HHS Public Access

Author manuscript

Proc SPIE Int Soc Opt Eng. Author manuscript; available in PMC 2014 May 07.

Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2014 March 21; 9034: 90341E–. doi:10.1117/12.2043182.

Statistical label fusion with hierarchical performance models

Andrew J. Asman[*,a], **Alexander S. Dagley**[a], and **Bennett A. Landman**[a,b]

[a]Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235

[b]Biomedical Engineering, Vanderbilt University, Nashville, TN, USA 37235

Abstract

Label fusion is a critical step in many image segmentation frameworks (e.g., multi-atlas segmentation) as it provides a mechanism for generalizing a collection of labeled examples into a single estimate of the underlying segmentation. In the multi-label case, typical label fusion algorithms treat all labels equally – fully neglecting the known, yet complex, anatomical relationships exhibited in the data. To address this problem, we propose a generalized statistical fusion framework using hierarchical models of rater performance. Building on the seminal work in statistical fusion, we reformulate the traditional rater performance model from a multi-tiered hierarchical perspective. This new approach provides a natural framework for leveraging known anatomical relationships and accurately modeling the types of errors that raters (or atlases) make within a hierarchically consistent formulation. Herein, we describe several contributions. First, we derive a theoretical advancement to the statistical fusion framework that enables the simultaneous estimation of multiple (hierarchical) performance models within the statistical fusion context. Second, we demonstrate that the proposed hierarchical formulation is highly amenable to the state-of-the-art advancements that have been made to the statistical fusion framework. Lastly, in an empirical whole-brain segmentation task we demonstrate substantial qualitative and significant quantitative improvement in overall segmentation accuracy.

Keywords

Label Fusion; Multi-Atlas Segmentation; STAPLE; Hierarchical Segmentation; Rater Performance Models

1. INTRODUCTION

Multi-atlas segmentation represents an extremely powerful generalize-from-example framework for image segmentation [1, 2]. In multi-atlas segmentation, multiple labeled examples (i.e., atlases) are registered to a previously unseen target-of-interest [3, 4], and the resulting voxelwise label conflicts are resolved using label fusion [5–10]. Herein, we focus on the problem of label fusion – a critical component of multi-atlas segmentation that has a substantial impact on segmentation accuracy.

[*]andrew.j.asman@vanderbilt.edu; http://masi.vuse.vanderbilt.edu; Medical-image Analysis and Statistical Interpretation Laboratory, Department of Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235.

Over the past decade, interest and research into the label fusion problem has exploded in popularity and significant improvement across an incredible range of applications has been shown. Broadly speaking, there are two primary perspectives on the problem of label fusion: (1) voting methods in which the underlying segmentation is modeled as some local, semi-local, or non-local weighted combination of the provided atlas information (e.g., [8–10]), and (2) statistical fusion methods in which the problem is cast from a Bayesian inference perspective and generative models of rater/atlas performance are maximized through an expectation-maximization (EM) [11] framework (e.g., [5–7, 12, 13]).

Regardless of the fusion approach, fusion algorithms typically treat all of the considered labels equally. As a result, the complex anatomical relationships that are often exhibited in multi-label segmentation problems are entirely neglected. To illustrate, consider a typical whole-brain segmentation problem in which there are often upwards of 100 unique labels that are estimated. Within those structures there are known anatomical and hierarchical relationships which could be leveraged – e.g., one such relationship might be medial frontal cortex → frontal cortex → cerebral cortex → cerebrum → brain (where "→" could be interpreted as "is part of"). While generalized hierarchical segmentation frameworks have been around for almost two decades (e.g., [14]) and recently considered for an application-specific voting fusion approach [15], a general hierarchical fusion framework has not been considered in the statistical fusion context.

We propose a generalized statistical fusion framework using hierarchical models of rater performance. Building on the seminal Simultaneous Truth and Performance Level Estimation (STAPLE) [7] algorithm, we reformulate the rater performance model to utilize hierarchical relationships through a multi-tier performance model (Figure 1). The proposed model is built on the simple concept that the performance of a rater at the higher levels of the hierarchical model (e.g., brain vs. non-brain or cerebrum vs. cerebellum) should propagate to the lower levels of the hierarchy (i.e., the individual labels-of-interest) in an informed manner. This work (1) provides an important theoretical advancement to the underlying theory of statistical fusion, (2) demonstrates superior performance in both simulated and empirical whole-brain data, and (3) shows that the proposed framework is amenable to many of the current advancements in the statistical fusion family.

This manuscript is organized in the following manner. First, the theory for the generalized hierarchical statistical fusion is derived. Second, we demonstrate superior performance on both simulated and empirical whole-brain segmentation data. Finally, we conclude with a brief discussion on the optimality of the approach and the potential for improvement.

2. THEORY

2.1 Problem Definition

Let $\mathbf{T} \in \mathbf{L}^{N\times 1}$ be the latent representation of the true target segmentation, where $\mathbf{L} = \{0, \ldots, L-1\}$ is the set of possible labels that can be assigned to a given voxel. Consider a collection of R raters (or registered atlases) with associated la decisions, $\mathbf{D} \in \mathbf{L}^{N\times R}$. The goal of any statistical fusion algorithm is to estimate the latent segmentation, $\boldsymbol{T}$, using the observed labels, $\boldsymbol{D}$ and the provided generative model of rater performance.

2.2 Hierarchical Performance Model

Consider a pre-defined hierarchical model with levels. At each level of the hierarchy, let $\mathscr{S}_m \in \mathscr{S} = \{\mathscr{S}_0, \cdots, \mathscr{S}_{M-1}\}$ be a mapping vector that maps a label in the original collection of labels, $s \in \boldsymbol{L}$, to the corresponding label at the be m^{th} level of the, hierarchy, $\mathscr{S}_{ms} \in \boldsymbol{L}^m$ is the collection labels at the m^{th} level of the hierarchy. Additionally, let the performance of the raters at hierarchical level m be parameterized by $\boldsymbol{\theta}^m \in \mathbb{R}^{R \times L^m \times L^m}$ (i.e., $L^m \times L^m$ *confusion matrix* for each rater). Specifically, $\theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}$ is the probability that rater j observes label s' given that the true label is m^{th} at the m^{th} level of the hierarchy. Thus, the generative model that must be defined is described by

$$f\left(D_{ij}=s' | T_i=s, \mathscr{S}, \left\{\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{M-1}\right\}\right) \quad (1)$$

which can be directly interpreted as the probability that rater j observes label s' given the true label, hierarchical model, and the corresponding confusion matrices. To directly estimate this distribution we propose a formulation in which the complete model of hierarchical performance (Eq. 1) is unified through a constrained geometric mean across the multitier estimate of rater performance.

$$\begin{aligned} f\left(D_{ij}=s' | T_i=s, \mathscr{S}, \left\{\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{M-1}\right\}\right) &= \left(\prod_{m=0}^{M-1} f\left(D_{ij}=\mathscr{S}_{ms'} | T_i=\mathscr{S}_{ms}, \boldsymbol{\theta}^m\right)\right)^{\alpha_{js}} \\ &= \left(\prod_{m=0}^{M-1} \theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right)^{\alpha_{js}} \end{aligned} \quad (2)$$

where, μ_{js} is an exponent that maintains the constraint that $\Sigma_{s'}\left(\Pi_m \theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right)^{\alpha_{js}} = 1$. In other words, μ_{js} ensures that the model in Eq. 1 is valid discrete probability mass function. Note, given the constraints on each individual θ^m (i.e., a valid confusion matrix) a unique value for μ_{js} is guaranteed to exist and can easily be found using a standard searching algorithm (e.g., binary search, gradient descent). Given the model in Eq. 2, it is now possible to utilize the provided hierarchical model within the EM-based statistical fusion framework.

2.3 E-Step: Estimation of the Voxelwise Label Probabilities

Let $\boldsymbol{W} \in \mathbb{R}^{L \times N}$ where $\boldsymbol{W}^{(k)}_{si}$ represents the probability that the true label associated with voxel i is label s at iteration k of the algorithm given the provided information and model parameters

$$W^{(k)}_{si} \equiv f\left(T_i=s | \boldsymbol{D}, \mathscr{S}, \left\{\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{M-1}\right\}^{(k)}\right). \quad (3)$$

Using a Bayesian expansion and the assumed conditional independence between the registered atlas observations, Eq. 3 can be re-written as

$$W_{si}^{(k)}=\frac{f\left(T_i=s\right)\Pi_j f\left(D_{ij}=s^{'}|T_i=s,\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\}^{(k)}\right)}{\Sigma_n f\left(T_i=n\right)\Pi_j f\left(D_{ij}=s^{'}|T_i=n,\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\}^{(k)}\right)} \quad (4)$$

where $f(T^i = s)$ is a voxelwise *a priori* distribution of the underlying segmentation. Note that the denominator of Eq. 4 is simply the solution for the partition function that enables $\boldsymbol{W}$ to be a valid probability mass function (i.e., $\Sigma_s W_{si} = 1$). Using the simplified generative model in Eq. 2, the final form for the E-step of the EM algorithm can be written as

$$W_{si}^{(k)}=\frac{f\left(T_i=s\right)\Pi_j\left(\Pi_m\theta^{m,(k)}_{j\mathscr{S}_{ms^{'}}\mathscr{S}_{ms}}\right)^{\alpha_{js}^{(k)}}}{\Sigma_{s''}f\left(T_i=s''\right)\Pi_j\left(\Pi_m\theta^{m,(k)}_{j\mathscr{S}_{ms^{'}}\mathscr{S}_{ms''}}\right)^{\alpha_{js''}^{(k)}}} \quad (5)$$

2.4 M-Step: Estimation of the Hierarchical Performance Level Parameters

The estimate of the performance level parameters (M-step) is obtained by finding the parameters that maximize the expected value of the conditional log likelihood function (i.e., using the result in Eq. 5). Unlike the traditional STAPLE approach, however, the parameters for each level of the hierarchy are maximized independently.

$$\begin{aligned}\boldsymbol{\theta}_j^{m,(k+1)} &= arg\ \underset{\theta_j^m}{max}\sum_i E\left[\ln f\left(D_{ij}=s^{'}|T_i=s,\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\}\right)|\boldsymbol{D},\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\}^{(k)}\right]\\ &= arg\ \underset{\theta_j^m}{max}\sum_i\sum_s W_{si}^{(k)}\ln f\left(D_{ij}=s^{'}|T_i=s,\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\}\right)\\ &= arg\ \underset{\theta_j^m}{max}\sum_i\sum_s W_{si}^{(k)}\ln\left(\prod_m\theta^m_{j\mathscr{S}_{ms^{'}}\mathscr{S}_{ms}}\right)^{\alpha_{js}^{(k)}}\\ &= arg\ \underset{\theta_j^m}{max}\sum_i\sum_s W_{si}^{(k)}\alpha_{js}^{(k)}\sum_m\ln\theta^m_{j\mathscr{S}_{ms^{'}}\mathscr{S}_{ms}}\end{aligned} \quad (6)$$

Noting the constraint that each row of the rater performance level parameters must sum to unity to be a valid probability mass function (i.e., $\Sigma_{s'}\theta^m_{j\mathscr{S}_{ms^{'}}\mathscr{S}_{ms}}=1$), we can maximize the performance level parameters at each level of the hierarchical model by differentiating with respect to each element and using a Lagrange Multiplier (λ) to formulate the constrained optimization problem. Following this procedure, we obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta^{m}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}} \left[\sum_i \sum_s W^{(k)}_{si} \alpha^{(k)}_{js} \sum_m \ln \theta^{m}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} + \lambda \sum_{s'} \theta^{m}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} \right] \\ -\lambda &= \frac{\Sigma_{i:\mathscr{S}_{mD_{ij}}=\mathscr{S}_{ms'}} \Sigma_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}} \alpha^{(k)}_{js''} W^{(k)}_{s''i}}{\theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}} \\ \theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} &= \frac{\Sigma_{i:\mathscr{S}_{mD_{ij}}=\mathscr{S}_{ms'}} \Sigma_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}} \alpha^{(k)}_{js''} W^{(k)}_{s''i}}{-\lambda} \\ \theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} &= \frac{\Sigma_{i:\mathscr{S}_{mD_{ij}}=\mathscr{S}_{ms'}} \Sigma_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}} \alpha^{(k)}_{js''} W^{(k)}_{s''i}}{\Sigma_i \Sigma_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}} \alpha^{(k)}_{js''} W^{(k)}_{s''i}} \end{aligned} \quad (7)$$

where $s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}$ is the collection of all labels that map to the true label of interest, $\mathscr{S}_{ms}$ and $i:\mathscr{S}_{mD_{ij}}=\mathscr{S}_{ms'}$ is the collection of all voxels in which the observed label, D_{ij} maps to the observed label of interest, $\mathscr{S}_{ms'}$. At this point, it is important to note that (1) that performance model formulation in Eq. 2 allows for each level of the hierarchy to be maximized independently when maximizing the log-likelihood function, and (2) the result in Eq. 7 uses $\alpha^{(k)}_{js}$ which can then be updated following the constraint:

$$\Sigma_{s'} \left(\Pi_m \theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} \right)^{\alpha^{(k+1)}_{js}} = 1.$$

2.5 Initialization and Detection of Convergence

Given an *a priori* hierarchical model, there are no additional parameters in the proposed approach when compared to the original STAPLE algorithm. As a result, the algorithm can be initialized in exactly the same manner as described in [7]. With that said, the detection of convergence is slightly different, as we utilize all levels of the hierarchy. Thus, convergence is detected when the normalized trace between consecutive iterations falls below some arbitrary threshold (herein, $\in = 10^{-4}$), where the normalized trace is given by

$$\frac{1}{LRM} \sum_j \sum_m tr\left(\boldsymbol{\theta}^m_j\right) \quad (8)$$

2.6 Application to state-of-the-art Statistical Fusion Approaches

Over the past decade, there have been several advancements to the statistical fusion framework, for instance (1) characterizing spatially varying performance – Spatial STAPLE [6], and (2) incorporation of non-local correspondence models – Non-Local STAPLE (NLS) [5]. In the interest of brevity, we only derive the hierarchical version of STAPLE in this manuscript. However, in the following experiments we demonstrate the amenability of the hierarchical approach to Spatial STAPLE, NLS, and the combination of the two Non-Local Spatial STAPLE (NLSS).

3. METHODS AND RESULTS

3.1 Motivating Simulation

Before assessing the empirical performance, we present a motivating simulation to demonstrate the manner in which hierarchical models can be integrated into the statistical fusion framework (Figure 2). Here, a single 2D slice model (300 × 300 voxels) was constructed to loosely approximate the types of relationships that are often exhibited in the brain. Given the provided truth model, a collection of 15 labeled observations were constructed by randomly applying boundary errors of varying strength (see Figure 2 for the best/worst observations). As baselines, the results of a majority vote, STAPLE, and Spatial STAPLE are presented. Additionally, we consider three different hierarchical models with depths ranging from 3 to 5. The presented simulation was performed with 20 Monte Carlo iterations in order to estimate the variance in the results. The results in Figure 2 demonstrate substantial qualitative and significant ($p < 0.01$, paired t – test) improvement exhibited by the hierarchical implementations of both STAPLE and Spatial STAPLE. The different hierarchical models provide important insight into the effect of differing perspectives on the hierarchical relationships exhibited in the data. Here, the 4-level model was statistically superior to both the 3-level model and the 5-level model. While the proposed formulation relies on an *a priori* hierarchical model, it is intriguing to quantify the impact of both neglecting the observed hierarchical relationships (i.e., the 3-level model) and over-modeling these relationships (i.e., the 5-level model).

3.2 Empirical Whole-Brain Data and Experimental Design

For the empirical whole-brain experiments, a collection of 45 MPRAGE images from unique subjects are considered as part of the Open Access Series of Imaging Studies (OASIS, http://www.oasis-brains.org) [16] with subjects ranging in age from 18 to 90. All images had a resolution of $1 \times 1 \times 1\text{mm}^3$. All images were labeled using the brainCOLOR protocol (http://www.braincolor.org/) [17] and provided by Neuromorphometrics, Inc. (Somerville, MA, www.neuromorphometrics.com). Each labeled image contained exactly 133 unique labels (including background). For the purposes of evaluation, 15 of these images were randomly selected as training data, and the remaining 30 were selected as testing data.

Herein, we consider two separate registration frameworks. First we consider an affine-only pairwise registration framework [3] (using “reg_aladin” as part of the “NiftyReg” package – http://sourceforge.net/projects/niftyreg/). Additionally, we consider a pairwise non-rigid registration framework in which the provided affine registrations are augmented with a non-rigid registration [4] (using the Advanced Normalization Tools (ANTs) package – http://stnava.github.io/ANTs/). For both registration frameworks, all 15 training atlases were independently registered to all 30 of the testing atlases – resulting in 450 registrations.

To evaluate fusion performance, we consider several algorithms. First, in order to provide a benchmark of algorithmic performance, we consider a majority vote and a locally weighted vote (as described in [8]). Additionally, we consider STAPLE, Spatial STAPLE, NLS, and NLSS as well as the hierarchical versions of each, referred to as Hierarchical STAPLE,

Hierarchical Spatial STAPLE, Hierarchical NLS, and Hierarchical NLSS, respectively. For the hierarchical algorithms, we constructed a 12-level hierarchical model (manually constructed by an experienced neuroimaging analyst). Note, for clarity of presentation only NLSS and hierarchical NLSS are considered for the non-rigid registration approach.

3.3 Empirical Whole-Brain Results

The quantitative results for the empirical whole-brain experiment (Figure 3) demonstrate consistent improvement by the hierarchical implementations of each of the considered statistical fusion algorithms. The quantitative results are broken up into 3 different categories (1) all labels, (2) non-cortical labels, and (3) cortical labels. As expected, the results for the cortical labels are considerably poorer than the results for the non-cortical labels. Regardless, the hierarchical implementations provide consistent improvement regardless of the differing label contexts. For the affine registration framework, the hierarchical implementations provided a mean improvement across the testing data of 0.0188, 0.0245, 0.0250, and 0.0237 for STAPLE, Spatial STAPLE, NLS, and NLSS, respectively. All improvements were statistically significant ($p < 0.01$, paired t – test). Note, the relatively poor performance by STAPLE and Spatial STAPLE are not surprising considering the fact that they do not utilize the atlas-target intensity differences when estimating the final segmentation.

For the non-rigid registration framework, Hierarchical NLSS provided a small, but significant mean improvement across the testing data of 0.0068 over NLSS. Despite this relatively small magnitude of improvement, Hierarchical NLSS provided statistically significant improvement ($p < 0.01$, paired t – test) over NLSS for labels (the remaining 54 were statistically indistinguishable).

The qualitative results (Figure 4) support the quantitative improvement. Using the affine registration framework, all of the considered statistical fusion algorithms exhibit substantial visual improvement for many of the considered labels. In particular, there appears to be marked improvement in the quality of the lateral ventricle labels and many of the cortical labels.

4. DISCUSSION

Herein, we propose a novel statistical fusion framework using a reformulated hierarchical performance model. Given an *a priori* model of the hierarchical relationships for a given segmentation task, the proposed generative model of performance provides a straightforward mechanism for quantifying rater performance at each level of the hierarchy. The primary contributions of this manuscript are: (1) we have demonstrated statistically significant improvement on both simulated and empirical whole-brain data, (2) we have shown that the proposed hierarchical formulation is highly amenable to many of the state-of-the-art advancements that have been made to the statistical fusion framework, and (3) we have provided a theoretical advancement to the statistical fusion framework that enables the simultaneous estimation of multiple (hierarchical) confusion matrices for each rater.

There are several potential advancements to this framework that require future exploration. First, all of the presented experiments have relied upon an *a priori* model of the hierarchical relationships within the data. The ability to infer these hierarchical relationships directly from a provided training set would dramatically increase the potential applications for this type of framework, and provide an underlying foundation for estimating the optimal hierarchical formulation for a given application. Second, we have derived this approach from the perspective of hierarchical relationships between labels. However, the same (or very similar) estimation framework could potentially be used to estimate rater performance using multiple labeling protocols. For example, if one had a collection of datasets that were labeled using two separate protocols (either manually or automatically) it may be possible to (1) estimate the relationships between the protocols, and (2) simultaneously estimate rater performance in terms of both protocols. This type of framework is fascinating and certainly warrants further investigation.

We have presented a powerful theoretical framework for leveraging the complex inter-structure relationships within the statistical fusion context. While traditional fusion approaches treat all labels equally, the proposed rater model more accurately infers the types of errors that raters (or atlases) make within a hierarchically consistent formulation.

Acknowledgements

We would like to thank Andrew Worth and Neuromorphometrics, Inc. for the whole-brain data. This work was supported in part by NIH 1R21NS064534, 1R03EB012461, 2R01EB006136, and R01EB006193. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

[1]. Rohlfing T, Russakoff DB, Maurer CR. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Transactions on Medical Imaging. 2004; 23(8):983–994. [PubMed: 15338732]

[2]. Heckemann RA, Hajnal JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage. 2006; 33(1):115–126. [PubMed: 16860573]

[3]. Ourselin S, Roche A, Subsol G, Pennec X, Ayache N. Reconstructing a 3D structure from serial histological sections. Image and vision computing. 2001; 19(1):25–31.

[4]. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage. 2011; 54(3):2033–44. [PubMed: 20851191]

[5]. Asman AJ, Landman BA. Non-Local Statistical Label Fusion for Multi-Atlas Segmentation. Medical Image Analysis. 2012; 17(2):194–208. [PubMed: 23265798]

[6]. Asman AJ, Landman BA. Formulating Spatially Varying Performance in the Statistical Fusion Framework. IEEE Transactions on Medical Imaging. 2012; 31(6):1326–1336. [PubMed: 22438513]

[7]. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging. 2004; 23(7):903–921. [PubMed: 15250643]

[8]. Sabuncu MR, Yeo BTT, Van Leemput K, Fischl B, Golland P. A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging. 2010; 29(10):1714–1729. [PubMed: 20562040]

[9]. Wang H, Suh JW, Das SR, Pluta J, Craige C, Yushkevich PA. Multi-Atlas Segmentation with Joint Label Fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012

[10]. Coupé P, Manjķn JV, Fonov V, Pruessner J, Robles M, Collins DL. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage. 2011; 54(2):940–954. [PubMed: 20851199]

[11]. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1977:1–38.

[12]. Commowick O, Akhondi-Asl A, Warfield SK. Estimating A Reference Standard Segmentation with Spatially Varying Performance Parameters: Local MAP STAPLE. IEEE transactions on medical imaging. 2012; 31(8):1593–1606. [PubMed: 22562727]

[13]. Cardoso MJ, Leung K, Modat M, Keihaninejad S, Cash D, Barnes J, Fox NC, Ourselin S. STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Medical Image Analysis. 2013; 17(6):671–684. [PubMed: 23510558]

[14]. Najman L, Schmitt M. Geodesic saliency of watershed contours and hierarchical segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1996; 18(12):1163–1173.

[15]. Wolz, R.; Chu, C.; Misawa, K.; Mori, K.; Rueckert, D. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2012. 2012. Multi-organ Abdominal CT Segmentation Using Hierarchically Weighted Subject-Specific Atlases; p. 10-17.

[16]. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. 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(9):1498–1507. [PubMed: 17714011]

[17]. Klein, A.; Dal Canton, T.; Ghosh, SS.; Landman, B.; Lee, J.; Worth, A. Open labels: online feedback for a public resource of manually labeled brain images. 16th Annual Meeting for the Organization of Human Brain Mapping; 2010.

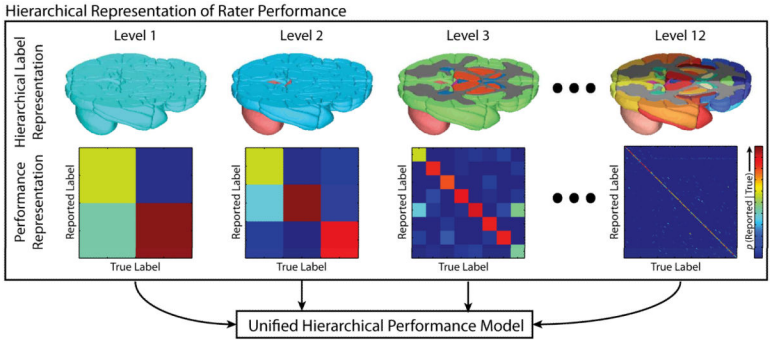

Figure 1.

Hierarchical representation of rater performance. A hierarchical model is developed for the brain, where, at each level, the performance of a rater is quantified. The overall quality of rater is then estimated through the unified hierarchical performance model.

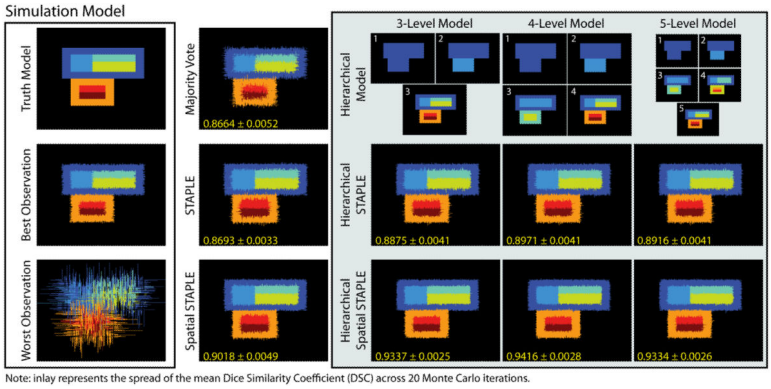

Figure 2.

Results on the motivating simulation. A simulated truth model was constructed to loosely model the types of relationships exhibited in the brain. The hierarchical formulations of STAPLE and Spatial STAPLE provide significant increases in overall segmentation accuracy. Here, the 4-level model results in statistically superior performance when compared to the 3- and 5-level models.

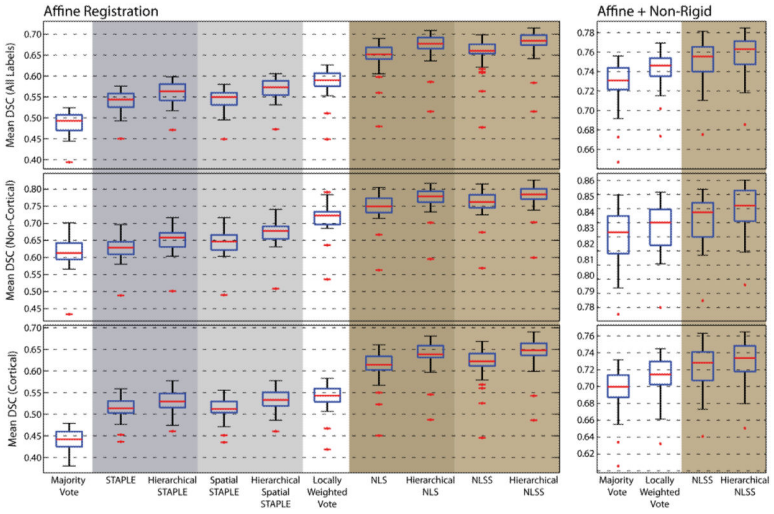

Figure 3.

Quantitative results on the empirical whole-brain segmentation experiment. The hierarchical implementations of STAPLE, Spatial STAPLE, NLS, and NLSS provide statistically significant accuracy improvements across each of the considered label sets for the affine registration framework. Similarly, Hierarchical NLSS provides substantial accuracy improvements for the non-rigid registration framework.

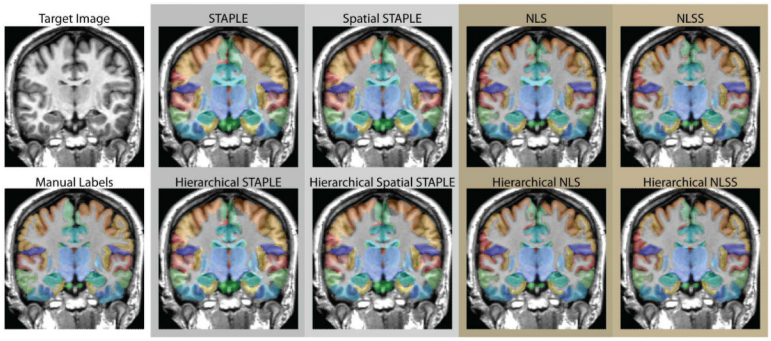

Figure 4.

Qualitative improvement exhibited by several state-of-the-art statistical fusion algorithms with the reformulated hierarchical performance model for the affine registration framework. For each of the considered statistical fusion algorithms we see substantial visual improvement for many of the considered labels. In particular, there appears to be marked improvement in the quality of the lateral ventricle labels and many of the cortical labels.