Agreement-Based Semi-supervised Learning for Skull Stripping

Juan Eugenio Iglesias[1], Cheng-Yi Liu[2], Paul Thompson[2], and Zhuowen Tu[2]

[1] Medical Imaging Informatics, University of California, Los Angeles
jeiglesias@ucla.edu
[2] Laboratory of Neuroimaging, University of California, Los Angeles
chengyiliu@ucla.edu, thompson@loni.ucla.edu, zhuowen.tu@loni.ucla.edu

Abstract. Learning-based approaches have become increasingly practical in medical imaging. For a supervised learning strategy, the quality of the trained algorithm (usually a classifier) is heavily dependent on the amount, as well as quality, of the available training data. It is often very time-consuming to obtain the ground truth manual delineations. In this paper, we propose a semi-supervised learning algorithm and show its application to skull stripping in brain MRI. The resulting method takes advantage of existing state-of-the-art systems, such as BET and FreeSurfer, to sample unlabeled data in an agreement-based framework. Using just two labeled and a set of unlabeled MRI scans, a voxel-based random forest classifier is trained to perform the skull stripping. Our system is practical, and it displays significant improvement over supervised approaches, BET and FreeSurfer in two datasets (60 test images).

1 Introduction

Supervised learning approaches have become increasingly popular and practical in brain MRI segmentation [1,2,3]. These algorithms produce classifiers that utilize a large number of features by applying modern learning algorithms. However, supervised learning often demands large amounts of training data with consistent manual labeling, which are difficult to obtain. Recent semi-supervised learning approaches [4,5,6,7] have provided new mechanisms to take advantage of the information in unlabeled data to train a better system.

In this paper, we propose a semi-supervised approach to skull stripping, which is the first element in most neuro image pipelines, and therefore critical for their overall performance. The goal of skull stripping is to segment the brain from non-brain matter in MRI in a robust manner. Skull stripping is expected to follow the major folds on the surface; if the deeper sulci are to be extracted for surface analysis, subsequent post-processing techniques can be used. Automated skull stripping is challenging due to the large variations in image intensity and shape in MRI scans. Expert systems exist in this domain (e.g. BET [8], FreeSurfer [9]), but none of them offer a fully satisfactory performance.

We propose taking advantage of these expert systems and unlabeled data to train a voxel classifier to segment the brain by: 1) training on the labeled data;

T. Jiang et al. (Eds.): MICCAI 2010, Part III, LNCS 6363, pp. 147–154, 2010.

and 2) iteratively re-training the classifier including samples from the unlabeled data for which the expert systems agree but the classifier is not confident yet. This approach is related to the tri-training algorithm [10] from the co-training family[7]. Co-training requires having two or more conditionally independent views (sets of features) for the data, which is often difficult [11], whereas tri-training works on single-view data by simultaneously training three classifiers. The system described in this study can be seen as a special case of tri-training in which two well-developed skull stripping algorithms (BET and FreeSurfer) play the role of two of the classifiers in the framework. The output for a test image is based solely on the trained classifier, so running BET and FreeSurfer on the image to be analyzed is not necessary.

2 A Semi-supervised Skull Stripping Algorithm

2.1 Proposed Method

Training: The trainining process (see top half of flowchart in Figure 1) is divided in four stages: registration, preprocessing of unlabeled data, feature extraction and learning. The first step is to coarsely align all the volumes, labeled and unlabeled, to a template brain scan. The first volume in the dataset was arbitrarily chosen to be the template. This alignment makes it possible to use position features in the posterior classification. ITK (*www.itk.org*) was used to optimize an affine transform using a mutual information metric and a multi-resolution scheme. Using a nonlinear registration method could make the classifier rely too much on the registration through location features, making the method less robust.

The next step is to preprocess the unlabeled volumes. The brain is first segmented using BET and FreeSurfer. The binary outputs of the two methods are then "softened" using a signed distance transform (positive if inside the brain, negative if outside). The distance map is mapped to the template space using the transforms from the registration step. The warped maps are used to calculate preliminary brain masks in the unlabeled scans by averaging the two maps for each volume and thresholding the result at zero, and they will also be used in the posterior semi-supervised learning step.

The third step in the training stage is feature extraction. A pool of 58 image features is used in this study: (x, y, z) position, Gaussian derivatives of order up to two at five different scales ($\sigma = \{1.0, 2.0, 4.0, 8.0, 16.0\}$, in mm), and gradient magnitudes at the same scales. A subset of voxels from the training volumes is randomly selected for training purposes under the constraints that: 1) all scans contribute the same number of voxels; 2) 50% of the voxels have to be positives according to the annotated boundary (for the labeled scans) or the preliminary mask from the previous step (for the unlabeled); and 3) 50% of the voxels have to lie within 5mm of the boundary and 75% within 25mm. The features are normalized to zero mean and unit variance.

Finally, a classifier can be trained using the labeled and unlabeled data. Breiman's random forests[12] were used as the base classifier because they

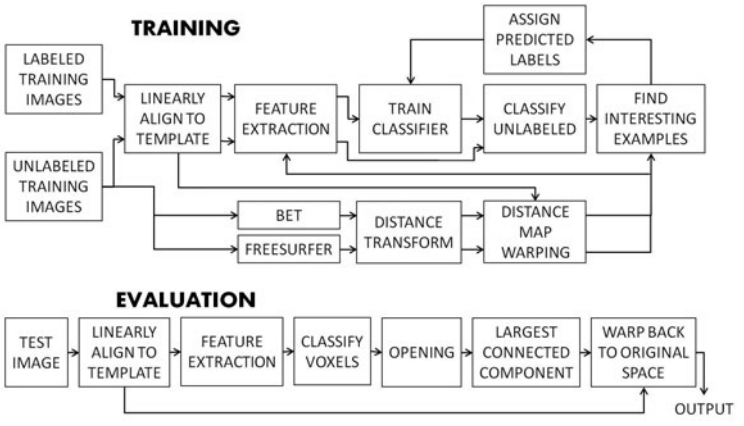

Fig. 1. Flowchart for the training and test stages of the proposed algorithm

compare favorably to other state-of-the-art algorithms[13]. Feature selection is performed through training a preliminary classifier with all the features and 20% of the available data and then retraining with the features that provide the highest mean decrease in accuracy in the out-of-bag data. For the semi-supervised learning, the random forest is first trained solely on the labeled scans, and then updated in an iterative fashion using "interesting" voxels from the unlabeled volumes i.e. those for which both distance maps (BET and FreeSurfer) are greater than a given positive threshold (5 mm in all the experiments in this study). Among those voxels, the ones for which the classifier predicts a negative label with highest probability (i.e. those with fewest trees voting for positive) are shifted from the unlabeled set to the labeled data with positive labels. Then the procedure is repeated with negative voxels using the opposite of the previous threshold. The iteration concludes with retraining the random forest.

Testing: The testing pipeline (see bottom half of flowchart in Figure 1) is similar to the training process, with the important difference that BET and FreeSurfer do not need to be run on the data in this stage. When a new volume is presented to the system, the first step is to register it to the template. The optimized transform is stored to warp the final mask back to the original space later on. From the aligned scan, features are extracted from every voxel and fed to the random forest. The output is a volume with the number of trees that have voted positive at each location. Upon division by the total number of trees, this volume can be interpreted as a probability map for the voxels being part of the brain. The map can then be smoothed and thresholded at 0.5 to binarize the data and obtain the preliminary mask. After applying a morphological opening operator to smooth this mask, the largest connected component is extracted, holes in it are filled, and then it is warped back to the original space using the inverse of the affine transform to obtain the final output.

Theoretical justification: Even though a formal theoretical analysis of the proposed method is out of the scope of this paper, in this section we try to provide a brief justification of why it works. Theoretically, the analysis is very similar to that of tri-training[10], which is in turn very related to [14] and mostly follows the PAC (probably approximately correct) learning theory. Let H^* and H be the ground-truth and our classifier, respectively. It can be shown that, for $Pr[d(H, H^*) \geq \epsilon] \leq \delta$, where $d(\cdot)$ is the difference between H and H^*, one needs to have a sequence of m samples where $m \geq \frac{2}{\epsilon^2(1-2\eta)^2} \ln(\frac{2N}{\delta})$. $\eta < 0.5$ is an upper bound on the error rate by the expert systems, N is the number of possible hypotheses and δ is the confidence in PAC learning.

The error bounds on the unlabeled data for BET and FreeSurfer can be directly estimated using labeled data. Our task is then to design a new rule by combining the experts to make a joint decision that achieves a small error rate η. This can be translated into a function that weights each unlabeled data point with a positive/negative label as:

$$w_i^{\pm}(X|H_i, F_1, F_2) \propto A^{\pm}[F_1(X), F_2(X)]$$

where i denotes the i-th iteration, H_i is the trained classifier at iteration i, $F_1(X)$ and $F_2(X)$ give a measure of how likely X is to be a positive according to the two expert systems, and $A^{\pm}$ measures the level of agreement and confidence of F_1 and F_2 for positive/negative samples. In the proposed method, $F_1(X)$ and $F_2(X)$ are the distance maps provided by BET and FreeSurfer. For the agreement function, success in pilot experiments led to the use of ($H[\cdot]$ is the Heaviside function): $A^{\pm} = H(\pm F_1(X) - 5mm) \cdot H(\pm F_2(X) - 5mm)$.

In this setting, one iteration would ideally suffice because two of the classifiers (BET and FreeSurfer) in the tri-training framework are fixed. The iterative scheme we propose to train the classifier can be seen as a bootstrapping process[15]: m is large but only a limited number of training samples are used. We define an empirical criterion to select samples which will potentially improve the classifier: agreement between BET and FreeSurfer but not with the classifier.

3 Experiments and Results

3.1 Data

Two different datasets are used in this study. The first one, henceforth dataset A, consists of 10 T1-weighted volumes from the LPBA40 dataset[16]. The brain surface was annotated by an expert radiologist in all the scans. The second dataset, henceforth dataset B, consists of 152 T1-weighted scans from healthy subjects acquired with an inversion recovery rapid gradient echo sequence on a Bruker 4T system. Manual delineations of the brain by an expert physiologist are available for 52 of the scans.

3.2 Setup of Experiments

Impact of the number of annotated scans (dataset A): In this experiment, the 10 volumes from dataset A are segmented using a cross-validation

scheme. For scan $i \in \{1, \ldots, 10\}$, and assuming N_{lab} annotated examples: 1) 10 random subsets of N_{lab} elements are extracted from the pool $P = \{1, \ldots, 10\} \setminus \{i\}$; 2) for each subset, a classifier is trained using the N_{lab} extracted elements as labeled data and the remaining $N_{unlab} = 9 - N_{lab}$ elements in P as unlabeled; 3) volume i is processed with the 10 resulting classifiers; and 4) the performance metric for scan i is the average of the performances of the 10 outputs for that scan. Four different metrics are used in this study: classification error rate, Jaccard index of the segmentation (related to the Dice coefficient as $D^{-1} = (1 + J^{-1})/2$), mean surface-to-surface distance, and the 95% quantile of this distance, which is a measure of robustness. The error rate and the Jaccard index are computed using only the voxels that are within 12.5 mm of the annotated boundary for easier interpretation of the results. The settings of the parameters were the following: 100,000 training voxels for the supervised classifier (independently of the number of images); 10 loops of the semi-supervised update with 2,500 voxels each; 500 trees for the random forest; $\sigma = 1$ mm to smooth the likelihood map; and a 2 mm radius spherical element for the opening.

Evaluation on a larger dataset and impact of the number of unlabeled scans (dataset B): In this experiment, two randomly selected scans of the 52 labeled volumes from dataset B are used to train the initial classifier, and the 100 scans without annotations play the role of unlabeled data. The remaining 50 scans are used for evaluation. Then, the experiment was repeated by randomly removing elements from the pool of unlabeled scans in order to asses the impact of the amount of available unlabeled instances N_{unlab}. The parameters were all set to the same values as in the previous experiment. No cross validation was performed, which is reasonable given the size of the dataset.

3.3 Results

The results from the first experiment (dataset A) are shown in Figure 2, which compares the semi-supervised strategy with a supervised version (i.e. same algorithm with no semi-supervised update), BET and FreeSurfer. Though FreeSurfer generally outperforms BET, the latter works better in this particular dataset. The semi-supervised approach outperforms all the others at every value of N_{lab}, whereas the supervised method requires $N_{lab} = 2 \sim 3$ to improve the results of BET, except for the robustness measure. As expected, the difference between the semi-supervised and supervised methods decreases as N_{lab} increases, since the number of unlabeled volumes $N_{unlab} = 9 - N_{lab}$ also becomes lower.

Table 1 shows the p-values for a paired t-test comparing the results given by the semi-supervised method and BET, and also for a test comparing the semi-supervised version against the supervised. Compared with BET, all the metrics improve significantly for $N_{ann} = 2$ at $\alpha = 0.05$ (also for $N_{ann} > 2$, not displayed due to lack of space). Compared with the supervised approach, the difference is significant at $N_{ann} = 1$ but not at $N_{ann} \geq 2$ except for the 95% quantile, due to the low robustness of the supervised method with a small number of

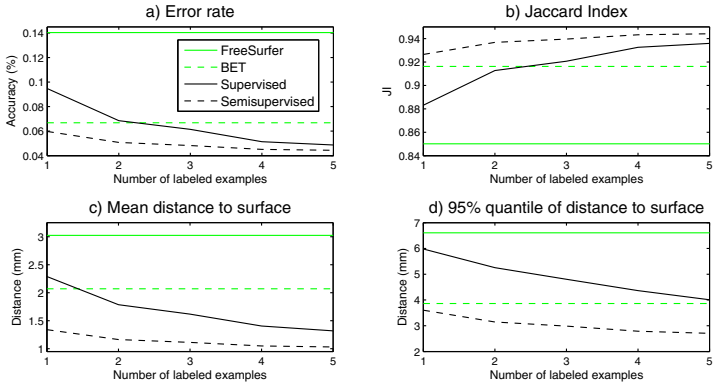

Fig. 2. Different performance metrics for dataset A depending on the number of labeled volumes used in the training N_{lab}, as well as results provided by BET and FreeSurfer

Table 1. Paired t-test comparing the performance metrics provided by the semi-supervised learning and the supervised version / BET

Method	Error rate	Jaccard index	Mean distances	95% quantile
BET ($N_{lab} = 1$)	$2.7{\cdot}10^{-1}$	$2.3{\cdot}10^{-1}$	$1.4{\cdot}10^{-3}$	$2.7{\cdot}10^{-1}$
BET ($N_{lab} = 2$)	$4.0{\cdot}10^{-5}$	$6.2{\cdot}10^{-5}$	$1.3{\cdot}10^{-19}$	$5.0{\cdot}10^{-2}$
Supervised ($N_{lab} = 1$)	$2.9{\cdot}10^{-2}$	$1.6{\cdot}10^{-2}$	$5.6{\cdot}10^{-3}$	$1.1{\cdot}10^{-2}$
Supervised ($N_{lab} = 2$)	$1.9{\cdot}10^{-1}$	$1.5{\cdot}10^{-1}$	$8.3{\cdot}10^{-2}$	$1.3{\cdot}10^{-2}$

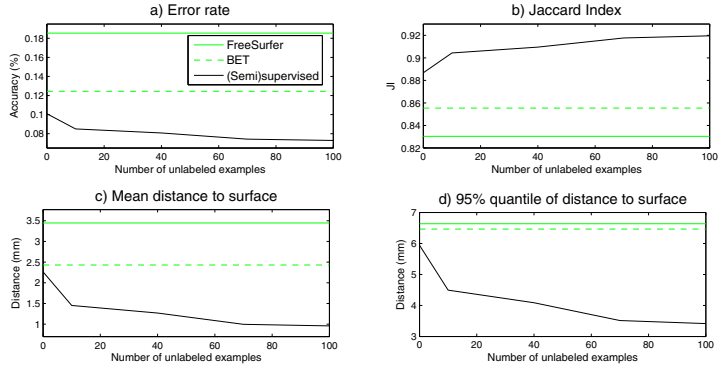

Fig. 3. Different performance metrics for dataset B depending on the number of unlabeled volumes N_{unlab} used in the training, as well as results by BET and FreeSurfer. The values of the metrics at zero give the performance for the supervised method.

training scans. This lack of significance is caused by the low number of available unlabeled volumes, as confirmed by the second setup.

Figure 3 shows the results for the experiments on dataset B, where the number of annotated scans is two. The performance keeps growing until $N_{unlab} = 100$,

Table 2. Paired t-test comparing the performance metrics provided by the semi-supervised learning and the supervised version / BET for $N_{unlab} = 100$

Method	Error rate	Jaccard index	Mean distances	95% quantile
BET	$2.8{\cdot}10^{-6}$	$3.6{\cdot}10^{-6}$	$1.4{\cdot}10^{-15}$	$2.3{\cdot}10^{-5}$
Supervised	$8.3{\cdot}10^{-3}$	$4.2{\cdot}10^{-3}$	$1.2{\cdot}10^{-11}$	$2.4{\cdot}10^{-4}$

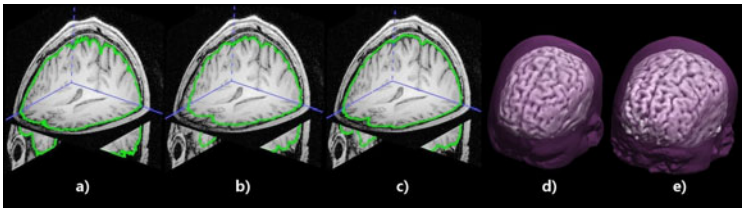

Fig. 4. Brain surfaces for a sample scan from dataset B: (a) ground truch, (b) BET, (c) semi-supervised. 3-D renderings for two scans: (d) dataset A, two labeled and nine unlabeled; (e) dataset B, two labeled and 50 unlabeled.

which, next to having a larger test data sample, provides a larger statistical significance for the improvement with respect to the supervised version (Table 2). The significance with respect to BET from the first experiment is preserved. Figure 4(a-c) shows three orthogonal slices of a fairly difficult scan from dataset B segmented with BET, the unsupervised algorithm (100 unlabeled examples) and the manual annotations. The output from BET is unacceptable in the frontal lobe, whereas the more robust semi-supervised version is still able to detect the correct boundary. Figure 4(d-e) displays renderings of two segmentations by the proposed semi-supervised algorithm, one from each dataset.

4 Discussion

A semi-supervised method for skull stripping which utilizes a small amount of labeled data has been presented in this paper. We take advantage of two factors to boost the performance of a classifier: the existence of expert systems and the abundance of unlabeled data. This situation is not uncommon in medical imaging, since large amounts of unlabeled scans and well-developed segmentation methods are often accessible. A key parameter of the system is the threshold of the distance transform at which unlabeled data are allowed to be sampled from. This depends on the performance of the expert systems and is justified in section 2.1: the success of a semi-supervised learning approach relies on the classification error by the other classifiers (experts in our case). If all the experts have high agreement but wrong predication, using unlabeled data may even degrade the performance. Future work related to this study includes: 1) testing the supervised algorithm on a large number of scans with more labeled examples; 2) designing more robust features that do not rely as much on the pure voxel intensities and

allow the algorithm to segment the brain precisely in cases where the gray levels are not very consistent with the training data; and 3) assessing the performance on scans from patients with pathology.

Acknowledgements

This work is funded by grants NSF IIS-0844566, N000140910099 and partly NIH U54 RR021813 and China863 2008AA01Z126. The authors would like to thank Cornelius Hojakshtani for the acquisition and annotation of dataset A, G. de Zubicaray, K. McMahon and M. Wright for the acquisition of dataset B, and M. Barysheva for the annotation of the dataset B. The first author would also like to thank the U.S. Department of State's Fulbright program for his funding.

References

1. Tu, Z., Narr, K., Dollar, P., Dinov, I., Thompson, P., Toga, A.: Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE T. Med. Imaging 27, 495–508 (2008)
2. Yi, Z., Criminisi, A., Shotton, J., Blake, A.: Discriminative, semantic segmentation of brain tissue in mr images. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 558–565. Springer, Heidelberg (2009)
3. Sabuncu, M., Yeo, B., Van Leemput, K., Fischl, B., Golland, P.: Supervised non-parametric image parcellation. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 1075–1083. Springer, Heidelberg (2009)
4. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, U. of Wisconsin-Madison (2005)
5. Bonwell, C., Eison, J.: Active learning: Creating excitement in the classroom. AEHE-ERIC Higher Education Report No.1 (1991) ISBN 1–87838–00–87
6. Zhu, X.: Semi-supervised learning with graphs. PhD thesis, CMU (2005)
7. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proc. COLT, pp. 92–100 (1998)
8. Smith, S.: Fast robust automated brain extraction. Hum. Brain Mapp. 17(3), 143–155 (2002)
9. Segonne, F., Dale, A., Busa, E., Glessner, M., Salat, D., Hahn, H., Fischl, B.: A hybrid approach to the skull stripping problem in MRI. Neuroimage 22(3)
10. Zhou, Z.H., Li, M.: Tri-training: Exploit unlabeled data using three classifiers. IEEE T. Knowl. Data En. 17(11), 1529–1541 (2005)
11. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. Proc. Info. Knowl. Manag., pp. 86–93 (2000)
12. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)
13. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proc. 23 Int. Conf. Mach. Learn., pp. 161–168. ACM, New York (2006)
14. Angluin, D., Laird, P.: Learning from noisy examples. Mach. Learn. 2(4)
15. Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. 7(1), 1–26 (1979)
16. Shattuck, D., Prasad, G., Mirza, M., Narr, K., Toga, A.: Online resource for validation of brain segmentation methods. NeuroImage 45(2), 431–439 (2009)