An Expectation Maximization-Like Algorithm for Multi-Atlas Multi-Label Segmentation

Torsten Rohlfing, Daniel B. Russakoff and Calvin R. Maurer, Jr.

Image Guidance Laboratories, Department of Neurosurgery, Stanford University,
Stanford, CA, USA

Abstract. We present in this paper a novel interpretation of the concept of an "expert" in image segmentation as the pairing of an atlas image and a non-rigid registration algorithm. We introduce an extension to a recently presented expectation maximization (EM) algorithm for ground truth recovery, which allows us to integrate the segmentations obtained from multiple experts (i.e., from multiple atlases and/or using multiple image registration algorithms) and combine them into a final segmentation. In a validation study with randomly deformed segmentations we demonstrate the superiority of our method over simple label voting.

1 Introduction

Segmentation by non-rigid registration to an atlas image is an established method for labeling of biomedical images [1]. We have recently demonstrated [3] that the choice of the atlas image has a big influence on the quality of the segmentation. Moreover, we demonstrated that by using multiple atlases the segmentation accuracy can be improved over approaches that use a single individual or even an average atlas.

As Warfield *et al.* [5] were able to show for binary segmentations (foreground vs. background), combining multiple expert segmentations by majority-based consensus methods does not in general produce the best results. Instead, they describe an expectation maximization (EM) algorithm that iteratively estimates each expert's quality parameters, i.e., sensitivity and specificity. The final segmentation is then computed with these parameters taken into account by weighting the decisions made by a reliable expert higher than ones made by a less reliable one.

We present in this paper an extension of the Warfield method to an arbitrary number of labels. Also, we propose a new interpretation of the term "expert" as the pairing of a non-rigid registration method with an individual atlas. Just as different human experts generate different segmentations, so do different registration methods using the same atlas, or the same registration method using different atlases. Regardless of whether one or the other applies, we can utilize our method to automatically distinguish good from bad, that is accurate from inaccurate, segmentations and incorporate this knowledge into the segmentation outcome.

2 Notation and Algorithm

Let $\mathcal{L} = \{0, \ldots, L\}$ be the set of (numerical) labels in the segmentation. Each element in $\mathcal{L}$ represents a different anatomical structure. Every voxel in a segmented image is assigned exactly one of the elements of $\mathcal{L}$ (i.e., we disregard partial volume effects), which defines the anatomical structure that this voxel is part of. For every voxel i, let $T(i) \in \mathcal{L}$ be the unknown ground truth, i.e., the a priori correct labeling. We assume that the prior probability $g(T(i) = \ell)$ of the ground truth segmentation of voxel i being ℓ is uniform (independent of i). During the course of the EM algorithm, we estimate weights $W(i, \ell)$ as the current estimate of the probability that the ground truth for voxel i is ℓ, i.e., $W(i, \ell) = P(T(i) = \ell)$.

Given segmentations by K experts, we denote by $D_k(i)$ the decision of "expert"[1] k for voxel i, i.e., the anatomical structure that, according to this expert, voxel i is part of. Each expert's segmentation quality, separated by anatomical structures, is represented by a $L + 1 \times L + 1$ matrix of coefficients λ. For expert k, we define

$$\lambda_k(m, \ell) := P(T(i) = \ell | D_k(i) = m), \tag{1}$$

i.e. the conditional probability that if the expert classifies voxel i as part of structure m, it is in fact part of structure ℓ. The diagonal entries ($\ell = m$) represent the *sensitivity* of the respective expert when segmenting structures of label ℓ, i.e.,

$$p_\ell^{(k)} = \lambda_k(\ell, \ell). \tag{2}$$

The off-diagonal elements quantify the crosstalk between the structures, i.e., the likelihoods that the respective expert will misclassify one voxel of a given structure as belonging to a certain different structure. The *specificity* of expert k for structure ℓ is easily computed as

$$q_\ell^{(k)} = \sum_{m \neq \ell} \lambda_k(m, \ell). \tag{3}$$

Estimation Step. In the "E" step of our EM-like algorithm, the (usually unknown) ground truth segmentation is estimated. Given the current estimate for λ, and given the known expert decisions D, the probability of voxel i having label ℓ is

$$W(i, \ell) = \frac{g(T(i) = \ell) \prod_k \lambda_k(D_k(i), \ell)}{\sum_j g(T(i) = j) \prod_k \lambda_k(D_k(i), j)}. \tag{4}$$

Maximization Step. The "M" step of our algorithm estimates the expert parameters λ to maximize the likelihood of the current ground truth estimate determined in the preceding "E" step. Given that previous ground truth estimate g,

[1] Note that in the context of the present paper, we use the term "expert" for the combination of a non-rigid registration algorithm with an atlas image.

the new estimates for the expert parameters are computed as follows:

$$\hat{\lambda}_k(\ell, m) = \frac{\sum_{i:D_k(i)=\ell} W(i,m)}{\sum_i W(i,m)}. \tag{5}$$

Obviously, since there is *some* label assigned to each voxel by each expert, the sum over all possible decisions is unity for each expert, i.e.,

$$\sum_l \hat{\lambda}_k(\ell, m) = \frac{\sum_l \sum_{i:D_k(i)=l} W(i,m)}{\sum_i W(i,m)} = \frac{\sum_i W(i,m)}{\sum_i W(i,m)} = 1. \tag{6}$$

Incremental Computation. We note that for the computation of the next iteration's expert parameters λ, we only need to know the *sums* of all weights W for all voxels as well as for the subsets of voxels for each expert that are labeled the same by that expert. In other words, only the values $W(i,j)$ for one fixed i and all j are needed at any given time. The whole field $W(i,j)$ *need not be present* at any time, thus relieving the algorithm from having to store an array of $N \cdot L$ floating point values. The weights W from Eq. (4) can instead be recursively substituted into Eq. (5), resulting in the incremental formula

$$\hat{\lambda}_k(\ell, m) = \frac{\sum_{i:D_k(i)=\ell} \prod_{k'} \lambda_{k'}(D_{k'}(i), m)}{\sum_i \prod_{k'} \lambda_{k'}(D_{k'}(i), m)}. \tag{7}$$

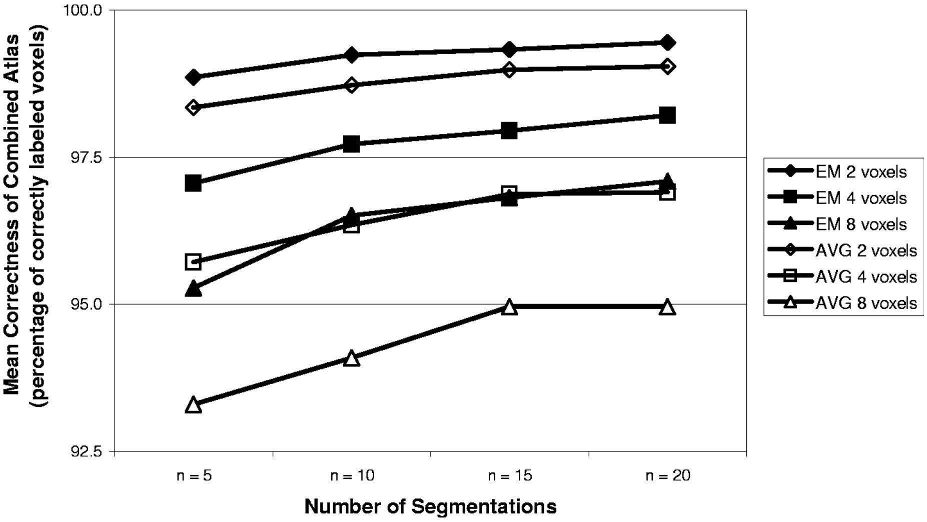
Domain Restriction. Mostly in order to speed up computation, but also as a means of eliminating image background, we restrict the algorithm to those voxels in the combined atlas for which at least one expert segmentation disagrees with the others. In other words, where all experts agree on the labeling of a voxel, that voxel is assigned the respected label and will not be considered during the algorithm.

3 Validation Study

We quantify the improvements of our algorithm over label averaging in a validation study. Three-dimensional biomedical atlases from 20 individuals [2] provide known ground truths. Simulated segmentations are generated by applying random deformations of varying magnitudes to the original atlases. For each ground truth, random B-spline-based free-form deformations [4] were generated by adding independent Gaussian-distributed random numbers to the coordinates of all control points. The variances of the Gaussian distributions corresponded to 2, 4, and 8 voxels. A total of 20 random deformations were generated for each individual and each σ.

The randomly deformed atlases were combined into a final atlas once by label voting, and once using our novel EM-like algorithm. Label voting simply counts for each voxel the number of atlases that assign a given label to that voxel. The label with most votes is assigned to the voxel in the final atlas.

Fig. 1. Mean correctness of combined segmentation over 20 individuals vs. number of random segmentations used. Results are shown for label voting (AVG) and EM algorithm, each applied to atlases after random deformations of magnitudes $\sigma = 10, 20, 30\,\mu\text{m}$.



4 Results

As a measure of segmentation quality, we compared the generated segmentation to the original atlas and computed the percentage of correctly labeled foreground voxels. Figure 1 shows a plot of the mean correctness over all 20 individuals versus the number of segmentations. The EM algorithm performed consistently better, i.e., produced more accurate combined segmentations, than simple label voting. The improvement achieved using the EM algorithm is larger for greater magnitudes of the random atlas deformations.

5 Discussion

This paper has introduced a novel method for combining multiple segmentations into one final segmentation. It can be used for example to combine segmentations generated using non-rigid registration with a population of atlas images. Our method is an extension of an algorithm described by Warfield *et al.* [5]. The equivalence of both techniques for binary segmentation ($\mathcal{L} = \{0, 1\}$) is easily proved by induction over the iterations of the algorithm.

Using a validation study with random segmentations and known ground truth we were able to demonstrate the superiority of our algorithm over simple label voting. Our algorithm particularly outperforms label voting for large variations in the input segmentations, in our case corresponding to large magnitudes of the random atlas deformations.

It is worth noting that, while seemingly similar, the situation we address with the validation study in this paper is fundamentally different from validation of non-rigid registration. A promising approach to validating non-rigid image registration is by simulating a known deformation using a biomechanical model. The simulated deformation is taken as the ground truth against which transformations computed using non-rigid registration can be validated. In that context it is vitally important that the simulated deformation be based on a different transformation model than the registration, e.g., a B-spline-based registration must not be validated using simulated B-spline deformations.

In our context, however, the opposite is true: in this paper, we have validated methods for combining different automatic segmentations generated by non-rigid registration. In this framework it makes sense (and is in fact necessary to correctly model the problem at hand) that the randomly deformed segmentations are generated by applying transformations from the class used by the registration algorithm. Only in this way can we expect to look at variations in the segmentations comparable to the ones resulting from imperfect non-rigid registration.

Acknowledgments

TR was supported by the National Science Foundation under Grant No. EIA-0104114. The authors gratefully acknowledge support for this research provided by CBYON, Inc., Mountain View, CA.

References

1. BM Dawant, SL Hartmann, JP Thirion, F Maes, D Vandermeulen, P Demaerel. Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations: Part I, methodology and validation on normal subjects. *IEEE Trans Med Imag*, 18(10):909–916, 1999.
2. T Rohlfing, R Brandt, CR Maurer, Jr, R Menzel. Bee brains, B-splines and computational democracy: Generating an average shape atlas. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 187–194, IEEE Computer Society, Los Alamitos, CA, 2001.
3. T Rohlfing, R Brandt, R Menzel, CR Maurer, Jr Segmentation of three-dimensional images using non-rigid registration: Methods and validation with application to confocal microscopy images of bee brains. In *Medical Imaging: Image Processing*, Proceedings of SPIE, 2003.
4. TW Sederberg, SR Parry. Free-form deformation and solid geometric models. *Comput Graph (ACM)*, 20(4):151–160, 1986.
5. SK Warfield, KH Zou, WM Wells. Validation of image segmentation and expert quality with an expectation-maximization algorithm. In *Proceedings of Fifth International Conference on Medical Image Computing and Computer-Assisted Intervention, Part I*, vol. 2488 of *LNCS*, pp. 298–306, Springer-Verlag, Berlin, 2002.