Published in final edited form as:

IEEE Trans Med Imaging. 2012 June ; 31(6): 1326–1336. doi:10.1109/TMI.2012.2190992.

Formulating Spatially Varying Performance in the Statistical Fusion Framework

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

To date, label fusion methods have primarily relied either on global (e.g. STAPLE, globally weighted vote) or voxelwise (e.g. locally weighted vote) performance models. Optimality of the statistical fusion framework hinges upon the validity of the stochastic model of how a rater errs (i.e., the labeling process model). Hitherto, approaches have tended to focus on the extremes of potential models. Herein, we propose an extension to the STAPLE approach to seamlessly account for spatially varying performance by extending the performance level parameters to account for a smooth, voxelwise *performance level field* that is unique to each rater. This approach, Spatial STAPLE, provides significant improvements over state-of-the-art label fusion algorithms in both simulated and empirical data sets.

Keywords

STAPLE; Spatial STAPLE; Rater Models; Statistical Fusion; Multi-Atlas Segmentation

I. Introduction

Segmentation of structures of interest in medical images is essential to understanding and quantifying clinically relevant biological conditions. The long-established gold standard for delineation or segmentation of medical images has been manual voxel-by-voxel labeling by an expert anatomist [1, 2]. Unfortunately, this process is fraught with both inter- and intra-rater variability [3, 4] (e.g. on the order of approximately 10% by volume [5, 6]). On the other end of the spectrum, it would be ideal if fully automated segmentation techniques resulted in highly accurate and robust segmentation estimations. For specific applications, fully automatic techniques (e.g., graph cuts, intensity based techniques) have been refined to provide highly robust and accurate segmentations. For example, the techniques provided by FreeSurfer represent the current state of the art for cortical [7] and subcortical [8] segmentation. Unfortunately, the research and development of new, fully automatic, application dependent, segmentation methods can take years due to a anatomical and imaging variability.

As a result, the *de facto* baseline techniques to provide accurate and robust segmentations revolve around the use of 1) multiple anatomical experts, 2) a database of labeled atlases to

estimate the underlying segmentation (i.e. multi-atlas based segmentation) or 3) the result of multiple segmentation algorithms. Label fusion lies at the heart of these techniques as it defines an optimal way in which information can be combined (or fused) into a single estimate.

There are two primary techniques to perform label fusion: (1) voting-based techniques and (2) statistically driven techniques to simultaneously estimate the underlying segmentation and rater accuracy. The simplest voting technique is to use a majority vote. More recent techniques have shown that utilizing level set based techniques (i.e. a log-odds majority vote [9]) improve segmentation accuracy. For multi-atlas segmentation, voting techniques that integrate intensity information into the fusion process using either global similarity [9-11] or local similarity [9, 12, 13] have shown dramatic improvement in segmentation accuracy (e.g. a Locally Weighted Vote (LWV) [9]). Alternatively, statistically driven techniques (e.g. Simultaneous Truth and Performance Level Estimation (STAPLE) [14]) provide an alternative approach in which rater performance is integrated into the estimation process. Several advancements to STAPLE have been proposed to advance the statistical fusion framework. For example, Rohlfing, *et al* [15] suggested ignoring “consensus voxels” (i.e., voxels where all raters agree) when constructing the performance level parameters, which was later theoretically generalized and formalized [16]. More recently, techniques for incorporating priors [17, 18] and meta-analyses to incorporate intensity have been considered [19]. In this manuscript we compare our algorithm to the algorithms that represent the foundations of the various fields of label fusion, including log-odds majority vote, STAPLE ignoring consensus voxels, and, when intensity information is available, a locally weighted vote.

Optimality of statistical fusion frameworks hinges upon the validity of the underlying stochastic model of how a rater errs (i.e., labeling process model). Existing methods to simultaneously estimate rater performance (e.g., STAPLE approaches) have used spatially invariant models (i.e. the probability of error does not change voxel-by-voxel) [10, 14-16, 20]. On the other end of the spectrum, voting based techniques (including global [9-11] and local [9, 12, 13] approaches) ignore the spatial relation between voxels and fuse each voxel independently. However, these models of observation behavior are at odds with an intuitive notion of rater performance. Regardless of the fusion context (e.g. human raters or multiple atlases), it would not be surprising if the manner in which the labels were observed resulted in spatially varying performance. For example, in a multi-atlas context (Figure 1), there exist regions where the quality of registration in a multi-atlas based approach is better than others, and, not surprisingly, this quality level is generally smooth on a semi-local level. As a result, we are left with several issues in the field of label fusion, primarily we lack, i) an alternative between the “global” STAPLE method and “local” voting techniques and ii) an understanding of the observed disparity between STAPLE and majority vote on multi-atlas segmentation. By addressing these two issues, a statistical fusion framework would more accurately model the manner in which labels were observed, and, hopefully, result in robustly fused segmentation estimates, regardless of the context.

Herein, we propose an extension to the STAPLE approach, “Spatial STAPLE,” to seamlessly account for spatially varying performance. This is accomplished by extending the performance level parameters present in STAPLE to a voxelwise *performance level field* that is unique to each rater. By estimating a field instead of global parameters, Spatial STAPLE captures the spatially varying behavior that is often present in 1) the fusion of human raters, 2) multiple segmentation algorithms and 3) multiple registered atlases. Additionally, as will become evident later in this manuscript, these performance level fields are guaranteed to be smooth, which allows for a seamless semi-local approach to the fusion of information. This approach is validated for both simulated and empirical datasets modeling

both the fusion of human raters and a multi-atlas context. The results suggest that Spatial STAPLE provides a valuable framework that can be utilized to construct an optimal framework for fusion of observed labels, regardless of the context.

It is important to note that, while Spatial STAPLE is demonstrated to perform very well on a variety of label fusion problems, it is, at its heart, a model of human observation behavior. Like STAPLE, Spatial STAPLE 1) models the raters as a group of collectively unbiased observers and 2) does not integrate intensity directly into the estimation process. As a result, a claim that Spatial STAPLE is consistently applicable to a multi-atlas context requires further validation. Herein, we demonstrate that Spatial STAPLE is capable of outperforming multi-atlas fusion methods (e.g., a locally weighted vote) for a CT segmentation application. However, it is unlikely that Spatial STAPLE would be able to outperform multi-atlas fusion techniques for problems in which intensity information provides highly valuable information about the complex relationships between labels and intensity (e.g., whole brain segmentation [9, 11, 21]).

Similar techniques to the ones presented in this paper have been proposed. Sabuncu, *et al* [9] proposed accounting for semi-local performance by augmenting locally weighted vote to include a Markov Random Field (MRF). However, this technique is particularly sensitive to using images whose intensities are normalized to one another. This manuscript extends an initial conference publication of the same underlying theory [21] with additional derivations, discussion, experiments and extensions. Subsequently, block-wise neighborhoods were proposed within a meta-algorithm for fusion of local classifiers [19] based upon a Maximum *a posteriori* (MAP) STAPLE framework [17]. Here, we focus specifically on the *performance level field* theory for statistical fusion rather than statistical boosting/meta-analysis.

This paper is organized in the following manner. Section II describes the Spatial STAPLE algorithm and discusses initialization and implementation. Section III compares Spatial STAPLE to traditional STAPLE, majority vote and locally weighted vote on a series of experiments and simulations. Lastly, Section IV provides additional discussion and brief concluding remarks.

II. Theory

The following derivation of Spatial STAPLE closely follows previous derivations of fusion frameworks [14, 16, 21].

A. Problem Definition

Consider an image of N voxels with the task of determining the correct label for each voxel in that image. Consider a collection of R raters that provide an observed delineation for each of N voxels exactly once. 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 possible 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 the traditional rater model presented in [14], the raters' quality of observation is characterized by $\boldsymbol{\theta}$, the performance level parameters for all raters. In this model, each element, θ_j (i.e., the performance level parameters for rater j) is an $L \times L$ confusion matrix, where each element in the matrix, $\theta_{js's}$ represents the probability that rater j would observe label s given that the underlying true label is s . These performance level parameters are

global parameters that are utilized at all voxels in order to obtain the estimate of the true segmentation.

Here, we extend these performance level parameters to characterize the performance of each rater with respect to spatial position. As a result, we estimate a *performance level field* for each rater, where each element of $\theta, \theta_j(\vec{x})$, represents the performance level (or confusion matrix) associated with rater j at spatial coordinate $\vec{x}$. Additionally, we define $B(\vec{x})$ to be the pooling region (i.e. the spatial region) over which $\theta_j(\vec{x})$ is influenced. We simplify this construct such that the performance level field is discretely defined at every voxel, the $\theta_j(\vec{x}) \rightarrow \theta_{ji}$ and the pooling regions is defined as a collection of voxels given by $B(\vec{x}) \rightarrow B_i$. Figure 2 illustrates the performance level field representation.

B.Spatial STAPLE Algorithm

The goal of Spatial STAPLE is to accurately estimate the performance level field of the R raters given the rater segmentation decisions and the estimation of the truth. The estimated performance level field will be calculated such that it maximizes the complete data log likelihood function

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

It is assumed that the segmentation decisions are conditionally independent given the true segmentation and the performance level parameters, that is $(D_{ij} | T_i \theta_j) \perp (D_{ij'} | T_i \theta_{j'}) \forall j \neq j'$. This model expresses the assumption that the raters derive their segmentations of the same truth model independently from one another and that the quality of the segmentations is captured by the estimation of the performance level field. The estimated performance level parameters for a given rater at differing voxels are not necessarily conditionally independent.

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 $\mathbf{T}$. The true segmentation $\mathbf{T}$ is regarded as the missing or hidden data, and is unobservable. Let θ_{ji} be the covariance, or confusion, matrix associated with rater j at voxel i and let

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{R1} \\ \theta_{12} & \theta_{22} & \dots & \theta_{R2} \\ \dots & \dots & \dots & \dots \\ \theta_{1N} & \theta_{2N} & \dots & \theta_{RN} \end{bmatrix} \quad (2)$$

be the complete set of unknown parameters for the R segmentations. Let $f(\mathbf{D}, \mathbf{T} | \theta)$ denote the probability mass function of the random vector corresponding to the complete data. The complete data log likelihood function is presented as $\ln L_c \{\theta\} = \ln f(\mathbf{D}, \mathbf{T} | \theta)$. We apply E-M to iteratively estimate and maximize the complete data log likelihood function. Let k denote the iteration for which all estimates were obtained. For more detail on E-M in the statistical fusion model see [14, 16].

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

We first derive an expression for the conditional probability density function of the true segmentation at each voxel given the raters decisions and the previous estimate of the performance fields, $W_{si}^{(k)} \equiv f(T_i=s|D_i, \theta_j^{(k-1)})$. Applying Bayes' rule and the fact that all of the observations are conditionally independent we obtain a MAP formulation of the underlying segmentation

$$W_{si}^{(k)} = \frac{f(T_i=s) \prod_j f(D_{ij}=s' | T_i=s, \theta_{ji}^{(k-1)})}{\sum_n f(T_i=n) \prod_j f(D_{ij}=s' | T_i=n, \theta_{ji}^{(k-1)})} \quad (3)$$

where $f(T_i=S)$ represents an *a priori* estimate of the true segmentation and $\theta_j^{(k-1)}$ is the prior estimate of the performance level fields. The distribution $f(T_i=S)$ will be discussed more thoroughly later in this manuscript. Finally, as will be seen in the calculation of the performance level fields, the distribution given by $f(D_{ij}=s' | T_i=s, \theta_j^{(k-1)})$ simplifies directly to $\theta_{jis's}^{(k-1)}$. Thus, the final equation for W is given by

$$W_{si}^{(k)} = \frac{f(T_i=s) \prod_j \theta_{jis's}^{(k-1)}}{\sum_n f(T_i=n) \prod_j \theta_{jis'n}^{(k-1)}} \quad (4)$$

This estimation of the conditional expectation of the complete data log likelihood function is almost identical to the STAPLE approach. The major difference is the utilization of the *performance level field* (i.e., $\theta_{jis's}^{(k-1)}$) as opposed to the global performance level parameters as seen in [14].

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

Given the estimated weight variable, $W_{si}^{(k-1)}$, which represents the conditional probability that the true segmentation of voxel i is equal to label S , it is now possible to estimate the rater performance field that maximizes the conditional expectation of the complete data log likelihood function. Considering each rater and voxel separately, we find the field estimates $\theta_{ji}^{(k)}$ by iterating only over the voxels given by pooling region B_i

$$\theta_{ji}^{(k)} = \arg \max_{\theta_{ji}} \sum_{i' \in B_i} E \left[\ln f(D_{i'j} | T_{i'}, \theta_j) | D, \theta_j^{(k-1)} \right] \quad (5)$$

where, $\theta_{ji}^{(k)}$ is only defined over voxels $i' \in B_i$. Carrying out the expectation yields

$$\theta_{ji}^{(k)} = \arg \max_{\theta_{ji}} \sum_{s'} \sum_{i' \in B_i: D_{i'j}=s'} \sum_s W_{si'}^{(k)} \ln \theta_{jis's}^{(k-1)} \quad (6)$$

At this point, we are left with the task of maximizing all of the elements in the performance level field using the constraint $\sum_s \theta_{jis's} = 1$ which can be solved with a Lagrange multiplier approach. The final solution for $\theta_{jis's}^{(k)}$ is given by

$$\theta_{jis's}^{(k)} = \frac{\sum_{i' \in B_i: D_{i'} = s'} W_{si'}^{(k)}}{\sum_{i' \in B_i} W_{si'}^{(k)}}. \quad (7)$$

E. Accounting for Limited Data and Computational Concerns

Particularly in multi-atlas based segmentation, the spatial quality of an observed segmentation can vary dramatically in a relatively small region (Figure 1). As a result, the number of voxels in a given pooling region, B_i should be relatively small in order to accurately characterize these semi-local performance variations. However, if the number of voxels contained in a given pooling region is too small, the performance level field will be unstable due to limited data. Additionally, if a given label is not observed (or only observed a handful of times) in the spatial region B_i then it will be impossible to estimate the related elements in the associated confusion matrix $\theta_{ji's}$. Thus, we introduce the idea of using a whole-image estimate of the performance level parameters for regularization.

For computational and stability concerns, we introduce an implicit prior in the following form. Let $\theta_j^{(0)}$ be the matrix associated with rater j estimated from an appropriate algorithm. The estimation of the performance level parameters seen in (13) would then be reformulated to be

$$\theta_{jis's}^{(k)} = \frac{\sigma_{ijs'} \theta_{jis's}^{(0)} + \sum_{i' \in B_i: D_{i'} = s'} W_{si'}^{(k)}}{\sigma_{ijs'} \sum_{s'} \theta_{jis's}^{(0)} + \sum_{i' \in B_i} W_{si'}^{(k)}} \quad (8)$$

where $\sigma_{ijs'}$ is a scale factor that is dependent upon the size of the pooling region, B_i , rater j and label s' . Our empirically derived expression for this scale factor is

$$\sigma_{ijs'} = I\left(N_{ijs'} < \frac{|B_i|}{L}\right) \left(\frac{|B_i|}{L} - N_{ijs'}\right) \kappa \quad (9)$$

where $I(\cdot)$ is the indicator function, $N_{ijs'}$ is the number of times rater j observed label s' in pooling region B_i , $|B_i|$ represents the cardinality of the pooling region (i.e. the number of voxels in the region) and κ is a scalar constant. Unless otherwise noted, the value of κ is unity for all presented experiments. This factor adjusts the impact of the implicit global estimate of performance on the estimate of performance for given rater j , label s' and voxel i .

Numerous approaches could be used to construct the performance level prior (e.g. STAPLE [14], COLLATE [16], Majority Vote, Locally Weighted Vote, etc..). In general, for the fusion of human raters (i.e. when no intensity information is available), we use STAPLE. For multi-atlas segmentation we use Majority Vote while ignoring consensus voxels [15].

The approaches presented in (8) and (9) utilize the observed segmentations to construct a non-parametric estimate of the underlying global performance level parameters. Alternatively, a parametric approach could be used to provide more stability in the performance level field, e.g., [17] used a component-wise beta distribution for the performance level parameters. To date, parametric methods have neglected the interdependence of the distribution of individual entries in the performance level matrix; hence, we advocate non-parametric approaches. Formulating the estimation of performance level parameters optimally in a maximum *a posteriori* framework remains an open problem.

Moreover, calculating confusion matrixes for all raters at all voxels is a daunting challenge from both a computational and resource perspective. Hence, we seek to sample and interpolate this field to reduce algorithm complexity. The sample locations are referred to as *seed points*. Herein, we apply linear interpolation using a rectilinear grid.

In this context, implementation of Spatial STAPLE's performance level field presented in this paper can be interpreted as a collection of sliding windows with varying levels of overlap. Herein, we use the same size window for all seed points and report this size as a fraction of the field of view in each cardinal direction. In other words, a window size of 0.1, would represent a window that is $0.1X \times 0.1Y \times 0.1Z$, where X , Y , and Z are the length of each of the dimensions on the input image. Lastly, the amount of overlap between windows is reported as the fraction linear overlap along each of the principle directions, i.e., an overlap of 0.5 would indicate that there is 50% overlap between consecutive windows along the X , Y , and Z directions.

F. Initialization Strategy, Convergence Detection, and the Prior Distributions

1) Initialization—Spatial STAPLE may be initialized by either providing an initial estimate of θ or W . Herein, Spatial STAPLE is initialized with an initial estimate of W as the results of a majority vote algorithm. Alternatively, an initial estimate of θ could be provided, however, if initialized in the manner suggested in [14] then it is essentially the same as initializing W to a majority vote estimate.

2) Convergence—Spatial STAPLE is guaranteed to converge given its use of E-M [22, 23]. In this implementation, we detect convergence by monitoring the change in the trace of the confusion matrix estimates at each of the seed points. We use a threshold of $\epsilon = 1 \times 10^{-5}$ for all simulations and empirical experiments presented in this paper. In the worst case empirical trials presented here (i.e. low number of raters, low quality raters) the algorithm converged in less than 20 iterations.

3) Prior Distribution—As with STAPLE, the *a priori* distribution, $f(T_i = S)$, must be defined. This can range from a global parameter to a spatially varying prior. As with the STAPLE implementation, we let $\gamma_s = f(T_i = S)$ and define it as the empirically observed label frequencies for a most general definition. When available, intensity information can be integrated into this prior. In our case, this spatially varying prior would be defined in a manner that is identical to a log-odds majority vote [9]. We let $\psi_{is} = f(T_i = S)$ and define:

$$\psi_{is} = \frac{1}{Z} \sum_{j=1}^R e^{p \tilde{D}_{ji}^l} \quad (10)$$

where Z is the partition function, $\tilde{D}_{ji}^l$ is the value of the signed distance transform of the observations of rater j and voxel i for label l . The parameter $p > 0$ defines the slope constant of the log-odds type approach (herein, we use $p = 1$). Finally, for simplicity we define the final value of $f(T_i = S)$ in terms of both the global prior and the spatially varying prior

$$f(T_i = s) = \alpha \psi_{is} + (1 - \alpha) \gamma_s \quad (11)$$

where $\alpha \in [0, 1]$ is a parameter that creates a continuum between local and global approaches to prior specification. In general, for STAPLE and Spatial STAPLE we use a prior governed by $\alpha = 1$.

III. Methods and Results

A. Implementation and Evaluation

We present four experiments: i) a simulation modeling human behavior, ii) an empirical fusion of human raters for the segmentation of malignant meningioma, iii) a simulation modeling the fusion of multiple whole-brain segmentation algorithms using multiple results from a locally weighted vote, and iv) an empirical, multi-atlas based experiment using expert-labeled head and neck CT scans. All experiments were implemented in MATLAB (Mathworks, Natick, MA); complete source code is available via the “MASI Label Fusion” project on the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC).

B. Simulation using a Boundary Model of Human Behavior

First, we compared the statistical fusion methods using a model of human behavior by simulating raters that only miss by inaccurately labeling boundary voxels (Figure 3A and 3B, similar to [16]). The truth model consisted of a 3-D volume ($80 \times 60 \times 60$) with 15 embedded labeled cubes. The truth model was observed by 16 different raters, where each rater was “perfect” in one sixteenth of the total volume and exhibited boundary error behavior in the remaining regions. In the regions exhibiting boundary errors, the level of boundary shift for each boundary voxel was chosen from a Gaussian distribution with zero mean and standard deviation — distributed $U(2.5, 3.5)$. When all 16 raters are considered, there exists a single rater that is perfect in each of the sub-regions.

1) Overall Accuracy Comparison—Accuracies of Spatial STAPLE, STAPLE and majority vote were assessed with 10 Monte Carlo iterations for each of a varying numbers of raters (ranging from 5 to 16). Spatial STAPLE was applied with: a 0.15 window size fraction (i.e. $12 \times 9 \times 9$), an overlap ratio of 0.5, linear interpolation, and a global performance estimate based on STAPLE with a value of $\kappa = 1$ (see Eq. 9).

The results from this simulation can be seen in Figure 3C-3F. The performance of Spatial STAPLE along the boundaries between the various regions is qualitatively superior to the other methods, as seen by comparing Figures 3C-3E. Quantitative results demonstrate that increasing the numbers of raters increases the benefit of Spatial STAPLE over both STAPLE and majority vote (Figure 3F).

2) Sensitivity to the Model Parameters—In addition to an overall accuracy comparison, we assess the sensitivity of Spatial STAPLE to the i) the global performance level bias amount (κ) ii) the size of each window in the calculation of the performance level field calculations and iii) the amount of overlap between the various windows in the field calculations (Figure 4). Other than the swept parameter of interest, Spatial STAPLE was applied with a 0.15 window size fraction, an overlap ratio of 0.5, and a value of $\kappa = 1$. The results from these experiments are reported as the percent improvement over STAPLE (in terms of fraction voxels correct).

The global performance level bias (κ , see Eq. 9) was swept in 21 logarithmic steps between 10^{-3} to 10^3 (Figure 4A). For both very low and very high bias amounts the accuracy of Spatial STAPLE degrades. However, as the amount of bias approaches unity we see that Spatial STAPLE is vastly superior to STAPLE. This level of improvement plateaus for approximately two orders of magnitude indicating that the accuracy of the algorithm is not particularly sensitive to the strength of the biasing prior. Note that as the strength of the bias increases, the accuracy of Spatial STAPLE converges to the accuracy level of STAPLE due to the fact that the prior estimate of the performance level parameters are estimated from the STAPLE algorithm.

The window size parameter was swept in 20 linear steps between 0.05 (~1-2 voxel windows) and 0.95 (essentially global fusion) (Figure 4B). For very small window sizes (i.e. 0.05) the accuracy of Spatial STAPLE is worse than STAPLE due to the inability to estimate accurate performance using such few voxels. However, for relatively small window sizes (i.e. between 0.15 and 0.3), the accuracy of Spatial STAPLE is at a maximum. For window sizes beyond 0.2, the accuracy slowly decreases and converges to an accuracy level that is approximately equal to that of STAPLE.

The window overlap parameter was swept in 20 linear steps between 0.05 (very little overlap) and 0.925 (~3-4 voxel difference) (Figure 4C). For increasing amounts of overlap the accuracy of Spatial STAPLE increases, as the regional performance level parameters more accurately reflect the area surrounding the voxel of interest. However, beyond 0.5 overlap the increase in accuracy is relatively small. Additionally, as the amount of overlap increases the computational time significantly increases as well.

C. Empirical Fusion of Minimally Trained Humans for Segmentation of Malignant Meningioma

Second, we analyze the accuracy of majority vote, STAPLE, and Spatial STAPLE on the fusion of multiple human raters for the segmentation of malignant meningioma. The cancer patients utilized in this paper are a collection of 15 pre-operative T1-weighted brain MRI scans based on varied (but standard of care) imaging protocols that were obtained in anonymous form under IRB approval. The resolution of each of the cancerous brains was $1 \times 1 \times 3 \text{ mm}^3$. The corresponding “ground truth” labels associated with each of the cancerous brains were obtained from an expert labeler. A collection of 60 minimally trained undergraduate students provided labels on a slice-by-slice basis (with the students providing anywhere between approximately 10 and 250 observed slices).

The results for the fusion of 8 of these cancerous brains are presented in Figures 5 and 6. The remaining 7 volumes were used as training data and were used to compute initial estimates of the performance level parameters (see [20] for details). Both STAPLE and Spatial STAPLE used these initial parameter estimates using a bias value of $\kappa = 1$. For both algorithms, consensus voxels were ignored so that the excessive background did not adversely affect the segmentation accuracy and a global prior was used ($\alpha = 1$, see Eq. 11). Lastly, instead of using regional performance level estimates on a window basis, unique performance level parameters were constructed for each observed slice for Spatial STAPLE. This allows us to model the way in which raters evolve over time (e.g. a degradation over time or a learning curve).

An overall accuracy comparison between the various models of human behavior is presented in Figure 5 with respect to increasing numbers of observations per slice (from 3 to 10). The accuracy is reported in terms of the Dice Similarity Coefficient (DSC) [24] and the boxplots represent the spread of accuracy across the 8 fused volumes. For all numbers of observations per slice Spatial STAPLE significantly outperforms STAPLE and majority vote (in a paired t-test, $p < 0.05$). Interestingly, beyond 5 observations per slice very little improvement in terms of overall accuracy is gained for any of the various algorithms. This improvement indicates that Spatial STAPLE is able to accurately characterize the spatially varying performance exhibited by the minimally trained undergraduate students. Lastly, qualitative results, using 8 observations per slice, are presented in Figure 6. For each of the presented examples Spatial STAPLE provides a significantly more accurate estimate of the underlying segmentation than STAPLE or majority vote. Particularly, for the problem of cancer segmentation, accurately localizing cancerous regions is of the utmost of clinical importance.

D. Simulation of Multi-Algorithm Fusion for Whole-Brain Segmentation

Next, we evaluate the process of fusing multiple algorithms for whole brain segmentation by using a collection of 5 estimates from a locally weighted vote algorithm (using a random collection of 5 atlases per estimate). A collection of 15 whole brain segmentations were utilized with 26 labels per brain. The labels range from large structures (e.g. cerebral gray matter) to small, deep brain structures (e.g. hippocampus). Each of the brains is part of the Open Access Series of Imaging Studies (OASIS) [25] data set and the labels were acquired using the brainCOLOR protocol (<http://www.braincolor.org/>). A representative segmentation can be seen in Figure 1A. The pairwise registrations were performed using an affine registration using FLIRT [26]. Since this experiment involves fusing the results of a labeling approach (the output of a $N=5$ locally weighted vote), only fusion techniques that don't utilize intensity differences are considered: majority vote, STAPLE and Spatial STAPLE. For both Spatial STAPLE and STAPLE a spatially varying prior was used (i.e., $\alpha = 1$, see Eq. 19) and consensus voxels were ignored.

The results from a leave-one-out cross-validation (LOOCV) study for all 15 of the brains of interest are presented in Figure 7. The accuracy of majority vote, STAPLE and Spatial STAPLE are considered for each of the labels individually. The boxplots represent the spread of results across the various atlases. The results indicate that Spatial STAPLE significantly outperforms (paired t-test, $p < 0.05$) both STAPLE and majority vote for nearly all of the considered labels. The only exceptions are for the left amygdala, left pallidum and the left putamen, where the results are statistically indistinguishable. Note that, particularly for the small labels, the accuracy of the estimates is less than what could be achieved with non-rigid registration and label fusion.

E. Empirical Experiments using Expert-labeled Head and Neck CT scans

Lastly, we analyze the accuracy of statistical fusion algorithms on an empirical multi-atlas based study. Computed tomography (CT) images were acquired from 15 patients who underwent intensity-modulated radiation therapy (IMRT) for larynx and base of tongue cancers and were expertly labeled by an interventional radiologist. For details see [27]. Briefly, each data set has in-plane resolution of $\sim 1\text{mm}$ and a slice thickness of 3mm (acquired on a Philips Brilliance Big Bore CT scanner with injection with 80mL of Optiray 320, a 68% ioversol-based nonionic contrast agent). Each volume contained four segmented structures: left parotid, right parotid, right lymph node regions and thyroid. Note that 15 atlases are fairly meager for many applications, but this represents a situation where the accuracy and limits of fusion algorithms are truly tested. All analyses were performed on the full 3D volume. Following an initial affine registration to a common template, the atlases were registered using the Vectorized Adaptive Bases Registration Algorithm (VABRA) [28] and cropped to isolate the neck ($\sim 170 \times 100 \times 80$ voxels).

Here, we compare the results between majority vote, locally weighted vote, STAPLE and Spatial STAPLE. Spatial STAPLE was used with a 0.2 window size fraction and an overlap ratio of 0.5. For both Spatial STAPLE and STAPLE a spatially varying prior was used (i.e., $\alpha = 1$, see Eq. 11) and consensus voxels were ignored. The global performance estimate utilized was from a majority vote estimate (ignoring consensus voxels) with a value of $\kappa = 1$. All accuracy comparisons were performed using the DSC. We analyze the overall accuracy across all labels as well as for each of the individual labels (i.e. the left and right parotid, the right lymph node regions, and the thyroid) using a (LOOCV) study.

The accuracy of majority vote, locally weighted vote, STAPLE and Spatial STAPLE were computed for each of the 15 LOOCV iterations and plotted in Figure 8. A paired two-sided t-test was performed to evaluate significance between the observed DSC of each approach

and that of Spatial STAPLE. Spatial STAPLE results in significantly higher DSC than the other algorithms for the mean DSC and all structures except for the thyroid (where it is statistically indistinguishable from a locally weighted vote). Furthermore Spatial STAPLE is significantly superior to majority vote and STAPLE for all structures. Note, unlike a locally weighted vote, Spatial STAPLE does not use intensity information when estimating the underlying segmentation.

It is important to note that in this scenario (i.e., consensus voxels are ignored and each of the labels is primarily separated from one another), the STAPLE result is nearly equivalent to a structure-wise fusion approach (i.e. fusing each of the labels separately). Thus, the improvement with Spatial STAPLE show that the biasing prior, small window sizes (i.e. smaller than the individual structures) and large overlap provide important accuracy benefits over the traditional multi-label and structure-wise STAPLE approaches.

On average, Spatial STAPLE improves upon locally weighted vote by a DSC value of 0.01. Thus, it is important to inspect whether or not this improvement is of qualitative relevance. Representative expert labels can be seen in Figure 9A and estimates from each of the algorithms can be seen in Figure 9B-E. Visual inspection shows that improvements provided by Spatial STAPLE result in superior label correspondence to anatomy — particularly in the right lymph node regions. Note that defining whether or not this is clinical relevance depends highly on the clinical application. Nevertheless, Spatial STAPLE significantly outperforms a locally weighted vote without using intensity information outlines the importance of regional performance level estimates in the field of label fusion.

IV. Discussion and Conclusion

Herein, we derive and present Spatial STAPLE — a new algorithm for statistically fusing rater label information using a spatially varying model of rater behavior. Spatial STAPLE i) provides significant improvement over the premier label fusion techniques, ii) more accurately reflects the way in which raters and atlases make mistakes than traditional global performance metrics and iii) provides a unified framework that can be used for the gamut of label fusion applications (i.e. the fusion of human raters, multi-atlas applications and the fusion of multiple algorithms). Additionally, Spatial STAPLE is not particularly sensitive to model parameters (Figure 4) which indicate a stable theoretical underpinning.

Like other statistical fusion algorithms [14, 16], Spatial STAPLE utilizes an E-M based approach in which the algorithm simultaneously estimates rater performance and the underlying segmentation. However, unlike its predecessors, Spatial STAPLE explicitly estimates the spatially varying performance of a collection of raters by estimating a voxelwise *performance level field* for each rater. The traditional model of rater behavior utilizes a single global confusion matrix that is utilized at all voxels. However, for many applications, global performance level parameters may fail to model the complexities of the observed labels (Figures 5-9). By introducing a smooth, spatially-varying performance level field, the inherent errors in registration, human performance and algorithmic performance are implicitly modeled. Thus, we dramatically relax the constricting assumptions of the typical rater models and allow for significantly more freedom when estimating the models by which raters make mistakes.

On the other end of the spectrum, voting based algorithms (e.g. majority vote, locally weighted vote) assume, unique and independent performance at all voxels. These types of models provide significant improvement over the traditional statistical fusion models in a multi-atlas context, but, at the same time, fail to model the inherent linkage structure between regional labels in a fusion context. Spatial STAPLE provides a valuable *semi-local* approach to the statistical fusion paradigm.

Perhaps surprisingly, for the head and neck CT application, Spatial STAPLE is able to outperform a locally weighted vote in a multi-atlas context despite the fact that it is inherently a model of human behavior. This is due to several factors including 1) the registrations of the data are actually quite poor, 2) the intensity profile for neighboring tissues are extremely similar, rendering intensity based segmentation difficult, and 3) the atlases exhibit largely spatially varying behavior due to the highly varying anatomy in the head and neck regions. While Spatial STAPLE cannot compete with intensity based techniques for many applications (e.g. whole brain segmentation), there exists a large collection of applications in a multi-atlas context for which intensity information is of limited use.

An important result presented in this paper is the ability of Spatial STAPLE to accurately estimate segmentation in the full gamut of label fusion applications. Previously, the accuracy of label fusion techniques has been highly dependent upon the application. For example, when fusing human raters, STAPLE was generally considered the best algorithm, while for other scenarios; the voting based algorithms generally provided more accurate estimations. By formulating a framework in which spatially varying behavior is characterized and semi-local performance captured, Spatial STAPLE has been shown to provide a robust and consistent framework for the fusion of human raters, multiple algorithms as well as multi-atlas fusion. Additionally, Spatial STAPLE is very versatile by modifying the way in which the semi-local performance is captured. For example, in the cancer segmentation example (Figures 5 and 6) the semi-local performance was considered on a slice-by-slice basis, while for the multi-atlas and algorithm fusion problems the semi-local performance was modeled on a small window basis (Figures 7-9). Given this versatility and robustness, we believe that Spatial STAPLE has the properties to cement itself as an algorithm to be considered regardless of the fusion application.

Spatial STAPLE, as presented, operates under the assumption of voxelwise independence. An alternative approach to explicitly modeling the spatial correlation between structures is to implicitly model them using a Markov Random Field (MRF). As has been done with several other algorithms [9, 14, 29-32], an MRF could readily be applied to Spatial STAPLE in order to regularize the label probability fields. There are several options for MRF approximations that could be utilized (e.g. the Potts [9] or Gibbs [14] models). However, so as to not obfuscate the presentation of results, we did not include a MRF and did not compare to algorithms that utilize MRF's either.

Although the focus of this paper is on utilizing Spatial STAPLE to model general spatially varying behavior, the techniques presented in this paper could be applied to more sophisticated models of human behavior (e.g. Figures 5 and 6) where the foibles and follies of minimally trained raters are taken into account. For example, the concept of crowd-sourcing the labeling problem using minimally trained raters has gained traction [20, 33]. Through simple manipulation of the seed points and window sizes, Spatial STAPLE provides a valuable resource that could, for instance, be used to *explicitly* model a rater's learning curve, degradation over time, and general human mistakes that are often present in the labeling process. Modeling this behavior using either a majority vote or STAPLE would require dramatic changes to the theoretical underpinning governing their optimality.

The implementation of Spatial STAPLE presented in this paper extends the concept of performance level parameters to the concept of a *performance level field* that it is varying on a voxel (local) level through analysis of regional windows and interpolation. In general there are 3 parameters that determine the manner in which the local label fusion is performed 1) the number of windows, 2) the size of the windows and 3) the type of interpolation between windows. Unfortunately, the optimal settings for these parameters depend largely on the

situation. For example, in multi-atlas based segmentation we would expect the registration-driven errors to be highly local, indicating a need for many, small windows in order to accurately characterize the performance of the raters (or registered atlases). On the other hand, for fusion of multiple minimally-trained humans who are observing labels on a slice-by-slice basis (i.e. the cancer segmentation example), the optimal parameters would allow for unique performance level parameters on each slice. This type of framework would allow the algorithm to *implicitly* capture natural human labeling phenomena such as degradation over time, or a learning curve. Spatial STAPLE provides a valuable framework for modeling the manner in which raters observe labels, which can provide optimal techniques for highly varying situations through simple parameter manipulation.

At this point, a discussion on the computational time of Spatial STAPLE compared to the other algorithms is warranted. In general voting-based algorithms are considerably faster than statistically driven algorithms such as STAPLE and Spatial STAPLE due to the fact that they do not require multiple iterations and the fact that there is no cohesive performance assessment integrated into the estimation process. The computational time required for Spatial STAPLE can be approximated as a function of the amount of overlap between the various windows. In general, Spatial STAPLE will take the reciprocal of the fraction overlap times longer than STAPLE due to the fact that the overlap determines whether or not a voxel is iterated over multiple times in the estimation process. With that said, if, for example, more sophisticated interpolation schemes were utilized then the computational time would obviously increase.

Herein, we have limited our consideration of the performance field to simple interpolations on a rectilinear grid. The implementation is computationally tractable and the performance is strong. Nevertheless, further investigation into alternative or data-driven placement of these sampling locations is certainly warranted. Particularly for a multi-atlas context, we believe that this investigation could provide valuable segmentation improvements.

Acknowledgments

The authors acknowledge Dr. Andrew Worth (Neuromorphometrics, Inc., <http://neuromorphometrics.com>) for the exquisitely labeled brain segmentations, Dr. Benoit Dawant for providing expertly labeled data from the IMRT studies and the various undergraduate students for their labels on the cancer study.

This work was supported in part by NIH/NINDS 1R01EB006136, NIH/NINDS 1R01EB006193, NIH/NINDS 1R03EB012461, and NIH/NINDS 1R21NS064534.

References

1. Crespo-Facorro B, Kim J, Andreasen N, O'Leary D, Wiser A, Bailey J, Harris G, Magnotta V. Human frontal cortex: an MRI-based parcellation method. *NeuroImage*. 1999; 10:500–519. [PubMed: 10547328]
2. Tsang O, Gholipour A, Kehtarnavaz N, Gopinath K, Briggs R, Panahi I. Comparison of tissue segmentation algorithms in neuroimage analysis software tools. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*. 2008;2008:3924–8. [PubMed: 19163571]
3. Fiez J, Damasio H, Grabowski T. Lesion segmentation and manual warping to a reference brain: intra- and interobserver reliability. *Human Brain Mapping*. Apr.2000 9:192–211. [PubMed: 10770229]
4. Filippi M, Horsfield M, Bressi S, Martinelli V, Baratti C, Reganati P, Campi A, Miller D, Comi G. Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis. *Brain*. 1995; 118:1593. [PubMed: 8595488]

5. Ashton E, Takahashi C, Berg M, Goodman A, Totterman S, Ekholm S. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. *Journal of Magnetic Resonance Imaging*. 2003; 17:300–308. [PubMed: 12594719]
6. Joe B, Fukui M, Meltzer C, Huang Q, Day R, Greer P, Bozik M. Brain Tumor Volume Measurement: Comparison of Manual and Semiautomated Methods¹. *Radiology*. 1999; 212:811. [PubMed: 10478251]
7. Fischl B, Van Der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D. Automatically parcellating the human cerebral cortex. *Cerebral Cortex*. 2004; 14:11. [PubMed: 14654453]
8. Fischl B, Salat D, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S. Whole Brain Segmentation:: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*. 2002; 33:341–355. [PubMed: 11832223]
9. Sabuncu M, Yeo B, Van Leemput K, Fischl B, Golland P. A Generative Model for Image Segmentation Based on Label Fusion. *IEEE Transactions on Medical Imaging*. 2010; 29:1714–1729. [PubMed: 20562040]
10. Langerak T, van der Heide U, Kotte A, Viergever M, van Vulpen M, Pluim J. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). *IEEE Transactions on Medical Imaging*. 2010; 29:2000–2008. [PubMed: 20667809]
11. Artaechevarria X, Muñoz-Barrutia A. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *Medical Imaging, IEEE Transactions on*. 2009; 28:1266–1277.
12. Isgum I, Staring M, Rutten A, Prokop M, Viergever M, van Ginneken B. Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans. *IEEE Transactions on Medical Imaging*. 2009; 28:1000–1010. [PubMed: 19131298]
13. Heckemann R, Hajnal J, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *Neuroimage*. 2006; 33:115–26. [PubMed: 16860573]
14. Warfield S, Zou K, Wells W. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*. Jul. 2004 23:903–21. [PubMed: 15250643]
15. Rohlfing T, Russakoff D, Maurer C. 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]
16. Asman A, Landman B. Robust Statistical Label Fusion through Consensus Level, Labeler Accuracy and Truth Estimation (COLLATE). *IEEE Transactions on Medical Imaging*. Apr 29.2011
17. 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.
18. Landman B, Asman A, Scoggins A, Bogovic J, Xing F, Prince J. Robust Statistical Fusion of Image Labels. *IEEE Transactions on Medical Imaging*. 2012 In Press.
19. Weisenfeld N, Warfield S. Learning Likelihoods for Labeling (L3): A General Multi-Classifer Segmentation Algorithm. *MICCAI 2011*. 2011:322–329.
20. Landman B, Asman A, Scoggins A, Bogovic J, Stein J, Prince J. Foibles, follies, and fusion: Web-based collaboration for medical image labeling. *NeuroImage*. Aug 2.2011
21. Asman A, Landman B. Characterizing Spatially Varying Performance to Improve Multi-Atlas Multi-Label Segmentation. *Information Processing in Medical Imaging (IPMI)*. 2011:81–92.
22. McLachlan, G.; Krishnan, T.; Corporation, E. *The EM algorithm and extensions*. Wiley; New York: 1997.
23. Moon T. The expectation-maximization algorithm,” *Signal Processing Magazine, IEEE*. 2002; 13:47–60.
24. Dice L. Measures of the amount of ecologic association between species. *Ecology*. 1945; 26:297–302.

25. Marcus D, Wang T, Parker J, Csernansky J, Morris J, Buckner R. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*. 2007; 19:1498–1507. [PubMed: 17714011]
26. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*. 2001; 5:143–156. [PubMed: 11516708]
27. Chen A, Niermann K, Deeley M, Dawant B. Evaluation of multi atlas-based approaches for the segmentation of the thyroid gland in IMRT head-and-neck CT images. *Physics in Medicine and Biology*. 2011; 57:93–11. [PubMed: 22126838]
28. Rohde G, Aldroubi A, Dawant B. The adaptive bases algorithm for intensity-based nonrigid image registration. *IEEE Transactions on Medical Imaging*. Nov.2003 22:1470–9. [PubMed: 14606680]
29. Van Leemput K, Maes F, Vandermeulen D, Suetens P. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*. Oct.1999 18:885–96. [PubMed: 10628948]
30. Yeo B, Sabuncu M, Desikan R, Fischl B, Golland P. Effects of registration regularization and atlas sharpness on segmentation accuracy. *Medical Image Analysis*. Oct.2008 12:603–15. [PubMed: 18667352]
31. Kapur T, Grimson W, Kikinis R, Wells W. Enhanced spatial priors for segmentation of magnetic resonance imagery. *Medical Image Computing and Computer-Assisted Intervention—MICCAI'98*. 1998:457.
32. Zhang Y, Smith S, Brady M. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Trans on Medical Imaging*. 2001; 20:45–57.
33. Asman A, Scoggins A, Prince J, Landman B. Foibles, Follies, and Fusion: Assessment of Statistical Label Fusion Techniques for Web-Based Collaborations using Minimal Training. *Proceedings - Society of Photo-Optical Instrumentation Engineers*. 2011; 7962:79623G.

Fig 1.

Registered atlases exhibit spatially varying behavior. Representative slices from an expertly labeled MR brain image and CT head and neck image are shown in (A). Example registered atlases with their local performance can be seen in (B) and (C). Note that atlases exhibit smooth spatially varying performance that is unique to each atlas.

Fig 2.

Demonstration of the Spatial STAPLE performance level field estimation procedure. An example expert segmentation can be seen in (A) with a collection of registered atlas observations seen in (B). Spatial STAPLE estimates local confusion matrices (C) in order to construct a whole-image estimate of performance that is smooth and spatially varying. The true performance for the atlas seen in (B) can be seen in (D) and the estimated performance from Spatial STAPLE presented in (E). Note that the intensity in (E) is an indication of average “performance” – i.e., the average diagonal element of Θ .

Fig 3.

Results for the human rater simulation. The cross-sectional view of the truth model used in this simulation can be seen in (A). An example observation utilizing the boundary model of human behavior can be seen in (B). Note the fact that there exists a unique region where each rater is perfect in these observations. The corresponding label estimate from Majority Vote, STAPLE and Spatial STAPLE can be seen in (C)-(E), respectively. All displayed estimates were constructed using 10 raters. Lastly, an accuracy analysis can be seen in (F), note that with increasing volumes, Spatial STAPLE continually outperforms both STAPLE and Majority Vote.

Fig 4.

Assessment of Spatial STAPLE sensitivity with respect to various model parameters for the human rater simulation. For each plot the percent improvement exhibited by Spatial STAPLE over STAPLE is assessed. The plot seen in (A) indicates the sensitivity of Spatial STAPLE to the impact of the global estimate of the performance level parameters. (B) indicates the sensitivity to the size of the pooling region (or window) associated with the voxelwise performance estimate. Lastly, plot (C) indicates the sensitivity to the amount of overlap between windows. The window overlap is a proxy for the number of seed points used in the estimation of the performance level field.

Fig 5.

Qualitative results for the human rater cancer label experiment. The accuracy of majority vote, STAPLE and Spatial STAPLE are considered with varying numbers of observations per slice (or “coverages”). For all number of observations per slice, Spatial STAPLE exhibits statistically significant improvement over both majority vote and STAPLE.

Fig 6.

Qualitative results for the human rater cancer labeling experiment. Four separate slices are shown, with the expert labels, majority vote, STAPLE and Spatial STAPLE presented for each example using 8 observations per slice. For all examples Spatial STAPLE is qualitatively superior to both majority vote and STAPLE. The arrows indicate areas of particular improvement exhibited by Spatial STAPLE.

Fig 7.

Quantitative results for the simulation of meta-analysis fusion for whole brain segmentation. The presented results represent the accuracy of majority vote, STAPLE and Spatial STAPLE for all 26 labels across the 15 atlases considered in this experiment. Spatial STAPLE significantly outperforms the other algorithms for nearly all labels (excluding the left amygdala, pallidum and putamen).

Fig 8.

Quantitative results for the segmented CT head and neck data. The mean DSC for all structures can be seen in (A). The DSC value for each of the individual algorithms can be seen in (B)-(E). Spatial STAPLE statistically outperforms locally weighted vote for all labels other than the thyroid despite the fact that Spatial STAPLE does not utilize intensity information.

Fig 9.

Qualitative results for the segmented CT head and neck data. The average mean DSC improvement exhibited by Spatial STAPLE was approximately 0.01 DSC (Fig. 8A). Thus, it is important to assess whether or not this improvement is qualitatively visible. The truth labels can be seen in (A), with the corresponding majority vote, locally weighted vote, STAPLE and Spatial STAPLE estimates seen in (B) – (E).