Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation

Simon K. Warfield*, *Member, IEEE*, Kelly H. Zou, and William M. Wells, *Member, IEEE*

***Abstract*—Characterizing the performance of image segmentation approaches has been a persistent challenge. Performance analysis is important since segmentation algorithms often have limited accuracy and precision. Interactive drawing of the desired segmentation by human raters has often been the only acceptable approach, and yet suffers from intra-rater and inter-rater variability. Automated algorithms have been sought in order to remove the variability introduced by raters, but such algorithms must be assessed to ensure they are suitable for the task.**

The performance of raters (human or algorithmic) generating segmentations of medical images has been difficult to quantify because of the difficulty of obtaining or estimating a known true segmentation for clinical data. Although physical and digital phantoms can be constructed for which ground truth is known or readily estimated, such phantoms do not fully reflect clinical images due to the difficulty of constructing phantoms which reproduce the full range of imaging characteristics and normal and pathological anatomical variability observed in clinical data. Comparison to a collection of segmentations by raters is an attractive alternative since it can be carried out directly on the relevant clinical imaging data. However, the most appropriate measure or set of measures with which to compare such segmentations has not been clarified and several measures are used in practice.

We present here an expectation-maximization algorithm for simultaneous truth and performance level estimation (STAPLE). The algorithm considers a collection of segmentations and computes a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation. The source of each segmentation in the collection may be an appropriately trained human rater or raters, or may be an automated segmentation algorithm. The probabilistic estimate of the true segmentation is formed by estimating an optimal combination of the segmentations, weighting each segmentation depending upon the estimated performance level, and incorporating a prior model for the spatial distribution of structures being segmented as well as spatial homogeneity constraints. STAPLE is straightforward to apply to clinical imaging data, it readily enables assessment of the performance of an automated image segmentation algorithm, and enables direct comparison of human rater and algorithm performance.

***Index Terms*—Accuracy, classifier fusion, expectation-maximization, gold standard, ground truth, Markov random field, precision, segmentation, sensitivity, specificity, STAPLE, validation.**

I. Introduction

MEDICAL image segmentation has long been recognized as a difficult problem. Many interactive and automated algorithms have been proposed, and in practice approaches specifically tuned to the important characteristics of the application are often successful. When selecting, designing, or optimizing particular algorithms for a segmentation task, the performance characteristics of the algorithms must be assessed.

Characterizing the performance of image segmentation approaches has also been a persistent challenge. Quantitative performance analysis is important since segmentation algorithms often have limited accuracy and precision. Interactive drawing of the desired segmentation by domain experts has often been the only acceptable approach, and yet suffers from intraexpert and inter-expert variability and is time consuming and expensive to carry out. Automated algorithms have been sought in order to remove the variability introduced by experts, but automated algorithms must be assessed to ensure they are suitable for the task.

Measurement tools are often characterized by assessment of their accuracy and precision. The accuracy of a human or an algorithm in creating a segmentation is the degree to which the segmentation corresponds to the true segmentation, and so the assessment of accuracy of a segmentation requires a reference standard, representing the true segmentation, against which it may be compared. Precision is determined by the reproducibility of the segmentations obtained repeatedly from the same image [1]. The precision may be assessed without comparison to a reference standard. High accuracy and high precision are both desirable properties.

An ideal reference standard for image segmentation would be known to high accuracy and would reflect the characteristics of segmentation problems encountered in practice. There is a tradeoff between the accuracy and realism of the reference standard. As the accuracy of the reference standard segmentation increases, the degree to which the reference standard

Manuscript received January 29, 2004; revised March 15, 2004. This work was supported in part by the Whitaker Foundation, in part by the National Institutes of Health (NIH) under Grant R21 MH67054, Grant R01 LM007861, Grant P41 RR13218, Grant P01 CA67165, Grant R01 AG19513, Grant R01 CA86879, Grant R01 NS35142, Grant R33 CA99015, and Grant R21 CA89449, and in part by an award from the Center for Integration of Medicine and Innovative Technology. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. A. Viergever. *Asterisk indicates corresponding author.*

*S. K. Warfield is with Harvard Medical School and the Department of Radiology of Brigham and Women's Hospital, 75 Francis St, Boston, MA 02115 USA. He is also with the Department of Radiology at Children's Hospital, Boston, and with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: warfield@bwh.harvard.edu).

K. H. Zou is with Harvard Medical School and the Department of Radiology of Brigham and Women's Hospital, Boston, MA 02115 USA.

W. M. Wells is with Harvard Medical School and the Department of Radiology at Brigham and Women's Hospital, Boston, MA 02115 USA. He is also with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Digital Object Identifier 10.1109/TMI.2004.828354

reflects segmentation problems encountered in practice tends to decrease. The accuracy of segmentations of medical images has been difficult to quantify in the absence of an accepted reference standard for clinical imaging data.

Synthetic images can be constructed with known true segmentation (high accuracy) but typically lack characteristics encountered in practice. Phantoms can be built and imaged, which increases the realism of the model by incorporating the imaging system characteristics, but also reduces the fidelity with which the true segmentation is known. Although such physical and digital phantoms [2], [3] have an important role to play in quantifying algorithm performance, such phantoms do not fully reflect clinical images due to the difficulty of constructing phantoms that reproduce the full range of imaging characteristics (such as partial volume artifact, intensity inhomogeneity artifact and noise). In addition, such phantoms typically cannot reproduce both the normal and pathological anatomical variability observed in clinical data. Therefore, alternative approaches have an important role, as it is not necessarily possible to generalize performance measurements on a phantom to the results expected in practice.

Cadavers provide a more realistic model of anatomy, but the true segmentation can only be estimated, and such models differ from *in vivo* data [4], [5]. Patient data provides the most realistic model for a given application task, but is the most difficult for which to identify a reference standard.

A common alternative to phantom studies has been to carry out behavioral comparisons: an automated algorithm is compared to the segmentations generated by a group of experts, and if the algorithm generates segmentations sufficiently similar to the experts, it is considered to be an acceptable substitute for the experts. In addition to acceptable accuracy, good automated segmentation algorithms also typically require less time to apply and have better precision than interactive segmentation by an expert.

The most appropriate way to carry out the comparison of a segmentation to a group of expert segmentations is so far unclear. A number of metrics have been proposed to compare segmentations. Simply measuring the volume of segmented structures has often been used [6], [7]. Two segmentation methods may be compared by assessing the limits of agreement [8] of volume estimates derived from the segmentations. However, volume estimates may be quite similar when the segmented structures are located differently, have different shapes or have different boundaries [9] and so alternative measures have been sought [10].

Other measures used in practice include measures of spatial overlap, such as the Dice and Jaccard Similarity Coefficients [11], [12]. For example, spatial overlap has been used to compare manual segmentations with segmentations obtained through nonrigid registration [13]. Furthermore, measures inspired by information theory have been applied [14], [15]. In many applications, assessment of boundary differences is useful and the Hausdorff measure and modifications have been used for this [16]. Agreement measures for comparing different experts, such as the kappa statistic, have also been explored [17].

A reference standard has sometimes been formed by taking a combination of expert segmentations. For example, a voting rule as often used in practice selects all voxels where some majority of experts agree the structure to be segmented is present [18], [19]. This simple approach unfortunately does not provide guidance as to how many experts should agree before the structure is considered to be present. Furthermore vote counting strategies treat each voter equally without regard to potential variability in quality or performance amongst the voters, and does not allow for *a priori* information regarding the structure being segmented to be incorporated. In the case of binary segmentations, a majority vote is at least unique in the case of an odd number of voters. However, in the case of multicategory segmentations, the category with the most votes may not be unique and may not reflect the overall preferred choice of the voters. Preferential voting strategies, operating on votes or class probabilities, have been examined in the context of classifier fusion [20]. Examples include the Borda count [21] for preferential vote ordering, class probability combining strategies (the Product Rule and the Sum Rule which can be used to express the Min Rule, the Max Rule, the Median Rule and the Majority Vote Rule as described by [22]), and strategies which assume each classifier has expertise in a subset of the decision domain [23]–[25]. Similar issues in combining decisions from multiple raters or studies arise in content-based collaborative filtering [26] and in meta-analysis of diagnostic tests [27], [28]. Estimating performance in the presence of an imperfect or limited reference standard has also been explored [29], [30]. The most appropriate vote ordering or decision combining strategy remains unclear.

We present here a new algorithm, simultaneous truth and performance level estimation (STAPLE), which takes a collection of segmentations of an image, and computes simultaneously a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation. The source of each segmentation in the collection may be an appropriately trained human rater or raters, or it may be an automated segmentation algorithm. Our algorithm is formulated as an instance of the expectation-maximization (EM) algorithm [31], [32] and builds upon our earlier work [33], [34]. In the formulation of our algorithm described here, the expert segmentation decision at each voxel is directly observable, the hidden true segmentation is a binary variable for each voxel, and the performance level, or quality, achieved by each segmentation is represented by sensitivity and specificity parameters. We also describe a generalization of the algorithm to unordered multicategory labels, suitable for when the hidden true segmentation at each voxel is one of an unordered finite set of labels.

The *complete* data consists of the segmentation decisions at each voxel, which are known, and the true segmentation, which is not known. If we did know the true segmentation, it would be straightforward to estimate the performance parameters by maximum-likelihood (ML) estimation. Since the complete data is not available, the complete data log likelihood cannot be constructed and instead must be estimated. Doing so requires evaluating the conditional probability density of the *hidden* true segmentation given the segmentation decisions and a previous estimate of the performance level of each segmentation generator. The expectation of the complete data log likelihood with respect to this density is then calculated, and from this estimate of the complete data log likehood, the performance parameters are found by ML estimation (or maximum *a posteriori* (MAP) estimation when a prior distribution for the parameters is considered). We iterate this sequence of estimation of the conditional probability of the true segmentation and performance parameters until convergence is reached. Convergence

to a local maximum is guaranteed. Since we obtain both an estimate of the true segmentation and performance parameters from a collection of segmentations and a prior model, the algorithm is straightforward to apply to clinical imaging data. The algorithm enables the assessment of the performance of an automated image segmentation algorithm, and provides a simple method for direct comparison of human and algorithm performance.

This paper is organized as follows. In Section II, we describe the STAPLE algorithm. We describe how STAPLE can be appropriately initialized both when there is no or limited prior information and when rich prior information is available. We describe how convergence of the estimation scheme is detected. In Section III, we describe the application of the STAPLE algorithm to digital phantoms where the true segmentation is known by construction and investigate the characterization of segmentations by synthetic raters with predefined sensitivities and specificities. We demonstrate that STAPLE accurately estimates the performance level parameters and the true segmentation. We demonstrate the use of STAPLE with data from several clinical applications in order to compare with previously reported validation measures and to illustrate how the algorithm may be used in practice. The computational complexity and the run time of the algorithm are reported, and indicate the algorithm converges rapidly in practical applications.

II. Method

We describe in this section an EM algorithm for estimating the hidden true segmentation and performance level parameters from a collection of segmentations and a prior model.

A. Description of STAPLE Algorithm

Consider an image of N voxels, and the task of segmenting a structure in that image by indicating the presence or absence of the structure at each voxel. Let $\mathbf{p} = (p_1, p_2, \ldots, p_R)^T$ be a column vector of R elements, with each element a sensitivity parameter characterising one of R segmentations, and $\mathbf{q} = (q_1, q_2, \ldots, q_R)^T$ be a column vector of R elements, with each element a specificity parameter characterising the performance of one of R segmentations. Let $\mathbf{D}$ be an $N \times R$ matrix describing the binary decisions made for each segmentation at each voxel of the image. Let $\mathbf{T}$ be an indicator vector of N elements, representing the hidden binary true segmentation, where for each voxel the structure of interest is recorded as present (1) or absent (0). Let the complete data be $(\mathbf{D}, \mathbf{T})$ and let the probability mass function of the complete data be $f(\mathbf{D}, \mathbf{T} \,|\, \mathbf{p}, \mathbf{q})$.

Our goal is to estimate the performance level parameters of the experts characterized by parameters $(\mathbf{p}, \mathbf{q})$ which maximize the complete data log likelihood function

$$(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \arg\max_{\mathbf{p},\mathbf{q}} \ln f(\mathbf{D}, \mathbf{T} \,|\, \mathbf{p}, \mathbf{q}). \tag{1}$$

If we knew the true segmentation, we could construct a 2×2 conditional probability table, by comparing the decision D_{ij} of segmentation j (for $j = 1, \ldots, R$) as to the presence or absence of a structure at voxel i with the true segmentation. Recall the definition of sensitivity and specificity, where p_j represent the "true positive fraction" or sensitivity (relative frequency of $D_{ij} = 1$ when $T_i = 1$),

$$p_j = \Pr(D_{ij} = 1 \,|\, T_i = 1) \tag{2}$$

and q_j represent the "true negative fraction" or specificity (relative frequency of $D_{ij} = 0$ when $T_i = 0$)

$$q_j = \Pr(D_{ij} = 0 \,|\, T_i = 0). \tag{3}$$

The parameters $p_j, q_j \in [0, 1]$ are assumed to be characteristic of the rater, and may be equal for different raters but in general are not. We assume that the segmentation decisions are all conditionally independent given the true segmentation and the performance level parameters, that is $(D_{ij} \,|\, T_i, p_j, q_j) \perp (D_{ij'} \,|\, T_i, p_{j'}, q_{j'}), \forall j \neq j'$. This model expresses the assumption that the raters derive their segmentations of the same image independently from one another and that the quality of the result of the segmentation is captured by the sensitivity and specificity parameters.

Implicit in this model is the notion that the experts have been trained to interpret the images in a similar way, the segmentation decisions may differ due to random or systematic rater differences, and that a probabilistic estimate of the true segmentation can be formulated as an optimal combination of the observed decisions and a prior model.

We now present the version of the EM algorithm that we have developed for this problem which enables us to estimate the solution of (1). Detailed descriptions of the EM algorithm and generalizations are available [31], [32] and our exposition closely follows those sources. The essence of the EM algorithm is the observation that certain ML estimation problems would be considerably simplified if some missing data were available, and this is the case for our problem. The observable data, the segmentation decisions at each voxel, is regarded as being incomplete and is regarded as an observable function of the complete data. Here, the complete data is the segmentation decisions $\mathbf{D}$ augmented with the true segmentation of each voxel $\mathbf{T}$. This true segmentation $\mathbf{T}$ is called the missing or hidden data, and is unobservable. Let

$$\boldsymbol{\theta}_j = (p_j, q_j)^T, \quad \forall j \in 1, \ldots, R \tag{4}$$

be the unknown parameters characterizing performance of segmentation j and

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_R] \tag{5}$$

the complete set of unknown parameters for the R segmentations that we wish to identify. Let $f(\mathbf{D}, \mathbf{T} \,|\, \boldsymbol{\theta})$ denote the probability mass function of the random vector corresponding to the complete data. We write the complete data log likelihood function as

$$\ln L_c\{\boldsymbol{\theta}\} = \ln f(\mathbf{D}, \mathbf{T} \,|\, \boldsymbol{\theta}). \tag{6}$$

The EM algorithm approaches the problem of maximizing the incomplete data log likelihood equation

$$\ln L\{\boldsymbol{\theta}\} = \ln f(\mathbf{D} \,|\, \boldsymbol{\theta}) \tag{7}$$

by proceeding iteratively with estimation and maximization of the complete data log likelihood function. As the complete data

log likelihood function is not observable, it is replaced by its conditional expectation given the observable data $\mathbf{D}$ using the current estimate of $\boldsymbol{\theta}$. Computing the conditional expectation of the complete data log likelihood function is referred to as the E-step, and identifying the parameters that maximize this function is referred to as the M-step.

In more detail, let $\boldsymbol{\theta}^{(0)}$ be some initial value for $\boldsymbol{\theta}$. Then, on the first iteration, the E-step requires the calculation of

$$Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(0)}\right) \equiv E\left[\ln f(\mathbf{D},\mathbf{T}\,|\,\boldsymbol{\theta})\middle|\mathbf{D},\boldsymbol{\theta}^{(0)}\right] = \sum_{\mathbf{T}} \ln f(\mathbf{D},\mathbf{T}\,|\,\boldsymbol{\theta}) f\left(\mathbf{T}\,|\,\mathbf{D},\boldsymbol{\theta}^{(0)}\right). \quad (8)$$

The M-step requires the maximization of $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(0)})$ over the parameter space of $\boldsymbol{\theta}$. That is, we choose $\boldsymbol{\theta}^{(1)}$ such that

$$Q\left(\boldsymbol{\theta}^{(1)}\,|\,\boldsymbol{\theta}^{(0)}\right) \geq Q\left(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(0)}\right) \quad (9)$$

for all $\boldsymbol{\theta}$. The E-step and M-step are then repeated as above where at each iteration k, the current estimate $\boldsymbol{\theta}^{(k)}$ and the observed segmentation decisions $\mathbf{D}$ are used to calculate the conditional expectation of the complete data log likelihood function, and then the estimate of $\boldsymbol{\theta}^{(k+1)}$ is found by maximization of $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$. The E- and M-steps are iterated until convergence. It has been shown [31] that the incomplete data log likelihood function is not decreased after an EM iteration, and so local convergence is guaranteed when the likelihood function has an upper bound.

The key requirements necessary to carry out the EM algorithm are to have a specification of the complete data, and to have the conditional probability density of the complete data given the observed data. We describe below how the E-step and M-step are formulated for our estimation problem.

The performance parameters at iteration k that maximize the conditional expectation of the log likelihood function are given by

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} E\left[\ln f(\mathbf{D},\mathbf{T}\,|\,\boldsymbol{\theta})\middle|\mathbf{D},\boldsymbol{\theta}^{(k-1)}\right] = \arg\max_{\boldsymbol{\theta}} E\left[\ln \frac{f(\mathbf{D},\mathbf{T},\boldsymbol{\theta})}{f(\boldsymbol{\theta})}\middle|\mathbf{D},\boldsymbol{\theta}^{(k-1)}\right]. \quad (10)$$

Hence, on expressing $\boldsymbol{\theta}^{(k)}$ as $(\mathbf{p}^{(k)},\mathbf{q}^{(k)})$

$$\left(\mathbf{p}^{(k)},\mathbf{q}^{(k)}\right) = \arg\max_{\mathbf{p},\mathbf{q}} E\Big[\ln \frac{f(\mathbf{D},\mathbf{T},\mathbf{p},\mathbf{q}) f(\mathbf{T},\mathbf{p},\mathbf{q})}{f(\mathbf{T},\mathbf{p},\mathbf{q}) f(\mathbf{p},\mathbf{q})}\Big| \mathbf{D},\mathbf{p}^{(k-1)},\mathbf{q}^{(k-1)}\Big] \quad (11)$$

and so we have

$$\left(\mathbf{p}^{(k)},\mathbf{q}^{(k)}\right) = \arg\max_{\mathbf{p},\mathbf{q}} E\Big[\ln(f(\mathbf{D}\,|\,\mathbf{T},\mathbf{p},\mathbf{q}) f(\mathbf{T}))\Big| \mathbf{D},\mathbf{p}^{(k-1)},\mathbf{q}^{(k-1)}\Big] \quad (12)$$

where $(\mathbf{p}^{(k)},\mathbf{q}^{(k)})$ is the estimate of the expert performance level parameters at iteration k, and the last follows under the assumption that $\mathbf{T}$ is independent of the performance level parameters so that $f(\mathbf{T},\mathbf{p},\mathbf{q}) = f(\mathbf{T}) f(\mathbf{p},\mathbf{q})$.

B. E-Step: Estimation of the Conditional Expectation of the Complete Data Log Likelihood Function

In this section, the estimator for the unobserved true segmentation is derived. We first derive an expression for the conditional probability density function of the true segmentation given the expert decisions (observed segmentations) and the previous estimate of the performance parameters

$$\begin{aligned} f\left(\mathbf{T}\middle|\mathbf{D},\boldsymbol{\theta}^{(k)}\right) &= \frac{f\left(\mathbf{D}\middle|\mathbf{T},\boldsymbol{\theta}^{(k-1)}\right) f(\mathbf{T})}{\sum_{\mathbf{T}'} f\left(\mathbf{D}\middle|\mathbf{T}',\boldsymbol{\theta}^{(k-1)}\right) f(\mathbf{T}')} \\ &= \frac{\prod_i \left[\prod_j f\left(D_{ij}\middle|T_i,\theta_j^{(k-1)}\right)\right] f(T_i)}{\sum_{T_1'}\cdots\sum_{T_N'} \prod_i \left[\prod_j f\left(D_{ij}\middle|T_i',\theta_j^{(k-1)}\right)\right] f(T_i')} \\ &= \frac{\prod_i \left[\prod_j f\left(D_{ij}\middle|T_i,p_j{}^{(k-1)},q_j{}^{(k-1)}\right)\right] f(T_i)}{\prod_i \left[\sum_{T_i'} \prod_j f\left(D_{ij}\middle|T_i',p_j{}^{(k-1)},q_j{}^{(k-1)}\right)\right] f(T_i')} \end{aligned}$$

such that at each voxel i we have

$$f\left(T_i\middle|\mathbf{D_i},\mathbf{p}^{(k-1)},\mathbf{q}^{(k-1)}\right) = \frac{\prod_j f\left(D_{ij}\middle|T_i,p_j{}^{(k-1)},q_j{}^{(k-1)}\right) f(T_i)}{\sum_{T_i'} \prod_j f\left(D_{ij}\middle|T_i',p_j{}^{(k-1)},q_j{}^{(k-1)}\right) f(T_i')} \quad (13)$$

where $f(T_i)$ is the prior probability of T_i, and the conditional independence of the segmentations allows us to write the joint probability as a product of rater-specific probabilities. A voxelwise independence assumption has been made here, and in Section II-E it is shown how voxelwise dependence can be efficiently modeled.

Since the true segmentation is treated as a binary random variable, the conditional probability $f(T_i = 0\,|\,\mathbf{D_i},\mathbf{p}^{(k-1)},\mathbf{q}^{(k-1)})$ is simply $1 - f(T_i = 1\,|\,\mathbf{D_i},\mathbf{p}^{(k-1)},\mathbf{q}^{(k-1)})$. We store for each voxel the estimate of the probability that the true segmentation at each voxel is $T_i = 1$. Section II-D describes a generalization to segmentations with unordered multicategory labels.

Now consider (13) for $T_i = 1$ and $T_i = 0$, respectively. Factoring over the raters and using the definition of p_j and q_j, we can write

$$\begin{aligned} a_i^{(k)} &\equiv f(T_i = 1) \prod_j f\left(D_{ij}\middle|T_i = 1,p_j{}^{(k)},q_j{}^{(k)}\right) \\ &= f(T_i = 1) \prod_{j:D_{ij}=1} p_j{}^{(k)} \prod_{j:D_{ij}=0} \left(1 - p_j{}^{(k)}\right) \end{aligned} \quad (14)$$

and

$$\begin{aligned} b_i^{(k)} &\equiv f(T_i = 0) \prod_j f\left(D_{ij}\middle|T_i = 0,p_j{}^{(k)},q_j{}^{(k)}\right) \\ &= f(T_i = 0) \prod_{j:D_{ij}=0} q_j{}^{(k)} \prod_{j:D_{ij}=1} \left(1 - q_j{}^{(k)}\right) \end{aligned} \quad (15)$$

where $j : D_{ij} = 1$ denotes the set of indexes for which the decision of rater j at voxel i (i.e., D_{ij}) has the value 1.

With these expressions, we can now write a compact expression for the conditional probability of the true segmentation at each voxel. Using the notation common for EM algorithms, we refer to this as the weight variable $W_i^{(k-1)}$

$$W_i^{(k-1)} \equiv f\left(T_i = 1 \left| \mathbf{D_i}, \mathbf{p}^{(k-1)}, \mathbf{q}^{(k-1)} \right.\right) = \frac{a_i^{(k-1)}}{a_i^{(k-1)} + b_i^{(k-1)}}. \quad (16)$$

The weight $W_i^{(k-1)}$ indicates the probability of the true segmentation at voxel i being equal to one. It is a normalized product of the prior probability of $T_i = 1$, the sensitivity of each of the experts that decided the true segmentation was one and (1—sensitivity) of each of the experts that decided the true segmentation was zero.

In order to complete the E-step, we require an expression for the conditional expectation of the complete data log likelihood function. This is dramatically simplified when we consider only the terms necessary for finding the parameters that maximize the function, and this is derived in the next section.

C. M-Step: Estimation of the Performance Parameters by Maximization

Given the estimated weight variables $W_i^{(k-1)}$, which represent the conditional probabilities of the true segmentation, we can find the values of the expert performance level parameters that maximize the conditional expectation of the complete data log likelihood function.

Considering (12), and noting that since $\ln \prod_i f(T_i)$ is free of $(\mathbf{p}, \mathbf{q})$ it follows that

$$\left(\mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right) = \arg\max_{\mathbf{p},\mathbf{q}} \sum_j \sum_i E\left[\ln f(D_{ij} \,|\, T_i, p_j, q_j) \left| \mathbf{D}, \mathbf{p}^{(k-1)}, \mathbf{q}^{(k-1)} \right.\right] \quad (17)$$

so for each expert j

$$\begin{aligned}
&\left(p_j{}^{(k)}, q_j{}^{(k)}\right) \\
&= \arg\max_{p_j,q_j} \sum_i E\left[\ln f(D_{ij} \,|\, T_i, p_j, q_j) \left| \mathbf{D}, \mathbf{p}^{(k-1)}, \mathbf{q}^{(k-1)} \right.\right] \\
&= \arg\max_{p_j,q_j} \sum_i \left[W_i^{(k-1)} \ln f(D_{ij} \,|\, T_i = 1, p_j, q_j) + \left(1 - W_i^{(k-1)}\right) \ln f(D_{ij} \,|\, T_i = 0, p_j, q_j) \right] \\
&= \arg\max_{p_j,q_j} \sum_{i:D_{ij}=1} W_i^{(k-1)} \ln p_j + \sum_{i:D_{ij}=1} \left(1 - W_i^{(k-1)}\right) \ln(1 - q_j) + \sum_{i:D_{ij}=0} W_i^{(k-1)} \ln(1 - p_j) + \sum_{i:D_{ij}=0} \left(1 - W_i^{(k-1)}\right) \ln q_j.
\end{aligned}$$

A necessary condition at a maximum of the above with respect to parameter p_j is that the first derivative equal zero. Differentiating $Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(k)})$ with respect to parameter p_j and equating to 0, yields (similarly for q_j)

$$p_j{}^{(k)} = \frac{\sum_{i:D_{ij}=1} W_i^{(k-1)}}{\sum_i W_i^{(k-1)}} \quad (18)$$

$$q_j{}^{(k)} = \frac{\sum_{i:D_{ij}=0} \left(1 - W_i^{(k-1)}\right)}{\sum_i \left(1 - W_i^{(k-1)}\right)}. \quad (19)$$

We can interpret the weight estimate $W_i^{(k-1)}$ as the strength of belief in the underlying true segmentation being equal to 1. In the case of perfect knowledge about the true segmentation, i.e., $W_i^{(k-1)} \in \{0.0, 1.0\}$, the estimator of sensitivity given by (18) corresponds to the usual definition of sensitivity as the true positive fraction. When the true segmentation estimate is a continuous parameter, i.e., $W_i^{(k-1)} \in [0, 1]$ as here, the estimator can be interpreted as the ratio of the jth expert true positive detections to the total amount of structure $T_i = 1$ voxels believed to be in the data, with each voxel detection weighted by the strength of belief in $T_i = 1$. Similarly, the specificity estimator of (19) is a natural formulation of an estimator for the specificity given a degree of belief in the underlying $T_i = 0$ state.

Summary: STAPLE estimates the performance parameters, and a probabilistic estimate of the true segmentation, by iterated estimation. The first step of each iteration is estimation of the conditional probability of the true segmentation given the expert decisions and previous performance parameter estimates as given by (16), and the second step is updated estimation of the performance parameters given by (18) and (19).

D. An Extension to Unordered Multicategory Labels

In this section, we provide a generalization of the estimation strategy to the case when the truth is a label from a finite set of unordered categories. For example, consider a brain magnetic resonance image (MRI) tissue segmentation application where we are interested to identify the true label as one of white matter, gray matter, or cerebrospinal fluid. Rather than being restricted to two labels only, here we allow $\mathbf{T}$ to take on one of $\mathbf{L}$ labels where $\mathbf{T_i} \in \{0, 1, \ldots, L-1\}$ and similarly the segmentation decisions may also be one of these L labels, $D_{ij} \in \{0, 1, \ldots, L-1\}$, and $\mathbf{D}$ is an $N \times R$ matrix describing the decisions made for each segmentation at each voxel of the image. In this case, we characterize each rater performance by θ_j, an $L \times L$ matrix of parameters where the element of the matrix characterizes the probability rater j will decide label s' is present when the true label is s. The perfect rater would have a performance matrix with 1 on the diagonal and 0 in each off-diagonal element, that is the rater performance matrix would be an identity matrix.

In order to evaluate the conditional expectation of the complete data log likelihood for this case, we need to compute the conditional probability that the true label at voxel i is s given the set of segmentations and the previous estimate of the performance characteristics. Similarly to the binary case, we find that

$$W_{si}^{(k)} \equiv f\left(T_i = s \left| \mathbf{D_i}, \boldsymbol{\theta}^{(k)} \right.\right) = \frac{f(T_i = s) \prod_j f\left(D_{ij} \left| T_i = s, \theta_j^{(k)} \right.\right)}{\sum_{s'} f(T_i = s') \prod_j f\left(D_{ij} \left| T_i = s', \theta_j^{(k)} \right.\right)}. \quad (20)$$

With this conditional probability, we can evaluate the expectation of the complete data log likelihood function, with the goal of identifying the performance parameters by ML estimation. Considering each rater separately, we find the new parameter estimates $\theta_j^{(k+1)}$ by

$$\begin{aligned}\theta_j^{(k+1)} &= \arg\max_{\theta_j} \sum_i E\left[\ln f(D_{ij}\,|\,T_i,\theta_j)\,\middle|\,\mathbf{D},\theta_j^{(k)}\right] \\ &= \arg\max_{\theta_j} \sum_i \sum_s W_{si}^{(k)} \ln f(D_{ij}\,|\,T_i=s,\theta_j) \\ &= \arg\max_{\theta_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_s W_{si}^{(k)} \\ &\quad \times \ln f(D_{ij}=s'\,|\,T_i=s,\theta_j) \\ &= \arg\max_{\theta_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_s W_{si}^{(k)} \ln \theta_{js's}. \end{aligned} \tag{21}$$

Noting the constraint that each row of the rater parameter matrix must sum to one to be a probability mass function

$$\sum_{s'} \theta_{js's} = 1 \tag{22}$$

we can find the parameters that maximize the above expression by formulating the constrained optimization problem

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_{jn'n}} \left[\sum_{s'} \sum_{i:D_{ij}=s'} \sum_s W_{si}^{(k)} \ln \theta_{js's} \right. \\ &\qquad \left. + \lambda \sum_{s'} \theta_{js's} \right] \\ &= \sum_{i:D_{ij}=n'} W_{ni}^{(k)} \frac{1}{\theta_{jn'n}} + \lambda. \end{aligned}$$

Therefore, we find

$$\theta_{js's}^{(k+1)} = \frac{\sum_{i:D_{ij}=s'} W_{si}^{(k)}}{-\lambda} \tag{23}$$

and on substituting the constraint from (22) we have

$$\theta_{js's}^{(k+1)} = \frac{\sum_{i:D_{ij}=s'} W_{si}^{(k)}}{\sum_i W_{si}^{(k)}}. \tag{24}$$

Note that when the number of possible labels is two ($L=2$), this estimator for unordered multicategory labels simplifies to the same expression as was derived above for the binary case.

E. A Model for Estimation of Spatially Correlated Structures

One mechanism to capture the spatial variation of a structure of interest is to provide a spatially varying probability for $f(T_i=1)$ to reflect prior information about the spatial distribution of the T_i. For example, when assessing brain segmentations of white matter, a probabilistic atlas of the distribution of white matter derived from a large group of subjects can provide the prior true segmentation probabilities at each voxel [35]–[37]. However, a probabilistic atlas is not available for all structures of interest. An alternative is to model the local spatial correlation and homogeneity of the structure.

The estimate of the posterior probability of T_i obtained from (13) is correct under the assumption that the true segmentation has no spatial correlation—that is, at any given voxel the probability of the true segmentation is independent of the true segmentation of the neighboring or adjacent voxels. In practical applications it is often the case that the true segmentation has an underlying spatial homogeneity.

A second mechanism for incorporating spatial homogeneity is to introduce a Markov random field (MRF) model. This allows milder assumptions than the strong voxelwise independence assumption used above. Constructing a fully general probability model on random fields such as $\mathbf{T}$ is a daunting prospect due to the huge size of such a model—just representing full covariation statistics would be difficult. A useful compromise that partially relaxes the independence assumption is the MRF model, in which the conditional dependence of a given voxel on all of the others is equal to its conditional dependence on the voxels in a local neighborhood.

Beginning with the Ising [38] models of ferromagnetism, these models have frequently been used to model phenomena that exhibit spatial coherence, including medical image segmentation problems [39]–[41].

While the estimators which use MRF models are usually more complex to implement, exact estimates may be obtained in reasonable (polynomial) time [42], [43] and efficient approximation schemes are also available [44], [45].

The Hammersley-Clifford theorem [46] establishes a one to one correspondence between MRFs and probability models written in Gibbs form, as follows:

$$\begin{aligned} f(\mathbf{T}) &= \frac{1}{Z} \exp\left(\frac{-E(\mathbf{T})}{\tau}\right) \\ Z &= \sum_{\mathbf{T}} \exp\left(\frac{-E(\mathbf{T})}{\tau}\right) \\ -E(\mathbf{T}) &= \sum_i U(T_i) + \sum_i \sum_j V(T_i,T_j). \end{aligned} \tag{25}$$

This is the stationary Gibbs form, where Z and τ are constants and E is called the energy function. Selecting $E(\mathbf{T})$ as above, we can factor this standard form into a term depending solely on the single site voxel prior and a term depending on pairwise interactions

$$\begin{aligned} f(\mathbf{T}) &= \frac{1}{Z} \exp\left(\frac{\sum_i U(T_i)}{\tau}\right) \\ &\quad \times \exp\left(\frac{\sum_i \sum_j V(T_i,T_j)}{\tau}\right) \\ &= \frac{1}{Z_1} \exp\left(\frac{\sum_i U(T_i)}{\tau}\right) \frac{Z_1}{Z} \\ &\quad \times \exp\left(\frac{\sum_i \sum_j V(T_i,T_j)}{\tau}\right) \\ &= \prod_i f(T_i) \frac{Z_1}{Z} \exp\left(\frac{\sum_i \sum_j V(T_i,T_j)}{\tau}\right) \\ \ln f(\mathbf{T}) &= \sum_i \ln f(T_i) + \ln Z_1 - \ln Z \\ &\quad + \frac{1}{\tau} \sum_i \sum_j V(T_i,T_j). \end{aligned} \tag{26}$$

The above natural model allows us to incorporate a model for spatial homogeneity of the true segmentation in our estimation scheme. One approach to MRF estimation that has low computational complexity and which is applicable here, is the mean field approximation. An extensive investigation of this has been reported by Elfadel [44] and it has been used to impose a spatial homogeneity constraint for image segmentation [39], [47] and

for motion estimation from images [48]. The mean field approximation has been used for constrained surface reconstruction, and its use has been motivated by the fact that it is the minimum variance Bayes estimator of the true field [49].

The mean value of the estimated true segmentation at iteration k (that is, the mean of $W_{si}^{(k)}$), can be found by an embedded iteration process. It is found by initializing with the voxelwise independent estimate, and then iterating the following relation to convergence [47, page 52]

$$\overline{W}_{si}^{(k)} \leftarrow \frac{1}{Z}\exp\Bigg(\ln f(T_i = s) + \ln \prod_j f\left(D_{ij}\left|T_i = s, \boldsymbol{\theta}_j^{(k-1)}\right.\right) + \sum_m \sum_n J_{sn}\,\overline{W}_{nm}^{(k)}\Bigg)$$

with

$$\sum_s \overline{W}_{si}^{(k)} = 1$$

where $\overline{W}_{si}^{(k)}$ is the mean field estimate of the conditional probability that spatial site i has label s, J_{sn} describes the degree of spatial compatibility between label s and label n, m indexes the spatial sites in the neighborhood of site i and Z is a normalizing constant. This MRF approximation is easily computed and is suitable for any number of labels, and allows a model of spatial homogeneity to be incorporated into the iterative estimation scheme of (16), (18), and (19).

In our experiments with binary true segmentations, we have investigated an alternative MRF representation which has twin virtues. First, it computes an exact solution to the MAP estimate of the true segmentation, and second, it is very computationally efficient. With this method a single estimate of the true segmentation incorporating the MRF is obtained following convergence of the EM iteration scheme described above.

We can express the prior probability of the field $\mathbf{T}$ by taking (26), omitting the constants, and letting $V(T_i, T_j) = \beta_{mn}(T_m T_n + (1 - T_m)(1 - T_n))$, so that we have

$$\ln f(\mathbf{T}) \propto \sum_i \ln f(T_i) + \sum_m \sum_n \beta_{mn}(T_m T_n + (1 - T_m)(1 - T_n)) \quad (28)$$

where $\beta_{mn} = 0$ for voxels that are not neighbors.

Let the voxels in the neighborhood of voxel i be denoted by ∂i, note as above the neighborhood dependence of the true segmentation is independent of the single site prior probability and recall $T_i \in \{0, 1\}$.

The true segmentation $T_i, \forall i$ will be estimated by maximizing the posterior probability

$$\begin{aligned} f\left(\mathbf{T}\left|\mathbf{D}, \mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right.\right) &\propto f\left(\mathbf{D}\left|\mathbf{T}, \mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right.\right) f(\mathbf{T}) \\ &\propto \prod_i \prod_j f\left(D_{ij}\left|T_i, {p_j}^{(k)}, {q_j}^{(k)}\right.\right) f(T_i) f(T_{\partial i}) \\ &\propto \prod_i f(T_{\partial i}) f(T_i = 1)^{T_i} f(T_i = 0)^{1-T_i} \\ &\quad \cdot \left(\prod_j f\left(D_{ij}\left|T_i = 1, {p_j}^{(k)}, {q_j}^{(k)}\right.\right)\right)^{T_i} \\ &\quad \cdot \left(\prod_j f\left(D_{ij}\left|T_i = 0, {p_j}^{(k)}, {q_j}^{(k)}\right.\right)\right)^{1-T_i} \\ &\propto \prod_i f(T_{\partial i}) \left(a_i^{(k)}\right)^{T_i} \left(b_i^{(k)}\right)^{1-T_i}. \end{aligned} \quad (29)$$

Taking the logarithm of this expression for the posterior probability and assuming a pairwise interaction neighborhood structure leads to

$$\begin{aligned} \ln f\left(\mathbf{T}\left|\mathbf{D}, \mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right.\right) &\propto \sum_{m=1}^{N}\sum_{n=1}^{N} \beta_{mn}(T_m T_n + (1 - T_m)(1 - T_n)) \\ &\quad + \sum_{i=1}^{N} \ln\left(\frac{a_i^{(k)}}{b_i^{(k)}}\right) T_i \end{aligned} \quad (30)$$

where a term in (29) not depending on T_i has been ignored in (30) as it does not alter the estimate of the true segmentation, and β_{mn} is an interaction weight between voxels m and n that encodes the strength of the prior probability of true segmentations. If $\beta_{mn} > 0$ then voxels m and n are neighbors, and $\beta_{mm} = 0\ \forall m$. If $\beta_{mn} = 0, \forall m, n$, then we have the first model presented above with no neighborhood interaction prior probability. The nonzero interaction weight is often chosen as $\beta_{mn} = \beta, \forall m, n$, in a given neighborhood, or an appropriately scaled modification of this depending on the distance between voxels along each coordinate axis, and acts such that as β grows larger, the estimated true segmentation is smoother and more spatially homogeneous.

This particular MRF model is attractive because it enables us to capture the intuitive notion of a spatially correlated true segmentation, and also because it has been shown by Greig *et al.* [42] that it can be solved exactly with an efficient network flow algorithm. Greig *et al.* demonstrated the equivalence between exact MAP estimation with the above binary MRF model and a maximum flow-minimum cut network flow problem. Since efficient (polynomial in number of edges and vertices) algorithms for the minimum cut problem are known, this equivalence enables a rapid estimation of the true segmentation $T_i\ \forall i$ to be found by solving

$$\sum_i \lambda_i T_i + \sum_m \sum_n \beta_{mn}(T_m T_n + (1 - T_m)(1 - T_n)) \quad (31)$$

where

$$\begin{aligned} \lambda_i &= \ln\left(\frac{a_i^{(k)}}{b_i^{(k)}}\right) \\ &= \ln\left(\frac{W_i^{(k)}}{1 - W_i^{(k)}}\right). \end{aligned} \quad (32)$$

The equivalent network flow problem consists of $\mathrm{N} + 2$ vertices, with a source s and a sink t and the N voxels of the image. For each voxel with $\lambda_i > 0$ there is a directed edge (s, i) from source s to voxel i with capacity $c_{si} = \lambda_i$, and for each voxel with $\lambda_i < 0$ there is a directed edge (i, t) from voxel i to sink t with capacity $c_{it} = -\lambda_i$. There is an undirected edge between pairs of voxels m, n with capacity $c_{mn} = \beta_{mn}$. A cut of this network is a partitioning of the vertices into two groups,

$B = s \cup i : T_i = 1$ and $W = t \cup i : T_i = 0$ consisting of edges with a vertex in B and a vertex in W. The capacity of the cut is $C(T)$, the sum of the capacities of the edges of the cut. A minimum cut is a cut of smallest capacity. Greig *et al.* [42] showed that the minimum cut of this network is equivalent to the exact MAP solution of the binary MRF problem, since

$$C(T) = \sum_{i=1}^{N} T_i \max(0, -\lambda_i) + \sum_{i=1}^{N} (1 - T_i) \max(0, \lambda_i) + \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_{ij} (T_i - T_j)^2 \quad (33)$$

differs from (31) only by the sign and a term that does not depend upon $\mathbf{T}$. Therefore, the minimum cut network is also the MAP MRF solution.

When the minimum cut of this network is found all the voxels for which $T_i = 0$ are on the sink side of the minimum cut and all the voxels for which $T_i = 1$ are on the other side. Our implementation uses the Edmonds–Karp maximum flow-minimum cut algorithm [50] with the multiresolution solution strategy suggested by Greig *et al.* [42].

This model allows us to proscribe a spatially correlated true segmentation in our estimation framework, and creates a more locally homogeneous true segmentation estimate. A recent extension of this maximum flow-minimum cut strategy removing the restriction to binary labels has been described [43]. Thus, to model local spatial homogeneity of the true segmentation for multicategory data, a MRF model can be used and both the mean field approximation [44] and maximum flow-minimum cut exact solver are suitable [43].

F. Initialization Strategy and Detection of Convergence

The STAPLE algorithm requires specification of some model parameters, and may be initialized with either performance level parameters or a probabilistic true segmentation. We describe suitable strategies for initialization and for detection of convergence of the estimation process below.

1) Initialization: STAPLE is most conveniently initialized by providing starting estimates for the sensitivity and specificity parameters. In the absence of other information regarding the relative quality of the experts or the true segmentation, initializing the sensitivity and specificity parameters to the same value and equal across all raters is recommended. This is equivalent up to a scaling factor to forming an initial estimate through a voting rule (such as assigning initial probabilities of voxels based on the frequency of selection by experts). In our experience with synthetic data and clinical experiments, successful estimation results were obtained by initializing all of the sensitivity and specificity parameters to the same values, with values close to but not equal to 1 : for example, selecting $p_j^{(0)} = q_j^{(0)} = 0.999\,99\ \forall j$.

An alternative strategy for initialization, useful when such information is available, is to provide an initial true segmentation estimate. An interesting strategy for certain problems is to use a probabilistic atlas to provide an initial true segmentation. For example, one may use a probabilistic atlas of the brain [37], [51] to provide an initial true segmentation estimate for the tissues of the brain.

We have used the estimated true segmentation from the voxelwise independent STAPLE to initialize the STAPLE algorithm when incorporating an MRF prior.

2) Convergence: The EM algorithm is guaranteed to converge to a local optimum. Detecting convergence of EM algorithms has been a topic of study, as have strategies for accelerating convergence [32]. STAPLE estimates both performance parameters and the true segmentation, and convergence may be detected by monitoring these. A simple and good measure of convergence is the rate of change of the sum of the true segmentation probability, obtained by summing the estimated weight at each voxel $S_k = \sum_{i=1}^{N} W_i$ evaluated at each iteration k. We have found that iterating until $S_k - S_{k-1} = 0$ using double precision IEEE arithmetic is both accurate and fast. In the case of L unordered multicategory labels and large three-dimensional (3-D) volumes, we have found it convenient to use a relative convergence based on the change in the normalized trace of the estimated expert performance parameters. We iterate until the normalized trace changes by less than some small amount, for example $\epsilon = 1 \times 10^{-7}$, where the normalized trace is given by

$$\frac{1}{\mathrm{LR}} \sum_{j=1}^{R} \mathrm{tr}(\theta_j). \quad (34)$$

Our experience has been that STAPLE convergence generally occurred in less than 20 iterations in several clinical segmentation validation problems.

3) Model Parameters: Selection of different global or spatially varying prior probability $f(T_i = 1)$ may alter the local maxima to which the algorithm converges. This is illustrated below in experiments with synthetic data.

A spatially varying prior $f(T_i = 1)$ is suitable for those structures for which a probabilistic atlas is available. In addition, we are interested in structures for which such an atlas is not available, and we have used a single global prior $\gamma = f(T_i = 1)\ \forall\, i$. The prior γ encodes information we have before seeing the segmentation decisions, regarding the relative probability of the structure we wish to segment in the data. In practice, this information may not be readily available, and we have found it convenient to estimate γ from the segmentations themselves as the sample mean of the relative proportion of the label in the segmentations

$$\gamma = \frac{1}{RN} \sum_{j=1}^{R} \sum_{i=1}^{N} D_{ij}. \quad (35)$$

In the case of multicategory labels, we have

$$\gamma_s = \frac{1}{RN} \sum_{j=1}^{R} \sum_{i=1}^{N} \delta(D_{ij}, s) \quad \forall s. \quad (36)$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta which takes the value 1 when D_{ij} and s are equal and 0 otherwise.

Similarly, when modeling a spatially homogeneous true segmentation with our MRF model, an interaction strength β_{mn} and neighborhood structure must be specified. This may depend strongly upon the nature of the structure, segmentations and noise pattern of the segmentations under consideration. In our experiments with synthetic data and strong uncorrelated random noise in the segmentations and a strongly homogeneous true

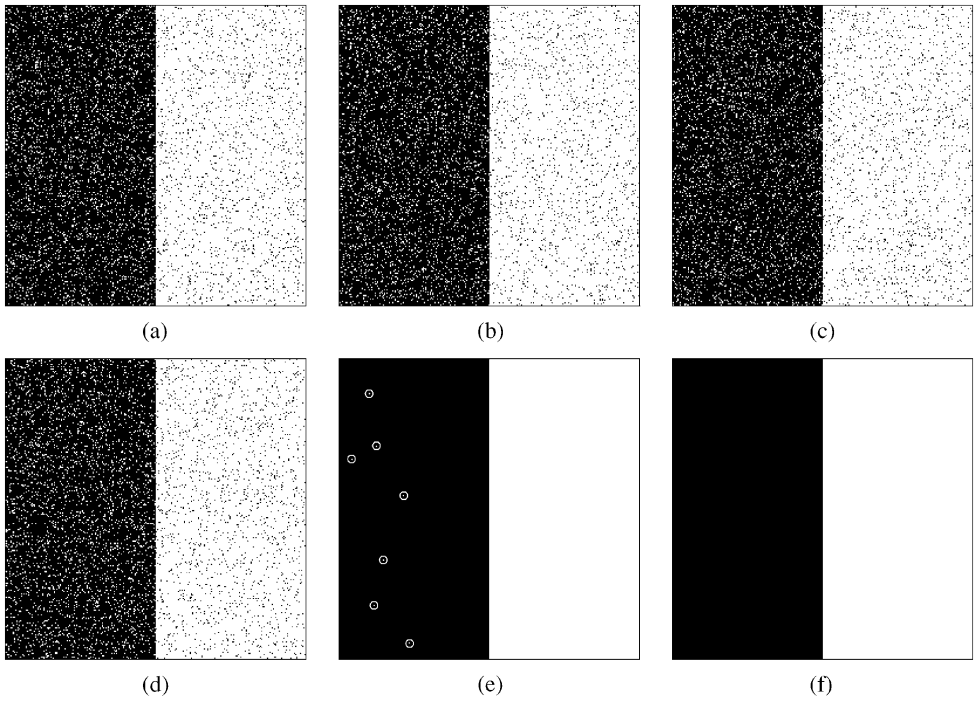


Fig. 1. A specified true segmentation was randomly sampled to generate $\mathrm{R} = 10$ segmentations with $(p_j, q_j) = (0.95, 0.90)\ \forall j \in 1, \ldots, 10$. STAPLE converged in less than 20 iterations. The estimated performance parameters were $\hat{p} = 0.950\,104 \pm 0.001\,201$ (mean $\pm$ standard deviation) and $\hat{q} = 0.900\,035 \pm 0.001\,685$, which closely matches the specified parameters. (a) Example of synthetic segmentation. (b) Example of synthetic segmentation. (c) Example of synthetic segmentation. (d) Example of synthetic segmentation. (e) STAPLE estimated true segmentation. The estimate contains seven isolated single pixel errors which are circled to enable clearer visualization. (f) Estimated true segmentation from STAPLE incorporating an MRF prior. The estimate exactly matches the known true segmentation.

segmentation, we found $\beta_{mn} = \beta = 2.5$ gave satisfactory results.

III. Results

We carried out experiments applying STAPLE to digital phantoms where the true segmentation is known by construction and investigated the characterization of segmentations by synthetic experts with predefined sensitivities and specificities. As described below, we found that STAPLE accurately estimates the performance level parameters and the true segmentation. We applied STAPLE to segmentations from several clinical applications to demonstrate the value of the approach in practice, and to compare to a previously reported validation measure. We assessed repeated segmentations by a single expert for the task of identifying the prostate peripheral zone from magnetic resonance images. We compared the STAPLE assessment of the segmentations to a measure of spatial overlap. A brain phantom was used to compare STAPLE to a voting rule. Synthetic segmentations and segmentations by medical students were evaluated.

The capacity of STAPLE to accurately estimate the true segmentation, even in the presence of a majority of raters generating correlated errors, was demonstrated. The quantitative assessment of segmentations of tissues from MRI of a newborn infant was carried out in order to demonstrate the use of STAPLE with 3-D volumetric unordered multicategory data. Empirical timing experiments were also carried out to demonstrate the practicality of the estimation scheme. In all of our experiments, the hardware platform for the assessment of execution time was a Dell Precision 650n workstation running Red Hat Linux 8.0, with the implementation executing on a 3-GHz Intel Xeon CPU.

A. Digital Phantom Experiments

1) Estimation From Noisy Segmentations With Specified Segmentation and Performance Parameters: Fig. 1 illustrates STAPLE estimation from digital phantoms drawn from randomly generated segmentations with specified sensitivity and specificity rates and specified true segmentation. The true segmentation was set to be a square image, 256×256 voxels, with the leftmost 128 columns having value 0 and the rightmost 128 columns having value 1. The prior probability of the foreground class is $f(T_i = 1) = 0.5$. $R = 10$ segmentations were generated by randomly sampling the specified true segmentation with $(p_j, q_j) = (0.95, 0.90)$. This generated ten random segmentations simulating experts with identical sensitivity and specificity rates. STAPLE required 0.18 s wallclock time to estimate the true segmentation and expert parameters, and the estimates closely matched those specified. Estimated sensitivity over the experts was $\hat{p} = 0.950\,104 \pm 0.001\,201$ (mean $\pm$ standard deviation) and sensitivity $q = 0.900\,035 \pm 0.001\,685$. The estimated true segmentation closely matched the specified true segmentation, with the foreground structure being exactly estimated from raters with a detection error rate of 5% and with $(7/32\,768)$ false positives in a region with a detection error rate of 10%. The incorporation of an MRF prior model with four-connected neighborhood structure and an interaction strength $\beta = 2.5$ resulted in a STAPLE estimate converging exactly to the known true segmentation. The MRF model was applied to

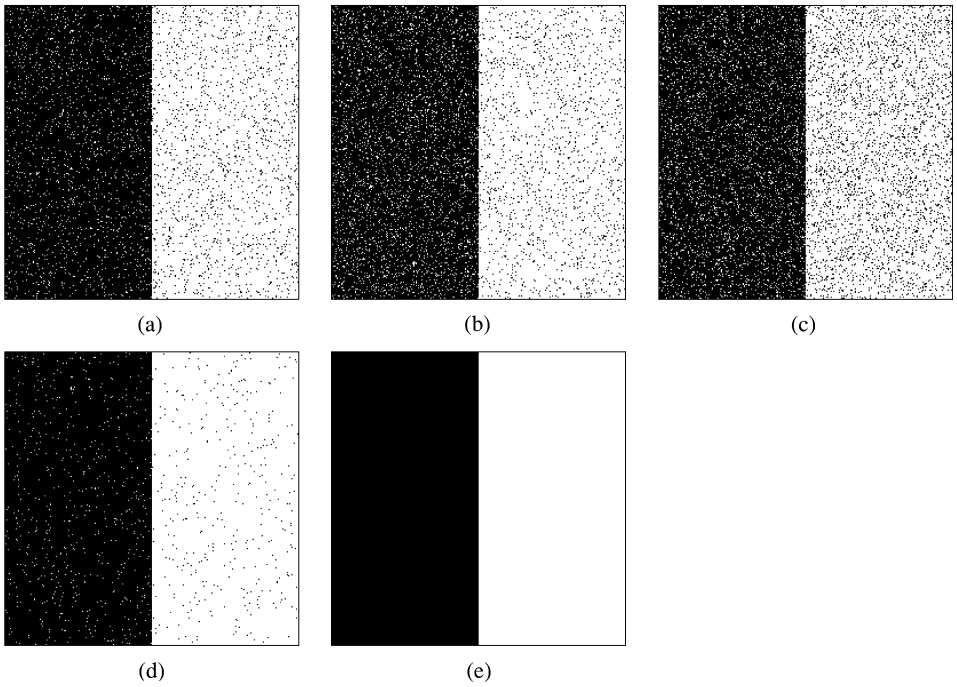


Fig. 2. Segmentations generated from $R = 3$ synthetic experts with parameters specified as $(p_1, q_1) = (0.95, 0.95), (p_2, q_2) = (0.95, 0.90)$, and $(p_3, q_3) = (0.90, 0.90)$. Only three observations of segmentations by these experts were generated, leading to a small and noisy set of data from which STAPLE was used to estimate the performance parameters and the true segmentation. STAPLE was executed to convergence, initialized with $f(T_i = 1) = 0.5$ and $(p_j^{(0)}, q_j^{(0)}) = (0.999\,99, 0.999\,99), \forall j$. Comparisons were made with and without a spatial homogeneity prior modeled with an MRF prior assuming four-nearest-neighbor pairwise interaction cliques and homogeneous interaction strength $\beta = 2.5$. The results indicate the estimated $\mathbf{T}$ found by STAPLE with the MRF prior exactly matches the specified true segmentation used for the simulations, whereas without this constraint the estimated $\mathbf{T}$ is somewhat noisier. In both cases, the estimated performance level parameters were very close to the parameters specified for the random segmentations. (a) Expert 1 segmentation. (b) Expert 2 segmentation. (c) Expert 3 segmentation. (d) STAPLE true segmentation estimate under voxelwise independence assumption. (e) STAPLE true segmentation estimate assuming spatially homogeneous true segmentation.

estimate the true segmentation by solving (33) once only, using the estimated true segmentation probabilities obtained from iterating STAPLE without an MRF model to convergence, and required 3.5 s to compute the exact MAP solution. The false positives that occur without the MRF prior are indicated with circles in Fig. 1(e) to enhance their visibility, as they are single isolated pixels which are difficult to see in the reproduction of the figure, and to illustrate that there are no false positives in Fig. 1(f).

Fig. 2 examines the application of STAPLE to noisy segmentations from unequal quality experts. Segmentations were generated from $R = 3$ synthetic experts with parameters specified as $(p_1, q_1) = (0.95, 0.95), (p_2, q_2) = (0.95, 0.90)$, and $(p_3, q_3) = (0.90, 0.90)$. Only three observations of segmentations by these experts were generated, leading to a small and noisy set of data from which STAPLE was used to estimate the performance parameters and the true segmentation. STAPLE was executed to convergence, requiring 0.22 s wallclock time. Comparisons were made with and without a spatial homogeneity prior modeled with an MRF prior assuming four-nearest-neighbor pairwise interaction cliques and a homogeneous interaction strength $\beta = 2.5$. The MRF model was applied to estimate the true segmentation by solving (33) once only, using the estimated true segmentation probabilities obtained from iterating STAPLE without an MRF model to convergence and required 53 s to compute the exact MAP solution. The increased run time is due to the greater amount of heterogeneity in the initial STAPLE estimate.

The results indicate the estimated $\mathbf{T}$ found by STAPLE with the MRF prior exactly matches the specified true segmentation used for the simulations, whereas without this constraint the estimated $\mathbf{T}$ is somewhat noisier. In both cases, the estimated performance level parameters were almost identical to the specified values of the parameters.

2) Influence of Prior Probability: Fig. 3 illustrates synthetic segmentations which differ by structural errors rather than random noise. The true segmentation was specified, as were synthetic expert segmentations, each having equal volume but different spatial location. STAPLE was initialized with $p_j^0 = q_j^0 = 0.99, \forall j$ and required 0.38 s to converge. The STAPLE true segmentation estimate is shown for different prior assumptions. The prior models information available before the data is seen. Three prior models are shown, first assuming global $f(T_i = 1) = 0.12$ which closely matches the segmentations, with $f(T_i = 1) = 0.5$ which corresponds to a prior belief that half of the field of view should be the foreground class, and with automatic estimation of the global prior using (35). As can be seen, the estimated true segmentation can be influenced by the global prior probability.

In the case of assuming $f(T_i = 1) = 0.12$, the estimated true segmentation exactly matches the specified true segmentation and one of the experts was identified as generating exactly the true segmentation $(\mathrm{p} = \mathrm{q} = 1.0)$, and the other two were estimated to have a sensitivity of 0.88 and a specificity of 0.99 reflecting the erroneous shift in these two segmentations. This result was also obtained with automatic estimation of the prior.

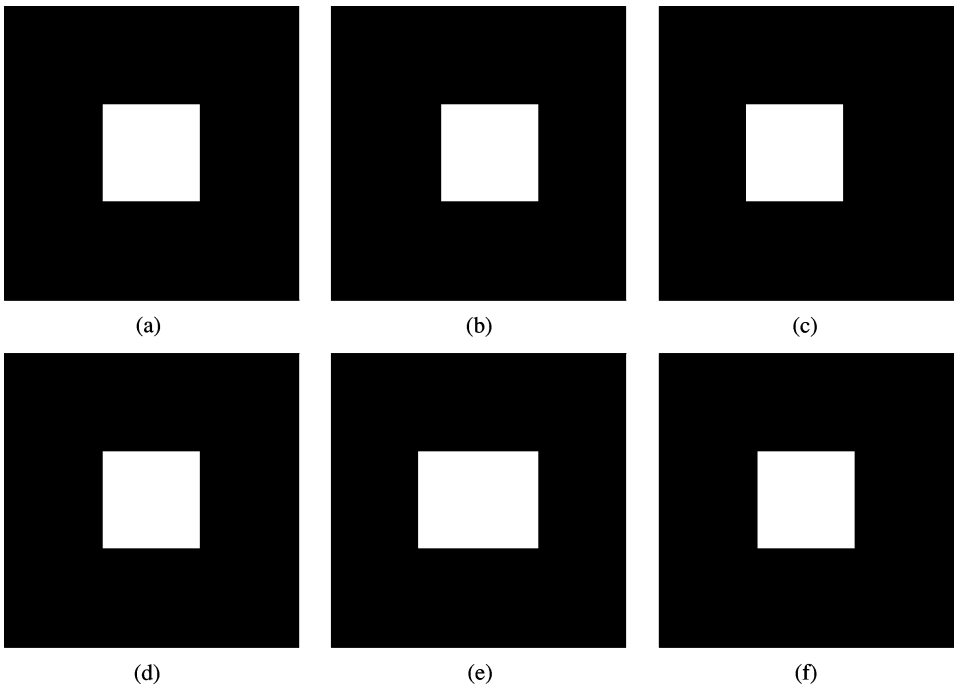


Fig. 3. $R = 3$ synthetic true segmentations with equal volume but different spatial locations. STAPLE was initialized with $(p_j^{(0)}, q_j^{(0)}) = (0.99, 0.99), \forall j$. The STAPLE true segmentation estimate is shown for the different prior probability assumptions: $f(T_i = 1) = 0.12$ which closely matches the segmentations, $f(T_i = 1) = 0.5$ which corresponds to a prior belief that half of the field of view should be the foreground class, and with automatic estimation of the prior via (35). (a) First synthetic segmentation. (b) Second synthetic segmentation, equal in size to the first but shifted to the left 10 voxels. (c) Third synthetic segmentation, equal in size to the first but shifted to the right ten voxels. (d) STAPLE true segmentation estimate for $f(T_i = 1) = 0.12\,\forall i$. The estimated performance parameters were $(\widehat{p_1}, \widehat{q_1}) = (1.0, 1.0), (\widehat{p_2}, \widehat{q_2}) = (\widehat{p_3}, \widehat{q_3}) = (0.88, 0.99)$. (e) STAPLE true segmentation estimate for $f(T_i = 1) = 0.5\,\forall\, i$. The estimated performance parameters were $(\widehat{p_j}, \widehat{q_j}) = (0.66, 1.0)\,\forall\, j$. (f) STAPLE true segmentation estimate with automatic prior estimation. The estimated performance parameters were $(\widehat{p_1}, \widehat{q_1}) = (1.0, 1.0), (\widehat{p_2}, \widehat{q_2}) = (\widehat{p_3}, \widehat{q_3}) = (0.88, 0.99)$.

In the case of assuming $f(T_i = 1) = 0.5$, the estimated true segmentation equals the union of each of the three segmentations (which still does not occupy 50% of the image) and each of the experts was identified as having $(p_j, q_j) = (0.66, 1.0), j = 1, 2, 3$. This reflects the fact that the segmentations are not in agreement with the assumed global prior. If this was discovered in the application of STAPLE to a clinical validation problem it would indicate either the need to re-evaluate the global prior assumption or the need for improved training of the experts generating the segmentations.

B. Brain Phantom Experiments

Consider estimation of a true segmentation and performance parameters from segmentations generated by two groups of raters, M of whom have significant expertise, and N of whom have limited expertise. We can consider two important cases: where $M > N$ and where $M < N$. If we have $M > N$ then the errors of the inexperienced raters do not exert significant influence and are overcome by the correct information from the more experienced raters, whereas by comparison, if $M < N$ then there is the possibility that the inexperienced raters may influence the estimation scheme to overcome the more experienced raters.

In the case, $M < N$, if the inexperienced raters make random errors, then as indicated by Fig. 2, accurate estimation can still be achieved. Similarly, if they make uncorrelated structural errors then, as for random errors, accurate estimation is possible. However, if they make correlated structural errors the impact of the errors could be significant. Without further external information, it is not possible to discriminate between the two groups, and to decide which constitutes the expert raters and which the inexperienced raters. For example, in a majority voting scheme, the inexperienced raters would collect the most votes for the regions with correlated errors and hence a wrong segmentation estimate would result. External information is incorporated in the STAPLE estimate through the prior probability for the structure being segmented, and through the initialization. It is also straightforward to introduce a prior probability density for the rater parameters, enabling expert raters to be distinguished from inexperienced raters via parameters of such a prior distribution.

We carried out an experiment to investigate the performance of the STAPLE algorithm when presented with $M = 1$ and $N = 3$ raters, where the task is brain cortical gray matter segmentation. We used a brain phantom created from a high-resolution high signal-to-noise ratio brain MRI scan, for which a consensus segmentation of the cortical gray matter was available. The segmentation was made by automatic segmentation and consensus correction of errors by experts [3]. We used the consensus segmentation to represent an exact ideal "expert" segmentation that we want STAPLE to identify, and we deleted all cortex labels from half of the rows of the exact segmentation and triplicated this in order to represent poor raters generating approximately half of the correct labeling of the cortex, and incorrectly labeling the other half of the cortex as background. Triplicating this data ensures that the inexperienced rater segmentations have correlated errors.

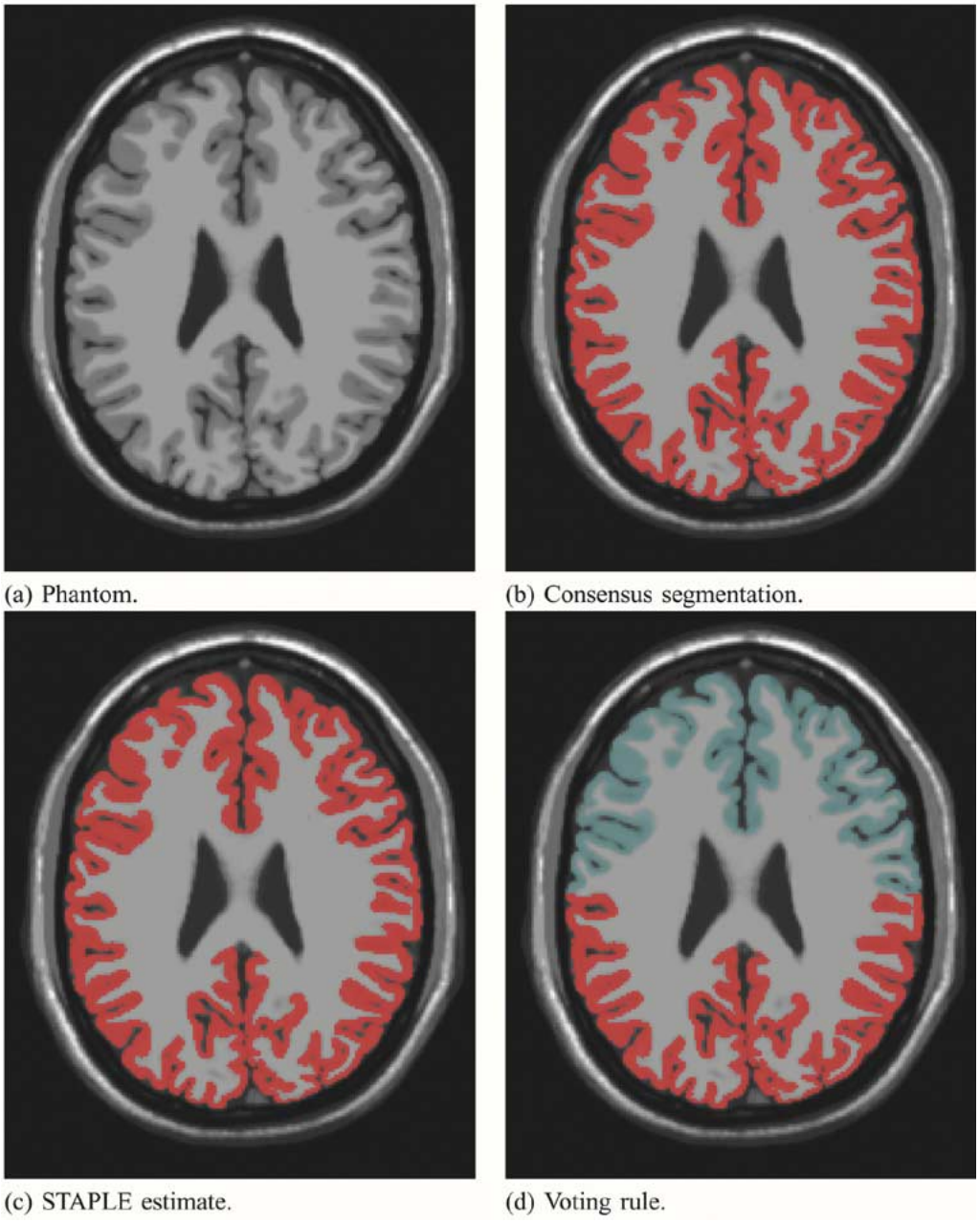


Fig. 4. Brain phantom and STAPLE estimated true segmentation with one expert and 3 inexperienced rater segmentations. Different initialization leads to different estimates. Exact estimation of the true segmentation is possible with STAPLE but not with a majority vote rule. The image that was segmented is shown in (a). The red color indicates the segmentation of the cortex from the consensus segmentation in (b), and from the STAPLE estimate in (c). In (d), the area of most frequent selection by the raters and expert is shown in red, and the light blue color represents the region selected only by the expert.

We ran the STAPLE algorithm on these four segmentations, to convergence, stopping when the change in normalized trace was less than 1×10^{-6}, and evaluated the effect of different initialization. We provided a range of different initial values for the sensitivity and specificity of each of the four input segmentations, setting the "expert" segmentation to have high initial sensitivity and specificity $(p_0^{(0)} = q_0^{(0)} = 0.999\,999\,999\,9)$, and then varying the initial parameter assignments of the other three raters (all equal) over the range of 0.05 to 0.95 in steps of 0.05, and automatic estimation of all other parameters with a uniform spatial prior for the cortex and background.

We found STAPLE identified one of two different estimated true segmentations, depending on the initialization. The first closely matched the exact segmentation, and the second closely matched the inexperienced rater segmentations. If the initialization indicated the "expert" segmentation was superior to the three inexperienced raters $(p_j^{(0)} = q_j^{(0)} \in [0.1, 0.9], j = 1, 2, 3)$ it converged to match the experienced rater, and if the inexperienced raters were initialized outside this range (for example, close in performance to the "expert"), it converged to the estimate close to that of the inexperienced raters. The phantom, consensus segmentation, STAPLE estimate of the true segmentation and voting rule estimate are illustrated in Fig. 4.

Since the parameter estimates are obtained by maximizing the conditional expectation of the complete data log likelihood function, $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$, some insight into the convergence be-

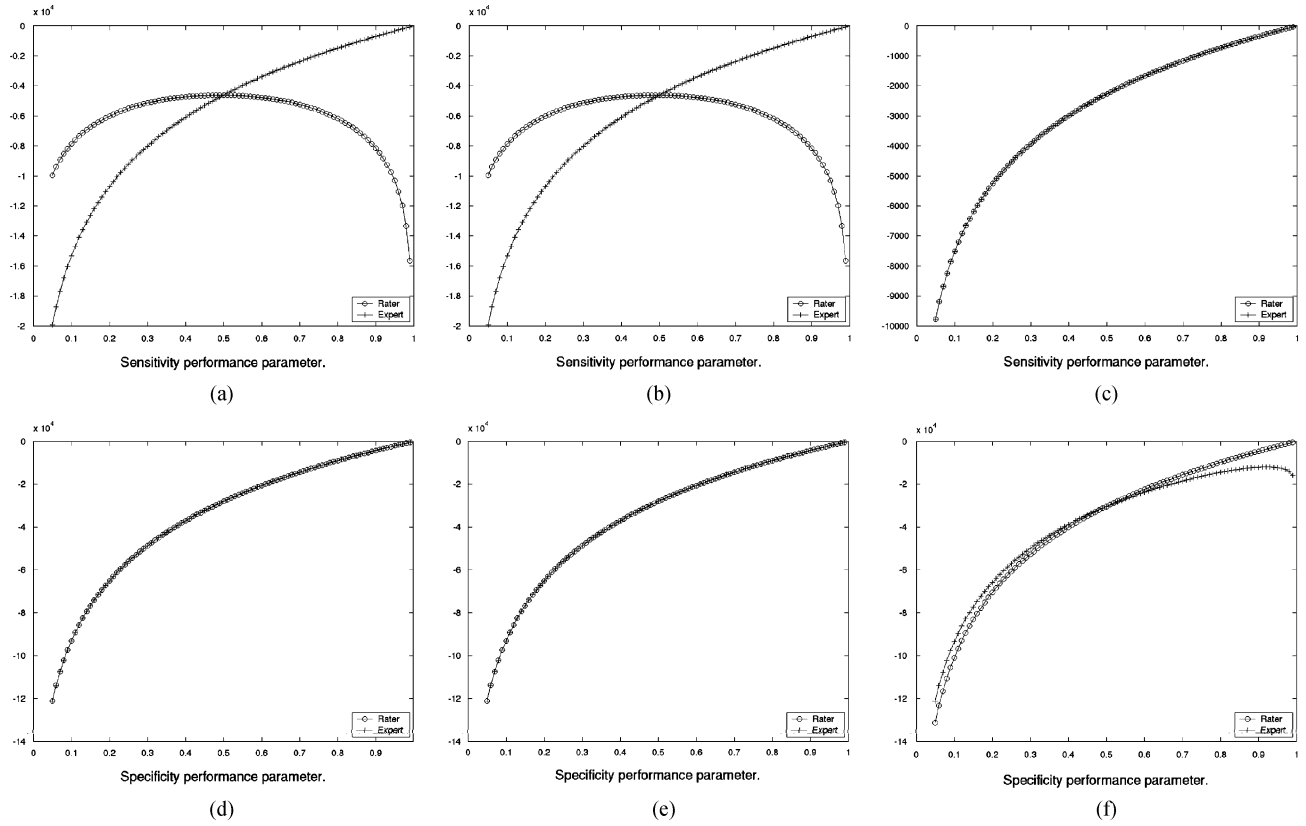


Fig. 5. Comparison of exact and estimated complete data log likelihood function for estimation of rater sensitivity and specificity, derived from the brain phantom with one expert and 3 inexperienced rater segmentations. (a) Exact $\ln f(\mathbf{D}, \mathbf{T} \mid \boldsymbol{\theta})$ dependence on sensitivity for inexperienced rater and expert. (b) Estimated function of sensitivity from $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}$ after convergence from unequal rater initialization. (c) Estimated function of sensitivity from $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ after convergence from equal rater initialization. (d) Exact $\ln f(\mathbf{D}, \mathbf{T} \mid \boldsymbol{\theta})$ dependence on specificity for inexperienced rater and expert. (e) Estimated function of specificity from $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ after convergence from unequal rater initialization. (f) Estimated function of specificity from $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ after convergence from equal rater initialization.

havior of STAPLE in this setting can be drawn by comparing the estimate of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ to the true complete data log likelihood function that can be calculated when the exact true segmentation is known. The estimate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ evolves until the iteration converges, and so in Fig. 5 the estimated function is shown as estimated at each of the two points where the algorithm converges. Shown in this figure is only the part of the estimated function that is necessary to find the sensitivity or the specificity parameters for one of the input segmentations at a time, as this depends on only one variable and is, therefore, simple to graphically represent. Such estimated functions are shown for both for the experienced rater and the inexperienced raters, together with the corresponding function from the exact complete data log likelihood. Fig. 5 shows that when the initialization indicates the experienced rater is superior to the inexperienced raters, the estimated complete data log likelihood exactly matches the true complete data log likelihood at convergence, and that if the initialization indicates all raters are equal then the estimated complete data log likelihood indicates all raters have equal sensitivity and that the experienced rater segmentation corresponds to an oversegmentation, with worse specificity, as compared to the estimated true segmentation. Hence, in the presence of a majority of raters generating repeated structural errors, appropriate initialization or further external information through prior distributions on rater parameters or the spatial distribution of the true segmentation may be used to ensure STAPLE converges to the correct true segmentation estimate. Straightforward vote counting cannot overcome the presence of a majority of raters generating repeated structural errors. It is expected that this situation is atypical, and represents a worst case for estimation of the true segmentation, and can be corrected by improved training or algorithmic modifications to alter the behavior of the raters.

TABLE I
COMPARISON OF STAPLE AND VOTING RULE ESTIMATES OF RATER PERFORMANCE. THE STAPLE ESTIMATE CORRECTLY IDENTIFIES THE EXPERT AS HAVING SUPERIOR PERFORMANCE TO THAT OF THE THREE MEDICAL STUDENTS

	Students			
Segmentations	1	2	3	expert
Estimate of rater performance from a voting rule.				
$\hat{p}$	0.88660	0.91366	0.86856	0.79768
$\hat{q}$	0.99940	0.99922	0.99946	0.99972
STAPLE estimate of rater performance.				
$\hat{p}$	0.88095	0.91667	0.88690	1.0
$\hat{q}$	0.99890	0.99872	0.99904	1.0
Exact rater performance from consensus segmentation.				
p	0.88095	0.91667	0.88690	1.0
q	0.99890	0.99872	0.99904	1.0

C. Illustration of Clinical Validation Applications

1) Comparison of STAPLE With Voting: Three medical students carried out interactive segmentation of the cortical gray

Fig. 6. Comparison of STAPLE and voting rule estimates of the true segmentation from segmentations of the cortex generated by one expert and three medical students. The color coding of the frequency of selection is as shown in Fig. 7.

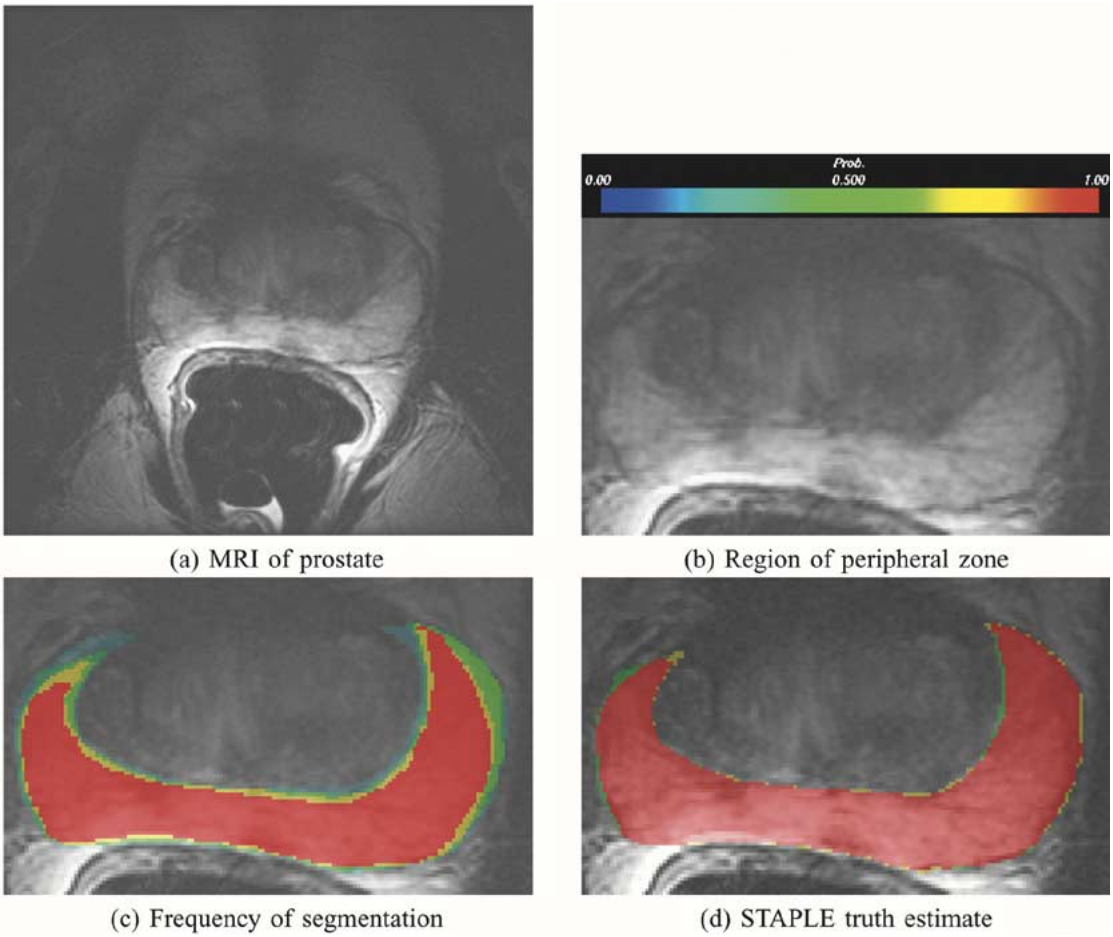


Fig. 7. This figure illustrates in (a), MRI of the prostate, in (b) the prostate peripheral zone, in (c) the frequency of assignment of voxels to the prostate peripheral zone in five repeated segmentations by the same rater, and in (d), the probabilistic true segmentation as estimated by STAPLE.

matter of a single slice of a brain phantom [3]. Upon review and visual comparison of the segmentations, it was found that over most of the cortex the segmentations were quite similar to the exact segmentation but in one region the students had all made a misjudgement and incorrectly labeled one region as cortical gray matter when it should be labeled white matter. We know that this is a misinterpretation on the part of the students since the region has signal intensity characteristics different from cortical gray matter and our prior knowledge of anatomy indicates the cortex does not extend through this region. The consensus segmentation correctly labels this region and we used this consensus segmentation to represent a perfect expert. We estimated the true segmentation from the expert and three medical student segmentations, by both vote counting and by STAPLE. Table I quantifies the difference in performance estimates for this region between a vote counting estimate and the STAPLE algorithm. The region is shown in Fig. 6 together with the exact segmentation, an image showing the frequency of selection of each voxel as cortex by the medical students, an image showing the frequency of selection of each voxel when combining the medical students and exact segmentation and the STAPLE estimate of the true segmentation. The figure illustrates that the students incorrectly label some of the white matter as cortical gray matter, and that when vote counting is used to estimate the true segmentation the frequent wrong selection overrules the expert. This does not occur with the STAPLE estimate. The color coding of the frequency of selection is as shown in Fig. 7.

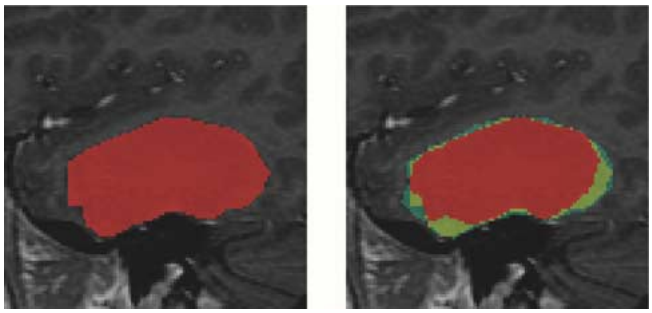
2) Segmentation of Prostate Peripheral Zone: Fig. 7 illustrates STAPLE applied to the analysis of five segmentations by one expert of the peripheral zone of a prostate as seen in a conventional MRI scan (T2w acquisition, $0.468\,750 \times 0.468\,750 \times 3.0$ mm^3). The goal of the operator was to segment the prostate peripheral zone to enable radiation dose planning in support of brachytherapy [52]. The STAPLE algorithm ran to convergence in 0.48 s of wallclock time.

TABLE II
QUALITY ESTIMATES FOR THE FIVE PROSTATE SEGMENTATIONS GENERATED BY ONE EXPERT, AND DICE SIMILARITY COEFFICIENT (DSC), $\mathrm{DSC} \equiv (2|A \cap B|/|A| + |B|)$ WITH $|A|$ THE AREA OF REGION A, COMPARING THE EXPERT SEGMENTATION WITH THE TRUE SEGMENTATION ESTIMATE THRESHOLDED WITH $W_i \geq 0.5$. RESULTS ARE SHOWN FOR TWO DIFFERENT ASSUMPTIONS OF THE PRIOR PROBABILITY OF PROSTATE PERIPHERAL ZONE AND DEMONSTRATE THAT THE PARAMETER ESTIMATES OBTAINED DO NOT DEPEND STRONGLY ON THE PRIOR PROBABILITY ASSUMPTION

Expert segmentations	1	2	3	4	5
$f(T_i = 1) = 0.10$					
$\hat{p}$	0.875090	0.987198	0.921549	0.907344	0.880789
$\hat{q}$	0.999163	0.994918	0.999435	0.999739	0.999446
DSC	0.927660	0.957527	0.954471	0.949058	0.932096
$f(T_i = 1) = 0.02$					
$\hat{p}$	0.878533	0.991261	0.936831	0.918336	0.894861
$\hat{q}$	0.998328	0.993993	0.999320	0.999359	0.999301
DSC	0.913083	0.951027	0.967157	0.954827	0.944756

For comparison with the STAPLE estimate of the true segmentation, an image color coded by probability of selection of each voxel is also presented. The colors range from blue to red corresponding to probability of selection ranging from 0 to 1 as indicated by the color bar in Fig. 7(b). It is interesting to consider a voting rule scheme and to compare that with the STAPLE result. One voting rule is to estimate the true segmentation by taking all voxels where a majority of the segmentations indicated the foreground was present. We achieved a similar binarization of the STAPLE probabilistic true segmentation estimate, by thresholding at $W_i = 0.5$, which gives a ML estimate of the binary true segmentation. An alternative is to apply the MAP estimation using the binary MRF model of (33) which also provides a binary estimate of the true segmentation. We found that the binarized STAPLE true segmentation estimate cannot be obtained with a simple voting rule estimate from the segmentations. In order to obtain the binarized STAPLE result directly from the manual segmentation, it would be necessary to identify all the voxels indicated as prostate peripheral zone by at least three of the five segmentations, and then all the voxels indicated by two of the high-quality segmentations but not those voxels indicated by only two of the lower quality segmentations. Without knowledge of the relative quality of the segmentations such a weighted combination cannot be constructed. Thus, the true segmentation estimate depends crucially upon the relative quality of the segmentations that is discovered by the STAPLE algorithm. The STAPLE result cannot be generated by a simple voting rule such as selecting voxels indicated by three out of five segmentations.

3) Comparison of STAPLE With Dice Similarity Coefficient: The performance level estimates for segmentation of the prostate peripheral zone are recorded in Table II. In order to enable comparison of these estimates with the Dice Similarity Coefficient (DSC) [11], a popular measure of spatial overlap defined as $\mathrm{DSC} \equiv (2|A \cap B|/|A| + |B|)$ where $|A|$ is the area of region A, $|B|$ is the area of region B, we formed a binary estimate of the true segmentation by thresholding at $W_i = 0.5$. We then computed the DSC between each of the individual segmentations and the binarized STAPLE true segmentation. Comparing segmentations 2 and 4, we observe the DSC is similar, approximately 0.95 in each case. However, the STAPLE performance parameter estimates indicate the two segmentations achieve this DSC differently—for example, segmentation 2 has higher sensitivity than segmentation 4, but lower specificity. This indicates that despite the similar DSC values, segmentation 2 is an over-estimate of the peripheral zone, and segmentation 4 is an under-estimate of the peripheral zone.



(a) (b)

Expert	$\hat{p}$	$\hat{q}$	PV
1	0.8951	0.9999	0.998
2	0.9993	0.9857	0.977
3	0.9986	0.9982	0.954
Program	0.9063	0.9990	0.963

(c)

Fig. 8. This illustrates the segmentation of a brain tumor from MRI. Three experts carried out segmentation of the brain tumor, and STAPLE was used to estimate the true segmentation of the tumor. The performance level assessment of each of three raters and a semi-automatic algorithm for tumor segmentation program based upon the estimated true segmentation is shown in (c). The estimated sensitivity, specificity, and tumor predictive value are reported. The performance assessment indicates that the program is performing with higher sensitivity than one rater but with lower sensitivity than the other raters, while exhibiting higher specificity than two of the raters and lower specificity than one of the raters. This illustrates that STAPLE can be used to evaluate the performance of segmentation algorithms through comparison to rater segmentations. (a) Estimated true segmentation. (b) Frequency of manual segmentation. (c) Performance level assessment.

4) Assessment of an Automated Segmentation Algorithm: Fig. 8 illustrates STAPLE applied to the analysis of expert segmentations of a brain tumor [53], [54]. The hidden true segmentation was estimated from three expert segmentations, requiring 0.15 s to compute. The segmentation generated by the program was then assessed by evaluating the sensitivity and specificity estimators of (18) and (19) with regard to this estimated true segmentation. The predictive value for tumor segmentation was also calculated. It is defined as $\mathrm{PV}(s) = \Pr(T_i = s \,|\, D_{ij} = s), \forall s \in 0, 1, \ldots, L-1$, the posterior probability that the structure s is present when the rater says it is present, and is easily computed with the STAPLE estimate of the rater performance parameters. This assessment indicates the program is generating segmentations similar to that of the raters, with higher sensitivity and predictive

TABLE III
ASSESSMENT OF VARIABILITY OF TISSUE CLASSIFICATION FROM NEONATE MRI. THE PREDICTIVE VALUE FOR EACH TYPE OF TISSUE AS DETERMINED FROM THE STAPLE ESTIMATES, AND THE MEAN PREDICTIVE VALUE, ARE REPORTED FOR EACH OF EIGHT SEGMENTATIONS. THE MEAN PREDICTIVE VALUE MAY BE USED TO RANK THE QUALITY OF THE SEGMENTATIONS. TISSUE TYPES ARE BACKGROUND (B), EXTRA-CRANIAL TISSUE (ECT), CORTICAL GRAY MATTER (CGM), CEREBROSPINAL FLUID (CSF), MYELINATED WHITE MATTER (MWM), UNMYELINATED WHITE MATTER (UWM), AND SUBCORTICAL GRAY MATTER (SCG)

Segmentation	B	ECT	CGM	CSF	MWM	UWM	SCG	mean PV
1	0.9999	0.6219	0.8013	0.7035	0.3648	0.9403	0.9097	0.7631
2	1.000	1.000	0.9990	1.000	0.6336	0.9456	0.9990	0.9396
3	1.000	1.000	0.9990	0.9870	0.6339	0.9888	0.9967	0.9436
4	1.000	0.8250	0.6493	0.8853	0.4974	0.9998	0.8076	0.8092
5	1.000	0.9586	0.9816	1.000	1.000	0.9967	0.9440	0.9830
6	1.000	1.000	0.9997	1.000	0.9997	0.9846	0.9692	0.9933
7	1.000	1.000	0.9186	1.000	0.9997	0.9977	0.5000	0.9166
8	1.000	1.000	0.7913	1.000	0.9930	0.9863	0.9889	0.9656

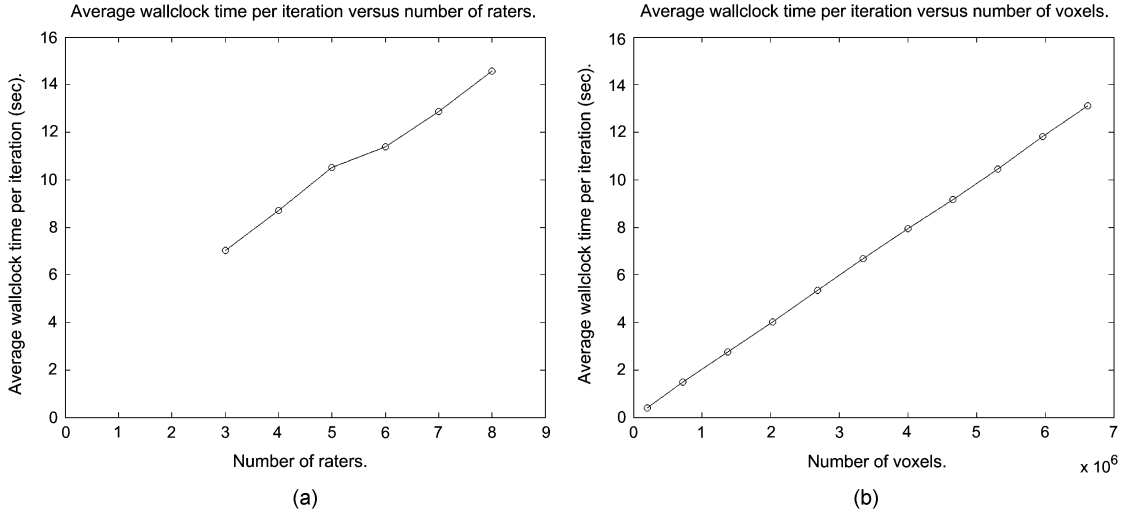


Fig. 9. STAPLE average wallclock time per iteration as a function of the number of raters, and average wallclock time per iteration as a function of the number of voxels. (a) Average time per iteration versus number of raters. (b) Average time per iteration versus number of voxels.

value than one of the raters, but with lower sensitivity and lower predictive value than the two other raters. This provides feedback indicating in what way the algorithm or algorithm parameters need to be modified to improve the performance of the algorithm.

D. Assessment of Supervised Tissue Classification of the Neonate Brain

The assessment of tissue types, including cortical gray matter, subcortical gray matter, myelinated white matter, unmyelinated white matter, and cerebrospinal fluid, from volumetric MRI of newborn infants [54]–[56] has been useful in understanding and characterising normal brain development and the processes of maturation [56], and the impact of injury upon brain development [57], [58].

We evaluated the impact of repeated selection of tissue type prototypes upon the classification of an MRI of a newborn infant at term equivalent age. A single operator repeatedly selected sample voxels from the image in order to carry out nonparametric estimation of the class-conditional probability density functions, which are then used to carry out segmentation as previously described [54]. We selected a 3-D volumetric MRI scan of the brain of a newborn infant which contained significant signal intensity artifact. We evaluated the amount of variability due to operator selection of training points in this particularly challenging scan, as the operator repeated the segmentation attempting to best identify each of background, extracranial tissue (such as skin, muscle, etc.), cortical gray matter, subcortical gray matter, myelinated white matter, unmyelinated white matter and cerebrospinal fluid as in [54], [56], [57]. These different tissues are distinguished by subtle signal intensity changes apparent from joint analysis of T1w, PDw, and T2w MRI. They also appear with unequal proportions and some are more difficult to identify than others (for example, myelinated white matter is more difficult to identify than cerebrospinal fluid).

Table III illustrates the assessment of neonate tissue classification, assessing the identification of each type of tissue using predictive value (posterior probability). The mean predictive value is a simple measure with which to rank the segmentations based on overall quality.

We also assessed the computational requirements of the STAPLE algorithm. The E-step of (20) is linear in the number of voxels N, linear in the number of raters R and linear in the number of labels L. The M-step of (24) is linear in the number of voxels N, linear in the number of raters R and quadratic in the number of labels L. The total number of iterations to convergence depends on the degree of agreement between the input segmentations. Fig. 9 illustrates the average computation time per iteration as a function of the number of raters.

This analysis illustrates the application of STAPLE to the assessment of 3-D unordered multicategory segmentations. The STAPLE algorithm ran to convergence in 300 s, computing on the eight segmentations of the dataset of $256 \times 256 \times 110$ voxels, with each voxel assigned one of seven labels. The algorithm is computationally efficient and provides valuable information that enables the ranking of the different segmentations. This may be used for selecting between different segmentations, or for adjusting objectively the parameters of enhancement or segmentation algorithms.

IV. Discussion and Conclusion

We have presented an algorithm for taking a collection of both binary and unordered multicategory segmentations and simultaneously constructing an estimate of the hidden true segmentation and an estimate of the performance level of each segmentation generator. This can be used to characterize any type of segmentation generator, including new segmentation algorithms or human operators, by direct comparison to the estimated true segmentation. When prior information regarding the expected anatomy is available, such as from a statistical anatomical atlas, STAPLE provides estimates of performance parameters accounting for this external standard reference. This enables the assessment of the accuracy of the segmentation generators. When such information is not available, STAPLE provides estimates of performance parameters with respect to a weighted combination of the input segmentations and the appropriate weighting is computed automatically. This enables the assessment of the precision of the segmentation generators. STAPLE adds new information regarding the difference between human and algorithm performance beyond what is available from previously reported spatial overlap measures such as the DSC, and generates a plausible "true" segmentation estimate from clinical data where alternative methods of obtaining ground truth estimates are often extremely difficult or unattractive to use.

We carried out experiments with a known true segmentation and raters of prespecified sensitivity and specificity, and demonstrated that STAPLE is able to correctly estimate the true segmentation and the performance parameters from observations of the segmentations. Sensitivity and specificity parameters depend on the relative proportion of the structure being segmented in the image, and it is useful to report the positive predictive value and negative predictive value as these account for the relative proportion of the foreground and background. The predictive values (posterior probabilities) can be derived from the sensitivity and specificity parameters and the proportion of foreground and background in each image. We compared a brain phantom for which the true segmentation is known to the STAPLE estimate of the true segmentation, derived from both synthetic segmentations with specified performance characteristics, and segmentations carried out by medical students. We demonstrated that STAPLE can identify the correct segmentation even when a majority of segmentations contain repeated errors, unlike a majority vote rule. We also demonstrated how to estimate performance parameters when the segmentations consist of unordered multicategory labels. We empirically verified the runtime performance of the algorithm and found it is rapid and easily applied in practice.

We applied STAPLE to illustrative clinically motivated segmentation problems, including the assessment of repeated segmentations of the prostate peripheral zone for brachytherapy radiation dose planning, the assessment of an algorithm for brain tumor segmentation and the analysis of 3-D multicategory unordered labeling of tissue type from MRI of a newborn infant. We compared the STAPLE analysis to commonly used alternative measures. We found the method computationally efficient and straightforward to apply in practice. We found the STAPLE estimated true segmentation cannot be replicated with simple voting rules, and that the performance parameters are important indicators of how segmentations differ, which aid in the interpretation of segmentation quality compared to standard measures such as the DSC.

We have described the STAPLE algorithm with several types of spatial constraints, including none at all, a statistical atlas of prior probabilities for the distribution of the structure of interest, and with MRF models for spatial homogeneity.

Recently the application of the STAPLE algorithm [34] to the evaluation of tumor segmentations was described [59]. The mean and standard deviation of sensitivity and specificity parameters, as estimated by STAPLE, were used to compare a rapid interactive level-set based segmentation algorithm to hand contouring by raters for tumor segmentation. Another application of the STAPLE algorithm [34] has been to identify an optimal combination of image processing algorithms for intracranial cavity segmentation of brain MRI [60].

The STAPLE algorithm [34] and generalizations have also been applied to the evaluation of segmentations derived from nonrigid registration [61], [62]. In that work, comparisons were made between a majority vote rule, pairwise binary and unordered multicategory label estimates of the hidden true segmentation and performance. The advantage of the STAPLE estimation scheme over a majority vote rule was again demonstrated.

A. Future Work

The representation we have chosen for our segmentations involves individual labeling of each voxel. This is particularly flexible and well suited to segmentations of structures of unknown topology, as is often the case when dealing with abnormal tissues. It may be interesting to consider a generalization of this to segmentations represented by enclosed surfaces, such as is often found in segmentations achieved by deformable models or by level set methods. In segmentations obtained with such surface-oriented techniques, the region of the surface boundary typically has the highest variability and is of most interest for the performance assessment. If it is the case that the interior of a region is easy to segment and the boundary is most variable, the STAPLE estimated true segmentation will reflect this and the performance level estimates will reflect the variability of the algorithms or raters in the region of the boundary.

Our algorithm estimates performance parameters and true segmentation probabilities, from which the posterior probability of the correct label being present when a segmentation indicates that label is present can be computed. Techniques for estimating bounds upon parameters estimated in the EM algorithm have been developed ([32, Chapter 4]] and it would be valuable to apply those techniques here also.

We have considered estimation of the true segmentation and performance parameters for an entire scene, and in specified regions of interest. In some applications it may be desirable to assess performance in subregions, for example, when high accuracy is required in some regions and lower accuracy may be tolerated in other regions. Analysis of this can be achieved by defining a region of interest and executing STAPLE only on this region of interest. Note also that the estimated sensitivity and specificity will depend upon the size of the region of interest, whereas the posterior probabilities (predictive values) will not.

Our model assumes the expert segmentations are conditionally independent. The notion is that the segmentations are achieved independently but with each rater having the same true segmentation goal in mind. If different raters differ in their conception of the ideal true segmentation which they are attempting to express in their decisions D_{ij}, this bias can be discovered and quantified with the STAPLE algorithm. Running STAPLE on repeated observations of a single raters' segmentations enables the estimation of the rater-specific true segmentation with the random variation of the rater removed. However, rater specific bias, such as structural errors, may remain. This could potentially be addressed by calibrating the rater through segmentation of a phantom, by standardizing the segmentation protocol or by supervised training of the rater. Alternatively, repeating this process for several raters, each with a potentially different bias, allows a spatial model for how raters differ to be constructed. Running STAPLE with the per-rater estimated true segmentations as input would then allow the estimation of the overall "true" segmentation—this effectively treats the bias of each rater as a perturbation away from the true segmentation that is to be recovered. Further investigation of this would be valuable.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of the experts and students who generated the segmentations of the clinical data.

REFERENCES

[1] D. Nicoll and W. Detmer, *Basic Principles of Diagnostic Test Use and Interpretation*. New York: McGraw-Hill, 2001, ch. 1, pp. 1–16.

[2] M. Styner, C. Brechbühler, G. Székely, and G. Gerig, "Parametric estimate of intensity inhomogeneities applied to MRI," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 153–165, Mar. 2000.

[3] D. Collins, A. Zijdenbos, V. Kollokian, J. Sled, N. Kabani, C. Holmes, and A. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 463–468, June 1998.

[4] V. Spitzer, M. J. Ackerman, A. L. Scherzinger, and D. Whitlock, "The visible human male: A technical report," *J. Amer. Med. Inform. Assoc.*, vol. 3, no. 2, pp. 118–130, 1996.

[5] T. S. Yoo, M. J. Ackerman, and M. Vannier, "Toward a common validation methodology for segmentation and registration algorithms," in *Proc. 3rd Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000)*, A. M. DiGioia and S. Delp, Eds., 2000, pp. 422–431.

[6] D. V. Iosifescu, M. E. Shenton, S. K. Warfield, R. Kikinis, J. Dengler, F. A. Jolesz, and R. W. McCarley, "An automated registration algorithm for measuring MRI subcortical brain structures," *NeuroImage*, vol. 6, pp. 12–25, 1997.

[7] S. K. Warfield, R. V. Mulkern, C. S. Winalski, F. A. Jolesz, and R. Kikinis, "An image processing strategy for the quantification and visualization of exercise induced muscle MRI signal enhancement," *J. Magn. Reson. Imag.*, vol. 11, no. 5, pp. 525–531, May 2000.

[8] J. M. Bland and D. G. Altman, "Applying the right statistics: Analyses of measurement studies," *Ultrasound Obstet. Gynecol.*, vol. 22, pp. 85–93, 2003.

[9] D. N. Levin, X. Hu, K. K. Tan, and S. Galhotra, "Surface of the brain: Three-dimensional MR images created with volume rendering," *Radiology*, vol. 171, pp. 277–280, 1989.

[10] M. Everingham, H. Muller, and B. Thomas, "Evaluating image segmentation algorithms using the pareto front," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2002, vol. 2353, ECCV, pp. 34–48.

[11] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[12] P. Jaccard, "The distribution of flora in the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, 1912.

[13] R. E. Hogan, K. E. Mark, L. Wang, S. Joshi, M. I. Miller, and R. D. Bucholz, "Mesial temporal sclerosis and temporal lobe epilepsy: MR imaging deformation-based segmentatino of the hippocampus in five patients," *Radiology*, vol. 216, pp. 291–297, 2000.

[14] F. Bello and A. C. F. Colchester, "Measuring global and local spatial correspondence using information theory," in *Proc. 1st Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 98)*, 1998, pp. 964–973.

[15] K. H. Zou, W. M. Wells, R. Kikinis, and S. K. Warfield, "Three validation metrics for automated probabilistic image segmentation in brain tumors," *Statist. Med.*, vol. 23, pp. 1259–1282, 2003.

[16] G. Gerig, M. Jomier, and M. Chakos, "Valmet: A new validation tool for assessing and improving 3D object segmentation," in *Proc. 4th Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001)*, M. A. Viergever, Ed., Utrecht, The Netherlands, 2001, pp. 516–523.

[17] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Med. Imag.*, vol. 13, pp. 716–724, Dec. 1994.

[18] S. Warfield, J. Dengler, J. Zaers, C. R. Guttmann, W. M. Wells III, G. J. Ettinger, J. Hiller, and R. Kikinis, "Automatic identification of grey matter structures from MRI to improve the segmentation of white matter lesions," *J. Image Guid. Surg.*, vol. 1, no. 6, pp. 326–338, 1995.

[19] R. Kikinis, M. E. Shenton, G. Gerig, J. Martin, M. Anderson, D. Metcalf, C. R. G. Guttmann, R. W. McCarley, W. E. Lorenson, H. Cline, and F. Jolesz, "Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging," *J. Magn. Reson. Imag.*, vol. 2, pp. 619–629, 1992.

[20] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento, "Reliability parameters to improve combination strategies in multi-expert systems," *Pattern Anal. Applicat.*, vol. 2, pp. 205–214, 1999.

[21] T. K. Ho, J. H. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 66–75, Jan. 1994.

[22] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 226–239, Mar. 1998.

[23] D. Windridge and J. Kittler, "A morphologically optimal strategy for classifier combination: Multiple expert fusion as a tomographic process," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 343–353, Mar. 2003.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, pp. 79–87, 1991.

[25] M. I. Jordan and R. A. Jacobs, "Hierarchical Mixtures of Experts and the EM Algorithm,", Tech. Rep. AIM-1440, 1993.

[26] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," in *Proc. 18th Nat. Conf. Artificial Intelligence*, vol. AAAI-2002, 2002, pp. 187–192.

[27] L. Irwig, P. Macaskill, P. Glasziou, and M. Fahey, "Meta-analytic methods for diagnostic test accuracy," *J. Clin. Epidemiol.*, vol. 48, no. 1, pp. 119–130, 1995.

[28] S.-L. Normand, "Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining, and reporting," *Statist. Med.*, vol. 18, pp. 321–359, 1999.

[29] T. A. Alonzo and M. S. Pepe, "Using a combination of reference tests to assess the accuracy of a new diagnostic test," *Statist. Med.*, vol. 18, pp. 2987–3003, 1999.

[30] S. V. B. SV, G. Campbell, K. L. Meier, and R. F. Wagner, "On the problem of ROC analysis without truth: The EM algorithm and the information matrix," *Proc. SPIE*, vol. 3981, pp. 126–134, 2000.

[31] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B.*, vol. 39, pp. 34–37, 1977.

[32] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley-Interscience, 1996.

[33] S. K. Warfield, K. H. Zou, M. R. Kaus, and W. M. Wells, "Simultaneous validation of image segmentation and assessment of expert quality," in *Proc. Int. Symp. Biomedical Imaging: Macro to Nano*, J. Fessler and M. Vannier, Eds., Washington, DC, 2002, pp. 1494:1–1494:4.

[34] S. K. Warfield, K. H. Zou, and W. M. Wells, "Validation of image segmentation and expert quality with an expectation-maximization algorithm," in *Proc. 5th Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2002),* , Tokyo, Japan, 2002, pp. 298–306.

[35] J. Rexilius, "Physics-Based Nonrigid Registration for Medical Image Analysis," masters thesis, Med. Univ. Luebeck, Luebeck, Germany, 2001.

[36] K. Friston, J. Ashburner, J. Poline, C. Frith, J. Heather, and R. Frackowiak, "Spatial registration and normalization of images," *Human Brain Mapping*, vol. 2, pp. 165–189, 1995.

[37] D. L. Collins, "3D Model-based segmentation of individual brain structures from magnetic resonance imaging data," Ph.D. dissertation, McGill Univ., Montréal, Canada, 1994.

[38] E. Ising, "Beitrag zur Theorie des Ferromagnetismus," *Zeitschrift f. Physik.*, vol. 31, pp. 253–258, 1925.

[39] T. Kapur, W. Grimson, R. Kikinis, and W. W. III, "Enhanced spatial priors for segmentation of magnetic resonance imagery," in *Proc. 1st Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 98)*, 1998, pp. 457–468.

[40] K. Van Leemput, F. Maes, D. Vanermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, pp. 885–895, Oct. 1999.

[41] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, pp. 45–57, Jan 2001.

[42] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum *a posteriori* estimation for binary images," *J. Roy. Statist. Soc. B*, vol. 51, no. 2, pp. 271–279, 1989.

[43] H. Ishikawa, "Exact optimization for Markov random fields with convex priors," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 1333–1336, Oct. 2003.

[44] I. M. Elfadel, "From fields to networks," Ph.D. dissertation, Massachussetts Inst. Technol., Cambridge, MA, 1993.

[45] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

[46] ——, "Spatial interaction and the statistical analysis of lattice systems," *J. Roy. Statist. Soc. Ser. B.*, vol. 36, pp. 192–235, 1974.

[47] T. Kapur, "Model-based three-dimensional medical image segmentation," Ph.D. dissertation, Massachussetts Inst. Technol., Cambridge, MA, 1999.

[48] J. Zhang and G. G. Hanauer, "The application of mean field theory to image motion estimation," *IEEE Trans. Imag. Processing*, pp. 19–33, Jan. 1995.

[49] D. Geieger and F. Girosi, "Parallel and deterministic algorithms from MRF's: Surface reconstruction," *IEEE Pattern Anal. Machine Intell.*, vol. 13, pp. 401–412, May 1991.

[50] J. Edmonds and R. M. Karp, "Theoretical improvements in algorithmic efficiency for network flow problems," *J. Assoc. Comput. Mach.*, vol. 19, pp. 248–264, 1972.

[51] J. Rexilius, S. K. Warfield, C. R. G. Guttmann, X. Wei, R. Benson, L. Wolfson, M. Shenton, H. Handels, and R. Kikinis, "A novel nonrigid registration algorithm and applications," in *Proc. 4th Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001)*, W. J. Niessen and M. A. Viergever, Eds., Utrecht,, The Netherlands, 2001, pp. 923–931.

[52] A. Bharatha, M. Hirose, N. Hata, S. K. Warfield, M. Ferrant, K. H. Zou, E. Suarez-Santana, J. Ruiz-Alzola, A. D'Amico, R. A. Cormack, R. Kikinis, F. A. Jolesz, and C. M. Tempany, "Evaluation of three-dimensional finite element-based deformable registration of pre- and intra-operative prostate imaging," *Med. Phys.*, vol. 28, pp. 2551–2560, Dec. 2001.

[53] M. Kaus, S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, and R. Kikinis, "Automated segmentation of MRI of brain tumors," *Radiology*, vol. 218, pp. 586–591, Feb. 2001.

[54] S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive, template moderated, spatially varying statistical classification," *Med. Image Anal.*, vol. 4, no. 1, pp. 43–55, Mar. 2000.

[55] ——, "Adaptive template moderated spatially varying statistical classification," in *Proc. 1st Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 98)*, W. M. Wells, A. Colchester, and S. Delp, Eds., MA, 1998, pp. 231–238.

[56] P. Hüppi, S. Warfield, R. Kikinis, P. D. Barnes, G. P. Zientara, F. A. Jolesz, M. K. Tsuji, and J. J. Volpe, "Quantitative magnetic resonance imaging of brain development in premature and mature newborns," *Ann. Neurol.*, vol. 43, no. 2, pp. 224–235, Feb. 1998.

[57] T. E. Inder, P. S. Huppi, S. K. Warfield, R. Kikinis, G. P. Zientara, P. D. Barnes, F. A. Jolesz, and J. J. Volpe, "Periventricular white matter injury in the premature infant is followed by reduced cerebral cortical gray matter volume at term," *Ann. Neurol.*, vol. 46, no. 5, pp. 755–760, 1999.

[58] B. P. Murphy, T. E. Inder, P. S. Huppi, S. Warfield, G. P. Zientara, R. Kikinis, F. A. Jolesz, and J. J. Volpe, "Impaired cerebral cortical gray matter growth after treatment with dexamethasone for neonatal chronic lung disease," *Pediatrics*, vol. 107, no. 2, pp. 217–221, Feb. 2001.

[59] A. Lefohn, J. Cates, and R. Whitaker, "Interactive, GPU-Based Level Sets for 3D Segmentation," in *Proc. 6th Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2003)*, 2003, pp. 564–572.

[60] C. Fennema-Notestine, I. B. Ozyurt, G. G. Brown, C. P. Clark, S. Morris, A. Bischoff-Grethe, M. W. Bondi, and T. L. Jernigan, (2003) Bias correction, pulse sequence, and neurodegeneration influence performance of automated skull-stripping methods. Program 863.23. 2003 Abstract Viewer/Itinerary Planner. Washington, CD: Soc. Neurosci., 2003. [Online].

[61] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, "Expectation maximization strategies for multi-atlas multi-label segmentation," in *Proc. Int. Conf. Information Processing in Medical Imaging*, 2003, pp. 210–221.

[62] ——, "Extraction and application of expert priors to combine multiple segmentations of human brain tissue," in *Proc. 6th Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2003)*, 2003, pp. 578–585.