Published in final edited form as:

IEEE Trans Med Imaging. 2011 October ; 30(10): 1779–1794. doi:10.1109/TMI.2011.2147795.

Robust Statistical Label Fusion through Consensus Level, Labeler Accuracy and Truth Estimation (COLLATE)

Andrew J. Asman[Student Member, IEEE] and

Department of Electrical Engineering, Vanderbilt University, Nashville, TN, 37235 USA (phone: 615-322-2338; fax: 615-343-5459; andrew.j.asman@vanderbilt.edu).

Bennett A. Landman[Member, IEEE]

Department of Electrical Engineering, Vanderbilt University, Nashville, TN, 37235 USA (bennett.landman@vanderbilt.edu)

Abstract

Segmentation and delineation of structures of interest in medical images is paramount to quantifying and characterizing structural, morphological, and functional correlations with clinically relevant conditions. The established gold standard for performing segmentation has been manual voxel-by-voxel labeling by a neuroanatomist expert. This process can be extremely time consuming, resource intensive and fraught with high inter-observer variability. Hence, studies involving characterizations of novel structures or appearances have been limited in scope (numbers of subjects), scale (extent of regions assessed), and statistical power.

Statistical methods to fuse data sets from several different sources (e.g., multiple human observers) have been proposed to simultaneously estimate both rater performance and the ground truth labels. However, with empirical datasets, statistical fusion has been observed to result in visually inconsistent findings. So, despite the ease and elegance of a statistical approach, single observers and/or direct voting are often used in practice. Hence, rater performance is not systematically quantified and exploited during label estimation.

To date, statistical fusion methods have relied on characterizations of rater performance that do not intrinsically include spatially varying models of rater performance. Herein, we present a novel, robust statistical label fusion algorithm to estimate and account for spatially varying performance. This algorithm, COnsensus Level, LAbeler Accuracy and Truth Estimation (COLLATE), is based on the simple idea that some regions of an image are difficult to label (e.g., *confusion regions*: boundaries or low contrast areas) while other regions are intrinsically obvious (e.g., *consensus regions*: centers of large regions or high contrast edges). Unlike its predecessors, COLLATE estimates the consensus level of each voxel and estimates differing models of observer behavior in each region. We show that COLLATE provides significant improvement in label accuracy and rater assessment over previous fusion methods in both simulated and empirical datasets.

Keywords

COLLATE; STAPLE; data fusion; delineation; labeling; parcellation; statistical analysis

I. Introduction

THE primary goal of any labeling process is to provide the most accurate and reliable labels possible with an efficient use of resources. Nevertheless, the most accurate labels that can possibly be achieved from a single rater are often drastically different from the “ground truth” due to the presence of ambiguous pixels, labeling confusion, and error. Thus, it is often desirable to allow several raters (either humans, machines, or both) to collaborate in

the segmentation processing. Finding an optimal way to combine the observations from several raters to form a single estimation of the “ground truth” is essential to this task. Here, we are primarily viewing this problem from the perspective of a collection of human raters, but the following discussion and method is equally applicable to semi- or fully-automated methods, such as intensity cluster, atlas registration, etc.

One simple approach that is often used in practice is the idea of a majority vote, where the “ground truth” estimate is obtained by taking the mode label decision at each voxel in the data set [1, 2]. Unfortunately, this approach does not provide any information about the likelihood of the “ground truth” estimate nor does it provide any information about the quality of the raters. In the case of multi-label segmentations, the majority vote rule does not even guarantee the presence of a majority at each voxel. Alternatives to the straightforward majority vote rule include the Borda count [3], strategies for combining classifiers [4], and strategies that take advantage of the fact that certain classifiers are experts in a subset of the full domain [5-7]. The concept of performance level estimation has also been investigated [8, 9].

The label fusion problem also arose in the context of statistical machine learning. Kearns and Valiant suggested that a collection of “weak learners” (raters that are only slightly better than chance) could be fused (“boosted”) to form a “strong learner” (a single rater with arbitrarily high accuracy) [10]. This proposal was first proven about a year later [11], and the process of “boosting” became widely practical and popular with the presentation of AdaBoost [12]. Statistical methods using automated results or complete data sets from several different human raters have been proposed to simultaneously estimate (1) the rater performance level and (2) the “ground truth” [13-16]. The algorithm presented by Warfield et al. provided a simultaneous estimation of both performance level parameters of expert segmentations and an estimation of the “ground truth” [13]. Extensions to this approach were introduced by Rohlfing et al [15]. These algorithms are based upon a maximum likelihood/maximum *a posteriori* approach (e.g. Simultaneous Truth and Performance Level Estimation, STAPLE [14]). When operating under the assumption that the raters performing the segmentations are collectively unbiased and independent, these algorithms increase the accuracy of a single labeling by probabilistically fusing multiple less accurate delineations. These statistical approaches have been widely used in atlas-fusion techniques [17-19] and have been extended to handle continuous (scalar or vector) images [20-22]

These proposed statistical methods require all raters to label all voxels in the complete data set. This means that each rater can observe each slice in the complete data set exactly once. In order to compensate for these limitations, extensions to the statistically based methodologies were introduced that allow for incomplete and over-complete data sets as well as the use of training data [23-26]. The approach proposed by Commowick *et al.* [25] introduced a parametric prior that has been shown to prevent label inversion and dramatically improve estimates using partial delineations. Other approaches have introduced a non-parametric prior to decrease the susceptibility to outlier delineations. This method (Simultaneous Truth and Performance Level Estimation with Robust Extensions (STAPLER) [23]) made significant advancements toward performing large scale label fusion. Additional work has been done to show that the robust extensions provided by STAPLER can be applied to accurately estimate the ground truth using minimally trained raters in a web-based collaboration environment [27].

Despite the recent advancements in the field of label fusion, there exists a fundamental limitation in the way that these algorithms compute performance level parameters of the raters and, thus, the estimation of the true segmentation: the observed model of rater behavior when dealing with human raters is not particularly accurately modeled by the

generative model of rater behavior used by STAPLE (and its descendents). Intuitively (and empirically – see Figure 1), raters tend to miss at a very small subset of the actual voxels present in a data set. These voxels tend to be boundary pixels and voxels where the value is ambiguous for one reason or another. This problem manifests itself, in many cases, by creating estimations of the rater performance parameters that are biased towards certain labels.

For example, imagine a truth model where there are only two labels present. One of the labels is the background, which composes a huge percentage of the total data set and the other is a label that is only present as a small circle in the middle of the truth model. Additionally, the only voxels where there is contention about the true label are the voxels that define the boundary between the background and the small label. If the observed model of rater behavior holds, the STAPLE estimate of rater behavior would estimate that the raters are very good at the background label and very bad at the small label, when, in actuality, the problems of the raters are directly related to their ability to delineate the boundary between the labels. Instead of the entire observation, there is a small subset of voxels that determine the quality of the raters, and the *consensus* voxels should not be as heavily weighted when determining the rater performance parameters. Herein, we present a robust statistical label fusion algorithm through *consensus level*, labeler accuracy and truth estimation (COLLATE). By simultaneously characterizing and estimating this additional *consensus*, we capture a more realistic model of rater behavior to more accurately estimate both rater performance and truth labels. The performance of COLLATE is characterized in simulation and with empirical data (i.e., labels provided by human raters).

Throughout this manuscript the terms *confusion* and *consensus* will be used to characterize the likelihood that a rater makes a mistake at a given voxel. These polar terms are used as a qualitative description of the quantitative *consensus level*. For example, a voxel that is determined to have a high consensus level is considered to be a voxel where there is high *consensus* and low *confusion* (i.e. it is unlikely that a rater would make a mistake at this voxel). Alternatively, a voxel that is determined to have a low consensus level is considered to be a voxel where this is high *confusion* and low *consensus* (i.e. there is a high probability that a rater would make a mistake at this voxel).

This paper is organized in the following manner. In Section II, the COLLATE algorithm is described. Techniques for initializing the algorithm, detecting convergence, and the recommended method of setting the model parameters is described. In Section III, the COLLATE algorithm is compared to traditional STAPLE on a series of experiments and simulations. One of the simulations demonstrates the sensitivity of the data-adaptive priors defined in Section II. Additional implementations include simulations using the modified COLLATE model of rater behavior, an approximation of a human model where raters miss at boundaries, the STAPLE model of rater behavior and an empirical experiment.

II. Theory

The following derivation of the COLLATE closely follows the approach of Warfield, *et al* [14].

A. Problem Definition

As in the Warfield approach, consider an image of N voxels with the task of determining the correct label for each voxel in that image. Also consider a collection of R raters that provide an observed delineation for each of N voxels exactly once. Herein, the index variable i will be used to iterate over the N voxels and the index variable j will be used to iterate over the R raters. The set of labels, $\mathbf{L}$, represents the set of possible values that a rater can assign to all

N voxels. Let $\mathbf{D}$ be an $N \times R$ matrix describing the labeling decisions of all R raters at all N voxels where $D_{ij} \in \{0, 1, \dots, L-1\}$. Let $\mathbf{T}$ be a vector of N elements that represents the hidden true segmentation for all voxels, where $T_i \in \{0, 1, \dots, L-1\}$.

In addition to the traditional model, consider a vector of N elements, $\mathbf{C}$, that represents a characterization of the *consensus* or *confusion* of each voxel at one of F level of possible consensus. All elements in this vector, $C_i \in \{0, 1, \dots, F-1\}$, indicate whether voxel i is a voxel of *confusion* ($C_i = 0$) or a voxel of some level of *consensus* ($C_i > 0$). It is important to note that the terms *consensus* and *confusion* are polar terms that are describing the same phenomenon from opposite perspectives. As the value of C_i increases the amount of *confusion* about voxel i decreases, while, conversely, the amount of *consensus* about voxel i increases. We present theory for a multi-consensus level framework. However, only a closed form solution for the binary consensus level solution is derived. This vector will subsequently be referred to as the “consensus level vector.” The E-M algorithm presented in this paper, will estimate the probability that voxel i belongs to each consensus level in the E-Step, and these estimated probabilities will be crucial in weighting each voxel when estimating the performance level parameters in the M-Step.

A characterization of the R raters' performance is characterized by $\mathbf{\Theta}$, where each element, $\mathbf{\Theta}_j$, is an $L \times L$ confusion matrix where each element in the matrix quantifies the probability that rater j will assign label s' to a voxel when the true label is s . For reference, the perfect rater would have a confusion matrix of the identity matrix. Let the complete data be $(\mathbf{D}, \mathbf{T}, \mathbf{C})$ and let the probability mass function of the complete data be $f(\mathbf{D}, \mathbf{T}, \mathbf{C}|\mathbf{\Theta})$.

B. COLLATE Algorithm

The goal of COLLATE is to accurately estimate the performance level parameters of the R raters given the rater segmentation decisions, the estimation of the truth, and the estimation of the consensus level vector. The estimated performance level parameters will be selected such that they maximize the complete data log likelihood function

$$\hat{\theta} = \arg \max_{\theta} \ln f(\mathbf{D}, \mathbf{T}, \mathbf{C}|\theta) \quad (1)$$

It is assumed that the segmentation decisions are all conditionally independent given the true segmentation and the performance level parameters, that is $(D_{ij}|C_i T_i \theta_j) \perp (D_{ij'}|C_i T_i \theta_{j'}) \forall j, j'$. 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 estimation of the performance level parameters.

Our version of the expectation-maximization (E-M) algorithm used to solve (1) is now presented. The complete data used to solve this E-M algorithm is the observed data, $\mathbf{D}$, and the true segmentation of each voxel $\mathbf{T}$ augmented with the consensus level vector, $\mathbf{C}$. The true segmentation $\mathbf{T}$ and the consensus level vector, $\mathbf{C}$, are regarded as the missing or hidden data, and are unobservable. Let $\mathbf{\Theta}_j$ be the covariance, or confusion, matrix associated with rater j and let

$$\theta = [\theta_1 \theta_2, \dots, \theta_R] \quad (2)$$

be the complete set of unknown parameters for the R segmentations. Let $f(\mathbf{D}, \mathbf{T}, \mathbf{C}|\mathbf{\Theta})$ denote the probability mass function of the random vector corresponding to the complete data. The complete data log likelihood function is presented as

$$\ln L_c \{\theta\} = \ln f(\mathbf{D}, \mathbf{T}, \mathbf{C}|\theta). \quad (3)$$

The E-M algorithm approaches the problem of maximizing the incomplete data log likelihood equation

$$\ln L\{\theta\} = \ln f(\mathbf{D}|\theta) \quad (4)$$

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 of the observable data $\mathbf{D}$ given the current estimate of θ . 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 $\theta^{(0)}$ be some initial value for θ . Then, on the first iteration, the E-step requires the calculation of

$$\begin{aligned} Q(\theta|\theta^{(0)}) &\equiv E \left[\ln f(\mathbf{D}, \mathbf{T}, \mathbf{C}|\theta) | \mathbf{D}, \theta^{(0)} \right] \\ &= \sum_{\mathbf{T}} f(\mathbf{D}, \mathbf{T}, \mathbf{C}|\theta) f(\mathbf{T}, \mathbf{C}|\mathbf{D}, \theta^{(0)}). \end{aligned} \quad (5)$$

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

$$Q(\theta^{(1)}|\theta^{(0)}) \geq Q(\theta|\theta^{(0)}) \quad (6)$$

for all θ . The E-step and the M-step are then repeated as above where at each iteration k , the current estimate $\theta^{(k)}$, the observed data $\mathbf{D}$ are used to calculate the conditional expectation of the complete data log likelihood function, and then the estimate of $\theta^{(k+1)}$ is found by maximizing $Q(\theta|\theta^{(k)})$. The E- and M- steps are repeated until convergence.

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

$$\begin{aligned} \theta^{(k)} &= \arg \max_{\theta} E \left[\ln f(\mathbf{D}, \mathbf{T}, \mathbf{C}|\theta) | \mathbf{D}, \theta^{(k-1)} \right] \\ &= \arg \max_{\theta} E \left[\ln \frac{f(\mathbf{D}, \mathbf{T}, \mathbf{C}, \theta)}{f(\theta)} | \mathbf{D}, \theta^{(k-1)} \right] \end{aligned} \quad (7)$$

Thus, on multiplying by $\frac{f(\mathbf{T}, \mathbf{C}, \theta)}{f(\mathbf{T}, \mathbf{C}, \theta)}$

$$\theta^{(k)} = \arg \max_{\theta} E \left[\ln \frac{f(\mathbf{D}, \mathbf{T}, \mathbf{C}, \theta) f(\mathbf{T}, \mathbf{C}, \theta)}{f(\mathbf{T}, \mathbf{C}, \theta) f(\theta)} | \mathbf{D}, \theta^{(k-1)} \right] \quad (8)$$

which yields

$$\theta^{(k)} = \arg \max_{\theta} E \left[\ln f(\mathbf{D}|\mathbf{T}, \mathbf{C}, \theta) f(\mathbf{T}, \mathbf{C}) | \mathbf{D}, \theta^{(k-1)} \right] \quad (9)$$

where $\theta^{(k)}$ is the estimate of the performance level parameters of the raters after the k^{th} iteration of the algorithm. The last step operates under the assumption that $\mathbf{T}$ and $\mathbf{C}$ are independent of the performance level parameters, i.e. $f(\mathbf{T}, \mathbf{C}, \theta) = f(\mathbf{T}, \mathbf{C})f(\theta)$.

C.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 and the consensus level vector at each voxel given the raters decisions, and the previous estimate of the performance parameters.

In order to maintain a compact representation of the result, the conditional probability of the true segmentation at each voxel is represented using a common notation for E-M algorithms.

$$\begin{aligned}
 W_{si\zeta}^{(k-1)} &\equiv f(T_i=s, C_i=\zeta | \mathbf{D}_i, \theta_j^{(k-1)}) \\
 &= \frac{f(T_i=s, C_i=\zeta)}{\sum_{s'} \sum_{\zeta'} f(T_i=s', C_i=\zeta')} \times \frac{\prod_j f(D_{ij}|T_i=s, C_i=\zeta, \theta_j^{(k-1)})}{\prod_j f(D_{ij}|T_i=s', C_i=\zeta', \theta_j^{(k-1)})} \\
 &= \frac{f(C_i=\zeta|T_i=s)f(T_i=s)}{\sum_{s'} \sum_{\zeta'} f(C_i=\zeta'|T_i=s')f(T_i=s')} \times \frac{\prod_j f(D_{ij}|T_i=s, C_i=\zeta, \theta_j^{(k-1)})}{\prod_j f(D_{ij}|T_i=s', C_i=\zeta', \theta_j^{(k-1)})}
 \end{aligned} \tag{10}$$

where $W_{si\zeta}^{(k-1)}$, the weight variable, indicates the probability of the true segmentation at voxel i being equal to label s , with consensus level value ζ . This representation is different from the traditional STAPLE representation of the weight variable due to the presence of the consensus level vector. For example, the value described by $W_{si0}^{(k)}$ represents the probability that voxel i is equal to label s for the k^{th} iteration and is a voxel that is likely to be confused by a given rater. The matrix constructed by considering this value at all N voxels and for all L labels is referred to later as the “consensus map.” The result of augmenting the weight variable with the consensus level vector is that *consensus* voxels are isolated so that they can be weighted less heavily when computing the rater confusion matrices. This results in an unbiased estimate of rater quality where the proportion of a given label in a truth model is significantly less influential than in the STAPLE algorithm.

D.M-Step: Estimation of the performance parameters by maximization

Given the estimated weight variable $W_{si\zeta}^{(k-1)}$, which represents the conditional probability that the true segmentation of voxel i is equal label s with consensus level value ζ , it is now possible to estimate the rater performance parameters that maximize the conditional expectation of the complete data log likelihood function. Considering each rater separately, we find the parameter estimates $\theta_j^{(k)}$ by

$$\begin{aligned}
 \theta_j^{(k)} &= \arg \max_{\theta_j} \sum_i E \left[\ln f(D_{ij}|T_i, C_i, \theta_j) | \mathbf{D}, \theta_j^{(k-1)} \right] \\
 &= \arg \max_{\theta_j} \sum_i \sum_s \sum_{\zeta} W_{si\zeta}^{(k-1)} \times \ln f(D_{ij}|T_i=s, C_i=\zeta, \theta_j) \\
 &= \arg \max_{\theta_j} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_s \sum_{\zeta} W_{si\zeta}^{(k-1)} \times \ln f(D_{ij}=s'|T_i=s, C_i=\zeta, \theta_j).
 \end{aligned} \tag{11}$$

We determined that

$$f(D_{ij}=s'|T_i=s, C_i=\zeta, \theta_j) = (1 - p(\zeta)) I(s=s') + p(\zeta) f(D_{ij}=s'|T_i=s, \theta_j) \tag{12}$$

where $I(s=s')$ is the indicator function. Plugging (11) into (12) yields

$$\begin{aligned}
\theta_j^{(k)} &= \arg \max_{\theta_j} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_s W_{si\zeta}^{(k-1)} \times \ln \left((1-p(\zeta)) I(s=s') + p(\zeta) f(D_{ij}=s' | T_i=s, \theta_j) \right) \\
&= \arg \max_{\theta_j} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_s W_{si\zeta}^{(k-1)} \times \ln \left((1-p(\zeta)) I(s=s') + p(\zeta) \theta_{js's} \right).
\end{aligned} \tag{13}$$

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

$$\sum_{s'} \theta_{js's} = 1. \tag{14}$$

The rater performance parameters can be maximized through the constrained optimization problem

$$\begin{aligned}
0 &= \frac{\partial}{\partial \theta_{jn'n}} \left[\sum_{s'} \sum_{i: D_{ij}=s'} \sum_s W_{si\zeta}^{(k-1)} \times \ln \left((1-p(\zeta)) I(s=s') + p(\zeta) \theta_{js's} \right) + \lambda \sum_{s'} \theta_{js's} \right] \\
&= \sum_{i: D_{ij}=n'} \sum_s W_{ni\zeta}^{(k-1)} \times \frac{p(\zeta)}{(1-p(\zeta))I(n=n') + p(\zeta)\theta_{jn'n}} + \lambda
\end{aligned} \tag{15}$$

where λ is a Lagrange multiplier.

In order to represent the solution for $\theta_{jn'n}$ there are two cases that need to be considered. First, the $n \neq n'$ case (off-diagonal) which can be shown to be equal to

$$\theta_{jn'n, n \neq n'} = \frac{\sum_{i: D_{ij}=n'} \sum_s W_{ni\zeta}^{(k-1)}}{-\lambda}. \tag{16}$$

Up until this point, the theory presented has been for the generic multi-consensus level approach for COLLATE. However, it is more involved to analytically solve for the $n = n'$ (on-diagonal) case for $\theta_{jn'n}$. Function optimization methods (e.g. simplex, annealing, etc) could be applied to numerically solve for the case of an arbitrary number of consensus levels.

For simplicity of representation in this paper the binary case, where $\zeta \in \{0,1\}$ and $p(1) = 1 - p(0)$, is solved below

$$\begin{aligned}
0 &= \sum_{i: D_{ij}=n'} \sum_s W_{ni\zeta}^{(k-1)} \frac{p(\zeta)}{1-p(\zeta)+p(\zeta)\theta_{jn'n, n=n'}} + \lambda \\
-\lambda &= \sum_{i: D_{ij}=n'} W_{ni0}^{(k-1)} \frac{p(0)}{(1-p(0))+p(0)\theta_{jn'n, n=n'}} + W_{ni1}^{(k-1)} \frac{1-p(0)}{p(0)+(1-p(0))\theta_{jn'n, n=n'}}
\end{aligned} \tag{17}$$

The solution for $\theta_{jn'n, n=n'}$ was obtained using Mathematica (Wolfram Research, Champaign, IL). For ease of representation, three dummy variables (a, b, and c) are declared below to solve for $\theta_{jn'n, n=n'}$

$$a = \lambda \prod_{\zeta} p(\zeta) \tag{18}$$

$$b = \lambda \left(\sum_{\varsigma} p(\varsigma)^2 \right) + \left(\prod_{\varsigma} p(\varsigma) \right) \sum_{i:D_{ij}=n', n=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} \quad (19)$$

$$c = \left(\sum_{i:D_{ij}=n', n=n'} \sum_{\varsigma} p(\varsigma)^2 W_{ni\varsigma}^{(k-1)} \right) + \lambda \prod_{\varsigma} (p(\varsigma)). \quad (20)$$

The final solution for $\theta_{jn' n, n=n'}$ is

$$\theta_{jn' n, n=n'} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \quad (21)$$

The remaining step is to solve for λ . This can be accomplished using the constraint defined in (16), which is redefined below due to the separation of the $n = n'$ and $n \neq n'$ cases

$$\theta_{jn' n, n=n'} + \sum_{n': n \neq n'} \theta_{jn' n, n \neq n'} = 1. \quad (22)$$

Given that both $\theta_{jn' n, n=n'}$ and $\theta_{jn' n, n \neq n'}$ are functions of λ it is possible, after some algebra, to represent the closed form solution for λ . As with the solution for $\theta_{jn' n, n=n'}$, three dummy variables are declared to ease the representation of the solution

$$\alpha = 1 \quad (23)$$

$$\beta = \sum_{n': n \neq n'} \sum_{i:D_{ij}=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} + \sum_{i:D_{ij}=n', n=n'} \sum_{\varsigma} p(\varsigma) W_{ni\varsigma}^{(k-1)} \quad (24)$$

$$\gamma = \left(\prod_{\varsigma} p(\varsigma) \right) \left(\sum_{n': n \neq n'} \sum_{i:D_{ij}=n'} \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} \right) \times \left(\sum_i \sum_{\varsigma} W_{ni\varsigma}^{(k-1)} \right). \quad (25)$$

The solution for λ can be written as

$$\lambda = \frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}. \quad (26)$$

With the solution for λ , the solution for the on-diagonal case is complete, resulting in a complete solution for $\theta_{jn' n}$. The solution presented integrates the consensus level vector directly into the estimation process by weighting the importance of each voxel based upon the estimated level of *confusion/consensus*. There are many benefits to this approach particularly the fact that it provides a performance level estimate that automatically dramatically decreases the importance of high consensus regions. However, because of the integration of the consensus level vector, the estimate for $\theta_{jn' n}$ no longer has a strict statistical interpretation as a measure of sensitivity/specificity.

At this point it is important to clarify the implications of the number of consensus levels. The consensus level estimation for each voxel is integrated into both the estimation of the true segmentation (E-Step) and the performance level parameters (M-Step). For the E-Step, the *probability* that a given voxel belongs to a given consensus level is estimated. Thus, even in the binary consensus level case, the actual state (i.e. *consensus* or *confusion*) exists on a spectrum. To illustrate this, consider a voxel where all raters agree, the probability that this voxel belongs to consensus level $\varsigma = F$ would be estimated to be 1.0 and 0.0 for all other consensus levels. On the other hand, for a voxel where most raters agree on the label but there is some disparity, then, in all likelihood, the probability associated with each consensus level would be greater than zero. For the M-Step, the number of consensus levels indicates the number of possible weighting factors that are taken into account when estimating θ . These weighting factors are then applied to the probability that a given voxel has label s and consensus level ς from the E-Step. The optimal number of consensus levels for a given task largely depends upon the difficulty of the labeling task. For a straightforward task where the only confusion about the true label would exist along the boundary between labels, then the binary consensus level case would be appropriate. For a more difficult problem, such as estimating full brain structure in a multi-atlas multi-label task, more than two consensus levels may be more appropriate and would make for an interesting area of future consideration.

E. Initialization Strategy, Convergence Detection, and Model Parameters

The theory presented above provides the framework for the implementation of the COLLATE algorithm. In order to fully implement the algorithm, however, an initialization strategy, method of detecting convergence and the model parameters must be set according to the needs of the application.

1) Initialization—COLLATE can be initialized by either providing an initial estimate of the performance level parameters (θ) or the true segmentation and consensus level vector (usually implemented as an initial estimation of $\mathbf{W}$). In this paper, COLLATE is initialized with an initial estimate of $\mathbf{W}$ as the results of a majority vote algorithm. If the data are available, a probabilistic atlas can be used to provide an initial estimate of the true segmentation [28, 29]. If an initial estimate of the true segmentation is provided, then the iterative process of the E-M algorithm begins by calculating the rater performance parameters from the initial estimate of the true segmentation and consensus level vector.

As opposed to providing an initial estimate of the true segmentation, initial estimates of the rater performance parameters can be provided to initialize COLLATE. Previous algorithms [14] have used this strategy for initialization of the E-M algorithm. If there is no prior information about the performance of the raters then the initialization strategy is generally to assume that all raters are of equally high quality. For example, this could be accomplished by setting $\theta_{jss} = 0.99 \forall j, \forall s$. It should be noted that if an initial estimate of the performance parameters is provided then the iterative COLLATE algorithm would begin with an estimation of the true segmentation.

In all of the simulations and empirical experiments presented in this paper, an initial estimation of the true segmentation is used to initialize the COLLATE algorithm.

2) Convergence—As with all E-M algorithms, the COLLATE algorithm presented in this paper is guaranteed to converge to a local maximum. The detection speedup of convergence is a topic that has been explored on multiple occasions [30, 31]. The COLLATE algorithm estimates the performance level parameters, the true segmentation and the consensus level given the input data provided by multiple raters. A close monitoring of any of these

parameters would provide a quality method of detecting convergence depending upon the application. In this paper, the desired method of convergence detection is through monitoring the change in the performance level parameters. As suggested in STAPLE, the change in the normalized trace of the estimated performance level parameters is the desired method of convergence detection. We use a threshold of $\epsilon = 1 \times 10^{-3}$ for all simulations and empirical experiments presented in this paper. The normalized trace calculation is given by

$$\frac{1}{LR} \sum_{j=1}^R \text{tr}(\theta_j). \quad (27)$$

The number of iterations required for convergence generally depends upon the number of coverages and the quality of the data passed to the COLLATE algorithm. In this worst case scenario (i.e. low number of coverages, low quality raters) the algorithm generally converges in around 20 iterations in our experience.

3) Data-adaptive Priors—There are three different data-adaptive priors that need to be determined in order to perform the COLLATE algorithm. The first prior that needs to be set is $p(C=0)$ which describes the probability of a rater giving the incorrect label for a *consensus* voxel. It is of note that in the binary consensus level case, the $p(C=1) = 1 - p(C=0)$ and describes the probability that the rater reports the correct label for a *consensus* voxel. In this paper, the values of 0.99 and 0.01 were used for $p(C=1)$ and $p(C=0)$, respectively.

The second prior that needs to be set is $f(T_i = s)$. This is equivalent to the prior in the STAPLE algorithm and can be either a global or a spatially varying prior. A spatially varying prior would be optimal in situations when prior knowledge, such as a probabilistic atlas, is available for the given segmentation. In general, this prior can be thought of as simply the probability that voxel i has associated label s . As a rule of thumb, if explicit spatial information about the true label is available (i.e. information above and beyond the observed data) it should be integrated into this prior. In all of the simulations presented in this experiment no explicit spatial information is provided, thus, a global prior is used, in which $f(T_i = s)$ is a vector where each element in the vector represents a prior probability for each available label in the segmentation. For notational consistency, we let $\gamma_s = f(T_i = s)$. In a situation where this quantity is not readily available, this value is found using the input segmentations provided by the labelers

$$\gamma_s = \frac{1}{NR} \sum_{j=1}^R \sum_{i=1}^N I(D_{ij}=s) \quad (28)$$

where $I(D_{ij}=s)$ is the indicator function which is equal to 1 when $D_{ij}=s$ and equal to 0 otherwise.

The final prior that needs to be set is $f(C_i = \zeta \mid T_i = s)$ which indicates the probability that a given voxel is *consensus* or *confusion* given that the true label is s . As with the parameter $f(T_i = s)$, this parameter could be a spatially varying prior or a global prior. Again, in all of the simulations and experiments in this paper a global prior is used. In the case of a global prior, let Ψ_i be a binary variable indicating whether or not voxel i has been estimated to be *confusion* ($\Psi_i = 0$) or *consensus* ($\Psi_i = 1$). In order to calculate the value of Ψ_i a threshold value of $\tau = 0.95$ is used, which indicates the fraction of raters that need to agree in order for a pixel to be estimated to be consensus. The value of Ψ_i is calculated at each voxel by

$$\Psi_i = I \left(\max_s \left(\frac{1}{R} \sum_{j=1}^R I(D_{ij}=s) \right) > \tau \right) \quad (29)$$

where s is a value in the set of $\{0, 1, \dots, L-1\}$.

In addition to the calculation of Ψ_i , the binary variable that estimates the status of each voxel ("consensus" or "confusion"), an estimation of the true label is garnered through the majority vote algorithm at each voxel. Let K_i represent the label estimated through a majority vote algorithm. The value of $\rho_{\zeta|s} = P(C_i = \zeta \mid T_i = s)$ is computed by

$$\rho_{\zeta|s} = \frac{1}{N_s} \sum_{i=1}^N I(\Psi_i = \zeta) I(K_i = s) \quad (30)$$

where

$$N_s = \sum_{i=1}^N I(K_i = s). \quad (31)$$

It is important to note that there is no guarantee that these methods of calculating the data-adaptive priors are optimal. Nevertheless, these parameters are meant to provide a basis by which COLLATE uses to compute the estimates of the rater performance parameters and the hidden data. Instead of constructing two separate priors (as presented above), it would accomplish a similar goal to use a single spatially varying prior as previous work has suggested [14]. In this case, the estimation of *consensus/confusion* would be integrated implicitly into the spatially varying prior. We felt that exposing two separate priors made the estimation of confusion level and the various opportunities for implementation more explicit. The implementation of the data adaptive priors was meant to be as simple as possible.

III. Methods and Results

A. Terminology

In the following results and methodologies presented in this section, several simulations and experiments are presented. These simulations range from a model that matches the COLLATE rater behavior model to an empirical experiment that uses data acquired from human raters. In order for the presentation to be as clear and consistent as possible it is necessary to define some terminology:

- A *label* is an integer valued category assigned to an anatomical location.
- A *rater* is an entity (real or simulated) that reports or observes labels.
- An *observation* is the result of a single rater observing all labels in a given slice (i.e. assigning an integer value to all pixels/voxels in an image).
- A *coverage* is the result of a single rater making exactly one observation of each available slice in the set.
- A *truth model* defines the true labels for all voxels/pixels in all of the available slices. If the slices are the output of a simulation then the truth model is generally known. For empirical data the truth model is the result of an anatomical expert carefully providing a label for each voxel.

- A *generative model of rater behavior* defines the way in which a label fusion algorithm (e.g. STAPLE, COLLATE) models the decision-making process of a given rater.
- The *consensus level vector* is the aspect of the true segmentation that is introduced by the COLLATE algorithm. This vector is fully integrated into the estimation process. The probability that a given voxel is in each consensus level is estimated and this estimated probability is used to determine the weighting of each voxel in the determination of the performance level parameters.
- A *consensus map* is a property specific to the COLLATE algorithm that defines the regions in a specific slice where consensus/confusion is present. This concept is a byproduct of that fact that we are using binary consensus levels, as it dramatically simplifies the representation of the estimated consensus levels. The consensus map is derived from the estimated consensus level vector in the weight variable, $W_{si}^{(k-1)}$.

Mathematically, this is defined as $\sum_{s=0}^{L-1} W_{si}^{(k)} \forall i$, where it defines the amount of weight in the in $C=0$ consensus level. All of the values of the consensus map are $\in [0,1]$, where 0 (black) is referred to as “full consensus” and 1 (white) is referred to as “full confusion.” The consensus map can be thought of as the continuous *probability* that each voxel is a voxel of *confusion*.

- A *confusion region* is the collection of high-valued pixels/voxels in the consensus map.
- A *confusion matrix* is an $L \times L$ matrix where each element in the matrix defines the probability that a rater would label a voxel with label s given that the true label is s' in high valued regions of the consensus map. As previously defined, the confusion matrix for rater j is θ_j .

This terminology will be used consistently in the following simulations and empirical experiments.

B. Implementation and Evaluation

COLLATE and all simulations were implemented in MATLAB (Mathworks, Natick, MA). The implementation used to produce the results seen in this paper is available via the “MASI Label Fusion” project on the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC, www.nitrc.org/projects/masi-fusion). For all simulations and experiments, both STAPLE and COLLATE use a global prior for the label probabilities (see Eq. (28)). COLLATE uses a global prior for the fraction of each label that are in consensus (see Eq. (30)). Additionally, for all simulations both COLLATE and STAPLE analyze the entire truth model when performing the estimation procedure (i.e. no regions of interest are considered [32]). The empirical data presented in this paper was gathered using the WebMill interface (<https://brassie.ece.jhu.edu/Home>) [23, 27, 33]. For data contributed by human raters detailed instructions about the labeling procedure were provided to the raters. All human raters were required to perform at least one practice labeling before proceeding to the actual labeling data. All studies were run on a 64 bit quad-core 3.07GHz desktop computer with 13GB of RAM, running Ubuntu 9.04.

When results are presented, the resulting truth model, performance level (confusion matrices) and consensus map estimations are presented. In situations where the true confusion matrix for the model is known, the accuracy of the estimations is presented. Otherwise, a comparison between the resulting STAPLE confusion matrix and COLLATE confusion matrix are presented for visual comparison. In all of the simulations the accuracy of the estimated truth labels is presented for varying numbers of coverages as the fraction of

pixels correct in the confusion region. For the empirical simulation the “ground truth” provided by an expert anatomist is provided for visual comparison with the COLLATE truth estimation.

C. Simulation 1: Simulation using COLLATE Model of Rater Behavior

The first simulation (Figure 3) used simulated raters that are nearly identical to the COLLATE model of rater behavior. This means that the consensus map is clearly defined to be low valued outside of the square in the middle of the truth model and high valued inside of the square. The truth model consists of 50 slices of size 100×100 pixels. A collection of 20 simulated raters were created that are described by confusion matrices with constant valued diagonals. The diagonal values for the raters were linearly spaced between 0.45 and 0.65. Thus, the raters from this simulation were slightly better than chance as there were 5 labels on the truth model. For all slices the size of the confusion region was held constant at 10%.

The purpose of this simulation was to assess the accuracy of COLLATE in a model where there exist well-defined regions of the image where raters are very accurate and other regions where they are only slightly more accurate than chance. The accuracy of truth estimation and the confusion matrices is assessed and compared to the accuracy of STAPLE as a reference point. The accuracy of these estimations was assessed by varying the number of coverages (from 3 to 20) passed to COLLATE. An estimate of the consensus map is provided for eight coverages with 25 Monte Carlo iterations.

Figure 3A represents an example slice from the truth model used in this simulation. Note that the ‘white’ label present in the confusion region (i.e the light-gray to white area of the consensus map) is not present in the consensus region (i.e. the dark-gray to black area of the consensus map). This was included because of the apparent problem with the STAPLE algorithm for small labels [18, 23, 27]. The reason for this problem is that the limited data used to estimate the performance level parameters for small labels tends to increase the likelihood for label inversion. It is of note that both the parametric prior approach proposed by Commowick et al. [25] and the non-parametric prior proposed by Landman et al. [23] have decreased the likelihood of witnessing label inversion on small labels. Figures 3B and 3C represent example observation made by the simulated raters with diagonal values of 0.45 and 0.65, respectively. Figure 3D represents the labels generated by STAPLE using eight coverages. As clearly evident, the presence of the small white label and the poor accuracy of the raters in the confusion region cause STAPLE to converge to an estimate that does not match the truth model. Figure 3E represents the COLLATE truth estimation which does not suffer from the same failures as the STAPLE estimation. Due to the introduction of the consensus map and the new generative model of rater behavior, COLLATE is able to converge to the correct answer despite the small label and poor labeling accuracy in the confusion region. Figures 3G and 3H show the accuracy of COLLATE and STAPLE for the simulation with varying numbers of coverages. The results shown in Figure 3G indicate that the COLLATE truth estimation is consistently more accurate in the confusion region than the STAPLE estimate. Additionally, due to label inversion, STAPLE converges to the incorrect truth estimate while COLLATE converges to the correct truth estimate. The results shown in Figure 3H show that COLLATE is able to converge to the confusion matrices that match the simulation model, while, not surprisingly, STAPLE converges to a significantly different approximation of the confusion matrices.

D. Simulation 2: Data Adaptive Prior Sensitivity

The next simulation (Figures 4 and 5) was constructed by creating a truth model that is equivalent to the model seen in Figure 3. The truth model consists of 50 slices of size

100×100 pixels. A collection of 20 simulated raters were created that were described by confusion matrices with constant valued diagonals. The diagonal values for the raters were linearly spaced between 0.55 and 0.75. Note that despite the low diagonal confusion matrices the raters are still significantly better than chance as there are five labels present on the truth model. Each rater observed one coverage of the truth model.

The purpose of this simulation was to quantify the sensitivity of COLLATE with respect to the $\mathcal{A}(C_i = \zeta \mid T_i = s)$ data adaptive prior. This simulation is broken up into two parts. The first part (Figure 4) maintains a constant confusion region (50%) and varies the data adaptive prior. The second part (Figure 5) focuses on the algorithm's ability to estimate the confusion prior by varying the size of the confusion region (5% to 95%). For both parts, the accuracy of the estimated labels is represented as a percent improvement over the STAPLE estimate of the same truth model. A truth model was created such that varying the fraction of the image that represents a confusion region is straightforward. For both parts of the simulation a random subset of six raters (six coverages) was chosen to construct the estimates with 10 Monte Carlo iterations.

Figures 4A and 4B represent the results for a constant 50% confusion region for the truth estimation and confusion matrix accuracy, respectively. These results show that, as expected, the ideal result is obtained when the estimate of the fraction confusion is equal to the actual fraction confusion. An underestimate has little effect on the accuracy of the results for the truth estimation, but causes the confusion matrix accuracy to decrease. An overestimate of the confusion region drastically affects the accuracy of both the truth estimation and the confusion matrix estimation. Nevertheless, regardless of the data adaptive prior estimate, the COLLATE estimates are consistently better than the estimates obtained by STAPLE. Additionally, confusion matrix accuracy of the STAPLE estimate is consistently different than the confusion matrices used in the simulation as the STAPLE rater behavior model does not take into account the consensus map introduced by the COLLATE algorithm. These results are present on Figure 3B to emphasize the differences between the two generative models of rater behavior.

Figures 5A and 5B represent the results for varying confusion region size. The size is varied between 5% and 95%. The results show that as the confusion region increases toward full confusion, the COLLATE truth estimate and STAPLE truth estimate (Figure 5A) converge to the same accuracy level. This is due to the fact that if the confusion region represents the entire image, then the COLLATE and STAPLE models of rater behavior are equivalent. Figure 5B shows that, regardless of the confusion region size, the accuracy of the COLLATE confusion matrix estimates remains approximately constant, while the STAPLE estimate increases in accuracy until a large confusion region is present, in which case the COLLATE and STAPLE confusion matrix estimates are of approximately equal accuracy.

E. Simulation 3: Simulation using Boundary Random Ratets

The third simulation (Figure 6) emulates a reliable model of rater behavior by simulating raters that only miss by inaccurately labeling the boundary between two adjacent label regions. This approach is slightly different than previous boundary random rater simulations [24] where each boundary pixel had a 50% chance of being chosen incorrectly. The truth model for this simulation consists of 50 slices of size 100×100 pixels. Once again, a collection of 20 raters were used to observe the truth model slices, however, the raters were designed to only miss at the boundaries and only assign the labels in the adjacent regions to each boundary. This was accomplished by identifying the boundary pixels and applying a “shift” amount (positive or negative) to each boundary pixel. A random number was drawn from a Gaussian distribution, where the standard deviation of the distribution was

determined by the quality of the rater. The standard deviations ranged from 1.2 to 3.26 for the best and worst raters respectively.

This simulation assesses the accuracy of the estimates returned by the COLLATE algorithm in a model that closely approximates the way that human raters observe truth models. The accuracy of the truth estimations was assessed for various numbers of coverages, ranging from 3 to 20. Due to the fact that the confusion matrices do not precisely correspond to the proposed generative model of rater behavior, a visual comparison is presented for the COLLATE and STAPLE estimations. Twenty-five Monte Carlo iterations were used. As with the second simulation, an estimated consensus map is provided for eight coverages.

Figure 6A represents the truth model used for this simulation. Figures 6B and 6C are representative observations made by a high quality rater and a low quality rater, respectively. Figures 6D and 6H show the STAPLE estimation of the true labels and an example confusion matrix after 8 coverages. Upon visual inspection it is evident that STAPLE has incorrectly placed the magenta label. The reason for this problem is due to label inversion, which can be seen in the 4th column (which corresponds to the magenta label) of the confusion matrix in Figure 6H. Label inversion occurs when the confusion matrix estimation indicates when a rater assigns a label other than the intended label [23]. This can have catastrophic effects on the truth estimation as it can cause large regions of the estimation to have incorrect label values. Recent work has focused on the label inversion problem and both parametric and non-parametric priors have been proposed that have been shown to prevent label inversion [23, 25]. On the other hand, the COLLATE estimates of the true labels, consensus map and example confusion matrix can be seen in Figure 6E, 6F, and 6I, respectively. The COLLATE estimate does not suffer from the same label inversion problem.

COLLATE estimates the rater to have a nearly constant diagonal confusion matrix. This makes logical sense as the raters were not designed to be biased towards certain labels. The reason for this benefit is due to the modified generative model of rater behavior, which serves to normalize the size of the labels in the confusion matrix estimation. In this simulation, the STAPLE confusion matrix estimations largely depend on the size of the region associated with a given label. The reason for this is the fact that larger regions have significantly more consensus voxels. Thus, due to the STAPLE model of rater behavior, the performance level estimations for a given label will be largely dependent on the size of the label in this simulation (see Figure 6H). On the other hand, COLLATE removes this dependence by weighting the voxels based upon the estimated consensus levels. It is of note that if modifications were made to the traditional STAPLE algorithm (i.e. spatially varying prior, or specifying a region of interest) this dependence may be reduced. The accuracy of the truth estimations with respect to number of coverages is presented in Figure 6G. As with the second simulation, the COLLATE estimates are of consistently higher accuracy and lower standard deviation than the estimates provided by STAPLE. Note that the y-axis on this plot is the fraction of pixels correct in the confusion region only. The fraction correct would be significantly higher if the consensus regions were included, however, the pixels of interest in the COLLATE model are the pixels where there is confusion about the true label, and thus, only the pixels in the confusion region are considered.

F. Empirical Experiment: COLLATE/STAPLE Comparison using Delineations by Human Raters

The empirical experiment presented in Figure 7 compares the accuracy of the COLLATE and STAPLE algorithms on data generated by human raters. The truth model for the empirical consisted of 10 slices of 70×110 pixels. These slices were selected from a whole-brain scan (182×218×182 voxels) of a healthy individual (after informed written consent)

that was cropped to isolate the posterior fossa. A specific region of the brain was isolated (i.e. the posterior fossa) to simplify the labeling process and allow a large collection of data to be collected in a relatively short amount of time. A collection of eight raters each performed a single coverage using the online WebMill system. The task defined for each rater was to label the sagittal cross-section of a cerebellum. Five different colors were assigned to five different regions of the cerebellum. The color blue was assigned to Lobules I-V (upper lobe), green was assigned to Lobules VI-VII (middle lobe), magenta was assigned to Lobules VIII-X (lower lobe), red was assigned to the Corpus Medulare White Matter and Yellow was assigned to the Vermis. All background pixels were assigned the color white. The raters observed the slices by applying the labels directly to the high resolution Magnetization Prepared Rapid Acquired Gradient Echo (MPRAGE) sequence. While performing each observation a reference image was placed in the top right corner to visually remind the raters of the task to be performed.

The spatial homogeneity and overlap are particularly important when comparing segmentations gathered using clinical data. Thus, the Dice and Jaccard similarity coefficients were used when comparing the accuracy of the truth label estimations acquired from the algorithms. The Dice Similarity Coefficient (DSC) [34] is an often used metric when comparing the spatial overlap between two vectors. The DSC is defined as $(2|A \cap B|) / (|A| + |B|)$ where $|A|$ and $|B|$ represent the area of regions A and B respectively. The Jaccard Similarity Coefficient [35] is another commonly used metric when defining the spatial overlap between two vectors. The Jaccard Similarity Coefficient (sometimes called the Jaccard Index) is defined as $|A \cap B| / |A \cup B|$ where $|A|$ and $|B|$ have the same meaning as seen in the DSC definition.

Figure 7 illustrates both that COLLATE can accurately fuse the labels from multiple raters, but also that it can outperform STAPLE. Figure 7A presents a representative slice from the truth model that was created by an expert neuroanatomist. Figures 7B and 7C represent the STAPLE and COLLATE estimations of the true labels by fusing the labels from eight raters. Figure 7B contains several mislabels around the outside boundary of the estimation and inaccurately extends the corpus medulare. It is important to note that these errors are due to partial label inversion. This is a highly studied problem with statistical fusion methods [24, 26] that results from when two labels are commonly confused. This problem manifests itself in the STAPLE confusion matrix estimate (see the second column of the STAPLE confusion matrix in Figure 7F).

Figure 7D represents the observed model of rater behavior and it demonstrates that the raters did not arbitrarily mis-label pixels on the slice. Instead, as previously described, the raters struggled with boundary pixels and regions of low contrast on the original slice. Figure 7E represents the estimated consensus map, which agrees with the observed model of rater behavior. The estimated consensus map presented here is certainly a little conservative. This could be adjusted by modifying the $p(C = \zeta)$ prior which is discussed in the theory section. The averaged confusion matrices for both the COLLATE and the STAPLE estimates can be seen in Figure 7F. Similar to the results seen in the third simulation, the COLLATE averaged estimated confusion matrix is very close to a constant diagonal matrix. This makes intuitive sense, because there is no reason why any given rater would be biased towards a given label value since all labels in this task share a boundary. The averaged STAPLE confusion matrix shows that nearly all raters are perfect at background simply because of the fact that the majority of the pixels are background in the truth model. Additionally, there is near label-inversion in the second column of the averaged STAPLE confusion matrix which explains the presence of the blue label on the perimeter of the green label on the STAPLE truth estimate.

Lastly, the range of Jaccard and Dice Similarity coefficient values for the 10 slices used in this experiment were computed for both the COLLATE estimates and the STAPLE estimates of the true labels. These ranges can be seen in the two plots in Figure 7G. A paired t-test was performed on both the Jaccard and Dice similarity coefficients and the resulting p-values were found to be less than .001 for both similarity metrics. This indicates that there is significant improvement gained by using COLLATE on empirical data.

G.Simulation 4: Simulation using STAPLE Model of Rater Behavior

The fourth, and final, simulation (Figure 8) presented in this paper assesses the accuracy of the COLLATE algorithm when using a model that matches the STAPLE generative model of rater behavior. Note that this is equivalent to a COLLATE truth model that has a true consensus map that is all confusion. The truth model consists of 50 slices of 100×100 pixels. The truth model used in this simulation is identical to the model used in Simulation 1 except that the raters miss uniformly throughout the volume. A collection of 20 simulated raters were created that are described by confusion matrices with constant valued diagonals. The diagonal values are linearly spaced from 0.45 to 0.65 for the worst and best rater, respectively.

The purpose of this simulation was to quantify the accuracy of COLLATE when the STAPLE model of rater behavior is fully accurate. As with the previous simulations, the accuracy of the truth estimations and confusion matrix estimations was assessed and compared to the results from the STAPLE algorithm. This was accomplished by varying the number of coverages used to perform the estimations (from 3 to 20 coverages). Ten Monte Carlo iterations were used to approximate the mean and standard deviation of the estimation accuracy. As with the previous simulations, an estimation of the consensus map is provided for eight coverages.

Figure 8A shows an example truth model used in the simulation. Figures 8B and 8C are representative observations of the truth model. Figures 8D and 8E represent STAPLE and COLLATE truth estimations. Upon visual inspection it appears that they are essentially equivalent. The estimated consensus map can be seen in Figure 8F. The true consensus map would be fully 'white,' indicating that all of pixels are in the confusion region. By chance, there are several pixels where raters agree, so the consensus map contains several isolated pixels where some level of consensus is present. Figures 8G and 8H represent the accuracy of the truth estimations and confusion matrix estimations for COLLATE and STAPLE. The average accuracy of the truth estimations (Figure 8G) by COLLATE is slightly (≈ 0.005) better than the STAPLE estimations for low numbers of coverages. However, by approximately seven coverages the STAPLE and COLLATE estimates converge to the same level of accuracy. Due to the inaccuracies of the consensus map estimation, the COLLATE estimations of the confusion matrix are less accurate than the STAPLE estimations for all numbers of coverages. Nevertheless, it should be considered that these differences are only on the magnitude of approximately 0.1% for seven or more coverages.

IV. Discussion and Conclusion

Herein, we presented an algorithm, COLLATE, for fusing a collection of rater label observations to estimate the consensus level, labeler accuracy and truth labels. COLLATE (1) provides significant improvement over previously developed algorithms, (2) more accurately reflects the realistic rater behavior as seen when human raters segment medical image data, and (3) results in nominal degradation when the rater assumptions are violated (Figure 7 and 8). Initialization parameters, detection of convergence and other model parameters are clearly defined.

Like its predecessors, COLLATE takes a collection of input observations from a group of raters (human or otherwise) and simultaneously estimates the truth labels and the rater performance parameters (“labeler accuracy”). However, COLLATE also estimates the consensus level of each voxel, which can be viewed as an inherent property of each voxel that determines the likelihood that a given rater would be confused about the label associated with a given voxel. The algorithm presented in this paper is formulated as an instance of the expectation-maximization (E-M) algorithm [30, 36]. As with STAPLE, the decisions of each rater are directly observable, the hidden true segmentation is an integer-valued array corresponding to the label decisions at each voxel. The hidden data in the COLLATE algorithm is augmented with a “consensus level vector” which describes the consensus level of each voxel. The labeler accuracy is iteratively estimated until convergence and is represented in the form of a confusion matrix for each rater. In often used E-M terminology, the *complete* data consists of the rater decisions, which is given, and the true segmentation and “consensus level vector” which are iteratively estimated. In order to perform this estimation, the conditional probability of the *hidden* true segmentation and the “consensus level vector” is evaluated given the rater decisions and the previous estimate of the rater confusion matrices. Convergence to a local maximum is guaranteed. As with previous algorithms, COLLATE is straightforward to apply to medical imaging data acquired from human raters or automated algorithms.

The sensitivity of the consensus-based priors was analyzed for varying confusion region sizes and estimates of the confusion region size. The sensitivity was captured by comparing the accuracy of both the COLLATE truth estimate and confusion matrix parameters to that of STAPLE. Optimal estimations are obtained when priors match the truth model; however, our analysis also indicates that when the confusion region size prior is underestimated the accuracy of the results are significantly better than when the confusion regions size prior is overestimated.

COLLATE is able to accurately estimate the quality of the raters and does not suffer from commonly encountered problems with STAPLE, such as label inversion. For small confusion region sizes, COLLATE significantly outperforms STAPLE for both the truth label accuracy and the confusion matrix accuracy. For large confusion region sizes, COLLATE and STAPLE converge to the same accuracy level. The benefits of the estimated “consensus level vector” are demonstrated in the differences in the estimated confusion matrices — which reflect a more “physical” characterization of failure likelihood. In the COLLATE estimates, likelihood of error is not biased by the area of the region being labeled.

In this paper, both COLLATE and STAPLE utilize strictly global priors when estimating the true segmentations and the performance level parameters. However, one could use a spatially varying prior instead of a global prior for $f(T_i = s)$. A spatially varying prior has the potential to prevent some of the poor STAPLE segmentation estimations that are presented in this paper (e.g. Figure 6). Nevertheless, simply implementing a spatially varying prior does not make COLLATE and STAPLE equivalent. COLLATE makes an iterative estimate of the consensus level of each voxel. Voxels that are in high consensus are de-weighted when calculating the performance level parameters. Regardless of whether a global or spatially varying prior was used, STAPLE would consider all voxels to have an equal impact on the calculation of the performance level parameters. Thus, if a given label is present in multiple consensus levels (such as Figure 3) the final STAPLE estimation of the performance level estimations would be artificially high in the regions where there is significant confusion.

Another modification to STAPLE, proposed by Rohlfing et al. [32], is to select only voxels that are not in consensus when performing the statistical fusion. In some cases, selecting a subset of the total voxels makes the STAPLE model of rater behavior significantly more appropriate and accurate. For the binary consensus level case seen in this paper, the dramatic improvement by COLLATE over STAPLE would certainly be lessened. In some cases, the regions of consensus and confusion are clearly defined and easily detected (e.g. the simulation presented Figure 3). In this scenario, COLLATE with binary consensus levels is essentially equivalent to performing STAPLE only over the confusion region. Nevertheless, COLLATE provides a framework for integrating any number of consensus levels directly into the estimation process without the need to perform any preprocessing to determine a reasonable region of interest for processing. This framework is the primary contribution of this paper and we feel that this new perspective on the problem will provide fascinating avenues for continuing research.

COLLATE, as presented, operates under the assumption of voxelwise independence. However, this premise could be relaxed in exactly the same manner as has been done with STAPLE using a Markov Random Field (MRF) model to regularize the label probability fields [14]. The MRF approach would model the conditional dependence of a given voxel on the voxels in a local neighborhood to be equal to the conditional dependence of that voxel on the rest of the voxels in the volume.

We suggest that COLLATE would be amenable to an application of the mean field approximation. The iterative mean field approximation has been widely examined [37] and has been used in diverse applications, e.g., [38, 39]. Essentially, the approximation is

achieved by estimating the mean value of the true segmentation (e.g., the mean of $W_{si\zeta}^{(k)}$) and adding local neighborhood conditional dependence to the exponentiated voxelwise independence estimate

$$W_{si\zeta}^{(k)} \leftarrow \frac{1}{Z} \exp \left(\ln f(C_i = \zeta | T_i = s) + \ln f(T_i = s) + \ln \prod_j f(D_{ij} | T_i = s, C_i = \zeta, \theta_j^{(k)}) + \sum_{\zeta'} \sum_m \sum_n J_{sn} W_{sm\zeta'}^{(k-1)} \right)$$

with

$$\sum_s \sum_{\zeta} W_{si\zeta}^{(k)} = 1 \quad (33)$$

where $W_{si\zeta}^{(k)}$ is the mean field estimate of the conditional probability that voxel i has label s given the voxel's surrounding label values. J_{sn} is an $L \times L$ matrix describing the spatial relation between label s and label n . Z is a normalizing constant for the estimate. This MRF can easily be appended to the voxelwise independent implementation that is seen in (10). The expression seen in (32) shows the calculation at iteration k of the mean field estimate.

For clarity in comparison with the STAPLE theory, the benefits of the mean field approximation MRF are not explored in this manuscript. However, the benefits of the MRF implementations have been widely documented [14], and the inclusion of spatial dependence will certainly be a fascinating topic of future research.

Acknowledgments

The authors acknowledge the numerous volunteer raters for their assistance in data acquisition, the creators of the WebMill online labeling system, and John Bogovic.

This work was supported in part by NIH/NINDS 1R01NS056307 and NIH/NINDS 1R21NS064534.

Biography

Andrew J. Asman (M'08) was born in Ft. Wright, Kentucky and studied electrical engineering (BS'10) and computer engineering (BS'10) at the University of Kentucky (Lexington, KY). He is currently pursuing a doctoral degree in electrical engineering from Vanderbilt University (Nashville, TN).

During the course of his undergraduate education, he worked at Lexmark International, Inc, as a student software developer in the Colorscience and Imaging Department. Additionally, he has done work as a research assistant at the University of Kentucky where his work focused on audio rendering, sound source localization and cocktail party problems. He is currently a graduate student at Vanderbilt University where his research interests include medical image processing, specifically, population characterization with magnetic resonance imaging (MRI).

Mr. Asman is a member of SPIE and the National Society of Collegiate Scholars (NSCS).

References

1. Warfield S, et al. Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. *J Image Guid Surg.* 1995; 1:326–38. [PubMed: 9080353]
2. Kikinis R, et al. Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging. *J Magn Reson Imaging.* Nov-Dec;1992 2:619–29. [PubMed: 1446105]
3. Ho TK, et al. Decision Combination in Multiple Classifier Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Jan.1994 16:66–75.
4. Kittler J, et al. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Mar.1998 20:226–239.
5. Windridge D, Kittler J. A morphologically optimal strategy for classifier combination: Multiple expert fusion as a tomographic process. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Mar.2003 25:343–353.
6. Jacobs R, et al. Adaptive mixtures of local experts. *Neural computation.* 1991; 3:79–87.
7. Jordan M, Jacobs R. Hierarchical mixtures of experts and the EM algorithm. *Neural computation.* 1994; 6:181–214.
8. Alonzo A, Pepe H. Using a combination of reference tests to assess the accuracy of a diagnostic test. *Statistics in Medicine.* Feb 28.2001 20:656–658. [PubMed: 11223907]
9. Beiden, S., et al. The problem of ROC analysis without truth: the EM algorithm and the information matrix; 2000. p. 126
10. Kearns M, Valiant LG. Learning boolean formulae or finite automata is as hard as factoring. *Harvard University Technical Report.* 1988; TR-14-88
11. Shapire RE. The strength of weak learnability. *Machine Learning.* 1990; 5:197–227.
12. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sciences.* 1997; 55
13. Warfield, SK., et al. Simultaneous validation of image segmentation and assessment of expert quality. presented at the International Symposium on Biomedical Imaging; Washington, DC. 2002;
14. Warfield SK, et al. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging.* Jul.2004 23:903–21. [PubMed: 15250643]
15. Rohlfing T, et al. Expectation maximization strategies for multi-atlas multi-label segmentation. *Inf Process Med Imaging.* Jul.2003 18:210–21. [PubMed: 15344459]
16. Udupa J, et al. A framework for evaluating image segmentation algorithms. *Comp Med Imag Graphics.* 2006; 30:75–87.

17. Heckemann RA, et al. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *Neuroimage*. 33:115–26. 1006. [PubMed: 16860573]
18. Langerak, TR., et al. IEEE International Symposium on Biomedical Imaging: From Nano to Micro. Boston, Massachusetts: 2009. Label Fusion Using Performance Estimation with Iterative Label Selection; p. 1186-9.
19. Lotjonen JM, et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage*. Feb 1.2010 49:2352–65. [PubMed: 19857578]
20. Warfield SK, et al. Validation of image segmentation by estimating rater bias and variance. *Philosophical transactions of the Royal Society*. 2008; 366:2361–2375.
21. Commowick O, War SK. A Continuous STAPLE for Scalar, Vector, and Tensor Images: An Application to DTI Analysis. *IEEE Trans. Med. Imag.* 2009; 28:838–846.
22. Xing, F., et al. SPIE Medical Imaging. Lake Buena Vista, FL: 2011. Statistical Fusion of Continuous Labels: Identification of Cardiac Landmarks.
23. Landman, B., et al. SPIE Medical Imaging Conference. San Diego, CA: 2010. Simultaneous Truth and Performance Level Estimation with Incomplete, Over-complete, and Ancillary Data.
24. Bogovic, JA., et al. SPIE Medical Imaging Conference. San Diego, CA: 2010. Statistical Fusion of Surface Labels provided by Multiple Raters, Over-complete, and Ancillary Data.
25. Commowick O, Warfield S. Incorporating Priors on Expert Performance Parameters for Segmentation Validation and Label Fusion: A Maximum a Posteriori STAPLE. *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. 2010;2010:25–32.
26. Landman, B., et al. International Society for Magnetic Resonance in Medicine. Honolulu, Hawaii: 2009. Efficient Anatomical Labeling by Statistical Recombination of Partially Label Datasets.
27. Asman, AJ., et al. SPIE Medical Imaging. Lake Buena Vista, FL: 2011. Foibles, Follies, and Fusion: Assessment of Statistical Label Fusion Techniques for Web-Based Collaborations using Minimal Training.
28. Collins, D. McGill University; 1994. 3D Model-based segmentation of individual brain structures from magnetic resonance imaging data.
29. Rexilius J, et al. A novel nonrigid registration algorithm and applications. 2010:923–931.
30. McLachlan, G., et al. The EM algorithm and extensions. Wiley; New York: 1997.
31. Moon T. The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*. 2002; 13:47–60.
32. Rohlfing T, et al. Performance-Based Classifier Combination in Atlas-Based Image Segmentation Using Expectation-Maximization Parameter Estimation. *IEEE Transactions on Medical Imaging*. 2004; 23:983–994. [PubMed: 15338732]
33. Stein, JA., et al. SPIE Medical Imaging. Lake Buena Vista, FL: 2011. Characterizing and Optimizing Rater Performance for Internet-based Collaborative Labeling.
34. Dice L. Measures of the amount of ecologic association between species. *Ecology*. 1945; 26:297–302.
35. Jaccard P. The distribution of flora in the alpine zone. *New Phytol*. 1912; 11:37–50.
36. Dempster AP, et al. Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological*. 1977; 39:1–38.
37. Zhang J. The mean field theory in EM procedures for Markov random fields. *Signal Processing, IEEE Transactions on*. 2002; 40:2570–2583.
38. Kapur T, et al. Enhanced spatial priors for segmentation of magnetic resonance imagery. *Medical Image Computing and Computer-Assisted Intervention—MICCAI'98*. 1998:457.
39. Zhang J, Hanauer G. The application of mean field theory to image motion estimation. *Image Processing, IEEE Transactions on*. 2002; 4:19–33.

Fig. 1.

The inaccuracies of the STAPLE model of rater behavior. A representative slice from the truth model is shown in (A). The expected STAPLE model of rater behavior can be seen in (B). STAPLE operates under the assumption that there is a uniform probability that any given rater would mis-label a given voxel. The observed model of rater behavior can be seen in (C). The primary difference between (B) and (C) is that the human raters showed a clear inclination to mislabel boundary pixels and other ambiguous regions.

Fig. 2.

The COLLATE model. The hidden data in the COLLATE E-M algorithm can be seen in (A), (B) and (C). These images (the true labels, rater confusion matrices and consensus map) represent the complete set of data that COLLATE attempts to estimate. The generative model of rater behavior can be seen in (D). This flowchart shows the path from an input voxel on some clinical data to a single observation. A flowchart demonstrating the way in which COLLATE takes input observations and estimates the hidden data can be seen in (E). Note the inclusion of priors in conditional probability that is estimated to generate the maximum *a posteriori* estimate of the hidden data. Example estimates of the hidden data can be seen in (F), (G) and (H). These images are meant for visual inspection and demonstration of the COLLATE algorithm.

Fig 3.

Results for simulation 1 using the COLLATE model of rater behavior. A representative slice from the truth model can be seen in (A). (B) and (C) represent example observations of the slice seen in (A). The STAPLE estimate using 8 coverages can be seen in (D). The COLLATE estimate using the same observations can be seen in (E). Note the improvement of the estimate seen in (E) over the estimate seen in (D). The estimated consensus map can be seen in (F). These are the expected results given the behavior of the raters seen in (B) and (C). The truth estimation accuracy comparison of the two algorithms in the confusion region for varying numbers of coverages can be seen in (G). The confusion matrix accuracy comparison for varying number of coverages can be seen in (H). The gray bars seen on (G) and (H) correspond to the number of coverages used in the estimations seen in (D), (E) and (F).

Fig 4.

Results for simulation 2, the COLLATE sensitivity with respect to the estimated confusion region size data-adaptive prior. The sensitivity of the confusion region size prior can be seen in (A) and (B). (A) represents the accuracy of the truth estimation with varying prior estimates from 0.05 to 0.95 for a given confusion region size of 0.5. The accuracy of the truth estimation is presented as a percent improvement over the STAPLE estimate for the same set of input observations. (B) represents the accuracy of the confusion matrix estimation with varying prior estimates from 0.05 to 0.95 for a given confusion region size of 0.5. All data presented in this Figure use six coverages for both COLLATE and STAPLE.

Fig 5.

Results for simulation 2, the accuracy of the COLLATE algorithm with respect to the confusion region size. This tests the ability of the algorithm to estimate the confusion region size. (A) represents the percent improvement for COLLATE over the STAPLE estimation for confusion region sizes varying from 0.05 to 0.95. (B) represents the average absolute error at each element in the confusion matrices for varying confusion region size. Note that the COLLATE estimate accuracy remains constant while the quality of the STAPLE estimate varies depending upon the size of the confusion region. All data presented in this Figure use six coverages for both COLLATE and STAPLE.

Fig 6.

Results for simulation 3 using boundary random raters. A representative slice from the truth model can be seen in (A). The numbers on (A) identify the numbers corresponding with the given labels so that the confusion matrix representations can be fully understood. (B) and (C) represent example observations of the slice seen in (A). The STAPLE estimate using eight coverages can be seen in (D). The COLLATE estimate using the same observations can be seen in (E). Note the improvement of the estimate seen in (E) over the estimate seen in (D). The estimated consensus map can be seen in (F). These are the expected results given the behavior of the raters seen in (B) and (C). The truth estimation accuracy comparison of the two algorithms in the confusion region for varying numbers of coverages can be seen in (G). The gray bar indicates the number of coverages corresponding to the estimates seen in (D), (E), (F), (H) and (I). An example confusion matrix from a single rater from the STAPLE estimate and the COLLATE estimate using eight coverages can be seen in (H) and (I), respectively.

Fig 7.

Empirical experiment using human raters. A representative slice from the 10-slice truth model can be seen in (A). The numbers on (A) identify the numbers corresponding with the given labels so that the confusion matrix representations can be fully understood. The STAPLE estimate using eight coverages can be seen in (B). The COLLATE estimate using the same observations can be seen in (C). Note the improvement of the estimate seen in (C) over the estimate seen in (B). The observed model of rater behavior can be seen in (D). The color value at each voxel corresponds to the fraction of raters that incorrectly labeled the given voxel. The estimated consensus map can be seen in (E). This is a rather conservative estimate, but it is the expected result given that $P(C=0) = 0.99$. The averaged confusion matrices for both the STAPLE and COLLATE estimations can be seen in (F). Note the fact that the COLLATE estimate appears to be a nearly constant valued diagonal matrix, while the STAPLE estimate is biased towards certain labels, particularly the background (first column). The range of Jaccard Similarity Coefficient values and Dice Similarity Coefficient values can be seen in (G). In both cases, a paired t-test results in a p-value of less than 0.001, indicating that the COLLATE estimates are significantly better than the STAPLE estimates.

Fig 8.

Results for simulation 4 using STAPLE model of rater behavior. A representative slice from the truth model can be seen in (A). (B) and (C) represent example observations of the slice seen in (A). The STAPLE estimate using eight coverages can be seen in (D). The COLLATE estimate using the same observations can be seen in (E). The estimated consensus map can be seen in (F). The true consensus map would indicate that the entire map is a confusion region, but due to the fact that some raters agree at various voxels, consensus is estimated at isolated voxels. The truth estimation accuracy comparison of the two algorithms for varying numbers of coverages can be seen in (G). The mean COLLATE estimate is slightly better for low numbers of coverages, but both COLLATE and STAPLE converge to the same accuracy level for seven or more coverages. The confusion matrix accuracy comparison for varying number of coverages can be seen in (H). The gray bars seen on (G) and (H) correspond to the number of coverages used in the estimations seen in (D), (E) and (F).