fed to an implementation [31] of the N3 bias field correction algorithm [32]. The method works much better when a brain mask is available, so we provide the algorithm with an eroded version of the mask of the reference volume (see Figure 2). Despite being a very rough approximation of the real mask, the erosion guarantees that the mask is highly specific. Even if the most outer part of the brain is left out of the mask, the extrapolation of the estimated correction field in that region is usually fair because the field is assumed to vary very slowly in space. The correction provided visually pleasing results for all the scans in the four datasets, even for the data acquired at 4T (bias field correction is known to sometimes falter at higher field strengths).

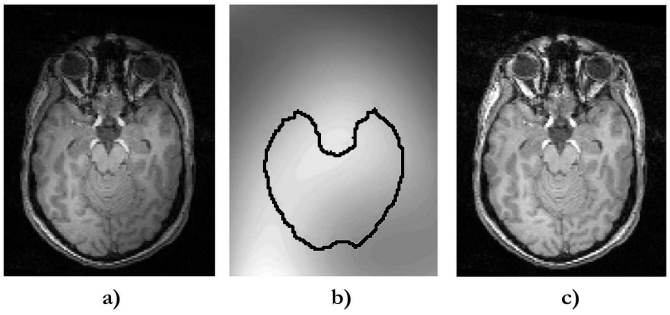

Fig. 2. Bias field correction. a) Saggital slice of a T1-weighted MRI from the training dataset (acquired at 4T). b) Corresponding slice of the estimated multiplicative bias field, with the registered, heavily eroded mask from the reference volume superimposed. c) Corresponding slice of the corrected volume.

The next step is to normalize voxel intensities. First, the robust minimum and maximum voxels intensities are estimated as the first and 99^{th} percentiles ($pc_{.01}, pc_{.99}$) of the histogram of the voxels the are located inside the eroded mask. Then, a linear grayscale transform that maps $pc_{.01}$ to 200 and $pc_{.99}$ to 800 is computed. The transform is then applied to the volumes, and the histogram cropped at 0 and 1000 i.e. all values below 0 and above 1000 are set to 0 and 1000 respectively. Finally, the contrast of the volume is enhanced with standard histogram equalization. Again, the histogram is computed using only the voxels inside the eroded mask. There are more complex intensity normalization techniques based on detecting the typical intensities of gray matter, white matter and cerebrospinal fluid [33], but they require skull-stripped data.

3) Feature pool: Once the intensities are standardized, features can be extracted for each voxel in the volume. A pool of 36 features is considered in this study. The (x, y, z) coordinates of each voxel in the space of the reference volume were used as features in order to capture the context. Since the goal is to detect boundaries, the gradient magnitudes at three different scales ($\sigma = 0.5, 2.0, 8.0$, in mm) were also used. Finally, Gaussian derivatives at the same scales complete the feature set. The Gaussian derivatives correspond to the truncated Taylor's expansion of the data around each point and therefore capture the local appearance of the volume at different scales.

4) Data sampling: Because the voxels in a scan are highly correlated, it is not necessary to use all the training data to build a strong classifier. Using fewer voxels lightens the computational load of the training stage. Preliminary experiments showed that the performance does not really improve much after 50,000 training voxels. In this study, 10,000 voxels were extracted randomly from each of the 92 training scans under the following two constraints: 1) to compensate for the extremely uneven prior probabilities of the two classes, half of the voxels are constrained to be positive examples i.e. they have to lie on the brain boundary, given by a mask that is calculated as the gold standard mask AND the negated of a minimally eroded version of itself; 2) 50% of

the negative examples are constrained to lie within 3 mm of the brain boundary, defined as above. This makes the classifier focus on the harder examples, decreasing the amount of required training voxels. The width of the band represents a compromise between how correlated and informative (i.e. near the boundary) the voxels are. Highly correlated voxels are not very useful for the classifier, but neither are voxels very far away from the boundary. Pilot experiments showed that 3 mm offered a good compromise. The described sampling scheme led to selecting ≈0.4% of the total number of voxels, ≈20% of the positive voxels, ≈2.5% of the voxels in the 3 mm band and ≈0.1% of the rest of negative voxels.

5) Classifier training and feature selection: The Random Forest classifier was trained using the following parameters. The number of trees was set to 200, a fairly large value that provides a good granularity for the class probabilities in testing ($1/200$). The number of features probed at each node in training was set to five. This is a relatively low value, which is justified by the large number of trees: it is not a problem that the trees are weaker if there are plenty of them. The minimum number of training voxels in a node was set to 20. The tree depth was not limited, but depths greater than 17 were never reached in training.

Random Forests do not require feature selection to be robust due to their intrinsic ability to disregard unimportant features. However, it is still useful to reduce the dimensionality of the data to lighten the computational load of computing features. In this study, we used backward feature elimination based on permutation importance. First, the classifier is trained with all the features. Then, the feature with the lowest permutation importance is dropped and the classifier retrained. The permutation importance of a feature is defined as the drop in accuracy in the out-of-bag (i.e., non-training) data of each tree caused by randomly permuting the values of that feature in the input data. The process is repeated as long as the accuracy of the classifier does not decrease noticeably. Even though the permutation importance is known to be biased [34] (more important features can eventually lead to lower accuracy), this method is considerably faster than combinatorially expensive approaches such as [35] and often provides comparable results.

During the feature selection, the accuracy of the classifier is evaluated cross validation. This is accomplished by randomly dividing the training data into two subsets at each elimination step: one for training (60 scans) and one for evaluation (32 scans). The classifier is trained using 100,000 randomly sampled voxels from the training subset, making sure each tree only uses voxels from 40 scans (with bagging purposes). The accuracy is computed upon all the the voxels in the testing subset. The evolution of the accuracy and the permutation importance of the least important feature are displayed in Figure 3.

Based on visual inspection of the curves, the final number of selected features was 10 (listed in Table I). The list of features reveals some interesting aspects of the problem. First, context features are very important: all three are selected, and the z coordinate has the largest permutation importance. Another observation is that only one gradient feature made it to the set. The scale of this gradient feature ($\sigma = 2.0$) must then be the most appropriate for our edge detection task. Finally, derivative features were only selected at the coarsest scale, which is reasonable given how noisy these features are at finer scales.

The final classifier was trained on 200,000 voxels which were selected randomly across the whole dataset. Figure 4 shows the probability volume for the first scan of each dataset. The probability volume is just the number of trees that have voted positive for each voxel. Upon division by the number of trees in the Random Forest, this volume can be interpreted as a "real" probability defined between zero and one. The maps display regions of false positives around the ethmoid sinuses, optic chiasm, and large portions of the scalp, but these will be easily discarded by the generative model described below.

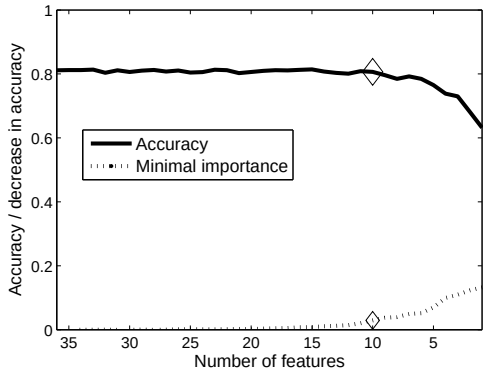

Fig. 3. Feature selection: accuracy in out-of-bag data (dashed) and minimal predicted loss accuracy for each drop. The diamond marks the operating point of the final classifier. Please note that the accuracy of the classifier is not very high in absolute terms because it is trained to solve the difficult problem of detecting voxels that are exactly located on the brain boundary.

TABLE I
LIST OF SELECTED FEATURES AND THEIR PERMUTATION IMPORTANCE IN THE LAST STEP OF FEATURE SELECTION. COORDINATE x CORRESPONDS TO LEFT/RIGHT, y TO ANTERIOR/POSTERIOR, AND z TO INFERIOR/SUPERIOR. THE NOTATION $abc(d)$ REPRESENTS THE ORDERS OF THE DERIVATIVES a, b AND c (CORRESPONDING TO THE x, y AND z AXES) AT SCALE $\sigma = d$ (IN MM). THE MAGNITUDE OF THE GRADIENT AT SCALE d IS REPRESENTED BY $|\nabla|(d)$. THE IMPORTANCE IS THE PREDICTED LOSS OF ACCURACY IN OUT-OF-BAG DATA WHEN THE VALUES OF A FEATURE IN THE TRAINING DATA ARE RANDOMLY PERMUTED.

Feature	z coord.	000(2.0)	001(8.0)	000(8.0)	$\|\nabla\|(2.0)$
Importance	0.1232	0.0881	0.0815	0.0844	0.0558
Feature	x coord.	010(8.0)	000(0.5)	200(8.0)	y coord.
Importance	0.0539	0.0301	0.0403	0.0375	0.0272

B. Generative model

The generative model ensures that the result of the segmentation is a plausible shape. In this study, a PDM is used to represent the set of possible brain shapes. PDMs are constructed from a set of training shapes (in 2D or 3D) which are represented by a set of corresponding landmarks. The landmarks can be manually placed or automatically determined from a continuous shape e.g. a parametrized curve or a polygonal line in 2D, or a parametrized surface or a binary mask in 3D [36]–[39]. Once the landmark representation of

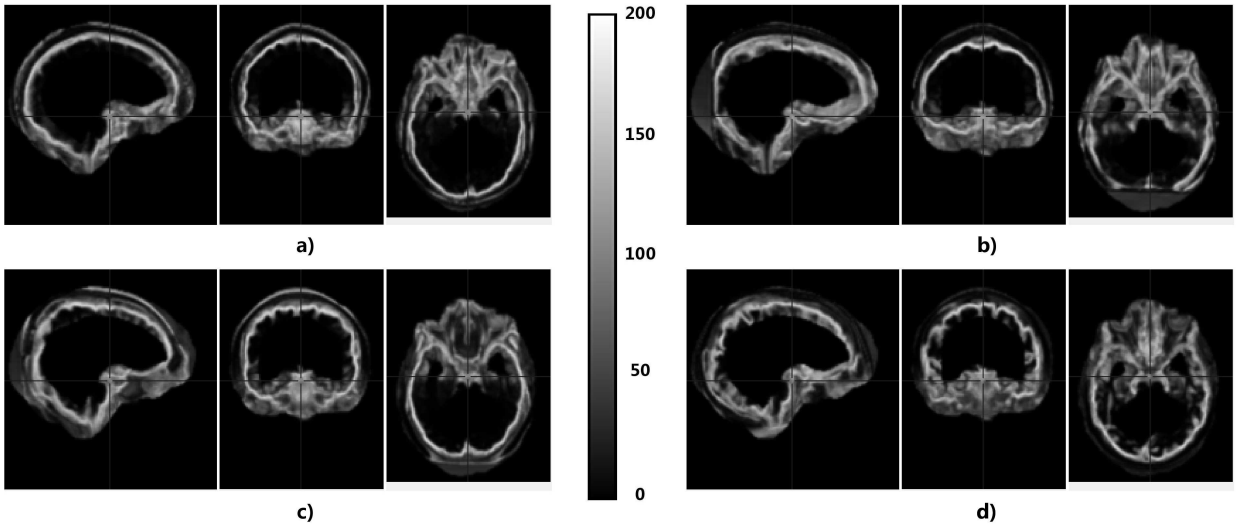

Fig. 4. Orthogonal slices of the probability volumes (number of trees from the 200 that have voted positive) of the first scan of each dataset: a) training dataset, b) IBSR, c) LPBA40, d) OASIS. The probability volumes have been smoothed with a Gaussian kernel of width σ=1 mm. As expected, the probability map looks cleaner for the training dataset. However, the image standardization does a good job and the map looks acceptable for the other datasets, especially OASIS. Please note that part of the head is out of the field of view in b) and c), therefore the smooth patches.

the training shapes is ready, all the shapes can be jointly aligned using Procrustes analysis [40] and the distribution of the aligned landmark coordinates can be fed to a principal component analysis (PCA) [41]. The PDM can be iteratively deformed, rotated and scaled to detect an instance of the shape in an image in a technique known as active shape models [22].

1) Point distribution model: Mathematically, if the Cartesian coordinates of the L (aligned) landmarks of training instance i are stacked into a $3L$-dimensional vector $\boldsymbol{s}_i = [x_{i,1}, y_{i,1}, z_{i,1}, ..., x_{i,L}, y_{i,L}, z_{i,L}]^t$, any shape in the aligned space can be approximated as $\boldsymbol{s} \approx \boldsymbol{\mu} + P\boldsymbol{b}$, where $\boldsymbol{\mu}$ is the $3L$-dimensional mean shape, $\boldsymbol{b} = P^t(\boldsymbol{s} - \boldsymbol{\mu})$ is the vector of p shape coefficients and P is a matrix whose columns are the normalized eigenvectors $\boldsymbol{e}_j$ of the empirical covariance matrix Cov corresponding to the p largest eigenvalues $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_p]^t$ (where it holds that $\lambda_j \geq \lambda_{j+1}$):

$$Cov = \frac{\sum_{i=1}^{N_{samples}} (\boldsymbol{s}_i - \boldsymbol{\mu})(\boldsymbol{s}_i - \boldsymbol{\mu})^t}{N_{samples} - 1} = \sum_{j=1}^{3L} \lambda_j \boldsymbol{e}_j \boldsymbol{e}_j^t$$

$$P = [\boldsymbol{e}_1 | ... | \boldsymbol{e}_p]$$

If all the eigenvectors are preserved, $\boldsymbol{b}$ follows a multivariate Gaussian distribution with zero mean and covariance matrix $\Sigma = diag(\lambda_1, \lambda_2, ..., \lambda_{3L})$ in which the shape coefficients are independent. The total variance of the model can be computed as $\sigma_{tot}^2 = \sum_{j=1}^{3L} \lambda_j$. Therefore, the number of components to keep p can be determined from η, the proportion of total variance to be preserved in the model ($\eta = 0.90$ in this study):

$$p = min\ p', \quad \text{subject to}: \ \frac{\sum_{j=1}^{p'} \lambda_j}{\sigma_{tot}^2} \geq \eta$$

2) Landmark extraction: In our case, the landmarks have to be extracted from a set of masks (the training dataset). This is accomplished by a method very similar to [37]. First, the landmarks for the reference volume are computed using a surface meshing algorithm [42]. Sometimes this type of landmarks is called pseudo landmark in the literature because they do not necessarily correspond to salient points. In this study, the maximal radius of the Delaunay sphere was set to 3mm, leading to $L = 3237$ landmarks. Then, the masks of the training scans are registered to the mask of the reference volume in order to obtain a transform that can be used to propagate the landmarks. Elastix was first used to optimize a translation transform that was subsequently refined by an affine transform and then a non-linear transform based on a deformable grid of points and B-spline interpolation (grid spacing: 15 mm). The κ agreement was used as registration metric.

Because the registration is not perfect, the propagated landmarks do not lie in general on the surface of the training masks. This inaccuracy was corrected by projecting them onto the surfaces, which were first meshed with high resolution (maximum radius: 1 mm) to increase the precision of the projection. The mesh of the reference scan and the maximum registration error for each landmark over the 92 training volumes are displayed in Figure 5. Most of the maximum errors are lower than 4mm.

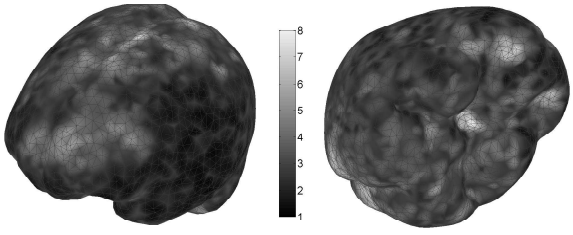

Fig. 5. Mesh of the mask corresponding to the reference brain. The maximum registration error for each landmark has been superimposed. The scale is in mm.

3) PCA and basis rotation: Once the landmarks have been extracted, the PDM can be built. In our case, since most

of the differences in pose are already filtered out by the affine registration, we bypass the Procrustes alignment. In the PCA, preserving 90% of the total variance led to $p = 19$ components. The first three modes of variation ($\boldsymbol{\mu} + k\sqrt{\lambda_j}\boldsymbol{e}_j$ for different values of k) are shown in Figure 6.

The optimization procedure in section III-C below explores the space of shape coefficients to find the model instance that best fits the output from the classifier. This search can be made more efficient by finding a rotation of the PCA basis P that minimizes the spatial overlap of its vectors; because rotations do not modify the spanned space, the set of shapes that can be represented by the model does not change. In the rotated basis, the shape parameters can be fitted independently with almost no interference, making coordinate descent (i.e. updating one shape coefficient at the time) very efficient.

We used the varimax criterion [43] to calculate a rotation matrix R such that the sparsity of rotated eigenvector matrix $Q = PR$ is maximized. The rotated shape coefficients are given by $\boldsymbol{b}_R = R^{-1}\boldsymbol{b}$, and they are not independent anymore. They still follow a multivariate Gaussian distribution, and the diagonal of the covariance matrix is $\boldsymbol{\lambda_R} = [R^{-1}]^2\boldsymbol{\lambda}$, where $[\cdot]^2$ denotes the element-wise squared matrix. Figure 7 shows the first three modes of variation for the rotated eigenvectors, which are (especially the first and third) highly localized compared with the original ones in Figure 6.

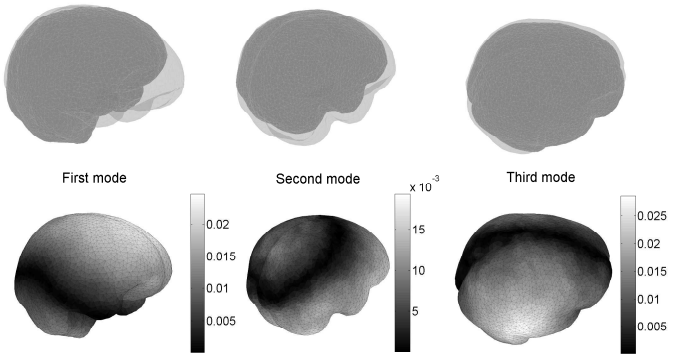

Fig. 6. First three modes of variation for the original PDM. Upper row: shapes corresponding to $\boldsymbol{\mu} \pm 3\sqrt{\lambda_j}\boldsymbol{e}_j$. Lower row: magnitude of the eigenvectors.

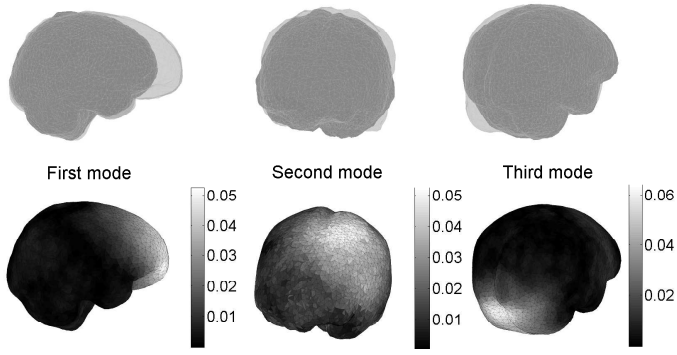

Fig. 7. First three modes of variation after rotating the eigenvectors. Upper row: shapes corresponding to $\boldsymbol{\mu} \pm 3\sqrt{\lambda_j}\boldsymbol{e}_j$. Lower row: magnitude of the eigenvectors.

C. Computing the mask for a test scan

The elements of the skull stripping pipeline are shown in Figure 8. The first steps include the same preprocessing that the training data went through: registration to template (as in section II), bias field correction, intensity normalization, feature calculation and voxel classification (as in section III-A). The following step is to fit the shape model to the probability volume. Because the shape model cannot exactly represent all plausible shapes, a small free deformation of the mesh is allowed to refine the output. Then, a brain mask is generated from the mesh by: 1. creating an empty volume; 2. setting to one all the voxels intersected by the triangles in the mesh; and 3. filling the hole in the mask. The resulting binary volume is warped back to the space of the original scan using the inverse of the affine transform from the registration, which is analytically invertible. The fitting of the shape model and the free deformation step are further discussed next.

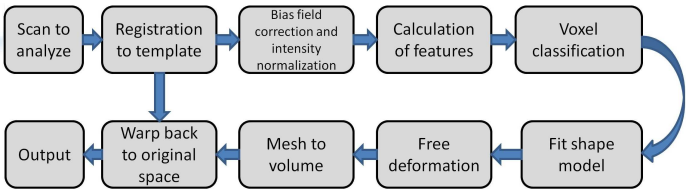

Fig. 8. Steps of the method to segment the whole brain from a scan.

1) Fitting the shape model: In active shape models [22], the pose (translation, rotation and scaling) and shape parameters of the model are iteratively updated to find an instance of a shape in an image. Given an initialization, the normal to the curve (2D) or surface (3D) at each landmark is first calculated. Then, a fitness function is evaluated along each normal. The optimal shift for each landmark is then added to the coordinates or the current shape, creating a "proposed" shape. Finally, the proposed shape is projected back onto the model and a new iteration begins.

This method is prone to getting stuck in local optima. Instead, we propose exploring the space of shape coefficients directly. Since the pose is not to be optimized (it is already taken care of by the affine registration) and the shape coefficients represent localized, minimally overlapping variations thanks to the varimax rotation of eigenvectors, the optimal shape can be efficiently computed using coordinate descent. If $I_p(x, y, z)$ is the normalized probability volume (i.e. the number of trees which have been voted positive over the total amount of trees), the cost C to minimize is:

$$C = \log \prod_{l=1}^{L} [\epsilon + (1-\epsilon)(1 - I_p(\boldsymbol{r}_l))] =$$

$$\sum_{l=1}^{L} \log[\epsilon + (1-2\epsilon)(1 - I_p(\boldsymbol{r}_l))] = \sum_{l=1}^{L} I_{q-log}(\boldsymbol{r}_l) \qquad (1)$$

where $\boldsymbol{r}_l$ is the (x, y, z) position vector of landmark l, which can be extracted from the current shape $\boldsymbol{s}(\boldsymbol{b}_R) = \mu + Q\boldsymbol{b}_R$. The constant ϵ avoids taking the logarithm of zero. If the probability volume is seen as set of independent Bernoulli

variables with parameters $I_p(x,y,z)$, then ϵ can be interpreted as a conjugate Beta prior on these distributions.

The cost function in Equation 1 can be evaluated very rapidly because it only requires interpolating $L = 3237$ points in the log-probability volume $I_{q-log}(\boldsymbol{r})$. Before computing this volume, it is convenient to smooth $I_p(x,y,z)$ with a Gaussian kernel (we used $\sigma = 1mm$) in order to increase the capture range of cost function in the optimization method.

The optimization algorithm starts from the mean shape, and iteratively loops along the $p = 19$ shape coefficients, optimizing one at the time. The first passes (10 in our implementation) do exhaustive search in the interval $\boldsymbol{b}_{R,j} \in [-3\sqrt{\boldsymbol{\lambda}_{R,j}}, 3\sqrt{\boldsymbol{\lambda}_{R,j}}]$ i.e three standard deviations. Later iterations use Newton's method to refine the output. At the end of each iteration, the algorithm verifies that the shape lies within a zero-mean ellipsoid that covers 99% of the probability mass of the distribution. If it does not, the shape is projected onto the surface of the ellipsoid. This ensures that the output of the optimization is a plausible shape. The steps of the algorithm are summarized in Table II.

TABLE II
STEPS OF THE ALGORITHM TO FIT THE SHAPE MODEL.

$\boldsymbol{b}_R \leftarrow 0$, $it = 0$
REPEAT
  LOOP along $j \in \{1,..,p\}$ in random order
    IF $it < 10$
      Full search in $\boldsymbol{b}_{R,j} \in [-3\sqrt{\boldsymbol{\lambda}_{R,j}}, 3\sqrt{\boldsymbol{\lambda}_{R,j}}]$ to minimize C.
    ELSE
      Use Newton's method to refine $\boldsymbol{b}_{R,j}$
    END
    Compute $\boldsymbol{b} = R\boldsymbol{b}_R$ and $D = \sqrt{\sum_{j=1}^{p} \frac{b_j^2}{\lambda_j}}$
    IF $D > D_{max}$ (calculated with the χ^2 distribution)
      $\boldsymbol{b} \leftarrow \frac{D_{max}}{D}\boldsymbol{b}$
      $\boldsymbol{b}_R \leftarrow R^{-1}\boldsymbol{b}$
    END
  END
  $it \leftarrow it + 1$
UNTIL $it > it_{max}$ OR convergence

2) Free deformation: Using the shape model has the advantage that the result is highly specific. However, the model cannot capture the full range of variations in brain shapes for two reasons. First, the set of training scans, though fairly large, does not represent the whole population. The second reason is that 10% of the captured variance was explicitly disregarded when building the model. Therefore, the method can often benefit from a smooth extra deformation outside the constraints of the shape model to improve the cost of the fit.

To solve this problem, we extend *Li et al.*'s method [24] to segment surfaces. The two differences with respect to their method are: 1. defining the graph for a triangular mesh, rather than tubular or terrain-like surfaces i.e. $(x,y) \rightarrow z(x,y)$; and 2. minimizing a explicitly learned cost function rather than a voxel intensity based criterion. The key of the approach is to allow the landmarks to shift along their corresponding normals to the surface in discrete steps. The smoothness of the deformation can be enforced by constraining the shifts of neighboring landmarks to be similar. *Li et al.* show in their study that the set of shifts that provides the global minimum of the metric can be found in polynomial time using graph cuts.

If the normals to the surface are calculated at each landmark location, and the cost function is sampled along the normals, a cost profile is defined along each normal at locations $\boldsymbol{r}_l + t_l\hat{n}_l$, where t_l is the (continuous) signed shift for landmark $l \in \{1, L\}$ and $\hat{n}_l$ is the normal vector at landmark location $\boldsymbol{r}_l$. Let us assume that that the shifts have to be bounded ($|t_l| < t_{max}$) and that neighboring shifts have to be similar ($|t_l - t_{\aleph(l)}| \leq \Delta$, where $\aleph(l)$ represents the neighbors of landmark l) and Δ is the bound. The problem is then finding the set of shifts $\{t_l\}$ under these constraints that minimizes the cost in equation 1:

$$C = \sum_{l=1}^{L} I_{q-log}(\boldsymbol{r}_l + t_l\hat{n}_l)$$

If $\boldsymbol{r}_l$ is outside the image field of view, a triangular profile is assumed for $I_p(\boldsymbol{r}_l + t_l\hat{n}_l) = 1 - |t_l/t_{max}|$ in order to: 1. encourage the landmarks to stay at the locations predicted by the shape model; and 2. ensure a smooth transition of the shifts from the landmarks which are inside the field of view to those which are not.

The problem of minimizing C can be discretized by assuming that the shifts must be multiples of a given step δ: $t_l = s_l\delta$, where s_l is an integer. Assuming that $\frac{t_{max}}{\delta}$ and $\frac{\Delta}{\delta}$ are integers (in our implementation $t_{max} = 19.5mm$, $\Delta = 1.5mm$ and $\delta = 0.75mm$), we must solve:

$$\underset{\{s_i\}}{\operatorname{argmin}} \sum_{l=1}^{L} I_{q-log}(\boldsymbol{r}_l + s_l\delta\hat{n}_l) \tag{2}$$

$$\begin{aligned} \text{subject to}: \quad & s_l \in \mathbb{Z} \\ & s_l \in \{-\frac{t_{max}}{\delta}, \frac{t_{max}}{\delta}\} \\ & |s_l - s_{\aleph(l)}| \leq \frac{\Delta}{\delta} \end{aligned}$$

Li et al. show in their study that solving equation 2 can be simplified to a problem of computing the minimum s-t cut in a related directed graph $G = (V, E)$. For the sake of completeness, we summarize their method here. First, the graph G must be built as follows. Each shift for each landmark represents a vertex. Each vertex is connected to the shift right below except for the vertices on the zero plane $s_l = -\frac{t_{max}}{\delta}$, i.e. $\{s_l = i\} \rightarrow \{s_l = i-1\}$, $s_l > -\frac{t_{max}}{\delta}$. Each vertex is also connected to the shifts exactly $\frac{\Delta}{\delta}$ levels below corresponding to neighboring landmarks, except for the vertices below $s_l <= -\frac{t_{max}}{\delta} + \frac{\Delta}{\delta}$, which are connected to the zero plane of the neighboring landmarks. Moreover, each vertex is given a weight which is equal to the difference between the cost of its shift and the cost of the shift right below, except for the zero plane, which is just assigned the cost of its corresponding shift. Figure 9 displays a typical "column" in the graph i.e. set of shifts for a landmark.

The problem of finding the optimal set of shifts is equivalent to finding a non-empty minimum closed set in G. This is a well-known problem in graph theory, and it can be solved by computing the minimum s-t cut in a related directed graph $G_{st} = (V_{st}, E_{st})$. V_{st} includes all the vertices in V plus a

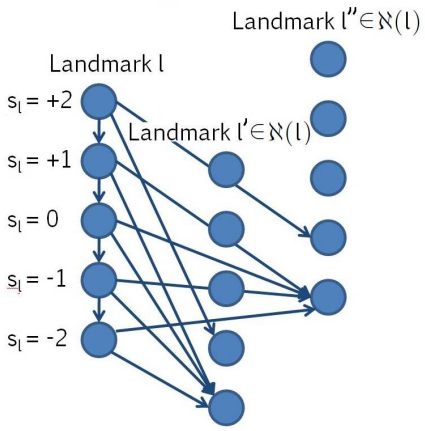

Fig. 9. Connections in the graph $G = (E, V)$ from the vertices corresponding to the shifts of a landmark, assuming $\frac{\Delta}{\delta} = 3$.

source and a sink. E_{st} includes all the edges in E, with weight infinity, plus a new set of edges: the source is connected to all the vertices that had negative weights in G, whereas every vertex that had positive weights in G is connected towards the sink. The weights of these new edges are equal to the absolute value of the weights of their corresponding vertices in G.

Once the graph is ready, the problem is to find the minimum cost cut that disconnects the source from the sink. This cut can be found by solving its dual problem (the maximum flow problem) using any of the multiple algorithms proposed in the literature. Here we used the Boykov-Kolmogorov algorithm [44], which is publicly available at Dr. Boykov's website. The vertices along the cut can be shown to correspond to the shifts that minimize equation 2. The reader is referred to the original papers for a more detailed explanation. Figure 10 shows the typical effect of this processing step on the detected surface.

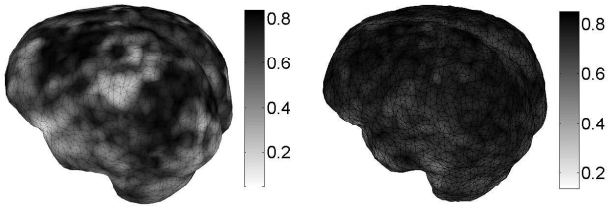

Fig. 10. Brain surface with the probability that each landmark is on the brain surface overlaid. a) Output from shape model. b) Refined with graph cuts.

IV. Experiments and Results

A. Setup

In order to compare the methods, the scans from the different databases were stripped using BET, BSE, FreeSurfer, AFNI, BridgeBurner, GCUT (with and without FreeSurfer) and the method proposed in this study. No preprocessing was carried out before feeding the scans to the methods. Our method requires no parameter adjustment. For the other methods, the corresponding authors were contacted and invited to provide parameter values for each of the three datasets:

- BET (version 2.1, in FSL 4.1.5): the authors suggested using the same sequence of commands for all three datasets. First, "bet -R" for a first pass with robust brain center estimation. Then, the preliminary stripped brain is fed to
  `standard_space_roi -roiNONE ...`
  `... -ssref MNI152_T1_1mm_brain.nii.gz`
  to align it to the MNI152 atlas from FSL. Finally, a second pass of BET (again, with the switch -R) yields the output. Both passes use the default value of the fractional intensity threshold: 0.5.
- BSE (2009 version): for IBSR, the author suggested using the default parameters: diffusion iterations = 3, diffusion constant = 25, edge constant = 0.64, erosion kernel width = 1, trim brainstem = true, remove neck = false. For LPBA40, he suggested using the following parameters: diffusion iterations = 5, diffusion constant = 15, edge constant = 0.65, erosion kernel width = 1, trim brainstem = true, remove neck = true. The author did not provide parameter settings for OASIS, so the default values were used.
- FreeSurfer (version 5.0.0): the authors answered that their software is robust and requires no tweaking. Therefore, default values were used: pre-weight the input image using atlas information = 0.82, use the basins merging atlas information = 0.32, presize the pre-flooding height = 10%, use the pre-weighting for the template deformation = true, use template deformation using atlas information = true, use seed points using atlas information = true.
- AFNI (version 2010-10-19-1028): the authors suggested that we used the following options for IBSR and LPBA40:
  `-shrink_fac_bot_lim .65 -shrink_fac .72`
  And for OASIS:
  `-shrink_fac_bot_lim .65 -shrink_fac .7`
- BridgeBurner (FireVoxel version 81B): the authors suggested using the default parameters: plane for seed search = axial, SI low = 0.528 relative to seed average, SI high = 1.35 relative to seed average, peel distance = 2.9 mm, grow distance = 6.4 mm, strict CoreSet surface = true, use edges = off. Subvoxel level = 1.
- GCUT (the only available version so far): the authors encouraged us to use the default values across the three datasets: threshold = 36, importance of intensity = 2.3. They also suggested that the increased threshold 40 should also be tested, since this is the value that will be used in the new version of FreeSurfer (5.1.0).
- FreeSurfer-GCUT: the intersection (AND) of the outputs from the two methods was taken. Default parameters were used in FreeSurfer, whereas the GCUT threshold was set to 40 (as in the new version of FreeSurfer).

The authors of BET also had a particular petition. Since their method performs better when the neck is not visible in the scan, they requested that the neck was removed from the IBSR volumes before running their software. Their petition is based on their claim that the neck could be easily removed from the scans automatically anyway. While this assumption

might be debatable, we also acknowledge that a skull stripping algorithm expects a volume centered on the brain, as opposed to a scan with a larger field of view. Therefore, BET was tested on the original IBSR dataset but also a version in which all voxels more than 30mm below the most inferior voxel in the ground truth brain mask were deleted. The authors of the other methods were given the chance of using these trimmed masks, but they all declined. The trimming process is illustrated in Figure 15.

The automatically segmented brains were compared with the gold standard using a number of metrics. The choice of metrics is motivated by two different reasons. First, the metrics must provide different perspectives of the results (e.g., precision vs. robustness). Second, the metrics must be similar to those used in other studies, for the sake of easy comparison. In this study, we used:

- The voxel-based Dice similarity coefficient $S(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|}$. This is arguably the most spread performance metric in the segmentation literature. It is related to the (also widely used) Jaccard index $J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$ by $S = \frac{2}{1+J^{-1}}$.
- The maximal surface-to-surface distance (Haussdorff distance): measures the robustness of the algorithm. As opposed to the Dice overlap, it penalizes cases in which two greatly overlapping objects have very different boundaries, e.g. 
- The mean symmetric surface-to-surface distance: measured from each voxel in the boundary of the ground truth to its nearest boundary voxel in the automated segmentation and the other way around (from each boundary voxel of the segmentation to its nearest boundary voxel in the ground truth). This metric is easier to interpret than the Dice coefficient.
- The 95% percentile of the surface-to-surface distance: this a more robust way of measuring robustness, since the Haussdorff distance is too sensitive to noise and outliers in the annotations and segmentations.

It is also interesting to study the distribution of the segmentation error around the different regions of the brain surface, i.e., in which brain regions is a given method more accurate? This can be done by computing the absolute difference volume between the automatic segmentation and the ground truth for each scan and method, warping it to the space of the reference volume and taking the average across each dataset / method pair. Henceforth, we call these "error volumes." In order to warp the difference volumes, Elastix was used to nonlinearly register each ground truth mask to the mask of the reference scan (using the same B-spline registration as in section III-B), and the resulting transform was then used in the warping.

Given that the age of each subject is given for IBSR and OASIS, it is also possible to assess the impact of age on the performance of our system. One would expect elder subjects from OASIS, whose brains are often atrophied, to be harder to segment for our method, since it is based on a shape model of healthy brain and can have trouble following deep sulci. Here, we use linear regression to relate age and performance using the slope of the fit and the correlation coefficient.

Another interesting experiment is to remove the free deformation step from our method (i.e., use the shape model directly) and study the impact on the segmentation. This allows us to quantify the importance of the graph-cut refinement.

Finally, we also studied the performance of the system when a different dataset is used to train the system. The aim of this experiment is to test whether the high performance of our system is due solely to the high quality of the data is is trained upon, and not due to the algorithm itself. The system was retrained using an identical configuration with the only difference being that our training dataset was replaced by OASIS. We used OASIS rather than IBSR or LPBA40 because it contains more scans and allows us to build a more accurate generative model, even though the fact that the delineations are not 100% manual compromises the quality of the discriminative model.

B. Results

1) Performance of the different methods: Figures 11, 12 and 13 show box plots with the different metrics for the evaluated methods on the three datasets. Tables III, IV and V display the means and standard deviations of the metrics, as well as Cohen's d and p-values for a one-tailed, paired t-test comparing our method with all the others. The p-value is the estimated probability that the hypothesis "ROBEX is better than method X according to metric Y" is false. Cohen's d is a type of effect size, which complements the information from the t-test: rather than assigning a significance level reflecting whether the relationship could be due to chance, it measures the strength of the apparent relationship between the variables. The larger d is, the stronger the relationship. Cohen's thresholds for small, medium and large are 0.2, 0.5 and 0.8. Figure 14 shows the error volumes, and Figure 15 displays sample outputs from each method and dataset. The box plots are rich in outliers, meaning samples (scans) for which a method produces segmentations much worse than for the other volumes in the same dataset. In the rest of the paper, we use the term "outlier" for such method-specific inferior segmentations.

BET provides good results in general across the datasets and generates very accurate segmentations around the superior region of the brain. It provides the best Dice overlap and mean surface-to-surface distance for LPBA40. However, it produces a number of outliers, especially in OASIS, in which the segmentation often leaks into the eyes (see OASIS-2 in Figure 15, in which the cerebellum is oversegmented). The two-pass approach indeed eliminates most of the false positives that a single call to BET is well-known to produce ([6] reports a $\sim$50% false positive rate in IBSR). Removal of the neck in the IBSR dataset has a large impact of the output (see top row of Figure 14), so preprocessing and/or controlled image acquisition are very important for the performance of this method.

BSE shows potential to produce very accurate segmentations when the parameters are carefully fine tuned. The default parameters work well with the IBSR dataset: except for a case in which the overlap is 0%, the segmentations are as accurate

and robust as those from BET. The parameters given by the authors for the LPBA40 dataset produced again excellent results except for two scans. However, BSE provides the worst results for OASIS when default parameters are used, failing to remove large parts of the neck and skull (see OASIS-2 in Figure 15).

AFNI produces extremely accurate and robust results in IBSR, as much as BET and BSE. However, the performance decreases slightly in LPBA40 (see minor under- and oversegmentations in Figure 15), and even further in OASIS, where the results are almost identical to those from BET, including the presence of outliers (see OASIS-2 in Figure 15 for an example).

BridgeBurner, despite not being a skull stripping algorithm, produces acceptable results for most brains. Its main problem is that it sometimes fails to burn some bridges, leaving in large chunks of skull that are not completely disconnected from the brain boundary. This happens particularly often in IBSR and LPBA40 (see LPBA40-2 in Figure 15 for an example).

FreeSurfer is very robust without any parameter tuning. The range of the metrics is in general small across the datasets, and it barely produces any outliers. However, it often undersegments the brain: it provides nearly 100% sensitivity but also the worst specificity for IBSR and LPBA40, and second-to-worst in OASIS. Moreover, FreeSurfer usually fails to remove the dura matter, which is a well-known flaw of the algorithm (see IBSR-1 in Figure 15). Another disadvantage of FreeSurfer is that it consistently crashes when trying to segment three of the scans in the IBSR dataset (these crashes are also reported in [6]).

The results provided by **GCUT** are very similar to those from FreeSurfer, with two differences: 1. The sensitivity is (on average) similar and the specificity better, resulting in an improvement of the metrics; and 2. they unfortunately produce more outliers, including a case in which there is no overlap between the ground truth and the segmentation (OASIS-1 in Figure 15). Regarding the value of the threshold parameter, 40 seems to give slightly better results in terms of Dice overlap and surface-to-surface distance. As FreeSurfer did in IBSR, GCUT consistently crashes when trying to segment two of the scans from OASIS.

Combining FreeSurfer and GCUT-40 improves the results from both methods. Because their sensitivities are near 100% for all datasets, the logical AND of their outputs has the effect of removing false positives with very little impact on the sensitivity. Moreover, the combined method displays very few outliers.

Finally, **ROBEX** produces extremely robust results in all three datasets, providing:

- The best Dice overlap for IBSR ($p \leq 0.02$ and $d \geq 0.5$ for all methods except for BSE and BridgeBurner, see Table III).
- The second to best Dice overlap in LPBA40 after BET ($p <$3e-5 and $d \geq 0.7$ for all the others except for BSE, see Table IV).
- The best Dice overlap in OASIS ($p \leq$2e-8 and $d \geq 0.7$ for all methods except for GCUT and its combination with FreeSurfer, see Table V).
- The best mean surface-to-surface distance for all datasets except for BET in LPBA40. Most of the differences are statistically significant at $p = 0.05$ and display medium or large effect size (again, see tables).
- At least medium effect size (i.e. $d \geq 0.5$) and significantly smaller (at $p = 0.05$) Haussdorff distances than any other method for all datasets, except for BSE in IBSR and LPBA40.
- The highest minimum Dice overlap across each dataset, which is another measure of robustness. Therefore, it also provides the highest minimum Dice overlap across all the scans at 93.3% (FreeSurfer is second at 82.6% and BET third at 77.3%).

Compared with the version without the free deformation step, the refined segmentation significantly improves all metrics ($p < 0.05$, $d \geq 0.6$) for every dataset except for the Haussdorff distance in LPBA40 and OASIS. The refinement captures obvious brain boundaries that are slightly outside of the model; see, for instance, case IBSR-1 in Figure 15.

The main disadvantage with ROBEX is that it does not produce segmentations as sharp as BET or BSE. In brains with very convoluted surfaces, gyri and sulci are oversmoothed, leading to inclusion of dura and/or gray matter loss. For the same reason, ROBEX fails to provide a very accurate segmentation at the posterior region of the cerebellum-cerebrum interface (see for example IBSR-2 in Figure 15 and the error volume in Figure 14).

The dependence of the performance on the datasets is also interesting to observe. IBSR is the dataset with lowest resolution, most anisotropic voxels ($1\times1\times3.1$mm) and most obvious artifacts. LPBA40 is also fairly anisotropic, but at a much better resolution ($0.86\times0.86\times1.5$mm) and with much less noise. Finally, OASIS is isotropic and has a good signal to noise ratio but it includes demented subjects with probable Alzheimer's disease. It is therefore not surprising that all methods perform worst in either IBSR or OASIS. BET, FreeSurfer, GCUT and their combination achieve the lowest Dice overlap in OASIS, whereas BSE, AFNI, BB and ROBEX perform worst in OASIS.

It is important to note that there are some discrepancies between the results presented here and in some other studies that used the same datasets. These discrepancies can be explained by differences in software version and parameter settings, especially for BSE. *Sadananthan et al.* [6] report a Dice overlap equal to 79% for BSE in IBSR (91% in this study), but they did not use the default parameters. They also report very poor results for BET (74% vs. 84% here), but the reason is that they used the single-pass version. *Zhuang et al.*, who achieve 96% overlap in IBSR with their proprietary method, report a 69% overlap for BET and 88% for BSE on this dataset, possibly due to using older versions. Finally, *Shattuck et al.* [45] report results that are quite consistent with ours in LPBA40. In that study, they also describe a website in which users can upload their segmentations of LPBA40 and the results are compared. There are several methods that have reported better results than ours, but they are not publicly

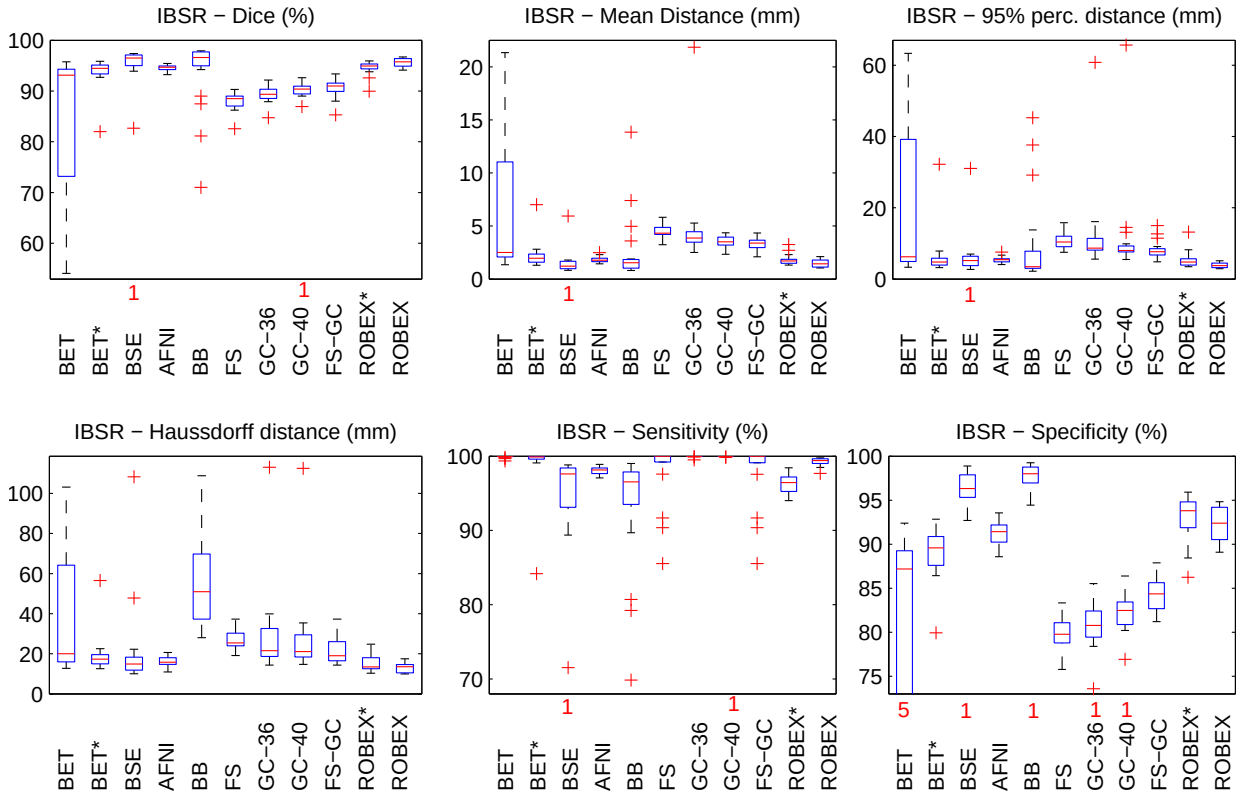

Fig. 11. Box plots of the results for the IBSR dataset. BET* refers to the results on the trimmed scans. ROBEX* corresponds to the results of our algorithm without the graph-cut refinement. On each box, the central mark is the median, the edges of the box are the 25^{th} and 75^{th} percentiles and the whiskers extend to the most extreme data points not considered outliers i.e. within three standard deviations from the mean. The outliers are plotted individually. The number of points that were left out of the plot (to allow a closer look and better interpretation of the rest of the data) is marked in red below the horizontal axis. The plots do not consider the three cases in which FreeSurfer (and therefore FreeSurfer+GCUT) crashed.

TABLE III
IBSR DATASET: MEANS AND STANDARD DEVIATIONS OF THE METRICS, COHEN'S d (EFFECT SIZE) AND P-VALUES OF PAIRED T-TESTS COMPARING THE DIFFERENT METHODS WITH ROBEX, AND NUMBER OF CASES FOR WHICH THE ALGORITHMS CRASH. BET* CORRESPONDS TO THE TRIMMED SCANS, AND ROBEX* TO OUR METHOD BEFORE THE FREE DEFORMATION.

Method	Dice	Av. dist.	Haussdorff	95% dist.	Sensitivity	Specificity	# crashes
BET	84.3±16.2	6.6±7.5	38.0±33.8	19.7±23.3	99.9±0.2	75.7±21.3	0
p-val / Cohen's d	3.4e-3/0.7	3.9e-3/0.7	3.0e-3/0.7	3.9e-3/0.7	1.0/-1.5	1.9e-3/0.7	n/a
BET*	93.8±2.9	2.2±1.2	19.1±9.2	6.2±6.2	99.0±3.5	89.1±2.8	0
p-val / Cohen's d	5.1e-3/0.6	5.4e-3/0.6	1.1e-2/0.6	5.0e-2/0.4	3.8e-1/0.1	1.7e-5/1.2	n/a
BSE	90.8±21.6	3.9±10.6	21.0±22.1	10.6±20.2	90.2±22.1	91.6±21.6	0
p-val / Cohen's d	1.6e-1/0.2	1.6e-1/0.2	6.8e-2/0.3	7.7e-2/0.3	4.2e-2/0.4	4.4e-1/0.0	n/a
AFNI	94.5±0.6	1.8±0.3	16.2±2.5	5.4±0.9	98.1±0.5	91.2±1.4	0
p-val / Cohen's d	8.1e-7/1.5	2.4e-7/1.7	4.7e-4/0.9	2.1e-8/2.0	2.5e-10/2.6	2.0e-3/0.7	n/a
BB	94.0±6.9	2.5±3.1	55.4±21.7	9.3±12.7	93.3±7.8	95.8±9.6	0
p-val / Cohen's d	1.6e-1/0.2	7.0e-2/0.3	3.1e-8/1.9	3.5e-2/0.4	1.1e-3/0.8	9.4e-1/-0.4	n/a
FS	87.9±1.8	4.4±0.6	26.6±4.4	10.7±2.2	97.9±4.4	79.8±1.9	3
p-val / Cohen's d	7.9e-12/4.0	8.3e-15/6.3	1.2e-8/2.5	3.8e-10/3.1	1.0e-1/0.3	4.0e-19/11.8	n/a
GC-36	87.5±8.8	4.8±4.1	29.0±21.2	12.2±11.8	100.0±0.1	78.6±10.8	0
p-val / Cohen's d	3.3e-4/0.9	9.6e-4/0.8	1.9e-3/0.7	2.6e-3/0.7	1.0/-1.3	1.2e-5/1.2	n/a
GC-40	85.8±20.2	4.5±4.6	27.3±20.9	11.4±12.9	95.0±22.3	78.2±18.5	0
p-val / Cohen's d	2.1e-2/0.5	4.0e-3/0.7	3.9e-3/0.7	9.1e-3/0.6	2.0e-1/0.2	1.6e-3/0.8	n/a
FS-GC	90.5±1.9	3.3±0.6	21.4±6.6	8.3±2.6	97.9±4.4	84.2±1.7	3
p-val / Cohen's d	1.3e-8/2.4	6.5e-11/3.5	2.8e-4/1.0	3.8e-6/1.6	9.7e-2/0.3	7.3e-17/8.5	n/a
ROBEX*	94.6±1.3	1.8±0.5	15.3±4.5	5.4±2.2	96.3±1.2	93.1±2.6	0
p-val / Cohen's d	1.9e-3/0.7	2.2e-3/0.7	1.1e-2/0.6	2.1e-3/0.7	2.4e-13/3.8	9.2e-1/-0.3	n/a
ROBEX	95.6±0.8	1.5±0.3	13.3±2.6	3.8±0.7	99.2±0.5	92.3±1.9	0

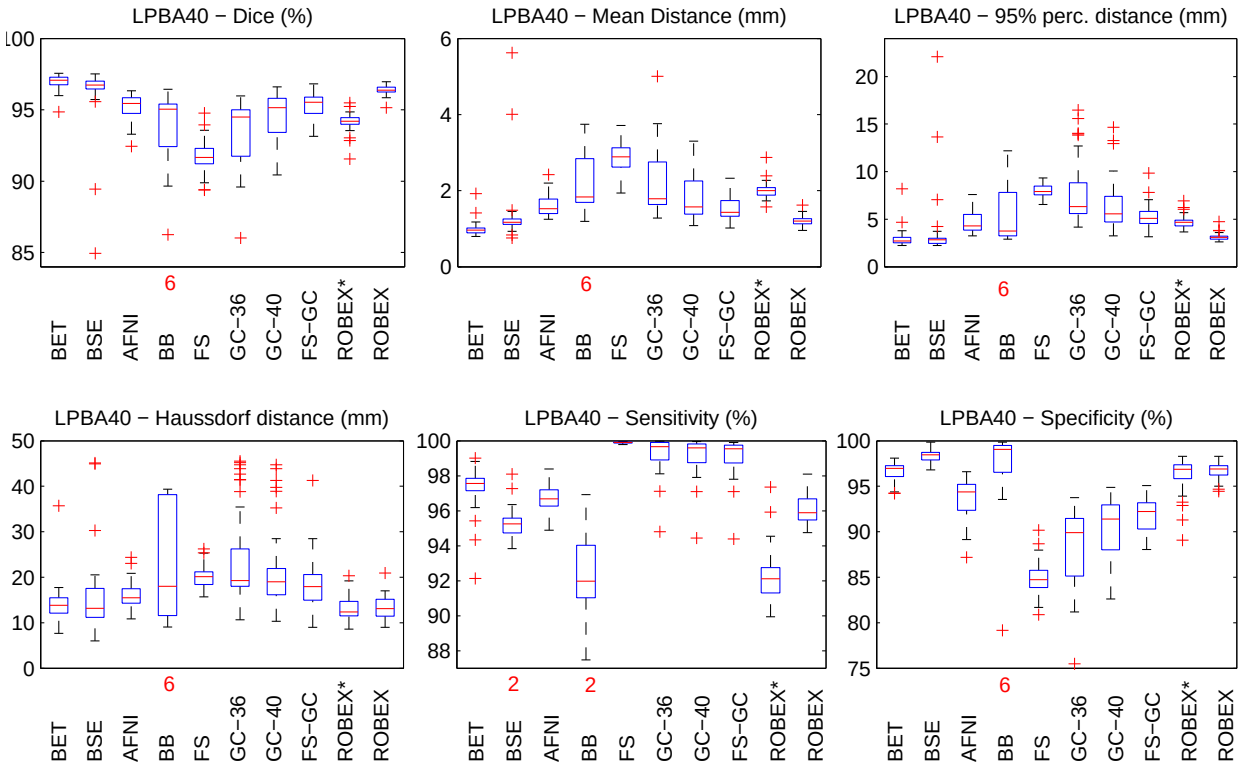

Fig. 12. Box plots of the results for the LPBA40 dataset (see caption of Figure 11).

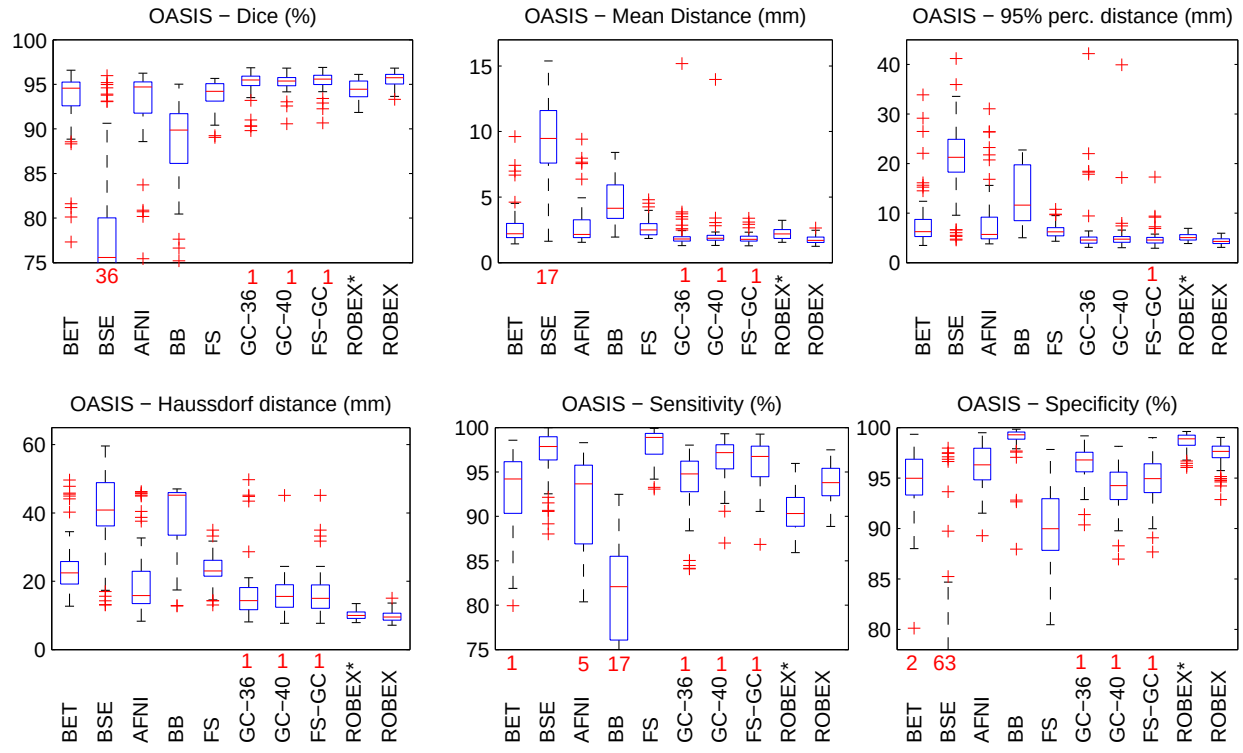

Fig. 13. Box plots of the results for the OASIS dataset (see caption of Figure 11). GCUT (and therefore FreeSurfer+GCUT) crashed in two cases.

TABLE IV
LPBA40 DATASET: MEANS AND STANDARD DEVIATIONS OF THE METRICS, COHEN'S d (EFFECT SIZE) AND P-VALUES (SEE CAPTION OF TABLE III).

Method	Dice	Av. dist.	Haussdorff	95% dist.	Sensitivity	Specificity	# crashes
BET	97.3±0.5	1.0±0.2	14.2±4.3	3.0±1.0	97.0±1.3	97.7±0.8	0
p-val / Cohen's d	1.0/-1.3	1.0/-1.1	3.9e-2/0.3	8.4e-1/-0.2	1.0/-1.1	4.7e-1/0.0	n/a
BSE	96.2±2.3	1.4±0.8	15.5±8.2	3.6±3.5	93.7±4.1	99.0±0.5	0
p-val / Cohen's d	1.1e-1/0.2	1.2e-1/0.2	6.1e-2/0.3	1.8e-1/0.1	1.9e-3/0.5	1.0/-1.8	n/a
AFNI	95.6±0.7	1.6±0.3	16.1±2.9	4.8±1.2	96.4±0.8	94.8±1.9	0
p-val / Cohen's d	1.1e-11/1.5	2.4e-12/1.5	4.0e-7/0.9	1.2e-12/1.6	1.0/-1.0	5.7e-14/1.8	n/a
BB	90.5±8.6	3.5±3.4	31.2±28.6	10.3±12.2	91.1±3.0	91.7±15.6	0
p-val / Cohen's d	2.5e-5/0.7	3.8e-5/0.7	1.4e-4/0.6	3.1e-4/0.6	5.6e-13/1.6	1.0e-2/0.4	n/a
FS	92.5±1.0	2.9±0.4	20.1±2.4	8.0±0.8	99.9±0.0	86.1±1.7	0
p-val / Cohen's d	5.0e-25/3.7	1.7e-26/4.1	2.6e-14/1.8	3.9e-32/5.8	1.0/-4.5	5.9e-37/7.7	n/a
GC-36	94.0±2.2	2.3±0.9	24.1±10.4	7.8±3.3	99.2±1.0	89.4±4.2	0
p-val / Cohen's d	1.4e-9/1.2	4.2e-10/1.3	8.7e-9/1.1	6.2e-12/1.5	1.0/-2.8	4.8e-16/2.1	n/a
GC-40	95.1±1.6	1.8±0.6	21.2±9.2	6.5±2.7	99.1±1.1	91.5±3.2	0
p-val / Cohen's d	1.1e-7/1.0	1.3e-8/1.1	3.8e-7/0.9	5.7e-11/1.4	1.0/-2.6	4.3e-16/2.1	n/a
FS-GC	95.9±0.8	1.5±0.3	18.2±5.6	5.3±1.2	99.0±1.1	93.1±1.8	0
p-val / Cohen's d	8.9e-6/0.8	1.5e-7/1.0	4.8e-8/1.0	1.7e-14/1.8	1.0/-2.6	8.5e-21/2.8	n/a
ROBEX*	94.3±0.6	2.0±0.2	13.0±2.5	4.7±0.7	91.6±1.5	97.1±1.7	0
p-val / Cohen's d	9.4e-35/6.8	6.3e-34/6.5	8.5e-1/-0.2	2.1e-24/3.6	1.1e-22/3.2	1.0e-2/0.4	n/a
ROBEX	96.6±0.3	1.2±0.1	13.3±2.5	3.1±0.4	95.6±0.9	97.7±0.7	0

TABLE V
OASIS DATASET: MEANS AND STANDARD DEVIATIONS OF THE METRICS, COHEN'S d (EFFECT SIZE) AND P-VALUES (SEE CAPTION OF TABLE III).

Method	Dice	Av. dist.	Haussdorff	95% dist.	Sensitivity	Specificity	# crashes
BET	93.1±3.7	2.7±1.4	23.7±8.3	8.2±5.5	92.5±5.4	94.2±5.0	0
p-val / Cohen's d	1.6e-8/0.7	3.9e-9/0.7	7.5e-25/1.7	8.2e-9/0.7	5.8e-3/0.3	8.4e-8/0.7	n/a
BSE	76.8±8.8	9.7±4.5	40.6±11.3	20.6±7.2	97.1±2.6	64.9±14.0	0
p-val / Cohen's d	4.7e-30/2.1	2.4e-25/1.8	5.9e-37/2.7	1.8e-31/2.2	1.0/-1.2	7.6e-33/2.3	n/a
AFNI	93.0±4.0	2.8±1.6	19.3±9.6	8.2±5.8	90.6±7.9	96.2±2.2	0
p-val / Cohen's d	1.2e-8/0.7	1.6e-9/0.8	3.9e-14/1.0	1.6e-8/0.7	3.6e-5/0.5	1.1e-8/0.7	n/a
BB	88.6±4.2	4.6±1.6	39.5±10.1	13.3±5.8	80.6±6.7	98.8±1.7	0
p-val / Cohen's d	7.6e-27/1.9	2.3e-28/2.0	1.2e-40/3.0	7.5e-24/1.6	2.2e-33/2.4	1.0/-0.9	n/a
FS	93.9±1.5	2.6±0.6	23.4±4.3	6.4±1.4	98.1±1.7	90.2±3.5	0
p-val / Cohen's d	1.8e-10/0.8	6.6e-14/1.0	1.9e-37/2.7	6.2e-17/1.2	1.0/-3.0	9.0e-37/2.7	n/a
GC-36	93.9±11.0	2.1±1.6	17.2±11.6	5.9±5.5	92.7±11.3	95.2±11.1	2
p-val / Cohen's d	9.7e-2/0.2	3.3e-2/0.2	1.0e-7/0.7	8.5e-3/0.3	2.1e-1/0.1	4.2e-2/0.2	n/a
GC-40	94.0±10.9	2.1±1.4	17.0±8.9	5.4±4.4	95.2±11.3	92.9±10.9	2
p-val / Cohen's d	1.2e-1/0.1	3.7e-2/0.2	1.5e-10/0.8	2.3e-2/0.2	8.8e-1/-0.1	3.0e-4/0.4	n/a
FS-GC	94.1±11.0	2.3±3.6	17.9±16.1	5.7±7.7	94.8±11.3	94.0±7.6	2
p-val / Cohen's d	1.4e-1/0.1	1.1e-1/0.1	1.5e-5/0.5	7.3e-2/0.2	7.9e-1/-0.1	9.7e-5/0.5	n/a
ROBEX*	94.4±1.0	2.2±0.4	10.1±1.2	5.2±0.8	90.6±2.3	98.6±0.9	0
p-val / Cohen's d	5.2e-25/1.7	6.2e-26/1.8	9.4e-2/0.2	6.0e-21/1.5	1.0e-42/3.3	1.0/-1.5	n/a
ROBEX	95.5±0.8	1.8±0.3	9.8±1.7	4.4±0.6	93.8±2.1	97.4±1.2	0

available and cannot be compared on the other datasets. [1]

2) Effect of age: Figure 16 displays a scatter plot of the Haussdorff distance achieved by ROBEX for each case against the age of the subject. We chose the Haussdorff distance because it shows more spread than any of the other metrics. The two datasets for which age data are available (IBSR and OASIS) are analyzed separately. The distance shows a positive correlation with age as expected, even though it is only significant for OASIS at $p = 0.05$. For IBSR, the 95% confidence interval of the correlation coefficient ρ extends beyond zero, and the p value for the hypothesis that the slope of the regression is greater than zero is p=6.4e-2. For OASIS, the lower bound of the 95% confidence interval of ρ is 3.6e-3 and the p value for the test on the slope is $p = 2.0$e-2, so the relation is significant but very weak: the range of the predicted Haussdorff distance across the dataset is just 1.25 mm.

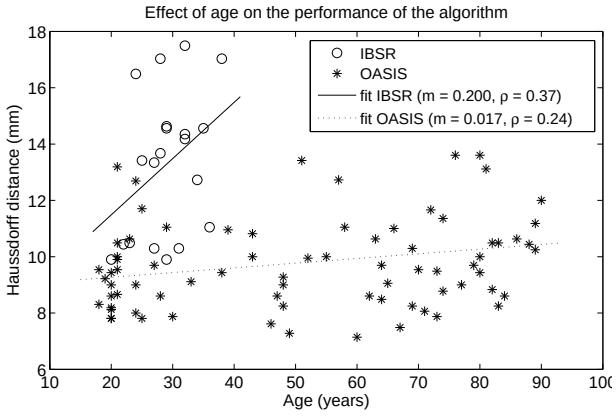

Fig. 16. Haussdorff distance vs. age for ROBEX in IBSR and OASIS. The 95% confidence interval of the correlation coefficient is [-0.08,0.70] for IBSR and [0.02,0.44] for OASIS. The p-value for the hypothesis that the slope is positive is 0.0644 for IBSR and 0.0195 for OASIS.

[1]Actually, some of the methods in the website have been trained on the test dataset (LPBA40) itself, which makes the comparison with other algorithms unfair.

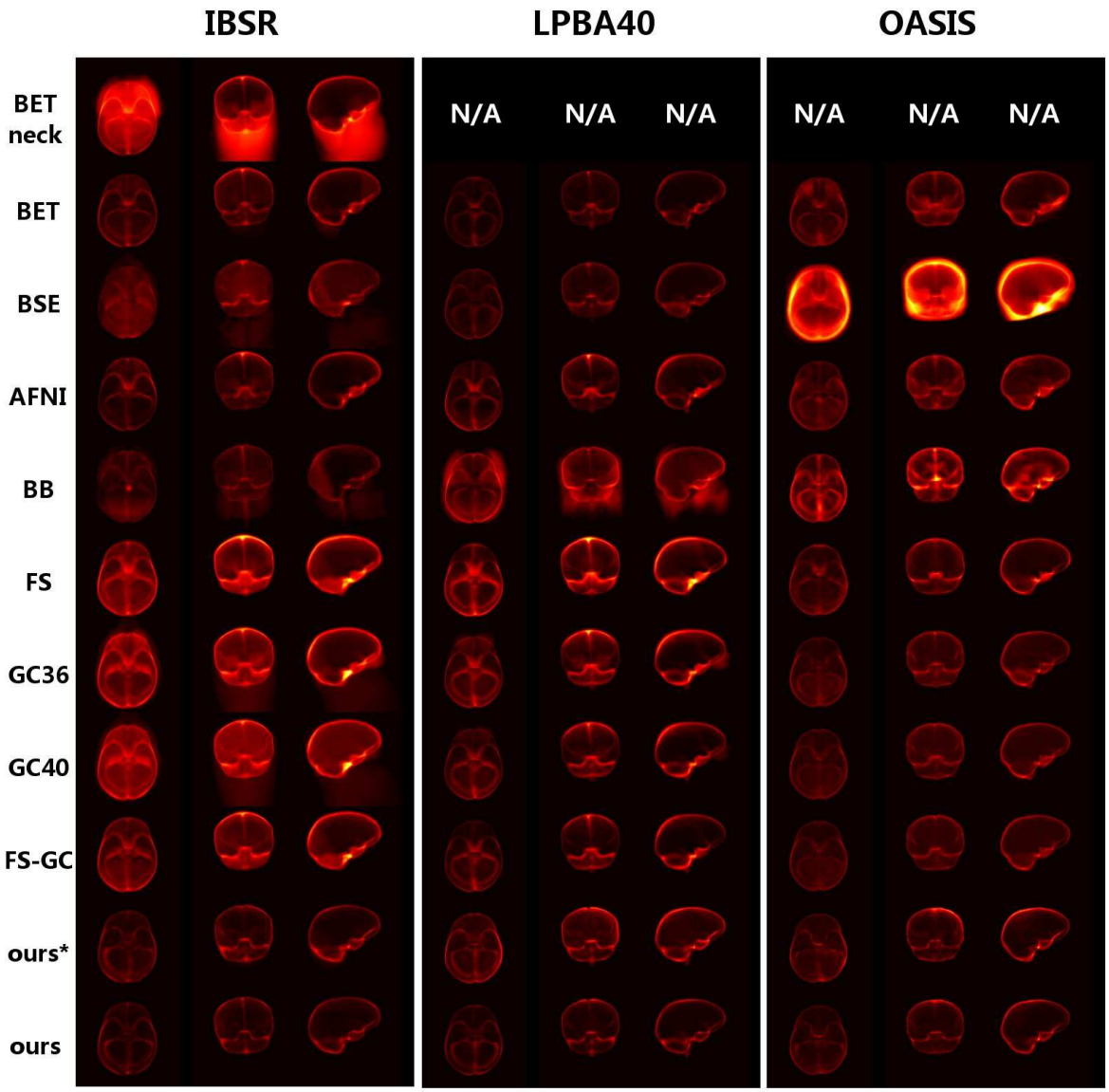

Fig. 14. Averages of the error volumes along the inferior/superior, anterior/posterior and left/right axes for each method and dataset.

3) Effect of training dataset: Table VI compares the different metrics when OASIS and the original training dataset are used to train the generative and discriminative models. The performance decreases slightly as expected given that the annotations of OASIS are not as accurate as the labeling of our proprietary dataset. The decrease is most noticed in the surface-to-surface distances, since the inaccurate training data makes finding the exact boundary difficult. However, the values of the robustness metrics are still better than those of the other methods, particularly the minimal Dice coefficient. We can thus conclude that the quality of the training data has some influence on the results, but it is clearly not the main reason why the system is robust.

V. Discussion and Conclusion

A new skull stripping method has been presented in this article. The main contribution of our study lies in two aspects: a learning-based hybrid generative/discriminative model for skull stripping and a thorough experimental evaluation using three different datasets and six competing methods.

The proposed algorithm uses a hybrid model in which a generative model of brain boundary is fitted to the output of a classifier that is trained to find the contour of the brain in the data. The use of a hybrid model is imperative in learning-based systems for MRI image analysis. Because of the lack of image intensity standardization in this modality (as opposed to other modalities such as computed tomography), analysis based solely on discriminative features is not sufficient to obtain good results, especially when the acquisition conditions change. However, the generative model in our framework guides the segmentation and guarantees than the output corresponds to a plausible brain shape. The two models complement one another very well because the classifier provides local precision whereas the shape model provides robustness.

The method has been compared with six popular, well-established methods that are commonplace in the literature and

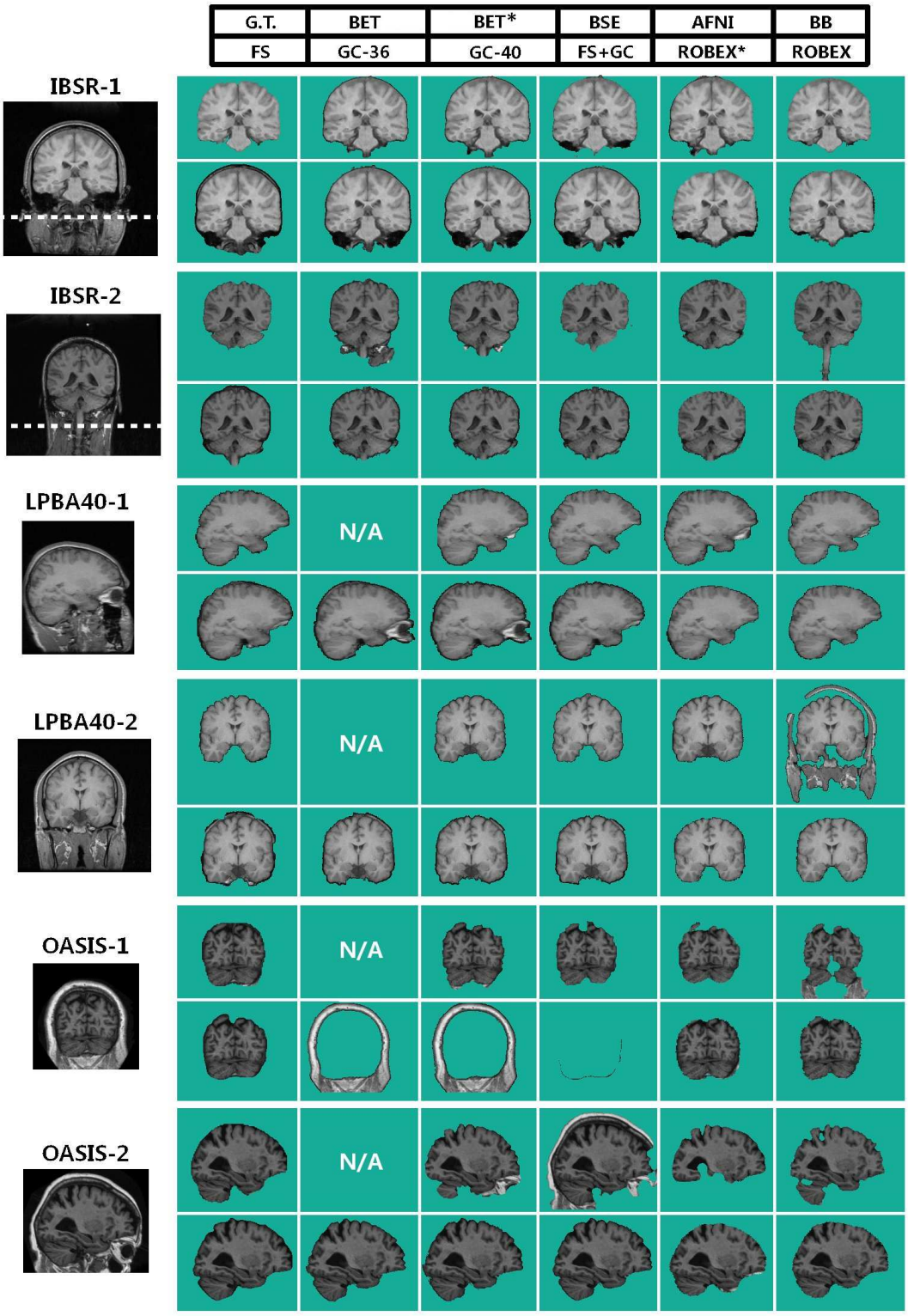

Fig. 15. Outputs for two scans from each dataset. Two orthogonal slices are shown for each volume. IBSR: coronal, coronal; LPBA40: saggital, coronal; and OASIS: coronal, saggital. Again, BET* refers to the output for the trimmed scans. The axial coordinate of the trimming is illustrated in cases IBSR-1 and IBSR-2.

TABLE VI
PERFORMANCE METRICS WHEN OASIS IS USED FOR TRAINING.

Metric	Dice (%)	Dice (range, in %)	Mean dist.(mm)	Haussdorff (mm)	95% dist.(mm)	Sensit.(%)	Specif.(%)
IBSR - our dataset	95.6±0.8	[94.1,96.7]	1.5±0.3	13.3±2.6	3.8±0.7	99.2±0.5	92.3±1.9
IBSR - OASIS	93.2±1.3	[91.4,96.1]	2.4±0.5	21.4±4.2	4.0±1.0	97.8±0.8	91.5±2.2
LPBA40 - our dataset	96.6±0.3	[95.2,97.0]	1.2± 0.1	13.3±2.5	3.1±0.4	95.6±0.9	97.7±0.7
LPBA40 - OASIS	95.7±0.5	[94.5,97.1]	1.7±0.2	15.7±3.1	3.4±0.6	96.0±1.0	94.8±1.1

publicly available. Three publicly available datasets were used for evaluation purposes. Our method outperforms all the others in almost every case, and it is much more robust: some of the other methods produce comparable results in certain datasets, but falter when the scan comes from a different source.

The lowest Dice overlap given by ROBEX in the three datasets is 93%. BET and AFNI, which follow similar principles, perform well in general with little or no parameter adjustment, but produce some outliers for which the Dice overlap is below 80%. BET also needs that little or no neck is visible in the input volume, which can require additional preprocessing. BSE produces very accurate results when its parameters are well tuned, but it is extremely sensitive to parameter values and small deviations can produce substantial loss of quality in the segmentation. For example, when default parameter values are used in LPBA40, the Dice overlap decreases almost 24 points [45]. BridgeBurner, which does brain tissue segmentation rather than skull stripping, is not robust at all but produces very sharp brain boundaries, which can be useful if the user is willing to manually edit the output. If that is not the case, FreeSurfer can be used instead, given its large sensitivity and robustness (and despite its relatively low specificity). Finally, GCUT is also provides high sensitivity and sharp brain boundaries, but when it makes mistakes, these are usually large e.g. leaving the cerebellum out or the eyes in. GCUT is best used in combination with FreeSurfer because they are both highly sensitive and cancel some of each other's false positives.

The experimental evaluation in this study is based only on publicly available datasets with publicly available ground truth. Hence, all the experiments in this paper can be reproduced. Even though the training dataset is not publicly available, the trained system can be downloaded from the first author's homepage: http://loni.ucla.edu/~jiglesia/ROBEX. Both the source code and executables in different platforms have been made available. To show that the performance of the system does not depend exclusively on the unavailable training data, the results were successfully reproduced using OASIS as training dataset. The use of publicly available datasets and methods is an increasing trend in the medical imaging community, for example in challenge workshops at conferences [46]–[49]. It is often the case that meta-algorithms that combine all the methods in the challenge provide the best results (see [50], [51] for meta-algorithms in skull-stripping combining some of the methods described above).

In this study we have focused on T1 MRI, but extending ROBEX to other modalities (T2, proton density, etc.) would be immediate: the only required modification would be to train the classifier with data acquired with the modality of interest; the fitting of the shape model would be the same. If images from more than one modality are available for a test case, which is the usual clinical scenario, it would certainly be possible to use all the channels simultaneously for the classification. This should in principle improve the results. However, there are no publicly available datasets (to the best of our knowledge) with multi-spectral information and manual delineations of the brain to test this approach.

Extending the method to other medical image segmentation problems would, in principle, be possible. The least general step of the algorithm is be the registration, not because registration is not general, but because one cannot expect the alignment to be as good as it usually is in the brain, which is relatively easy to register. This is particularly true for articulated or highly anatomically variable structures.

One of the disadvantages of the presented approach is that it tends to oversmooth the contour of the brain. In some extreme cases, ROBEX can leave out some gray matter, which can represent a problem if the next step in the image analysis pipeline is estimating the cortical thickness or measuring the gray matter volume. However, it would be possible to add a second refinement stage to ameliorate this problem, perhaps increasing the density of the mesh or using some other approach. It would also be interesting to study the behavior of the algorithm in cases with pathologies that alter the brain structure more severely than dementia and Alzheimer's disease, for example, brain tumors.

Another aspect of the system that could be improved is the image intensity standardization step. The proposed system uses a combination of robust histogram stretching and equalization, but the segmentation could benefit from more sophisticated, brain MRI-specific approaches. The better the intensity matching, the higher the quality of the boundary probability volumes (Figure 4) and the better the final segmentation.

Finally, it is important to discuss the computational requirements and execution time of the algorithm. Most of the methods discussed in this paper run in approximately one minute on a modern desktop. The two exceptions are BridgeBurner and BSE, which run in just two or three seconds. The original BET algorithm is also extremely fast, but the two-pass version used in this study requires registration to an atlas, which is the bottleneck of the algorithm. Our single-threaded implementation of ROBEX runs in two or three minutes. Half of that time is spent on the registration. Making ROBEX faster, refining the output mask and improving the intensity standardization remain as future work.

VI. ACKNOWLEDGEMENTS

This work was funded by grants NSF 0844566, NIH U54 RR021813, NIH P41 RR013642 and ONR N000140910099. The authors would like to thank Dr. Stephen Smith, Dr. David

Shattuck, Dr. Ziad Saad, Dr. Henry Rusinek, Dr. Bruce Fischl, Dr. Vitali Zagorodnov and Artem Mikheev for tuning their algorithms to the different datasets. The first author would also like to thank the U.S. Department of State's Fulbright program for the funding.

REFERENCES

[1] S. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.
[2] D. Shattuck, S. Sandor-Leahy, K. Schaper, D. Rottenberg, and R. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, no. 5, pp. 856–876, 2001.
[3] "FreeSurfer," online at http://surfer.nmr.mgh.harvard.edu.
[4] "AFNI," online at http://afni.nimh.nih.gov.
[5] A. Mikheev, G. Nevsky, S. Govindan, R. Grossman, and H. Rusinek, "Fully automatic segmentation of the brain from T1-weighted MRI using Bridge Burner algorithm," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 6, pp. 1235–1241, 2008.
[6] S. Sadananthan, W. Zheng, M. Chee, and V. Zagorodnov, "Skull stripping using graph cuts," *NeuroImage*, vol. 49, no. 1, pp. 225–239, 2010.
[7] "IBSR," online at http://www.cma.mgh.harvard.edu/ibsr.
[8] D. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. Narr, R. Poldrack, R. Bilder, and A. Toga, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, 2008.
[9] "OASIS," online at http://www.oasis-brains.org.
[10] D. Greig, B. Porteous, and A. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society, series B.*, vol. 51, no. 2, pp. 271–279, 1989.
[11] F. Segonne, A. Dale, E. Busa, M. Glessner, D. Salat, H. Hahn, and B. Fischl, "A hybrid approach to the skull stripping problem in MRI," *Neuroimage*, vol. 22, no. 3, pp. 1060–1075, 2004.
[12] L. Lemieux, G. Hagemann, K. Krakow, and F. Woermann, "Fast, accurate, and reproducible automatic segmentation of the brain in T1-weighted volume MRI data," *Magnetic Resonance in Medicine*, vol. 42, no. 1, pp. 127–135, 1999.
[13] H. Hahn and H. Peitgen, "The skull stripping problem in mri solved by a single 3d watershed transform," in *Proceedings of MICCAI*, 2000, pp. 134–143.
[14] X. Zeng, L. Staib, R. Schultz, and J. Duncan, "Segmentation and measurement of the cortex from 3-d mr images using coupled-surfaces propagation," *IEEE Transactions onMedical Imaging*, vol. 18, no. 10, pp. 927–937, 1999.
[15] D. MacDonald, N. Kabani, D. Avis, and A. Evans, "Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI," *NeuroImage*, vol. 12, no. 3, pp. 340–356, 2000.
[16] C. Baillard, P. Hellier, and C. Barillot, "Segmentation of brain 3D MR images using level sets and dense registration* 1," *Medical Image Analysis*, vol. 5, no. 3, pp. 185–194, 2001.
[17] A. Zhuang, D. Valentino, and A. Toga, "Skull-stripping magnetic resonance brain images using a model-based level set," *NeuroImage*, vol. 32, no. 1, pp. 79–92, 2006.
[18] Z. Shan, G. Yue, and J. Liu, "Automated histogram-based brain segmentation in T1-weighted three-dimensional magnetic resonance head images," *NeuroImage*, vol. 17, no. 3, pp. 1587–1598, 2002.
[19] Z. Tu, K. Narr, P. Dollar, I. Dinov, P. Thompson, and A. Toga, "Brain anatomical structure segmentation by hybrid discriminative/generative models," *IEEE Transactions on Medical Imaging*, vol. 27, pp. 495–508, 2008.
[20] M. De Bruijne and M. Nielsen, "Shape particle filtering for image segmentation," *Proceedings of MICCAI*, pp. 168–175, 2004.
[21] B. Van Ginneken, A. Frangi, J. Staal, B. ter Haar Romeny, and M. Viergever, "Active shape model segmentation with optimal features," *IEEE Transactions on Medical Imaging*, vol. 21, no. 8, pp. 924–933, 2002.
[22] T. Cootes, C. Taylor, D. Cooper, J. Graham *et al.*, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.
[23] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[24] K. Li, X. Wu, D. Chen, and M. Sonka, "Optimal surface segmentation in volumetric images-a graph-theoretic approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 119–134, 2006.
[25] S. Klein, M. Staring, K. Murphy, M. Viergever, and J. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2010.
[26] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage*, vol. 23, pp. 151–S160, 2004.
[27] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings 23rd International Conference on Machine learning, ACM*, 2006, pp. 161–168.
[28] Z. Yi, A. Criminisi, J. Shotton, and A. Blake, "Discriminative, semantic segmentation of brain tissue in mr images," in *Proc. of MICCAI*, 2009, pp. 558–565.
[29] B. Andres, U. Köthe, M. Helmstaedter, W. Denk, and F. Hamprecht, "Segmentation of sbfsem volume data of neural tissue by hierarchical classification," in *Proceedings of Pattern recognition*. Springer, 2008, pp. 142–152.
[30] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
[31] N. Tustison and J. Gee, "N4itk: Nick's n3 itk implementation for mri bias field correction," http://hdl.handle.net/10380/3053.
[32] J. Sled, A. Zijdenbos, and A. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998.
[33] P. Hellier, "Consistent intensity correction of MR images," vol. 1, pp. 1109–1112, 2003.
[34] C. Strobl, A. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.
[35] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.
[36] R. Davies, T. Cootes, and C. Taylor, "A minimum description length approach to statistical shape modelling," in *Proceedings of IPMI*, 2001, pp. 50–63.
[37] A. Frangi, D. Rueckert, J. Schnabel, and W. Niessen, "Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling," *IEEE Transactions on Medical Imaging*, vol. 21, no. 9, pp. 1151–1166, 2002.
[38] H. Thodberg, "Minimum description length shape and appearance models," in *Proceedings of IPMI*, 2003, pp. 51–62.
[39] R. Davies, C. Twining, T. Cootes, and C. Taylor, "Building 3-d statistical shape models by direct optimization," *IEEE Transactions on Medical Imaging*, vol. 29, no. 4, pp. 961–981, 2010.
[40] I. Dryden and K. Mardia, *Statistical shape analysis*. Wiley New York, 1998.
[41] I. Jolliffe, *Principal component analysis*. Springer verlag, 2002.
[42] Q. Fang and D. Boas, "Tetrahedral mesh generation from volumetric binary and gray-scale images," in *Proceedings of ISBI*, 2009, pp. 1142–1145.
[43] H. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.
[44] Y. Boykov and V. Kolmogorov, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.
[45] "Online resource for validation of brain segmentation methods," *NeuroImage*, vol. 45, no. 2, pp. 431 – 439—, year = 2009.
[46] B. van Ginneken, T. Heimann, and M. Styner, "3d segmentation in the clinic: A grand challenge," in *Proceedings of MICCAI workshop with the same title*, 2007, pp. 7–15.
[47] C. Metz, M. Schaap, T. van Walsum, van der Giessen A.G, A. Weustink, N. Mollet, G. Krestin, and W. Niessen, "Editorial: 3d segmentation in the clinic: A grand challenge ii - coronary artery tracking, miccai workshop proceedings," 2008.
[48] X. Deng and G. Du, "3D Liver Tumor Segmentation Challenge, MICCAI workshop proceedings," 2008.
[49] P. Lo, B. van Ginneken, J. Reinhardt, and M. de Bruijne, "Extraction of Airways from CT (EXACT09), MICCAI workshop proceedings," 2009.
[50] K. Rehm, K. Schaper, J. Anderson, R. Woods, S. Stoltzner, and D. Rottenberg, "Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes," *Neuroimage*, vol. 22, no. 3, pp. 1262–1270, 2004.
[51] D. Rex, D. Shattuck, R. Woods, K. Narr, E. Luders, K. Rehm, S. Stolzner, D. Rottenberg, and A. Toga, "A meta-algorithm for brain extraction in MRI," *NeuroImage*, vol. 23, no. 2, pp. 625–637, 2004.