Expectation Maximization Strategies for Multi-atlas Multi-label Segmentation

Torsten Rohlfing[1], Daniel B. Russakoff[1,2], and Calvin R. Maurer[1]

[1] Image Guidance Laboratories, Department of Neurosurgery,
[2] Department of Computer Science,
Stanford University, Stanford, CA, USA
{rohlfing,dbrussak}@stanford.edu, calvin.maurer@igl.stanford.edu

Abstract. It is well-known in the pattern recognition community that the accuracy of classifications obtained by combining decisions made by independent classifiers can be substantially higher that the accuracy of the individual classifiers. In order to combine multiple segmentations we introduce two extensions to an expectation maximization (EM) algorithm for ground truth estimation based on multiple experts (Warfield *et al.*, MICCAI 2002). The first method repeatedly applies the Warfield algorithm with a subsequent integration step. The second method is a multi-label extension of the Warfield algorithm. Both extensions integrate multiple segmentations into one that is closer to the unknown ground truth than the individual segmentations. In atlas-based image segmentation, multiple classifiers arise naturally by applying different registration methods to the same atlas, or the same registration method to different atlases, or both. We perform a validation study designed to quantify the success of classifier combination methods in atlas-based segmentation. By applying random deformations, a given ground truth atlas is transformed into multiple segmentations that could result from imperfect registrations of an image to multiple atlas images. We demonstrate that a segmentation produced by combining multiple individual registration-based segmentations is more accurate for the two EM methods we propose than for simple label averaging.

1 Introduction

One way to automatically segment an image is to perform a non-rigid registration of the image to a labeled atlas image; the labels associated with the atlas image are mapped to the image being segmented using the resulting non-rigid transformation [1]. This approach has two important components that determine the quality of the segmentations, namely the registration method and the atlas. Just as human experts typically differ slightly in their labeling decisions, different registration methods produce different segmentations when applied to the same raw image and the same atlas. Likewise, different segmentations typically result from using different atlases. Therefore, each combination of a registration algorithm with an atlas effectively represents a unique classifier for the voxels in the target image.

C.J. Taylor and J.A. Noble (Eds.): IPMI 2003, LNCS 2732, pp. 210–221, 2003.

The atlas can be an image of an individual or an average image of multiple individuals. Our group recently showed [2] that the choice of the atlas image has a substantial influence on the quality of a registration-based segmentation. Moreover, we demonstrated that by using multiple atlases, the segmentation accuracy can be improved over using a single atlas (either an image of an individual or an average of multiple individuals). Specifically we showed that a segmentation produced by combining multiple individual segmentations is more accurate than the individual segmentations.[1] This finding is consistent with the observation that a combination of classifiers is generally more accurate than an individual classifier in many pattern recognition applications.

Typically among the individual segmentations there are more accurate ones as well as less accurate ones. This is true for human experts, due to different levels of experience, as well as for automatic classifiers, due, for example, to differences in similarities between the image to be segmented and different atlases. In this paper we present and evaluate methods that automatically estimate the classifiers' segmentation qualities and take these into account when combining the individual segmentations into a final segmentation. For binary segmentations (object vs. background), Warfield *et al.* [3] recently introduced an expectation maximization (EM) algorithm that derives estimates of segmentation quality parameters (sensitivity and specificity) from segmentations of the same image performed by several experts. Their method also enables the generation of an estimate of the unknown ground truth segmentation. This ground truth estimate can provide a way of defining a combined segmentation that takes into account all experts, weighted by their individual reliability. We introduce two extensions of the Warfield method to non-binary segmentations with arbitrary numbers of labels. We also perform an evaluation study to quantitatively compare different methods of combining multiple segmentations into one. Our study is specifically designed to model situations where the segmentations are generated by non-rigid registration of an image to atlas images.

2 Binary Multi-expert Segmentation

This section briefly reviews the Warfield algorithm [3] and introduces the fundamental notation. Our notation differs slightly from that used by the original authors in order to simplify notation for the multi-label extension proposed below.

In binary segmentation, every voxel in a segmented image is assigned either 0 or 1, denoting background and object, respectively. For any voxel i, let $T(i) \in \{0, 1\}$ be the unknown ground truth, i.e., the *a priori* correct labeling. It is assumed that the prior probability $g(T(i) = 1)$ of the ground truth segmentation of voxel i being 1 is uniform (independent of i). During the course of the EM

[1] Each individual registration was produced by non-rigid registration of an image to a different atlas that is a labeled image of a reference individual. The combination was performed by simple label averaging.

algorithm, weights $W(i)$ are estimated, which denote the likelihood that the ground truth for voxel i is 1, i.e., $W(i) = P(T(i) = 1)$.

Given segmentations by K experts, we denote by $D_k(i)$ the decision of "expert"[2] k for voxel i, i.e., the binary value indicating whether voxel i has been identified as object voxel by expert k. Each expert's segmentation quality is represented by values p_k and q_k. While p_k denotes the likelihood that expert k identifies an *a priori* object voxel as such (*sensitivity*), q_k is the likelihood that the expert correctly identifies a background voxel (*specificity*).

2.1 Estimation Step

Given estimates of the sensitivity and specificity parameters for each expert, the weights for all voxels i are calculated as

$$W(i) = \frac{g(T(i)=1)\alpha}{g(T(i)=1)\alpha + (1-g(T(i)=0))\beta} \tag{1}$$

where

$$\alpha = \Bigg(\prod_{k:D_k(i)=1} p_k\Bigg)\Bigg(\prod_{k:D_k(i)=0} (1-p_k)\Bigg) \text{ and } \beta = \Bigg(\prod_{k:D_k(i)=0} q_k\Bigg)\Bigg(\prod_{k:D_k(i)=1} (1-q_k)\Bigg). \tag{2}$$

2.2 Maximization Step

From the previously calculated weights W, the new estimates $\hat{p}_k$ and $\hat{q}_k$ for each expert's parameters are calculated as follows:

$$\hat{p}_k = \frac{\sum_{i:D_k(i)=1} W(i)}{\sum_i W(i)} \quad \text{and} \quad \hat{q}_k = \frac{\sum_{i:D_k(i)=0}(1-W(i))}{\sum_i (1-W(i))}. \tag{3}$$

2.3 Application to Multi-label Segmentation

An obvious way to apply Warfield's algorithm (described above) to multi-label segmentation is to apply it repeatedly and separately for each label. In each run, one of the labels is considered as the object in the sense of the algorithm. This strategy, however, may lead to inconsistent results, i.e., some voxels can be assigned multiple labels (in other words, voxels can be classified as object voxels in more than one run of the algorithm). To address this issue, we propose to combine the results of all runs as follows: each application of the algorithm provides sensitivity and specificity estimates for all experts for one label (the label that is considered the object of interest in this run of the algorithm). These values are used to compute the weights $W(i)$ according to Eq. (1) separately for

[2] 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. However, the framework we propose is also appropriate for human experts or any other kind of classifier.

each label. The voxel i is then assigned the label that has the highest weight W. One could instead use the weights W calculated during the last EM iteration for each label, but this requires storing all weights. It is more memory efficient and only slightly more computationally expensive to compute the weights once more after all EM iterations have been completed.

3 Multi-label Multi-expert Segmentation

This section describes a multi-label extension to Warfield's EM algorithm that simultaneously estimates the expert parameters for all labels. This extension contains Warfield's algorithm as a special case for one label ($\mathcal{L} = \{0, 1\}$). This is easily proved by induction over the iterations of the algorithm.

For a multi-label segmentation let $\mathcal{L} = \{0, \dots, L\}$ be the set of (numerical) labels in the atlas. 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 $\ell \in \mathcal{L}$ is uniform (independent of i). During the course of the 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" 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

$$\lambda_k(m, \ell) := P(T(i) = \ell \mid D_k(i) = m), \tag{4}$$

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. We note that this matrix is very similar to the normalized confusion matrix of a Bayesian classifier [9]. The diagonal entries or our matrix ($\ell = m$) represent the *sensitivity* of the respective expert when segmenting structures of label ℓ, i.e.,

$$p_k^{(\ell)} = \lambda_k(\ell, \ell). \tag{5}$$

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 computed as

$$q_k^{(\ell)} = 1 - \sum_{m \neq \ell} \lambda_k(m, \ell). \tag{6}$$

3.1 Estimation Step

In the "E" step of our EM algorithm, the (usually unknown) ground truth segmentation is estimated. Given the current estimate for λ and the known expert decisions D, the likelihood of the ground truth for voxel i being label ℓ is

$$W(i,\ell) = \frac{g(T(i)=\ell)\prod_k \lambda_k(D_k(i),\ell)}{\sum_m \left[g(T(i)=m)\prod_k \lambda_k(D_k(i),m)\right]}. \tag{7}$$

The likelihoods W for each voxel i are normalized and, over all labels, add up to unity:

$$\sum_\ell W(i,\ell) = \sum_\ell \frac{g(T(i)=\ell)\prod_k \lambda_k(D_k(i),\ell)}{\left[\sum_m g(T(i)=m)\prod_k \lambda_k(D_k(i),m)\right]} \tag{8}$$

$$= \frac{\sum_\ell \left[g(T(i)=\ell)\prod_k \lambda_k(D_k(i),\ell)\right]}{\sum_m \left[g(T(i)=m)\prod_k \lambda_k(D_k(i),m)\right]} = 1. \tag{9}$$

3.2 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 the previous ground truth estimate g, the new estimates for the expert parameters are computed as follows:

$$\hat{\lambda}_k(\ell,m) = \frac{\sum_{i:D_k(i)=\ell} W(i,m)}{\sum_i W(i,m)}. \tag{10}$$

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.,

$$\sum_\ell \hat{\lambda}_k(\ell,m) = \frac{\sum_\ell \sum_{i:D_k(i)=\ell} W(i,m)}{\sum_i W(i,m)} = \frac{\sum_i W(i,m)}{\sum_i W(i,m)} = 1. \tag{11}$$

The proof that the update rule in Eq. (10) indeed maximizes the likelihood of the current weights W is tedious, but largely analogous to the proof in the binary case (see Ref. [3]).

4 Implementation

Incremental Computation. Warfield *et al.* state in their original work [3] that for each voxel they store the weight W, which expresses the current confidence estimate for that voxel being an object voxel. When considering 3-D instead of 2-D images, however, the memory required to store the (real-valued) weights W for each voxel becomes a problem. For the multi-label algorithm introduced in Section 3, the situation is even worse, since it would require storing as many weights per voxel as there are labels in the segmentation. Fortunately, it is possible to perform the EM iteration without storing the weights, instead propagating the expert parameters estimated in the M-step of the previous iteration directly to the M-step of the next iteration.

Inspection of Eq. (3) for the binary algorithm and Eq. (10) for the multi-label algorithm reveals that the computation of the next iteration's expert parameters requires only 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, the value $W(i)$ (the values $W(i,j)$ for all j in the multi-label case) is needed only for one fixed i at any given time. The whole field $W(i)$ ($W(i,j)$ in the multi-label case) need not be present at any time, thus relieving the algorithm from having to store an array of N floating point values ($N \cdot L$ in the multi-label case). The weights W from Eq. (1) can instead be recursively substituted into Eq. (3), resulting in the incremental formulas

$$\hat{p}_k = \frac{\sum_{i:D_k(i)=1} \frac{g(T(i)=1)\alpha}{g(T(i)=1)\alpha+(1-g(T(i)=0))\beta}}{\sum_i \frac{g(T(i)=1)\alpha}{g(T(i)=1)\alpha+(1-g(T(i)=0))\beta}}, \tag{12}$$

$$\hat{q}_k = \frac{\sum_{i:D_k(i)=0}(1-\frac{g(T(i)=1)\alpha}{g(T(i)=1)\alpha+(1-g(T(i)=0))\beta})}{\sum_i(1-\frac{g(T(i)=1)\alpha}{g(T(i)=1)\alpha+(1-g(T(i)=0))\beta})}, \tag{13}$$

where α and β are defined as in Eq. (2) and depend *only* on the parameters p and q from the previous iteration and the (invariant) expert decisions. Analogously, in the multi-label case the weights W from Eq. (7) can be recursively substituted into Eq. (10), resulting in the incremental formula

$$\hat{\lambda}_k(\ell,m) = \frac{\sum_{i:D_k(i)=\ell} \prod_{k'} \lambda_{k'}(D_{k'}(i),m)}{\sum_i \prod_{k'} \lambda_{k'}(D_{k'}(i),m)}. \tag{14}$$

Restriction to Disputed Voxels. Consider Eqs. (1) and (7) and let us assume that for some voxel i, all experts have made the same labeling decision and assigned a label ℓ. Let us further assume that the reliability of all experts for the assigned label is better than 50%, i.e., $p_k > 0.5$ for all k during the ℓ-application of the repeated binary method, or $\lambda_k(\ell,\ell) > 0.5$ in the multi-label method. It is then easy to see that voxel i will *always* be assigned label ℓ. We refer to such voxels as *undisputed*. Conversely, we refer to all voxels where at least one expert disagrees with the others as *disputed*.

Mostly in order to speed up computation, but also as a means of eliminating image background, we restrict the algorithm to the disputed voxels. In other words, where all experts agree on the labeling of a voxel, that voxel is assigned the respective label and is not considered during the iterative optimization procedure. In addition to the obvious performance benefit, it is our experience that this restriction actually improves the quality of the segmentation outcome. To understand this phenomenon, consider application of the binary EM algorithm to an image with a total of N voxels that contains a structure n voxels large. Take an expert who correctly labeled the n foreground voxels, but mistakenly labeled m additional background voxels as foreground. This expert's specificity is therefore $q = \frac{(N-n)-m}{N-n}$. By increasing the field of view, thus adding peripheral background voxels, we can increase N arbitrarily. As N approaches infinity, q approaches 1, regardless of m. Therefore, we lose the ability to distinguish between specific and unspecific experts as the amount of image background increases. Due to limited floating-point accuracy this is a very real danger, and it explains why, in our experience, it is beneficial to limit consideration to disputed voxels only.

5 Volume-Weighted Label Averaging

As a reference method for the two EM algorithms above, a non-iterative label averaging algorithm is implemented. The fundamental function of this method is to assign to each voxel in the final segmentation the label that was assigned to this voxel by the (relative) majority vote of the experts [4]. However, the situation we are interested in is slightly different. Instead of presenting an image to a human expert, each expert in our context is merely a non-rigid coordinate transformation from an image into an atlas. Since the transformation is continuous, while the atlas is discrete, more than one voxel in the atlas may contribute to the labeling of each image voxel. The contributing atlas voxels can (and will near object boundaries) have different labels assigned to them.

The simplest way to address this situation is to employ nearest-neighbor interpolation. However, it is our experience that it is a better idea to use Partial Volume Integration (PVI) as introduced by Maes *et al.* [5] in order to properly consider fractional contributions of differently labeled voxels. For a quick review of PVI, consider a voxel i to be segmented. From each of the k expert segmentations, looking up the label for this voxel under some coordinate transformation yields an 8-tuple of labels ℓ from a $2 \times 2 \times 2$ neighborhood of voxels in the atlas, numbered 0 through 7. Each voxel is also assigned a weight w based on its distance from the continuous position described by the non-rigid image-to-atlas coordinate mapping. Therefore, each expert segmentation for each voxel produces an 8-tuple $X_k(i)$ of label-weight pairs:

$$X_k(i) = ((w_k^{(0)}, \ell_k^{(0)}), \dots, (w_k^{(7)}, \ell_k^{(7)})). \tag{15}$$

For each expert, all weights of atlas voxels with identical labels are added:

$$W_k(\ell) = \sum_{\substack{j=0\dots7, \\ \ell_k^{(j)}=\ell}} w_k^{(j)}. \tag{16}$$

In what is commonly referred to as "Sum fusion" [4], the image voxel is finally assigned the label with the highest total weight summed over all experts, i.e.,

$$\arg\max_{\ell} \sum_k W_k(\ell). \tag{17}$$

6 Validation Study

The goal of the algorithms described above is to improve the accuracy of segmentation results by taking into account estimates of all experts' segmentation qualities. We are particularly interested in the case where each expert is an instance of a non-rigid registration method combined with an atlas image. Unlike statistics-based methods, atlas-based segmentation is by nature capable of, and typically aims at, labeling *anatomical structures* rather than *tissue types*. As an atlas is usually comprised of continuously defined objects, multiple independent atlas-based segmentations differ by *deformation* of these objects, rather than by *noise* (sparse pixels of different labels within a structure). The validation study described below is designed accordingly.

An increasingly popular non-rigid registration method was originally introduced by Rueckert *et al.* [6]. It applies free-form deformations [7] based on B-spline interpolation between uniform control points. We implemented this transformation model and simulate imperfect segmentations by applying random deformations to a known atlas. Each randomly deformed atlas serves as a model of an imperfect segmentation that approximates the original atlas. Several of these deformed atlases are combined into one segmentation using the methods described in the previous sections. Since the original (undeformed) atlas is known, it provides a valid ground truth for the results of all three methods.

6.1 Atlas Data

In order to ensure that the underlying undeformed atlas is meaningful and relevant, we did not generate a geometric phantom. Instead, we used real three-dimensional atlases derived from confocal microscopies of the brains of 20 adult foraging honey bees (see Ref. [8] for details). Each volume contained 84–114 slices with thickness $8\,\mu$m and each slice had 610–749 pixels in x direction and 379–496 pixels in y direction with pixel size $3.8\,\mu$m. In each individual image, 22 anatomical structures were distinguished and labeled.

For each ground truth, random B-spline-based free-form deformations were generated by adding independent Gaussian-distributed random numbers to the coordinates of all control points. The control point spacing was $120\,\mu$m, corresponding to approximately 30 voxels in x and y direction and 15 voxels in z direction. The variances of the Gaussian distributions were $\sigma = 10$, 20, and $30\,\mu$m, corresponding to approximately 2, 4, and 8 voxels in x and y direction (1, 2, and 4 voxels in z direction). Figure 1 shows examples of an atlas after application of several random deformations of different magnitudes. 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 averaging, and once using each of our novel algorithms.

6.2 Algorithm Parameters

Initialization. The expert parameters were initialized as follows. In the binary case, p and q were set to 0.9 for all experts. In the multi-label case, $\lambda_k(\ell,\ell)$

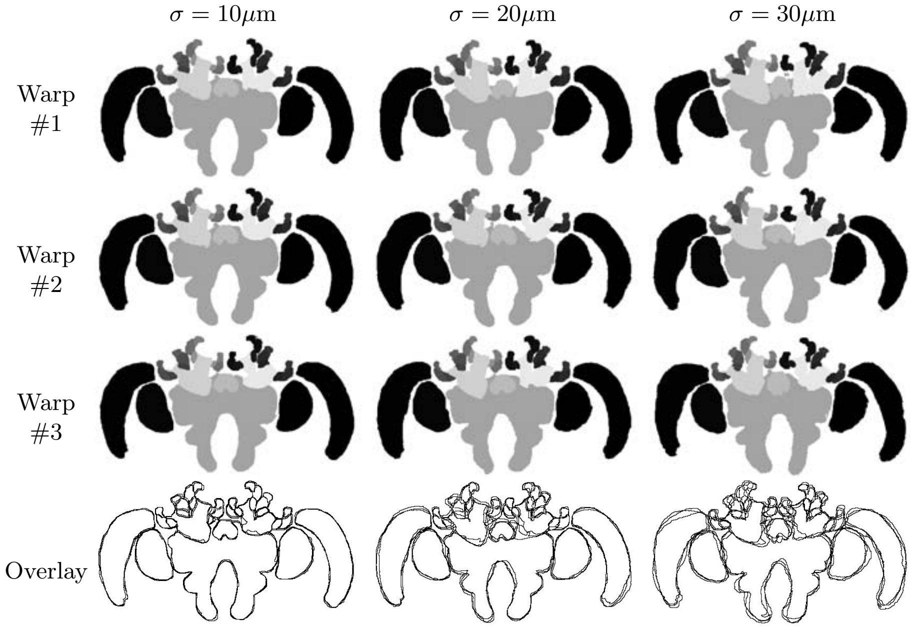

Fig. 1. Examples of a randomly deformed atlas. Each image shows a central axial slice from the same original atlas after application of a different random deformation. Within each column, the magnitudes of the deformations (variance of random distribution of control point motion) were constant. The images in the bottom row show overlays of the isocontours from the three images above to emphasize the subtle shape differences.

was initialized as 0.9 for all k and all ℓ. The off-diagonal elements were set to $(1 - \lambda_k(\ell,\ell))/L$.

Convergence Criterion. We are interested in processing large amounts of image data with many labels. In order to keep computation times somewhat reasonable, we do not wait for actual convergence of the results. Instead, we perform a fixed number of iterations, typically 7. In the validation study described below, our experience was that in the final iteration typically only one out of 10,000 voxels changed its value.

6.3 Evaluation

For every registration, the registration-based segmentation is compared with the manual segmentation. As one measure of segmentation quality we compute the global segmentation correctness measure C, which we define as the fraction of voxels for which the automatically generated registration-based segmentation matches the manually assigned labels:

$$C = \frac{\sum_s \left| V_{\text{comb}}^{(s)} \cap V_{\text{GT}}^{(s)} \right|}{\sum_s |V_{\text{GT}}^{(s)}|}. \tag{18}$$

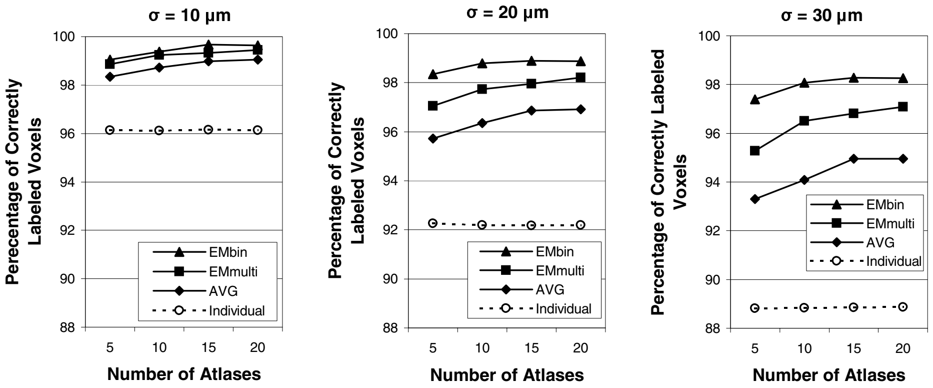

Fig. 2. Mean correctness C of combined segmentation over 20 individuals vs. number of random segmentations used. Results are shown for PVI label averaging (AVG), repeated application of the binary EM algorithm (EMbin), and the multi-label EM algorithm (EMmulti). Each method was applied to atlases after random deformations of magnitudes $\sigma = 10\,\mu$m (*left diagram*), $\sigma = 20\,\mu$m (*center*), and $\sigma = 30\,\mu$m (*right*). The dashed line in each graph shows the average correctness achieved by the respective set of individual atlases with no combination method.

where $V^{(s)}_{\mathrm{GT}}$ and $V^{(s)}_{\mathrm{comb}}$ denote the sets of indices of the voxels labeled as belonging to structure s in the undeformed ground truth (GT) and the combined estimated segmentation (comb), respectively.

6.4 Results

Figure 2 shows a plot of the mean correctness over all 20 individuals versus the number of segmentations. Both EM algorithms performed consistently better, i.e., produced more accurate combined segmentations, than simple label averaging. The improvement achieved using the EM strategies was larger for greater magnitudes of the random atlas deformations. Between the two EM methods, repeated application of the binary algorithm outperformed the multi-label method. For all algorithms, adding additional segmentations increased the accuracy of the combined segmentation. The incremental improvement obtained by adding an additional segmentation decreased as the number of atlases increased. The figure also nicely illustrates the superiority of using multiple atlases over using just one: in all cases, the individual correctnesses are substantially lower than any of the combined results. Again, the difference increases as the magnitude of the random deformations is increased.

7 Discussion

This paper has several new ideas. First, based on a novel interpretation of the term "expert", we propose to combine multiple registration-based segmentations into one in order to improve segmentation accuracy. Second, we introduce two multi-label extensions to an EM algorithm [3] for ground truth estimation in binary segmentation. Finally, we evaluate the segmentation quality of the two methods and a combined segmentation method based on simple label averaging. Effectively, this paper introduces the principle of combining multiple classifiers [4,9] to atlas-based image segmentation. In fact, the multi-label EM algorithm presented here can be understood as a learning method for the confusion matrix of a Bayesian classifier [9].

The quantitative evaluation of segmentation accuracy using random deformations of a known atlas demonstrated that both methods introduced in this paper produce better segmentations than simple label averaging. This is true despite the natural advantage that label averaging has by being able to consider fractional label contributions using PVI. Both EM algorithms described here more than make up for this inherent disadvantage. This finding is particularly significant as our previous research showed that combining multiple registration-based segmentations by label averaging already produces results that are better than the individual segmentations [2]. This finding, which corresponds to the experience of the pattern recognition community that multiple classifier systems are generally superior to single classifiers [4], was also confirmed by the validation study performed in this paper.

Between the two EM methods, the repeated application of a binary EM algorithm was superior to a dedicated multi-label algorithm, but at substantially increased computation cost. However, this may be different for different atlas topologies. Assume, for example, that there is an adjacency relationship between two anatomical structures in the form that one encloses the other. In this case, the crosstalk between classifications of both structures may be beneficial to consider, which is precisely what our novel multi-label EM algorithm does.

It should be mentioned that, like the original Warfield algorithm, our methods and their validation are based on several assumptions regarding the nature of the input data. Most notably, we assume that the errors of the individual segmentations are somewhat independent. In the presence of systematic errors made by all or at least a majority of the experts, the same error will very likely also appear in the final ground truth estimate. This problem, however, is not restricted to the machine experts that we focused on in this paper. In fact, since the individual training and experience of human experts are not mutually independent (in fact, similarity in training and expertise is what makes us consider someone an expert with respect to a certain problem), the same is true for manual segmentations.

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 involves 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 important that the simulated deformation be based on a different transformation model than the registration, for example, a B-spline-based registration should not be validated using simulated B-spline deformations.

In our context, however, the opposite is true. In this paper, we 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. DBR was supported by the Interdisciplinary Initiatives Program, which is part of the Bio-X Program at Stanford University, under the grant "Image-Guided Radiosurgery for the Spine and Lungs."

References

1. BM Dawant, SL Hartmann, JP Thirion, *et al.* 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, R Menzel, *et al.* 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, Feb. 2003. In print.
3. 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 (MICCAI)*, pp. 298–306, Springer-Verlag, Berlin, 2002.
4. J Kittler, M Hatef, RPW Duin, *et al.* On combining classifiers. *IEEE Trans Pattern Anal Machine Intell*, 20(3):226–239, Mar. 1998.
5. F Maes, A Collignon, D Vandermeulen, *et al.* Multimodality image registration by maximisation of mutual information. *IEEE Trans Med Imag*, 16(2):187–198, 1997.
6. D Rueckert, LI Sonoda, C Hayes, *et al.* Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Trans Med Imag*, 18(8):712–721, 1999.
7. TW Sederberg, SR Parry. Free-form deformation and solid geometric models. *Comput Graph (ACM)*, 20(4):151–160, 1986.
8. T Rohlfing, R Brandt, CR Maurer, Jr., *et al.* Bee brains, B-splines and computational democracy: Generating an average shape atlas. In *Proceedings of IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 187–194, 2001. IEEE Computer Society, Los Alamitos, CA.
9. L Xu, A Krzyzak, CY Suan. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans Syst Man Cybern*, 22(3):418–435, 1992.