Guiding Automatic Segmentation with Multiple Manual Segmentations

Hongzhi Wang and Paul A. Yushkevich*

Department of Radiology, University of Pennsylvania

Abstract. Most image segmentation algorithms are designed to estimate a single segmentation for each image, where the gold standard segmentation is often labeled by a human expert. However, it is common that multiple manual segmentations are available for some images, e.g. independently labeled by different experts. For efficient usages of manual segmentations, we propose to simultaneously produce automatic estimations for each expert. The key advantage of this proposal is that it allows to incorporate the correlations between different experts to improve the accuracy of automatic segmentation. In a brain image segmentation problem, where for each image six manual segmentations are available, we show that jointly estimating several manual segmentations produces significant improvement over independently estimating each of them.

1 Introduction

Image segmentation is the primary mechanism for quantifying the properties of anatomical structures and pathological formations using imaging data. Given the often prohibitive cost of manual segmentation, accurate automatic segmentation is highly desirable. To mimic manual segmentation, automatic segmentation is often guided and evaluated against manual segmentations. However, segmentations labeled by different experts are often inconsistent.

Existing inconsistent manual segmentations not only reveals the significant difficulty in performing manual segmentation, but also poses challenges on how to develop automatic segmentation algorithms. Most automatic algorithms produce a single solution for each image. When evaluated against inconsistent manual segmentations, the automatic solution is either separately compared with each of the manual segmentations or directly compared with the consensus manual segmentation, e.g. derived by STAPLE [11]. Either way the automatic algorithm is biased to produce solutions close to the consensus of all manual segmentations.

Employing consensus manual segmentation simplifies the evaluation process, therefore makes the task of developing automatic methods more straightforward. However, it also sacrifices the rich information contained in the original set of manual segmentations. Our contribution is to propose a novel scheme to

* We thank John Pluta, Sylvia Orozco and Salmon Kadivar for producing the manual segmentations used in this study. This work was supported by the Penn-Pfizer Alliance grant 10295 and NIH awards K25 AG027785, R01 AG010897.

N. Ayache et al. (Eds.): MICCAI 2012, Part II, LNCS 7511, pp. 429–436, 2012.

incorporate multiple manual segmentations to guide automatic segmentation. To maximize the usage of the valuable manual segmentations, we propose to simultaneously produce automatic estimations for all manual segmentations. The key advantage of this proposal is that the label correlations between different human experts can be incorporated to improve automatic segmentation.

We apply our method to segment the hippocampus in magnetic resonance images (MRI) and show significant improvement over independently producing estimations for each manual segmentation segmentation.

2 Jointly Estimating Multiple Manual Segmentations

Image segmentation can be addressed via estimating the conditional probability $p(S_F|F)$, where F is an image and S_F is a segmentation for F. Assuming that labeling different voxels is conditionally independent given the image patches located on the voxels, we have $p(S_F|F) = \prod_i p(S_F(i)|F(\mathcal{N}(i)))$, where i indexes through image voxels. $\mathcal{N}(i)$ represents a neighborhood centered at i. $F(\mathcal{N}(i))$ is the intensity patch located on the region. To estimate this probability, discriminative learning techniques learn the label distribution $p(l|F(\mathcal{N}(i)))$ from training data, e.g. [7], [8], which can be addressed by most classification algorithms. l indexes through all possible labels. The segmentation is then obtained via maximum a posterior inference, i.e. $S_F(i) = \text{argmax}_l p(l|F(\mathcal{N}(i)))$.

Motivation for jointly estimating multiple manual segmentations. In the context of clinical imaging studies involving segmentation, it is common to generate repeat manual segmentations by multiple raters in order to establish inter-rater and intra-rater reliability for a manual segmentation protocol. To handle the inconsistency between multiple manual segmentations, one approach attempts to infer the "ground truth" segmentation with the consideration of the reliability of each rater [11]. However, the inferred "ground truth" loses the rich information in the original manual segmentations and the errors in deriving the hard decision of "ground truth" will affect the performance of automatic segmentation.

We advocate an alternative solution that produces a separate estimation for each manual segmentation[1]. The key advantage of this strategy is that it allows to incorporate the correlations between manual segmentations to improve the accuracy of automatic segmentation. In our experiment, we observed that some raters consistently produced larger volumes than others when segmenting the hippocampus in MRI (see section 3.1). With such correlations, observing the segmentation labeled by one rater provides meaningful information to estimate the segmentation labeled by the other. Even if only one most reliable manual segmentation is selected for one study, which is common is practice, as we show below, incorporating manual segmentations labeled by less reliable raters helps improving the automatic segmentation accuracy for the selected rater.

[1] A unique manual segmentation is defined as consistently labeled by one human expert in one segmentation trial.

Formulation of jointly estimating multiple manual segmentations. Jointly estimating multiple manual segmentations can be solved via estimating the following joint conditional probability $p(S_F^1, ..., S_F^m|F) \propto p(F|S_F^1, ..., S_F^m)p(S_F^1, ..., S_F^m)$

$$\propto \left[\prod_{j=1}^{m} p(F|S_F^j)\right] p(S_F^1, ..., S_F^m) \propto \left[\prod_{j=1}^{m} p(S_F^j|F)\right] p(S_F^1, ..., S_F^m) \quad (1)$$

where $S_F^1, ..., S_F^m$ estimate m manual segmentations, respectively. Given any manual segmentation for an image, we assume conditional independence between the image and any other manual segmentations for the image. The last equation is obtained by dropping the term $p(F)/p(S_F^j)$, where $p(S_F^j)$ is the prior for observing a segmentation labeled by rater j. Since it is hard to approximate this prior, we treat it as a constant and focus on optimizing the remaining terms. The first term in (1) can be estimated by separately applying discriminative learning to estimate each manual segmentation. The second term is the joint probability of observing all manual segmentations for one image, which captures their correlations. Estimating this term is difficult as well, but a good approximation can be obtained by applying pseudolikelihood [2]. We have:

$$p(S_F^1, ..., S_F^m|F) \propto \prod_{j=1}^{m} p(S_F^j|F)p(S_F^j|\{S_F^1, ..., S_F^m\}\backslash S_F^j) \quad (2)$$

In summary, each manual segmentation is estimated based on two constraints: 1) image information, which directly captures the correlation between the manual segmentation and an image; and 2) the segmentations estimated for the remaining manual segmentations, which enforces the estimated segmentations to respect the mutual correlations between different raters. As in (2), assuming assigning labels to different voxels are conditionally independent given the patches located on the voxels, we have the final approximation as $p(S_F^1, ..., S_F^m|F)$

$$\propto \prod_{j=1}^{m}\prod_{i} p(S_F^j(i)|F(\mathcal{N}(i)))p(S_F^j(i)|\{S_F^1(\mathcal{N}(i)), ..., S_F^m(\mathcal{N}(i))\}\backslash S_F^j(\mathcal{N}(i))) \quad (3)$$

2.1 Discriminative Learning

Here, we describe in detail how we estimate the conditional probabilities in (3).

Learning to approximate one manual segmentation. For each manual segmentation, to estimate $p(l|F(\mathcal{N}(i)))$, i.e. the first term in (3), we train one *segmentation classifier* using the modified AdaBoost algorithm [4],[9] for each label l to identify voxels assigned to label l in the target manual segmentation.

For better performance, we apply the *corrective learning* technique [8]. This method applies learning as an error correction tool to improving the segmentation produced by a host segmentation method. It was shown that it significantly improved the performance of the learning algorithm and the host segmentation method [8]. In our experiments, we apply multi-atlas label fusion as the host method (see detail in section 3). Note that multi-atlas segmentation can be applied alone to estimate each manual segmentation. Applying corrective learning improves the performance. The region of interest and the features used in [8], including spatial, appearance and contextual, joint spatial-appearance and joint spatial-contextual features, are applied to train the classifiers, where the contextual features are extracted from the initial segmentation produced by the host method. To transfer the output of an AdaBoost classifier to a probability, we apply the logistic transform, i.e. $p(x) = \frac{e^x}{e^x+e^{-x}}$.

Learning the correlations between manual segmentations. To estimate the second term in (3), we train *correlation classifiers* for each manual segmentation to capture the correlation between this manual segmentation and the remaining manual segmentations. For this task, we apply spatial, contextual and joint spatial-contextual features, as in [8], to train one classifier for each label l to identify the voxels assigned to label l in the target manual segmentation. The contextual features are extracted from all the remaining manual segmentations. To effectively handle the contextual features provided by multiple manual segmentations, we merge the contextual features from different manual segmentations into one label distribution D_l^j for each label l, $D_l^j(i) = \frac{1}{m-1}\sum_{k \neq j} I(S_F^k(i) = l)$, where $I(\cdot)$ is an indicator function. The contextual features used in the correlation classifier for rater j and label l are constructed based on D_l^j.

Segmentation Algorithm. The algorithm is summarized below:

1. Given a test image F, apply a host method to produce an initial segmentation S for it. When applicable, produce one initial segmentation for each rater.
2. For $j = 1, ..., m$
   - Apply the segmentation classifier(s) learned for j^{th} manual segmentation to produce an improved estimation, S_F^j, based on image F and the initial segmentation S produced for the rater.
3. For $j = 1, ..., m$
   - Apply the correlation classifier(s) learned for j^{th} manual segmentation to update S_F^j such that (3) is maximized, i.e. selecting the label with the largest probability produced by the classifiers at each voxel.
4. If none of the automatic estimations is changed or the maximal iteration is reached, then output the estimations. Otherwise, goto 3.

3 Experiments

3.1 Imaging Data and Experiment Setup

Data and manual segmentations. For 10 images (5 controls and 5 patients with mild AD) from the Open Access Series of Imaging Studies (OASIS) [5], we produced six manual hippocampal segmentations for each image. These manual segmentations were labeled by three trained experts in two trials.

Table 1. Left: Inter-rater and intra-rater segmentation overlaps (in Dice $\frac{2|A \cap B|}{|A|+|B|}$) between the three raters. The intra-rater overlaps are between the two segmentation trails labeled by the same rater. The inter-rater overlaps are averaged over the two segmentation trials. Right: Hippocampal volume (in voxel) produced by each rater.

raters	R^1	R^2	R^3
R^1	0.902±0.020	0.872±0.024	0.847±0.032
R^2	0.872±0.024	0.915±0.020	0.846±0.043
R^3	0.847±0.032	0.846±0.043	0.836±0.046

raters	R^1	R^2	R^3
trial 1	1615±267	1787±316	1731±319
trial 2	1683±285	1811±261	1461±232

Table 1(left) summarizes the inter-rater and intra-rater reproducibility of the manual segmentation. Table 1(right) shows the hippocampal volume labeled by each rater. Note that R^2 consistently labeled larger hippocampi than R^1 in both trials. The segmentations labeled by R^3 in the second trial are significantly smaller than those produced by the same rater in the first trial. Such strong correlation can be easily seen in most individual subjects, as shown in Fig. 1.

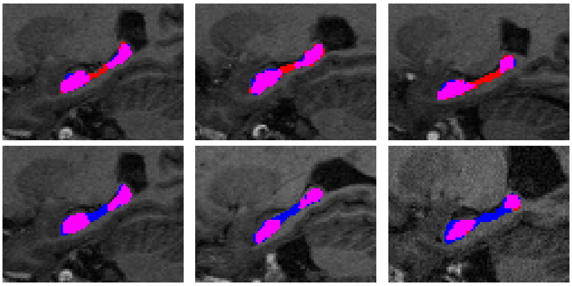

Fig. 1. Illustration of correlation between raters. First row: segmentations produced by R^1(blue) and R^2(red) in the first trial. Pink is the overlapped region. Second row: segmentations produced by R^3 in the first (blue) and second trial (red).

For each image, we derived one consensus segmentation using STAPLE [11] for the three manual segmentations produced in each trial. To incorporate the correlations between the two trials, we jointly estimate the 6 manual segmentations and the two inferred segmentations by STAPLE.

Experiment setup. For cross-validation, we randomly selected five images for training and the remaining 5 images for testing. The experiment was repeated 10 times. In each cross-validation, a different set of training and testing images were selected. The results reported below are averaged over the 10 experiments.

Details on learning segmentation classifiers. Since the state-of-the-art hippocampus segmentation are all produced by multi-atlas label fusion (MALF), e.g. [6],[3],[10], we applied MALF as the host segmentation method to produce the

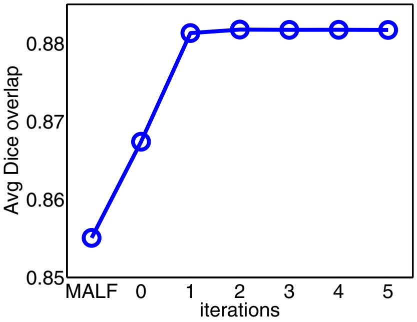

Fig. 2. Segmentation accuracy (in Dice) at each iteration for all raters. The results are averaged over 10 cross-validation experiments. The performance of independently applying error correction to estimate each manual segmentation is given at iteration 0.

initial hippocampus segmentation for corrective learning to learn the segmentation classifier for each manual segmentation.

Through deformable registration, MALF warps multiple atlases, i.e. pre-labeled images, to a target image, and uses a “label fusion” strategy to derive a consensus segmentation. To implement MALF, image guided registration is performed by the Symmetric Normalization (SyN) algorithm implemented by ANTS [1] between each pair of the atlas image, i.e. the training image, and the test image. For label fusion, we apply image similarity based local weighted voting technique, which is shown to be the most effective label fusion techniques in recent studies [6],[10]. The voting weights were computed based on image patches of size $5 \times 5 \times 5$ by using the joint label fusion algorithm [10].

To produce the initial segmentation used in corrective learning for one manual segmentation, we use the segmentation labeled by the corresponding rater to define the atlas. For each cross-validation, we also apply MALF to produce an initial segmentation for each training image by using the remaining training images as atlases and use the segmentation produced by MALF for training images to train the segmentation classifiers for each manual segmentation. For each cross validation, learning all segmentation classifiers and all correlation classifiers took about 1 hour and 30 minutes on a 2GHz CPU, respectively.

3.2 Results

Convergence. Fig. 2 shows the average segmentation performance produced by MALF, MALF + corrective learning (iteration 0), and our joint segmentation algorithm at different iterations. Typically, the iterative optimization converges within only a few iterations, with the first iteration producing the maximal performance improvement and dramatic diminishing performance gains in later iterations. In our experiment, we set the maximal iteration to be 10.

Quantitative comparison. Table 2 compares the performance between our method with separately estimating each manual segmentation. Corrective learning substantially improved the accuracy produced by MALF. Our method further improved the accuracy to the level greater than inter-rater accuracy for each rater. The improvements for each rater are statistically significant, with $p < 0.001$ on the paired Students t-test. Fig. 3 shows some segmentation results produced by applying corrective learning alone and by our method, respectively.

Table 2. Segmentation accuracy (in Dice) with respect to each manual segmentation. R^j_k is the segmentation produced by rater R^j in the k_{th} segmentation trial. STP_k is the consensus segmentation produced by STAPLE for the k_{th} segmentation trial.

rater	MALF	MALF+learning	MALF+jointSeg
R^1_1	0.865±0.021	0.878±0.014	**0.890**±0.015
R^2_1	0.852±0.020	0.869±0.018	**0.881**±0.018
R^3_1	0.833±0.032	0.837±0.025	**0.859**±0.022
STP_1	0.869±0.023	0.888±0.016	**0.900**±0.014
R^1_2	0.859±0.023	0.864±0.018	**0.880**±0.021
R^2_2	0.861±0.018	0.877±0.017	**0.887**±0.016
R^3_2	0.829±0.035	0.840±0.024	**0.857**±0.023
STP_2	0.871±0.022	0.886±0.016	**0.900**±0.017

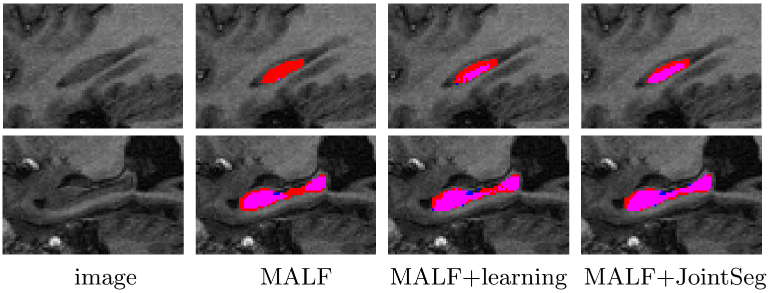

Fig. 3. Sagittal views of hippocampus segmentation results. Red: one of the manual segmentations for the image; Blue: automatic segmentation; Pink: overlap between manual and automatic.

Our results compare well to the state-of-the-art hippocampus segmentation performance. For example, [6] reported average ∼0.87 (Dice) for hippocampus using 29 atlases. [3] reported average 0.887 (Dice) using 79 atlases. Our final results for R^1 and R^2 are >0.880 (Dice)[2], but we only used 5 training images, which is only a small fraction of those used by the competing work.

4 Conclusion

As an important evaluation target, manual segmentation is crucial in the development of automatic segmentation algorithms. We developed a technique to

[2] Our results for R^3 are lower due to the poor intra-rater performance.

incorporate multiple inconsistent manual segmentations to improve the performance of automatic segmentation. Via experiments on hippocampus segmentation in MRI, we showed the advantage of our method over traditional approaches. Note that including the segmentations produced by less reliable raters helped to better estimate the segmentations by more reliable raters. Our work offers a new perspective on how to more effectively use the valuable manual segmentations.

References

1. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12(1), 26–41 (2008)
2. Besag, J.: Statistical analysis of non-lattice data. J. R. Statist. Soc. B 24(3), 179–195 (1975)
3. Collins, D., Pruessner, J.: Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. NeuroImage 52(4), 1355–1366 (2010)
4. Freund, Y., Schapire, R.: A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–27. Springer, Heidelberg (1995)
5. Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. Journal of Cognitive Neuroscience 22(12), 2677–2684 (2010)
6. Sabuncu, M., Yeo, B., Leemput, K.V., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE TMI 29(10), 1714–1720 (2010)
7. Tu, Z., Zheng, S., Yuille, A., Reiss, A., Dutton, R., Lee, A., Galaburda, A., Dinov, I., Thompson, P., Toga, A.: Automated extraction of the cortical sulci based on a supervised learning approach. IEEE TMI 26(4), 541–552 (2007)
8. Wang, H., Das, S., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.: A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. NeuroImage 55(3), 968–985 (2011)
9. Wang, H., Suh, J.W., Das, S., Pluta, J., Altinay, M., Yushkevich, P.: Hippocampus segmentation using a stable maximum likelihood classifier ensemble algorithm. In: 2011 IEEE International Symposium on Proceeding of: Biomedical Imaging: From Nano to Macro (2011)
10. Wang, H., Suh, J., Pluta, J., Altinay, M., Yushkevich, P.: Optimal Weights for Multi-atlas Label Fusion. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 73–84. Springer, Heidelberg (2011)
11. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE TMI 23(7), 903–921 (2004)