Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Multiprotocol, multiatlas statistical fusion: theory and application

Andrew J. Plassard
Bennett A. Landman

Multiprotocol, multiatlas statistical fusion: theory and application

Andrew J. Plassard^{a,*} and Bennett A. Landman^{a,b}

^aVanderbilt University, Department of Computer Science, Nashville, Tennessee, United States

^bVanderbilt University, Department of Electrical Engineering, Nashville, Tennessee, United States

Abstract. Multiatlas segmentation offers an exceedingly convenient process by which image segmentation tools can be created from a series of labeled atlases (i.e., raters). However, creation of the atlases is exceedingly time consuming and prone to shifts in clinical/research demands as anatomical definitions are refined, combined, or subdivided. Hence, a process by which atlases from distinct, but complementary, anatomical “protocols” could be combined would allow for greater innovation in structural analysis and efficiency of data (re)use. Recent innovation in protocol fusion has shown that propagation of information across distinct protocols is feasible. However, how to effectively include this information in simultaneous truth and performance level estimation (STAPLE) has been elusive. We present a generalization of the STAPLE framework to account for multiprotocol rater performance (i.e., accuracy of registered atlases). This approach, multiset STAPLE (MS-STAPLE), provides a statistical framework for combining label information from atlases that have been labeled with distinct protocols (i.e., whole brain versus subcortical) and is compatible with the current local, nonlocal, probabilistic, log-odds, and hierarchical innovations in STAPLE theory. Using the MS-STAPLE approach, information from a broad range of datasets can be combined so that each available dataset contributes in a spatially dependent manner to local labels. We evaluate the model in simulations and in the context of an experiment where an existing set of whole-brain labels (14 structures) is refined to include parcellation of subcortical structures (26 structures). In the empirical results, we see significant improvement in the Dice similarity coefficient when comparing MS-STAPLE to STAPLE and nonlocal MS-STAPLE to nonlocal STAPLE. © 2017 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.4.3.034002]

Keywords: multiatlas segmentation; label-fusion; multiprotocol fusion.

Paper 17067R received Mar. 20, 2017; accepted for publication Jul. 26, 2017; published online Aug. 28, 2017.

1 Introduction

The multiatlas technique has become an essential medical image processing approach and been adopted widely for applications ranging from the brain¹ to the abdomen² and pelvic structures.³ The promise of generalizing robust algorithms from limited collections of labeled data without needing to posit specific structural models is highly appealing for clinical applications and rapid prototyping. However, manual labeling of medical images can be extraordinarily resource intensive. For each new application (or even refinement of an existing application), multiatlas methods require labeling a new atlas set.

Atlas-based segmentation is a widely studied topic in the medical image processing literature. In atlas-based segmentation approaches, one or more atlases, or sets of image and expert delineated labels, are used to provide an estimate of the segmentation of a given target image. In the simplest case, one atlas is nonrigidly registered to a target image.⁴ The labels atlas's labels are then deformed following the registration. The resulting labels are then treated as the segmentation of the target image. Since registration is a challenging process and results are generally not guaranteed to be optimal, multiple atlases are commonly used to provide a better estimate of the target

segmentation. When multiple atlases are considered in the segmentation, this approach is known as multiatlas segmentation.⁵

Since multiple atlases are considered in multiatlas segmentation, a function to produce a consensus representation of the atlases is needed. One of the simplest approaches for this is majority vote (MV). In MV, the mode label at each voxel is selected as the label decision for the target image.¹ A secondary approach is the simultaneous truth and performance level estimation (STAPLE) algorithm.⁶ STAPLE estimates a confusion matrix for each atlas that corresponds to the probability of the true label being s given that a particular atlas observes some label s' . The confusion matrix is estimated through expectation maximization (EM) and after the algorithm has converged, the true labels for the target can be estimated. The STAPLE algorithm has since been expanded to incorporate spatially varying⁷ and hierarchically consistent performance.⁸

Many modern segmentation approaches utilize the intensity information to improve the segmentation accuracy. Globally weighted vote calculates a similarity metric between each registered atlas and the target and uses that weight to calculate the weighted mode label at each voxel.⁹ Other techniques expand the weighted voting principle at the voxel level.¹ Since registration is an imperfect process, Coupé et al.¹⁰ proposed incorporating nonlocal correspondence as a weighting metric for

*Address all correspondence to: Andrew J. Plassard, E-mail: andrew.j.plassard@vanderbilt.edu

nonlocal weighted voting. This principal has also been incorporated into the STAPLE framework.¹¹

Consider segmentation of the hippocampus. Numerous protocols (e.g., Ref. 12) exist to delineate the hippocampus in MR images. These protocols vary on the basis of hippocampal white matter, the border between the hippocampus and amygdala, the hippocampal tail, and various other markers. With current techniques, individuals interested in studying the hippocampus are limited to either ignoring all other label sets or using a coarse protocol as an anatomical “stamp” and are left to rectify the space between the protocols. This is clearly suboptimal as it does not allow for joint inference between protocols and does not specifically estimate any one protocol.

Recently, Ref. 13 illustrated that protocols need not be considered fully independent and that a generative latent model could be used to exploit the dependence among protocols. However, Ref. 13 required a specific human-provided mapping function to join protocols. Herein, we revisit the need to combine atlases of different protocols and show that dependence structure among atlases can be learned without human intervention. Our approach leads to natural generalization of MV and locally weighted vote (while maintaining computational efficiency). The dependence structure can also be estimated within the STAPLE perspective through jointly modeling the protocol behaviors.

This paper is organized as follows. Section 2 presents notation and derives the theory for fusion. Section 3 develops and characterizes the performance of the methods in a simulated example with a well-known error model and in an empirical example of deep brain labeling. Section 4 concludes with a brief discussion.

2 Theory

We first present the underlying framework connecting generalization of voting fusion approaches and MS-STAPLE. Consider a target image with N voxels and intensity values $\mathbf{I} \in \mathbb{R}^{N \times 1}$ with a set of R registered atlases, associated intensities $\mathbf{A} \in \mathbb{R}^{N \times R}$, and registered labels $\mathbf{D} \in \mathcal{L}^{N \times R}$. Let D_{ij} be the decision of atlas j at voxel i . In the case of multiple sets of labeling protocols, $\mathcal{L}$ corresponds to an arbitrary set of labels, that is to say that the numerical label denotations among the atlases do not necessarily correspond to the same anatomical structure. Lastly, let $\mathbf{L} = \{L_1, L_2, \dots, L_\psi\}$, where L_j corresponds to the number of labels in atlas class j and ψ corresponds to the total number of different labeling protocols used with a vector $\mathbf{d} \in \mathbb{I}^{R \times 1}$ mapping each rater to its atlas class. The objective of label fusion is to estimate $\hat{\mathbf{S}}$, a voxel discrete mapping of the target image. Figure 1 shows the rater models from the following sections. Many of the derivations in the MS- and SMS-STAPLE sections

Fig. 1 Modeling rater performance with the STAPLE, MS-STAPLE, and SMS-STAPLE algorithms. (a) A target image can be labeled using atlases of multiple protocols. (b) In STAPLE only atlases with a 1:1 correspondence can be used in segmentation, (c) in MS-STAPLE atlases of multiple protocols can be used to jointly segment all of the protocols, and (d) in SMS-STAPLE atlases of multiple protocols can be used to jointly segment one target protocol.

follow the notation originally set out in the STAPLE and follow-up papers.

2.1 Generalized Majority, Locally Weighted, and Nonlocally Weighted Vote

In standard MV, the probability of label s at voxel i , $p_{MV,i}(s|D)$, is empirically determined as the fraction of atlases that observe s at voxel i . We may generalize MV to include atlases (raters) of different protocols by

$$p_{GMV,i}(s|D) = \frac{1}{R} \sum_j p_i(s|D_{ij}, d_j), \quad (1)$$

where $p_i(s|D_{ij}, d_j)$ is discrete probability of the co-occurrence across labeling protocols of registered atlases (which can be empirically estimated as seen in Sec. 2.4). Henceforth, for convenience, we simplify the co-occurrence probability to be spatially invariant and drop the subscript. Generalized majority vote (GMV) is thus $S_{GMV,i} = \arg \max_s p_{GMV,i}(s|D)$. Similarly, locally weighted vote¹ can be framed around the co-occurrence probability as

$$\begin{aligned} p_{LWV,i}(s|D, d, A, I) &= \sum_j p_i(s|D_{ij}, d_j, A_{ij}, I_i) \\ &= \frac{\sum_j p(A_{ij}|I_i) p(s|D_{ij}, d_j)}{\sum_j \sum_{s'} p(A_{ij}|I_i) p(s'|D_{ij}, d_j)} \end{aligned} \quad (2)$$

assuming marginal independence of image intensity values and observed labels, where Z is a partition function normalizing distribution to a valid probability distribution and $p(A_{ij}|I_i)$ is the likelihood of observing the intensity value of atlas j at voxel i . Hence, $S_{LWV,i} = \arg \max_s p_{LWV,i}(s|D, d, A, I)$.

Moreover, nonlocally weighted voting (NLWV)^{10,11} can be reframed with $p_{NLWV,i}(s|D, d, A, I) = \frac{1}{N} \sum_j \sum_{s' \in L_{d_j}} p_{i,j}(s'|D, d, A, I) p(s|s', d_j)$ with s' as the latent correspondence and $p_{i,j}(s'|D, d, A, I)$ as the likelihood of atlas j observing s' at i . Briefly, $p_{i,j}(s'|D, d, A, I)$ can be defined as

$$p_{i,j}(s'|D, d, A, I) = \frac{1}{Z} \sum_{i' \in N} p(I_i|A_{i'}) p(s'|D_{i'}, d) \quad (3)$$

with

$$p(I_i|A_{i'}) = \exp\left(-\frac{\|p(A_{i'}) - p(I_i)\|^2}{\sigma}\right) \quad (4)$$

as to the nonlocal correspondence and Z as a normalization constant to maintain a valid probability distribution. A full derivation of nonlocally weighted vote is presented in Ref. 10. Note that the latent corresponding label, s' , occupies the role of the atlas labels in the NLWV model, and $S_{NLWV,i} = \arg \max_s p_{NLWV,i}(s|D, d, A, I)$.

These formulations reduce to their classical definitions if all of the atlases are of the target class (i.e., co-occurrence matrix is 1:1) or the co-occurrence probabilities of nontarget classes are uniform (i.e., other label protocols are uninformative and, hence, ignored).

2.2 Multiset STAPLE

STAPLE label fusion maintains a confusion matrix θ for each atlas.⁶ Each confusion matrix entry $\theta_{js's} \equiv p(D_{ij} = s'|T_i = s)$ presents a discrete probability distribution where s' corresponds to the label observed and T corresponds to a latent true segmentation. Consider segmentation with ψ labeling protocols, each with its own confusion matrix such that the likelihood of observing a label is $f(D_{ij} = s'|T_i = l, \{\theta_j\})$, where s' is the observed later by rater j at voxel i , l is a set of true labels of size $1 \times \psi$ observed at i , and $\{\theta_j\}$ is a set of ψ confusion matrices for atlas j where $\theta_{jps's}$ corresponds to the likelihood that label s' observed by rater j given that label s of set ρ is the true label. Note that $\theta_{j\rho}$ is a possibly nonsquare matrix where $\theta_{j\rho}$ is of size $L_{d_j} \times L_\rho$. This is the core of MS-STAPLE.

To estimate the data likelihood, we follow Ref. 8 to capture dependence between protocols through a geometric mean

$$\begin{aligned} f(D_{ij} = s'|T_i = l, \{\theta_j\}) &= \left(\prod_{\rho \in \psi} f(D_{ij} = s'|T_i = l_\rho, \theta_{j\rho}) \right)^{a_{jl}} \\ &= \left(\prod_{\rho \in \psi} \theta_{j\rho s' l_\rho} \right)^{a_{jl}}, \end{aligned} \quad (5)$$

where a_{jl} maintains $\sum_{s' \in L_{d_j}} \left(\prod_{\rho \in \psi} \theta_{j\rho s' l_\rho} \right)^{a_{jl}} = 1$, thus maintaining a valid probability distribution. Note that Ref. 8 captured joint information across specific hierarchical protocols, while here we are using it to normalize across relationships that must be estimated assuming conditional independence between sets of labels.

Given this model, we can apply EM. Briefly, let $W \in \mathbb{R}^{\prod_{\rho \in \psi} L_\rho \times N}$ where $W_{li}^{(k)} \equiv f(T_i = l|D, \{\theta\}^{(k)})$ is the probability that the true label set observed at voxel i during iteration k is l . Using a Bayesian expansion and the assumed conditional independence between atlases

$$W_{li}^{(k)} = \frac{f(T_i = l) \prod_{j \in R} f(D_{ij} = s'|T_i = l, \{\theta_j^{(k)}\})}{\sum_{l'} f(T_i = l') \prod_{j \in R} f(D_{ij} = s'|T_i = l', \{\theta_j^{(k)}\})}, \quad (6)$$

where $f(T_i = l)$ is a voxelwise *a priori* distribution of the underlying segmentation. The denominator corresponds to a partition function normalizing W to a valid probability distribution. Substituting Eq. (3) in Eq. (4) yields the MS-STAPLE E-step

$$W_{li}^{(k)} = \frac{f(T_i = l) \prod_{j \in R} \left(\prod_{\rho \in \psi} \theta_{j\rho s' l_\rho}^{(k)} \right)^{a_{jl}}}{\sum_{l'} f(T_i = l') \prod_{j \in R} \left(\prod_{\rho \in \psi} \theta_{j\rho s' l'_\rho}^{(k)} \right)^{a_{jl'}}}. \quad (7)$$

To estimate the performance parameters, maximize the expected value of the conditional log-likelihood function. We follow the traditional M-step expansion

$$\begin{aligned}
 \theta_{jp}^{(k+1)} &= \operatorname{argmax}_{\theta_{jp}} \sum_i E[\ln(f(D_{ij}|Ti, \{\theta_j\})|D, \{\theta_j\})] \\
 &= \operatorname{argmax}_{\theta_{jp}} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_l W_{li}^{(k)} \ln \left(\prod_{\rho \in \Psi} \theta_{jps'l_p}^{(k)} \right)^{\alpha_{jl}^{(k)}} \\
 &= \operatorname{argmax}_{\theta_{jp}} \sum_{s'} \sum_{i: D_{ij}=s'} \sum_l W_{li}^{(k)} \alpha_{jl}^{(k)} \sum_{\rho \in \Psi} \ln \left(\theta_{jps'l_p}^{(k)} \right), \quad (8)
 \end{aligned}$$

where $i: D_{ij} = s'$ corresponds to the voxels where $D_{ij} = s'$. To constrain each row of the confusion matrix to be a valid probability distribution ($\sum_{s'} \theta_{jps's} = 1$), we differentiate each element and use a Lagrange multiplier (λ)

$$\begin{aligned}
 0 &= \frac{\partial}{\partial \theta_{jps's}} \left[\sum_{s'} \sum_{i: D_{ij}=s'} \sum_l W_{li}^{(k)} \alpha_{jl}^{(k)} \sum_{\rho \in \Psi} \ln(\theta_{jps'l_p}^{(k)}) \right. \\
 &\quad \left. + \lambda \sum_{s'} \theta_{jps's}^{(k)} \right], \\
 0 &= \sum_{i: D_{ij}=s'l: l_p=s} \left(\frac{W_{li}^{(k)} \alpha_{jl}^{(k)}}{\theta_{jps's}^{(k)}} \right) + \lambda, \\
 -\lambda \theta_{jps's}^{(k+1)} &= \left(\sum_{i: D_{ij}=s'l: l_p=s} W_{li}^{(k)} \alpha_{jl}^{(k)} \right), \\
 \theta_{jps's}^{(k+1)} &= \frac{\left(\sum_{i: D_{ij}=s'l: l_p=s} W_{li}^{(k)} \alpha_{jl}^{(k)} \right)}{\left(\sum_i \sum_{l: l_p=s} W_{li}^{(k)} \alpha_{jl}^{(k)} \right)}, \quad (9)
 \end{aligned}$$

where $l: l_p = s$ corresponds to label sets where the voxel of atlas set ρ is s .

2.3 Simplified MS-STAPLE and Nonlocal-SMS-STAPLE

Empirically, we have found the fully parameterized MS-STAPLE model (Sec. 2.2) less numerically stable than one would desire. Here, we present a simplified model to improve stability with limited data. In place of $\{\theta_j\}$ (with $\sum_j L_{d_j} \Pi_{\rho} L_{\rho}$ degrees of freedom), consider $\tilde{\theta}_j$, an $L_{d_j} \times L_t$ confusion matrix with t to be the index of the target label. Each element of $\tilde{\theta}_{js's}$ corresponds to the probability j observes label $s' \in L_{d_j}$ given that the true label is $s \in L_t$. Note that each atlas has a confusion matrix with the number of rows dependent on the labeling protocol, but with the number columns matching the target protocol. The degree of freedom of $\tilde{\theta}$ is $\sum_j L_{d_j} L_t$, a possibly dramatic reduction from the Sec. 2.2 model by assuming conditional independence.

With a simplified multiset STAPLE model (i.e., SMS-STAPLE), the EM update equations are found to be

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \tilde{\theta}_{js's}^{(k)}}{\sum_{s''} f(T_i = s'') \prod_j \tilde{\theta}_{js's''}^{(k)}} \quad (10)$$

for the E-step, and

$$\tilde{\theta}_{js's}^{(k+1)} = \frac{\left(\sum_{i: D_{ij}=s'} W_{si}^{(k)} \right)}{\left(\sum_i W_{si}^{(k)} \right)} \quad (11)$$

for the M-step.

Following Ref. 11, we derive the EM update equations to incorporate nonlocal correspondence in the SMS-STAPLE model (herein SMS-nonlocal STAPLE) as

$$W_{si}^{(k)} = \frac{f(T_i = s) \prod_j \sum_{i' \in s(i)} \tilde{\theta}_{js's}^{(k)} c_{ji'i}}{\sum_{s''} f(T_i = s'') \prod_j \sum_{i' \in s(i)} \tilde{\theta}_{js's''}^{(k)} c_{ji'i}} \quad (12)$$

for the E-step where $s(i)$ is the spatial neighborhood around voxel i and $c_{ji'i}$ is the likelihood of correspondence between atlas j at voxel i' and the target at voxel i . The M-step follows as

$$\tilde{\theta}_{js's}^{(k+1)} = \frac{\sum_i \left(\sum_{i' \in s(i): D_{i'j}=s'} c_{ji'i} \right) W_{si}^{(k)}}{\sum_i W_{si}^{(k)}}. \quad (13)$$

2.4 Modeling Co-Occurrence Probability and Initialization

For the voting approaches (Sec. 2.1), the co-occurrence probability definitions were calculated empirically as

$$p(s|s', \rho) = \frac{\sum_{j: d_j=\rho} \sum_{h: d_h=t} \sum_{i: D_{ij}=s'} \delta(D_{ih}, s)}{\sum_{j: d_j=\rho} \sum_i \delta(D_{ij}, s')}, \quad (14)$$

where s is the latent label, s' is the observed label, and ρ is the index of the label set from which s' is drawn. Note that this model is conditionally independent of the true label. Intuitively, this corresponds to the likelihood that the true label for some protocol, x , is s given that an atlas of protocol y observes label s' , calculated based on co-registered examples labeled with protocols x and y . The MS-STAPLE approaches (Sec. 2.2) were initialized by

$$\theta_{jps's}^{(0)} = \theta_{d_jps's}^{(0)} = \frac{\sum_{k: d_k=\rho} \sum_{h: d_h=t} \sum_{i: D_{ik}=s'} \delta(D_{ih}, s)}{\sum_{k: d_k=t} \sum_i \delta(D_{ik}, s)}. \quad (15)$$

For the SMS-STAPLE approaches (Sec. 2.3), initial confusion matrixes were

$$\theta_{js's}^{(0)} = \theta_{d_js's}^{(0)} = \frac{\sum_{j: d_j=\rho} \sum_{h: d_h=t} \sum_{i: D_{ij}=s'} \delta(D_{ih}, s)}{\sum_{j: d_j=t} \sum_i \delta(D_{ij}, s)}. \quad (16)$$

Convergence of EM was detected when the average change in $\tilde{\theta}$ from k to $k+1$ was less than $\epsilon = 10^{-6}$, specifically

$$\frac{1}{M} \sum_j \sum_{s' \in L_{d_j}} \sum_{s \in L_t} abs(\tilde{\theta}_{js's}^{(k+1)} - \tilde{\theta}_{js's}^{(k)}), \quad (17)$$

where M is the total number of elements in all of the confusion matrixes (i.e., $\sum_j L_{d_j} \times L_t$).

3 Methods and Results

3.1 Simulation of Distinct Protocol

We first consider labeling an idealized square object within an 80×80 pixel background where the object consists of four distinct quadrants. There were four possible labeling protocols, and each protocol labels one quadrant as distinct from the other three [see Fig. 2(a)]. The simulated raters independently made errors in their respective protocols in terms of randomly shifted boundaries, which has been commonly used since.¹¹ Boundary shift errors were selected uniformly at random from between -10 and $+10$ pixels for boundary point [Fig. 2(b)]. Here, we evaluated fusion of between 3 and 10 raters for each protocol (i.e., between 12 and 40 total simulated rater label sets) for a single target image. The experiment was Monte Carlo repeated with new simulated rater label sets 20 times to establish model variability.

Each of the fusion algorithms (Secs. 2.1 through 2.3) was applied to each unique set of simulated rater observations [Fig. 2(c)]. Note that the distinct algorithms were applied to the same simulated data for each number of raters and Monte Carlo iteration. Performance was evaluated by the Dice similarity coefficient and Wilcoxon sign-rank test. MS-STAPLE showed significant improvement over STAPLE with four to six training examples of each class available ($p < 10^{-5}$). SMS-STAPLE showed significant improvement over all models with all numbers of simulated observations ($p < 10^{-5}$). GMV was significantly worse than all models ($p < 10^{-5}$).

3.2 Empirical Deep Brain Segmentation

To evaluate empirical performance, we constructed a dataset with two distinct whole-brain protocols (a fine protocol and a coarse protocol). To ensure that true results were well known, we studied 40 T1-weighted MRIs expertly labeled with 14 fine deep brain structures and 12 coarse labels for the remainder of the brain (derived from the BrainCOLOR protocol; Neuromorphometrics, Inc., Somerville, Massachusetts). For 10 randomly selected subjects, we reduced the deep brain structures to two lateralized labels (i.e., the coarse protocol), and the remaining 30 subjects were randomly split into two groups of 15, one for training and one for testing (both with the fine protocol). All pairs of images were coregistered with ANTS-Syn with default parameters,¹⁴ and labels were deformed to match the target images with nearest-neighbor interpolation.

We evaluated a situation where the 10 coarse atlases we assumed to be pre-existing and between 2 and 15 of the fine training atlases were made available. For each number of new fine atlases (Q), we randomly selected Q atlases of the 15 images in the training set to construct a simulated available dataset of 10 coarse and Q fine atlases. For each of the 15 images in the testing set, we performed each of label fusion algorithms and evaluated performance with the Dice similarity coefficient and statistically evaluated with the Wilcoxon sign-rank test (Fig. 3).

For nondeep brain structures [Fig. 3(a)], GMV outperforms MV with all numbers of atlases, and generalized nonlocally weighted vote, SMS-STAPLE, and SMS-nonlocal STAPLE

Fig. 2 Simulation results based on algorithm and number of atlases used. (a) A set of simulated truth models were generated to model a relationship with interprotocol spatial dependence. (b) Simulated observations with increasing numbers of observations were fused. (C and D) For STAPLE versus MS-STAPLE and SMS- versus MS-STAPLE: $*p < 10^{-5}$.

(a) Non-deep brain structures (12 structures)

(b) Deep brain structures (14 structures)

Fig. 3 Results from empirical deep brain segmentation experiment. Both SMS-STAPLE and SMS-nonlocal STAPLE show statistically significant improvement ($*p < 0.05$, $**p < 0.01$) for many of the segmentation tasks particularly when few atlases of the target class are available. † indicates segmentation task shown in Fig. 4.

outperform their counterparts when few (<6) target atlases are available ($p < 0.05$). For deep-brain structures [Fig. 3(b)], the MV and local weighted vote outperformed GMV and generalized locally weighted vote ($p < 0.05$). Meanwhile, SMS-STAPLE outperformed STAPLE when few target atlases

were available and SMS-nonlocal STAPLE outperformed non-local STAPLE in all the experiments ($p < 0.01$). Note that MS-STAPLE results are not shown and performance was worse than all methods (as discussed in Sec. 2). These results are shown qualitatively in Fig. 4.

Fig. 4 Qualitative results comparing eight segmentation algorithms (identified by † in Fig. 3). Contour lines indicate manual truth target labels (shown lower left).

4 Discussion and Conclusion

This paper presents generalizations of MV and locally weighted vote to incorporate multiple labeling protocols into the segmentation of a target label set. We also present a generalization of the STAPLE algorithm to jointly incorporate and segment multiple protocols and a simplification of this model for an individual target protocol. To achieve tractable models, we assume conditional independence among the labeling protocols. As an aside, an alternative approach would have been to design a “wide” confusion matrix θ_j of size $L_{d_j} \times \prod_{\rho \in \Psi} L_\rho$ per rat to capture all potential relationships, but the degrees of freedom in such an approach quickly exceed the number of atlases likely to be available in practice and the model would reduce to classic STAPLE separately for each protocol unless atlases labeled with multiple protocols were included.

In Ref. 13, the authors developed an alternate formulation of voting-based fusion and STAPLE for multiple protocols by defining a matching between the observed labels of the individual protocols, an intermediate protocol specific coarse labeling, and a set of fine labels that they wish to segment. To achieve the fusion, the authors build a mapping of the vector relationship between an individual protocol and the target, assuming there is a clear relationship that can be defined between the protocols, which in many cases may not be achievable.

In comparison, the work presented here assumes a softer relationship among labels. In our empirical study, we found that 46.7% of pairwise label relationships contained nonzero co-occurrence probabilities. In contrast, only 6.4% of the joint label relationships would be captured by Ref. 13. By using pairwise label relationships, we are both simplifying the initialization of the model and potentially capturing more robust relationships found within the data.

Lastly, the work of Ref. 13 presents a natural generalization of STAPLE, which follows directly from the approaches of Ref. 6. The generalizations of Ref. 13 assume a hierarchically defined relationship between the observed labels of each atlas and the target labels. Conversely, the work presented in this paper provides an alternate generalization where we jointly segment multiple protocols and learn the relation between them as a

function of the observations. We then simplify the proposed model to a more tractable solution for the scale of training examples available. Furthermore, the work proposed here is amenable to other advancements from STAPLE such as spatially varying performance parameters, which may be particularly useful when some labels are generalizations of groups of labels in the target protocols. Other improvements to the co-occurrence matrix presented in Sec. 2.4 may be improved by applying an L_2 -norm optimization to the matrix, such that labels are equally informative for the subset of labels that it has information for, and null for other subjects.

This work has shown how atlases with complementary labeling protocols can be used to improve segmentation of deep brain structures. These techniques hold substantial promise for improving fusion with existing protocols using related atlas sets from disparate research groups (e.g., from the extensive list of available atlases of distinct protocols). Alternatively, this approach could form the basis of a bootstrapping technique where a new protocol is developed by extending/refining one or more existing protocols, as is illustrated in the empirical deep brain explanation. Finally, the technique could be used to work toward consensus protocols while quantifying the joint information between the protocols. For example, hippocampal subfield segmentation¹² is an active area of research, but groups providing high-resolution protocols are rarely concerned with detailed manual whole-brain labeling. This approach could be used to jointly segment full brain¹⁵ and hippocampus subfield labels without having one set of humans label both.

In conclusion, we have presented algorithms to use jointly multiple labeling protocols for multitlas segmentation. We have compared these approaches both in simulation and an empirical study. Our results show statistically significant improvements in comparison to previously published gold-standard techniques when evaluated with defined truth models for the simulation and manually labeled examples for the empirical data.

Disclosures

The authors have no relevant financial interests in this work.

Acknowledgments

This research was supported by NSF CAREER 1452485 (Landman), R01NS095291 (Dawant), T32LM012412 (Malin). This study was also supported by NIH 5R01NS056307, 5R21NS082891, and in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, Tennessee, USA. This project was supported in part by the National Center for Research Resources, Grant UL1 RR024975-01, and is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. M. R. Sabuncu et al., "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imaging* **29**, 1714–1729 (2010).
2. R. Wolz et al., "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Trans. Med. Imaging* **32**, 1723–1730 (2013).
3. T. R. Langerak et al., "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *IEEE Trans. Med. Imaging* **29**, 2000–2008 (2010).
4. J. C. Gee, M. Reivich, and R. Bajcsy, "Elastically deforming 3D atlas to match anatomical brain images," *J. Comput. Assisted Tomogr.* **17**, 225–236 (1993).
5. J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Med. Image Anal.* **24**, 205–219 (2015).
6. S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Trans. Med. Imaging* **23**, 903–921 (2004).
7. A. J. Asman and B. A. Landman, "Formulating spatially varying performance in the statistical fusion framework," *IEEE Trans. Med. Imaging* **31**, 1326–1336 (2012).
8. A. J. Asman and B. A. Landman, "Hierarchical performance estimation in the statistical label fusion framework," *Med. Image Anal.* **18**, 1070–1081 (2014).
9. X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solórzano, "Combination strategies in multi-atlas image segmentation: application to brain MR data," *IEEE Trans. Med. Imaging* **28**, 1266–1277 (2009).
10. P. Coupé et al., "Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation," *NeuroImage* **54**, 940–954 (2011).
11. A. J. Asman and B. A. Landman, "Non-local statistical label fusion for multi-atlas segmentation," *Med. Image Anal.* **17**, 194–208 (2013).
12. C. Konrad et al., "Defining the human hippocampus in cerebral magnetic resonance images—an overview of current segmentation protocols," *NeuroImage* **47**, 1185–1195 (2009).
13. J. E. Iglesias et al., "An algorithm for optimal fusion of atlases with different labeling protocols," *NeuroImage* **106**, 451–463 (2015).
14. B. B. Avants et al., "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.* **12**, 26–41 (2008).
15. K. Oishi et al., "Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants," *NeuroImage* **46**, 486–499 (2009).

Andrew J. Plassard graduated with a bachelors of science (2013) in biomedical engineering from the University of Cincinnati, Cincinnati, Ohio. After graduation, he immediately began his doctoral studies at Vanderbilt University under the advisement of Bennett Landman PhD. His research focuses on advancements to segmentation theory focused on translation to diseased populations.

Bennett A. Landman graduated with a bachelors of science (2001) and masters of engineering (2002) in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, Massachusetts. After graduation, he worked in an image processing startup company and a private medical imaging research firm before returning for a doctorate in biomedical engineering (2008) from Johns Hopkins University School of Medicine, Baltimore, Maryland. Since 2010, he has been with the Faculty of the Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, Tennessee, where he is currently an associate professor. His research concentrates on applying image-processing technologies to leverage large-scale imaging studies to improve understanding of individual anatomy and personalize medicine.