Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Med Image Anal. Author manuscript; available in PMC 2015 October 01.

Published in final edited form as:
Med Image Anal. 2014 October ; 18(7): 1070–1081. doi:10.1016/j.media.2014.06.005.

Hierarchical Performance Estimation in the Statistical Label Fusion Framework

Andrew J. Asman[1] and **Bennett A. Landman**[1,2]

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

[2] Institute of Imaging Science, 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. The proposed 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, the primary contributions of this manuscript are: (1) we provide a theoretical advancement to the statistical fusion framework that enables the simultaneous estimation of multiple (hierarchical) confusion matrices for each rater, (2) we highlight the amenability of the proposed hierarchical formulation to many of the state-of-the-art advancements to the statistical fusion framework, and (3) we demonstrate statistically significant improvement on both simulated and empirical data. Specifically, both theoretically and empirically, we show that the proposed hierarchical performance model provides substantial and significant accuracy benefits when applied to two disparate multi-atlas segmentation tasks: (1) 133 label whole-brain anatomy on structural MR, and (2) orbital anatomy on CT.

Graphical abstract

© 2014 Elsevier B.V. All rights reserved.

Corresponding author Andrew J. Asman Vanderbilt University EECS 2301 Vanderbilt Pl. PO Box 351679 Station B Nashville, TN 37235-1679 **Work:** (615) 322-2338 andrew.j.asman@vanderbilt.edu.

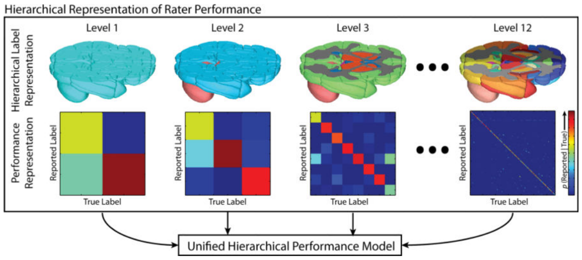
Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript


Hierarchical Representation of Rater Performance
Level 1
Level 2
Level 3
Level 12
Hierarchical Label Representation
Performance Representation
Reported Label
True Label
p (Reported | True)
Unified Hierarchical Performance Model

Keywords

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

Introduction

Multi-atlas segmentation represents a powerful generalize-from-example framework for image segmentation ***(Heckemann et al., 2006; Rohlfing et al., 2004c)***. In multi-atlas segmentation, multiple labeled examples (i.e., atlases) are registered to a previously unseen target-of-interest (Avants et al., 2008; Klein et al., 2009; Ourselin et al., 2001), and the resulting voxelwise label conflicts are resolved using label fusion (Asman and Landman, 2012a; Asman and Landman, 2012c; Coupé et al., 2011; Sabuncu et al., 2010; Wang et al., 2012; Warfield et al., 2004). Since its inception, multi-atlas segmentation has exploded in popularity and has been used across a wide range of potential applications – including, but not limited to, whole-brain (Aljabar et al., 2009; Artaechevarria et al., 2009; Asman and Landman, 2011; Asman and Landman, 2012a, b; Heckemann et al., 2006; Klein and Hirsch, 2005; Sabuncu et al., 2010; Weisenfeld and Warfield, 2011; Wolz et al., 2010), hippocampus (Cardoso et al., 2011; Coupé et al., 2011; Wang et al., 2012), head and neck ***(Asman and Landman, 2012a, b; Chen et al., 2011)***, cardiac (Bai et al., 2013; Depa et al., 2010; Isgum et al., 2009), prostate ***(Langerak et al., 2010)***, and abdomen ***(Wolz et al., 2012)***. Herein, we focus on the problem of label fusion – a critical component of multi-atlas segmentation that has a substantial impact on segmentation accuracy.

Over the past decade, interest and research into the label fusion problem has grown in popularity and significant improvement across a vast range of applications has been shown. Broadly speaking, there are two primary perspectives on the problem of label fusion: The first perspective builds on voting-based methods in which the underlying segmentation is modeled through the selection of appropriate atlases (e.g., (Aljabar et al., 2009; Cao et al., 2011; Rohlfing et al., 2004a)) or, through a local, semi-local, or non-local weighted combination of the provided atlas information (e.g., (Coupé et al., 2011; Iglesias et al., 2013; Sabuncu et al., 2010; Wang et al., 2012)). The second perspective, based on the Simultaneous Truth and Performance Level Estimation (STAPLE) framework ***(Warfield et al., 2004)***, is commonly referred to as statistical fusion – an approach in which the problem is cast from a Bayesian inference perspective and generative models of rater/atlas performance are maximized through expectation-maximization (EM) ***(Dempster et al.,***

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

1977) (e.g., (Akhondi-Asl and Warfield, 2013; Asman and Landman, 2012a; Asman and Landman, 2012c; Cardoso et al., 2013; Commowick et al., 2012; Rohlfing et al., 2004b)).

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 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., ***(Beucher, 1994; Najman and Schmitt, 1996)***) and recently considered for an application-specific voting fusion approach ***(Wolz et al., 2012)***, a generalized 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 STAPLE 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) is indicative of the rater's performance at the lower levels of the hierarchy (i.e., the individual labels-of-interest). Thus, the performance at the higher levels of the hierarchy should propagate to lower levels of the hierarchy in a theoretically and probabilistically consistent manner.

This manuscript is organized in the following manner. First, the theory for the generalized hierarchical statistical fusion framework is derived and the pertinent details for extension to state-ofthe-art statistical fusion are provided. Second, we demonstrate superior performance on both simulated and empirical multi-atlas segmentation data – herein, whole-brain and orbital data. Finally, we conclude with a brief discussion on the optimality of the approach and the potential for improvement. The research presented in this manuscript is an extension of a previously published conference paper ***(Asman et al., 2014)***. Herein, we (1) provide additional theoretical derivations for the hierarchical model, (2) explicitly define the extension to state-of-the-art statistical fusion algorithms, (3) provide additional insights through a reformulated simulation, and (4) include two distinct empirical experiments to more clearly highlight the benefits of hierarchical performance estimation.

Theory

Problem Definition

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

estimate the latent segmentation, $\boldsymbol{T}$, using the observed labels, $\boldsymbol{D}$, and the provided generative model of rater performance.

Hierarchical Performance Model

Consider a pre-defined hierarchical model with M levels. At each level of the hierarchy, let $\boldsymbol{S_m} \in \boldsymbol{S} = \{\boldsymbol{D}_0 , \ldots , \boldsymbol{D}_{\mathrm{M}-1}\}$ be a mapping vector that maps a label in the original collection of labels, $s \in \boldsymbol{L}$, to the corresponding label at the m^{th} level of the hierarchy, $\boldsymbol{S_{ms}} \in \boldsymbol{L}^m$, where $\boldsymbol{D}^m = \{0, \ldots , \boldsymbol{D}^m - 1\}$ 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 $\boldsymbol{\theta} \in \mathbb{R}^{\mathbf{R}\times\mathbf{L^m}\times\mathbf{L^m}}$ (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 s at the m_{th} level of the hierarchy. Additionally, let $\boldsymbol{\beta} \in \mathbb{R}^{\mathbf{R}\times\mathbf{L}}$ be a collection of exponential normalization values that ensure that the generative model is properly normalized. Thus, the generative model is described by

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

which can be directly interpreted as the probability that rater j observes label s' given the true label, hierarchical model, and the corresponding model parameters. 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 multi-tier estimate of rater performance.

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

where, β_{js} is an exponent that maintains the following constraint:

$$\sum_{s'}\left(\prod_m \theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right)^{\beta_{js}}=1 \quad (3)$$

In other words, β_{js} ensures that the model in Eq. 1 is a valid discrete probability mass function. Note, given the constraints on each individual $\boldsymbol{\theta}^m_j$ (i.e., that it is a valid confusion matrix) a unique value for β_{js} is guaranteed to exist and can easily be found using a standard searching algorithm (e.g., binary search, gradient descent).

In summary, the constrained geometric mean model of hierarchical performance provides a mechanism for enforcing consistent performance across the pre-defined hierarchical model. Specifically, in order for a given rater (or atlas) to make a positive impact on the final segmentation, the performance parameters for that rater must be indicative of high quality performance at all levels of the hierarchy. Alternatively, if a given rater performs poorly at the highest levels of the hierarchy, then this poor performance will automatically propagate to the lower levels of the hierarchy through the multiplicative model. As a result, the

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

hierarchical performance model provides two primary advantages over the traditional performance model: (1) the final estimate of performance is guaranteed to be consistent with the provided hierarchical representation (i.e., a high quality rater exhibits this quality throughout the hierarchy), and (2) poor performance at high levels of the hierarchy will automatically propagate to all sub-labels – providing a natural framework for consistently penalizing raters who exhibit globally poor performance. Regardless of the interpretation, given the model in Eq. 2 and constraint in Eq. 3, it is possible to utilize the provided hierarchical model within the statistical fusion EM framework. See Figure 1 for a graphical representation of the newly proposed generative model of hierarchical performance.

E-Step: Estimation of the Voxelwise Label Probabilities

Let $\boldsymbol{W} \in \mathbb{R}^{L \times N}$, where $W_{si}^{(k)}$ is 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

$$W_{si}^{(k)} \equiv f\left(T_i{=}s|\boldsymbol{D}, \mathscr{S}, \left\{\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{M-1}\right\}^{(k)}, \boldsymbol{\beta}^{(k)}\right). \quad (4)$$

Using a Bayesian expansion and the assumed conditional independence between the registered atlas observations, Eq. 4 can be re-written as

$$W_{si}^{(k)} = \frac{f\left(T_i{=}s\right) \prod_j f\left(D_{ij}{=}s'|T_i{=}s, \mathscr{S}, \left\{\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{M-1}\right\}^{(k)}, \boldsymbol{\beta}^{(k)}\right)}{\sum_n f\left(T_i{=}n\right) \prod_j f\left(D_{ij}{=}s'|T_i{=}n, \mathscr{S}, \left\{\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{M-1}\right\}^{(k)}, \boldsymbol{\beta}^{(k)}\right)} \quad (5)$$

where $f(T_i = s)$ is a voxelwise a *priori* distribution of the underlying segmentation. Note that the denominator of Eq. 5 is simply the solution for the partition function that enables $\boldsymbol{W}$ to be a valid probability mass function (i.e., $\sum_s W_{si}{=}1$). Using the simplified generative model in Eq. 2, the final form for the E-step of the EM algorithm can be written as

$$W_{si}^{(k)} = \frac{f\left(T_i{=}s\right) \prod_j \left(\prod_m \theta^{m,(k)}_{j\mathscr{S}_{ms'}, \mathscr{S}_{ms}}\right)^{\beta_{js}^{(k)}}}{\sum_{s''} f\left(T_i{=}s''\right) \prod_j \left(\prod_m \theta^{m,(k)}_{j\mathscr{S}_{ms'}, \mathscr{S}_{ms''}}\right)^{\beta_{js''}^{(k)}}} \quad (6)$$

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. 6). Unlike the traditional STAPLE approach, however, the parameters for each level of the hierarchy are maximized independently.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$$\begin{aligned}\boldsymbol{\theta}_j^{m,(k+1)} &= arg \underset{\boldsymbol{\theta}_j^m}{max} \sum_i E\left[ln\, f\left(D_{ij}=s'|T_i=s,\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\},\boldsymbol{\beta}^{(k)}\right)|\boldsymbol{D},\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\}^{(k)},\boldsymbol{\beta}^{(k)}\right]\\ &= arg \underset{\boldsymbol{\theta}_j^m}{max} \sum_i\sum_s W_{si}^{(k)} ln \quad f\left(D_{ij}=s'|T_i=s,\mathscr{S},\left\{\boldsymbol{\theta}^0,\ldots,\boldsymbol{\theta}^{M-1}\right\},\boldsymbol{\beta}^{(k)}\right)\end{aligned} \tag{7}$$

We can then perform a simple substitution using the hierarchical performance model defined in Eq. 2

$$\boldsymbol{\theta}_j^{m,(k+1)}=arg \underset{\boldsymbol{\theta}_j^m}{max}\sum_i\sum_s W_{si}^{(k)} ln\left(\prod_m \theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right)^{\beta_{js}^{(k)}} \tag{8}$$

Finally, using the properties of logarithms we obtain the final form of the conditional log likelihood function that we need to maximize in order to find the updated hierarchical performance parameters

$$\boldsymbol{\theta}_j^{m,(k+1)}=arg \underset{\times_j^m}{max}\sum_i\sum_s W_{si}^{(k)}\beta_{js}^{(k)}\sum_m ln\,\theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} \tag{9}$$

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., $\sum_{S'}\theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}=1$), 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

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}}\left[\sum_i\sum_s W_{si}^{(k)}\beta_{js}^{(k)}\sum_m ln\theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}+\lambda\sum_{s'}\theta^m_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right]\\ -\lambda &= \frac{\sum_{i:\mathscr{S}_{mD_{ij}}=\mathscr{S}_{ms'}}\sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}}\beta^{(k)}_{js''}W^{(k)}_{s''i}}{\theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}}\\ \theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} &= \frac{\sum_{i:\mathscr{S}_{mD_{ij}}=\mathscr{S}_{ms'}}\sum_{s'':\mathscr{S}_{ms''}=^{ms}}}{\beta}{}^{(k)}_{js''}W^{(k)}_{s''i}-\lambda\\ \theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}} &= \frac{\sum_{i:\mathscr{S}_{mD_{ij}}=\mathscr{S}_{ms'}}\sum_{s'':\mathscr{S}_{ms'}=\mathscr{S}_{ms}}\beta^{(k)}_{js''}W^{(k)}_{s''i}}{\sum_i\sum_{s'':\mathscr{S}_{ms}}\beta^{(k)}_{js''}W^{(k)}_{s''i}}\end{aligned} \tag{10}$$

where s'': $S_{ms''} = S_{ms}$ is the collection of all labels that map to the true label of interest, and i: $S_{mD_{ij}} = S_{ms'}$ is the collection of all voxels in which the observed label, D_{ij}, maps to the observed label of interest, $S_{ms'}$. At this point, it is important to note: (1) the 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. 10 uses $\beta_{js}^{(k)}$ which can then be updated, $\beta_{js}^{(k)} \rightarrow \beta_{js}^{(k+1)}$ following the constraint:

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$$\sum_{s'}\left(\prod_{m}\theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right)^{\beta^{(k+1)}_{js}}=1 \quad (11)$$

Extension to state-of-the-art Statistical Fusion Approaches

Recently, there have been several advancements to the statistical fusion framework, for instance (1) characterizing spatially varying performance – Spatial STAPLE ***(Asman and Landman, 2012a)***, (2) incorporation of non-local correspondence models – Non-Local STAPLE (NLS) ***(Asman and Landman, 2012c)***, and (3) a combination of the two – Non-Local Spatial STAPLE (NLSS). In the interest of brevity, we only fully derive the hierarchical version of STAPLE in this manuscript. However, we will briefly describe the extension to each of the above advancements to the statistical fusion framework. Note, while this is certainly not an exhaustive collection of advancements to the statistical fusion framework, the point is demonstrating the amenability of the proposed hierarchical reformulation to the new advancements to the STAPLE framework.

Hierarchical Spatial STAPLE

Background—Spatial STAPLE (Asman and Landman, 2011; Asman and Landman, 2012a) is an extension to the original STAPLE formulation to allow for smooth voxelwise estimates of rater performance. As a result, the confusion matrix describing rater performance, $\boldsymbol{\theta_j}$, becomes a function of the location in the image, $\boldsymbol{\theta_{ij}}$, defined over a pre-defined window surrounding the voxel of interest, $\boldsymbol{B_i}$. Where $\boldsymbol{B_i}$ is the set of voxels that are part of the window (or "pooling region") for the current voxel of interest, *i*.

E-Step:

$$W^{(k)}_{si}=\frac{f\left(T_i=s\right)\prod_j\left(\prod_m\theta^{m,(k)}_{ij\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right)^{\beta^{(k)}_{ijs}}}{\sum_{s''}f\left(T_i=s''\right)\prod_j\left(\prod_m\theta^{m,(k)}_{ij\mathscr{S}_{ms'}\mathscr{S}_{ms''}}\right)^{\beta^{(k)}_{ijs''}}} \quad (12)$$

M-Step:

$$\theta^{m,(k+1)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}=\frac{\sigma_{is}\theta^{m,(k)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}+\sum_{i'\in\boldsymbol{B}_i:\mathscr{S}_{mD_{i'j}}=\mathscr{S}_{ms'}}\sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}}\beta^{(k)}_{i'js''}W^{(k)}_{s''i'}}{\sigma_{is}\sum_s\theta^{m,(k)}_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}+\sum_{i'\in\boldsymbol{B}_i}\sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}}\beta^{(k)}_{i'js''}W^{(k)}_{s''i'}} \quad (13)$$

$$\sum_{s'}\left(\prod_{m}\theta^{m,(k+1)}_{ij\mathscr{S}_{ms'}\mathscr{S}_{ms}}\right)^{\beta^{(k+1)}_{ijs}}=1 \quad (14)$$

where σ_{is} is a scale factor that regularizes the local performance estimate to be closer to the global estimate of rater performance (Eq. 8). We formulate σ_{is} to be equal to relative amount

Author Manuscript

that the current label of interest, *s*, is estimated to be the correct answer over the pooling region, $\boldsymbol{B_i}$:

$$\sigma_{is}=|\boldsymbol{B}_i| - \sum_{i\in\boldsymbol{B}_i} W_{si}^{(k)} \quad (15)$$

where $|\boldsymbol{B_i}|$ is the number of elements in the pooling region centered at voxel *i*. Using this formulation, σ_{is} allows consistent estimates of rater performance despite the fact that only certain labels may be estimated in a given local region of the image – see ***(Asman and Landman, 2012a)*** for further details.

Note, while Spatial STAPLE uses a non-parametric regularizing function to prevent instabilities in the local performance estimates, alternative techniques could be used. For instance, one could use a maximum *a posteriori* (MAP) approach and assume a prior Beta distribution on the performance parameters – e.g., (Commowick et al., 2012; Commowick and Warfield, 2010). It is straightforward to see that the hierarchical formulation using a MAP formulation would remain valid. Regardless, the optimal framework for characterizing spatially varying performance remains an open problem and outside the scope of this manuscript.

Hierarchical Non-Local STAPLE (NLS)

Background—NLS (Asman and Landman, 2012b; Asman and Landman, 2012c) is another alternative to the original STAPLE algorithm in which the rater performance model is reformulated from a non-local means perspective. Briefly, using the intensity image provided for the target image, $\boldsymbol{I} \in \mathbb{R}^N$, and the corresponding registered intensity images from the atlases, $\boldsymbol{A} \in \mathbb{R}^{N\times R}$, the goal is the estimate the likelihood that $A_{i\hat{j}}$ is the true corresponding voxel to the target image I_i, where $i'$is an element in the search neighborhood defined for voxel $i - \mathscr{N}(i)$. Mathematically, we estimate this likelihood as $f(A_{i\hat{j}}|I_i)$

$$f\left(A_{i'j}|I_i\right) \equiv \alpha_{ji'i}=\frac{1}{Z_\alpha}\Delta\left(A_{i',j}, I_i\right) exp\left(-\frac{\mathscr{E}_{ii'2}}{2\sigma_d^2}\right) \quad (16)$$

where $\Delta(A_{i\hat{j}}, I_i)$ is a generic similarity model, $\left(-\frac{\varepsilon^2_{ii'}}{2\sigma_d^2}\right)$ is the spatial compatibility model, and Z_a is a partition function that ensures that $\sum_{i'\in\mathscr{N}(i)} \alpha_{ji'}=1$. For the similarity model, there are many different techniques that could be used (e.g., Gaussian difference model (Asman and Landman, 2012c; Isgum et al., 2009; Sabuncu et al., 2010), locally normalized correlation coefficient ***(Asman et al., 2013b; Cardoso et al., 2013)***, mutual information ***(Artaechevarria et al., 2009)***). In the spatial compatibility model, $\varepsilon_{ii'}$is the Euclidean distance between voxels *i* and i'in image space and σ_d is the corresponding standard deviation. For additional information on NLS see ***(Asman and Landman, 2012c)***.

E-Step:

Author Manuscript

$$W_{si}^{(k)}=\frac{f\left(T_i=s\right)\prod_j\sum_{i'\in\mathscr{N}(i)}\alpha_{ji'i}\left(\prod_m\theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m,(k)}\right)^{\beta_{js}^{(k)}}}{\sum_{s''}f\left(T_i=s''\right)\prod_j\sum_{i'\in\mathscr{N}(i)}\alpha_{ji'i}\left(\prod_m\theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms''}}^{m,(k)}\right)^{\beta_{js''}^{(k)}}}\qquad(17)$$

M-Step:

$$\theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m,(k+1)}=\frac{\sum_{i:\mathscr{S}_{m}D_{ij}=\mathscr{S}_{ms'}}\overline{\alpha_{ji}}\sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}}\beta_{js''}^{(k)}W_{s''i}^{(k)}}{\sum_i\sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}}\beta_{js''}^{(k)}W_{s''i}^{(k)}}\qquad(18)$$

where $\overline{\alpha_{ji}}=\left(\sum_{i'\in\mathscr{N}_S(i):\mathscr{S}_{mD_{i'j}}=\mathscr{S}_{ms'}}\alpha_{ji'i}\right)$, and the update of $\beta^{(k)}\rightarrow\beta^{(k+1)}$ is the same as Eq. 9.

Hierarchical Non-Local Spatial STAPLE (NLSS)

Background—NLSS is a unified statistical fusion algorithm that combines the formulations of Spatial STAPLE and NLS into a single unified framework that (1) allows for smooth spatially varying estimates of rater performance, and (2) reformulates the local performance estimates from a non-local means perspective.

E-Step:

$$W_{si}^{(k)}=\frac{f\left(T_i=S\right)\prod_j\sum_{i'\in\mathscr{N}(i)}\alpha_{ji'i}\left(\prod_m\theta_{ij\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m,(k)}\right)^{\beta_{ijs}^{(k)}}}{\sum_{s''}f\left(T_i=S''\right)\prod_j\sum_{i'\in\mathscr{N}(i)}\alpha_{ji'i}\left(\prod_m\theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms''}}^{m,(k)}\right)^{\beta_{ijs''}^{(k)}}}\qquad(19)$$

M-Step:

$$\theta_{ij\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m,(k+1)}=\frac{\sigma_{is}\theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m,(k)}+\sum_{i'\in\boldsymbol{B}_i:\mathscr{S}_{m}D_{i'j}=\mathscr{S}_{ms'}}\overline{\alpha_{ji'}}\sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}}\beta_{i'js''}^{(k)}W_{s''i'}^{(k)}}{\sigma_{is}\sum_s\theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m,(k)}+\sum_{i'\in\boldsymbol{B}_i}\sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}}\beta_{i'js''}^{(k)}W_{s''i'}^{(k)}}\qquad(20)$$

where all of the mathematical formulations are the same as described above for Spatial STAPLE and NLS.

Initialization, Detection of Convergence and Implementation

Given an *a priori* hierarchical model, there are no additional parameters in the proposed approach when compared to the non-hierarchical implementations of the statistical fusion framework. As a result, the hierarchical statistical fusion algorithms can be initialized in exactly the same was their traditional counterparts. Specifically, for all of the statistical

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

fusion approaches, the performance parameters were initialized by setting the on-diagonal elements to 0.95 and randomly setting the off-diagonal elements to fulfill the required constraints. The voxelwise label prior, $f(T_i = s)$, was initialized using the label probabilities from a “weak” log-odds majority vote (i.e., decay coefficient set to 0.5 voxels) ***(Sabuncu et al., 2010)***. For Spatial STAPLE, NLSS, and their hierarchical implementations, the pooling region, $\boldsymbol{B}_i$, was set with a half window radius of 5*mm* along all of the principal directions. For NLS, NLSS, and their hierarchical formulations, a Gaussian difference metric (using an intensity standard deviation of 0.1) was used with a half-window radius of 2*mm* along all of the principal directions for both the patch neighborhood and search neighborhood and the spatial standard deviation, σ_d, was set to 1.5*mm*

Detection of convergence in the hierarchical statistical fusion framework is slightly different than the traditional approach as we utilize all levels of the hierarchy. Thus, convergence is detected when the normalized trace of the raters’ performance parameters at each level of the hierarchy falls below some arbitrary threshold (herein, $\varepsilon = 10^{-4}$) between consecutive iterations of the EM algorithm.

$$\frac{1}{LRM}\sum_j\sum_m tr\left(\boldsymbol{\theta}_j^m\right) \quad (21)$$

For all of the presented experiments, the hierarchical implementations of the statistical fusion algorithms consistently converged in fewer than 15 iterations.

In terms of computational complexity, the hierarchical performance estimation has a marginal effect on the overall computation time of the algorithms. For example, for the whole-brain multi-atlas experiment presented below, estimation of the hierarchical performance parameters consistently added between 2-3 minutes to each of the corresponding statistical fusion algorithms. Moreover, this difference is largely negated by the substantial computational impact of estimating spatially-varying performance models (e.g., Spatial STAPLE), and/or non-local correspondence models (e.g., NLS).

Finally, the implementation of all of the considered statistical fusion algorithms presented in this paper are publicly available as part of the Java Image Science Toolkit (JIST) -- ***(Lucas et al., 2010)***, http://www.nitrc.org/projects/jist.

Methods and Results

For all of the presented simulations and experiments, the segmentation accuracy is measured using the Dice similarity coefficient (DSC) ***(Dice, 1945)***. Additionally, any claims of statistical significance refer to the results of a Wilcoxon signed-rank test ***(Wilcoxon, 1945)*** with a *p*-value threshold of 0.01.

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).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Experimental Design

A single 2D slice model (300 × 300 voxels, with 7 unique labels) was constructed to loosely approximate the types of relationships that are 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 2A for the best/worst observations). Additional details on the simulation model can be found in (Asman and Landman, 2012a; Asman and Landman, 2012c). As a baseline, a representative STAPLE result is presented. For incorporating hierarchical models into the statistical fusion framework, a single reference hierarchical structure was established (Figure 2B). Given, this structure, all unique trees (630 in total) were constructed via label permutation, and the resulting segmentation was estimated using hierarchical STAPLE.

Experimental Results

The quantitative results, measured by the mean DSC across 10 Monte Carlo iterations, for each of the considered hierarchical representations can be seen in Figure 2C. Here, it is evident that the hierarchical label representation plays a substantial role in determining overall segmentation accuracy. For reference, the accuracy of the traditional STAPLE framework and the accuracy using a “logical” hierarchical representation (Figure 2D) are highlighted. The qualitative results (Figure 2D-2F) support the quantitative assessment of accuracy. Specifically, the accuracy of the “logical” (Figure 2D), “best” (Figure 2E), and “worst” (Figure 2F) hierarchical label representations are presented. While not the absolute optimal representation in terms of overall mean DSC, the “logical” hierarchical representation: (1) results in a substantial qualitative improvement over the traditional STAPLE estimate and (2) results in a quantitatively superior segmentation estimate than more than 99.5% of the considered hierarchical representations. The “best” hierarchical representation is extremely similar to and results in a very minor improvement over the “logical” representation. Meanwhile, the “worst” representation completely ignores the underlying relationships exhibited in the truth model, and, not surprisingly, results in a very poor estimate of the final segmentation.

Whole Brain Multi-Atlas Segmentation

Data

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) ***(Marcus et al., 2007)*** with subjects ranging in age from 18 to 90. All images had a resolution of 1 × 1 × 1mm^3. All images were labeled using the BrainCOLOR protocol (http://www.braincolor.org/) ***(Klein et al., 2010)*** 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.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Experimental Design

We consider two separate registration frameworks. First we consider an affine-only pairwise registration framework ***(Ourselin et al., 2001)*** (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 ***(Avants et al., 2011)*** (using the Advanced Normalization Tools (ANTs) package – http://stnava.github.io/ANTs/). For the non-rigid registration, the deformations were computed using the symmetric normalization (SyN) transformation model computed using a local cross-correlation cost metric with a 3mm isotropic radius. 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 label fusion algorithms. First, in order to provide a benchmark of algorithmic performance, we consider a majority vote ***(Heckemann et al., 2006)***, a locally weighted vote (as described in ***(Sabuncu et al., 2010)***), and joint label fusion ***(Wang et al., 2012)*** (using the parameters described in ***(Wang and Yushkevich, 2013)***). Additionally, we consider STAPLE ***(Warfield et al., 2004)***, Spatial STAPLE ***(Asman and Landman, 2012a)***, ***NLS (Asman and Landman, 2012c)***, 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 representation, we constructed a 12-level hierarchical model (manually constructed by an experienced neuroimaging analyst). The goal of this hierarchy was to group labels together that provide similar contextual information about the quality of the atlas observations. Specifically, the first level of the hierarchy grouped all non-background labels together to form a “brain” label. Next, the “brain” label was partitioned into cerebrum, cerebellum and brain stem labels. In the third level, the gray matter, white matter, and cerebrospinal fluid labels in both the cerebrum and cerebellum were grouped together. This process was then carried out until all of the 133 labels were uniquely represented. For additional information on the hierarchical representation used for the brainCOLOR labels see: https://masi.vuse.vanderbilt.edu/index.php/TR-MASI-14-01.

Experimental Results

To summarize the improvements exhibited through the use hierarchical performance estimation, the results of the whole-brain multi-atlas segmentation experiment are presented in Figures 3-7. First, to quantitatively summarize the overall improvement for both registration frameworks, the mean DSC (across the 132 non-background labels) is presented in Figure 3. It is evident that the registration framework has a substantial impact on overall segmentation accuracy; yet, regardless of the registration model, the hierarchical reformulation of the performance parameters provides significant improvement in overall accuracy across each of the considered statistical fusion algorithms. For the affine registration framework, the hierarchical implementations provided a mean improvement across the testing data of 0.0070, 0.0118, 0.0199, and 0.0152 for STAPLE, Spatial STAPLE, NLS, and NLSS, respectively. All improvements were statistically significant. Additionally, hierarchical NLSS provided significant improvement over joint label fusion (JLF), the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

current state-of-the-art label fusion algorithm, for the affine registration framework. For the non-rigid registration framework, the hierarchical implementations provided a mean improvement across the testing data of 0.0104, 0.0049, 0.0048, and 0.0048 for STAPLE, Spatial STAPLE, NLS, and NLSS, respectively. Again, all improvements were statistically significant. Given the overall improvement in registration quality, the drop in improvement exhibited by the hierarchical implementations for the non-rigid registration framework is expected. Moreover, unlike the affine registration results, hierarchical NLSS and JLF reported statistically indistinguishable results when using the non-rigid registration framework. For both registration frameworks, there are two subjects that consistently appear as outliers. These subjects represent the two oldest subjects in the testing set (a 93 year old female and an 86 year old male) and are suboptimally represented by the demographics provided in the training set. In terms of fusion performance, the relatively low accuracy 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. Additionally, it is important to note that STAPLE and Hierarchical STAPLE are both out performed by majority vote for the non-rigid registration framework. This highlights the limitations of using a single global performance metric when estimating the final segmentation.

In addition to the overall results, the per-label accuracy for the non-cortical labels using the affine and non-rigid registration frameworks is presented in Figures 4 and 5, respectively. Here, for both registration frameworks, only the results using the NLS and NLSS and their hierarchical implementations are presented to avoid obfuscating the improvement provided by their corresponding hierarchical implementations. For the affine registration (Figure 4), Hierarchical NLS resulted in statistically significant improvement over NLS for 22 of the considered 34 non-cortical labels. Similarly, Hierarchical NLSS resulted in statistically significant improvement for 26 of the 34 non-cortical labels. NLS and NLSS were not significantly superior to their corresponding hierarchical implementations for any of the considered labels. For the non-rigid registration (Figure 5), Hierarchical NLS resulted in statistically significant improvement over NLS for 12 of the 34 considered labels, while Hierarchical NLSS resulted in statistically significant improvement for 19 of the 34 non-cortical labels. As with the affine registration, NLS and NLSS were not significantly superior to their corresponding hierarchical implementations for any of the considered labels.

The quantitative improvement (in terms of the DSC) in the cerebral cortex is summarized in Figure 6. As with before, only the results using the NLS and NLSS and their hierarchical implementations are presented. Here, it is evident that the hierarchical implementation of each algorithm provides substantial improvement in cortical segmentation accuracy, particularly for the affine-only registration framework. To summarize, for the affine registration, Hierarchical NLS resulted in statistically significant improvement over NLS for 57 of the 98 considered cortical labels and was statistically outperformed by NLS on 2 of the 98 cortical labels. Similarly, Hierarchical NLSS resulted in statistically significant improvement for 52 of the 98 cortical labels; however, it was not statistically outperformed by NLSS for any of the considered cortical labels. For the non-rigid registration, Hierarchical NLS resulted in statistically significant improvement over NLS for 28 of the considered 98 cortical labels and was statistically outperformed by NLS on 3 of the 98

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

cortical labels; while Hierarchical NLSS resulted in statistically significant improvement for 22 of the 98 cortical labels and, again, was not statistically outperformed by NLSS for any of the cortical labels.

The qualitative results (Figure 7) 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. While this improvement would almost certainly be smaller if a more complex, highly deformable registration model was used instead of a global affine registration, these qualitative results demonstrate the ability of the hierarchical performance parameters to enforce hierarchically consistent label estimations through constrained parameter estimation.

CT Orbit Multi-Atlas Segmentation

Data

A collection of 31 clinically acquired computed tomography (CT) images of the orbital region were retrieved in anonymous form under IRB supervision. The voxel size of the various images varied wildly, with in-plane resolution of approximately 0.5 mm and slice thickness ranging from 0.4 mm to 5 mm for the various target images. The “ground truth” labels were obtained from an experienced rater and were verified by multiple additional raters. In total, there were 5 considered labels on each dataset: background, left and right optic nerves, and left and right globe/orbital muscles.

Experimental Design

Using a leave-one-out cross-validation (LOOCV) we performed a multi-tier multi-atlas segmentation framework – see ***(Asman et al., 2013a)*** for additional details. Briefly, the images were affinely registered ***(Ourselin et al., 2001)*** and then cropped to form a reasonable region of interest surrounding the orbital area. After cropping, the images were non-rigidly registered ***(Avants et al., 2011)***, and the resulting label conflicts were resolved using label fusion. For the non-rigid registration, the deformation parameters were identical to the ones used for the previous whole-brain multi-atlas experiment.

Unlike the whole-brain segmentation experiments, the goal of this experiment was two-fold. First, we want to demonstrate the impact of reasonable and logical hierarchical representations of the orbital anatomy on the hierarchical statistical fusion accuracy. To accomplish this, we constructed three logical hierarchical representations (see Figure 8A). Each of these hierarchical representations could be considered a reasonable representation of the orbital anatomy (e.g., *left optic nerve* → *optic nerves* → *non-background* [“Hierarchy 2” in Figure 8A] or *left optic nerve* → *left orbit* → *non-background* [“*Hierarchy 3” in Figure 8A]*).

Second, we assess the accuracy of the statistical fusion model compared to the “ideal” segmentation estimate (i.e., the segmentation estimate obtained using the “ideal” performance parameters that are directly calculated using the desired manual segmentation). Obviously, in a typical empirical study, these ideal performance parameters are unknown,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and we rely on EM to optimally estimate them. Regardless, given the desired segmentation, obtaining the "ideal" performance parameters is straightforward. For STAPLE, the "ideal" performance parameters are computed as:

$$\theta_{js's}^{(ideal)} = \frac{\sum_{i:D_{ij}=s'} \delta\left(T_i, s\right)}{\sum_i \delta\left(T_i, s\right)} \quad (22)$$

where $\delta\left(\cdot,\cdot\right)$ is the Kronecker delta function which is equal to 1 if $T_i = s$ and 0 otherwise. Likewise, for Hierarchical STAPLE the computation of the "ideal" performance parameters is:

$$\theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m(ideal)} = \frac{\sum_{i:\mathscr{S}_m D_{ij}=\mathscr{S}_{ms'}} \sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}} \delta\left(T_i, s''\right)}{\sum_i \sum_{s'':\mathscr{S}_{ms''}=\mathscr{S}_{ms}} \delta\left(T_i, s''\right)} \quad (23)$$

where the corresponding exponential normalization factors, $\beta_{js}^{(ideal)}$, are then chosen based upon the following constraint

$$\sum_{s'} \left(\prod_m \theta_{j\mathscr{S}_{ms'}\mathscr{S}_{ms}}^{m,(ideal)} \right)^{\beta_{js}^{(ideal)}} = 1. \quad (24)$$

Experimental Results

The results from the LOOCV experiment for multi-atlas segmentation of the orbital region are summarized in Figure 8. The considered logical hierarchical label representations for this segmentation task are presented in Figure 8A. The quantitative comparison of STAPLE and the corresponding Hierarchical STAPLE estimates are presented in Figure 8B. Here, for each of the considered hierarchical representations, Hierarchical STAPLE estimates result in statistically significant improvement over the traditional STAPLE framework. Additionally, the "ideal" Hierarchical STAPLE estimates result in statistically significant improvement over the corresponding "ideal" STAPLE estimate. As a result, it can be directly inferred that *empirically and theoretically*, using a reasonable and logical hierarchical representation for Hierarchical STAPLE results in substantial improvement in overall accuracy. Interestingly, "Hierarchy 3" which utilizes the relationships between the left and right orbital regions, results in best overall performance when estimated using EM and using the ideal parameters. While not definitive, this illustrates the importance of hierarchically grouping labels that are (1) likely to be confused with one another, and (2) indicative of one another's performance.

The "ideal" performance parameters represent an upper bound on potential performance achieved by the statistical fusion framework. In addition to providing statistically significant improvement in overall accuracy over the other considered approaches (Figure 8B – left), Hierarchical STAPLE using "Hierarchy 3" results in segmentation estimates that are closest the "ideal" performance estimate (Figure 8B – right). This strongly implies that utilizing a logical representation of the hierarchical relationships exhibited in the data results in an

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

increased likelihood of converging to a local optimum that is closer to the ideal global optimum.

The qualitative results (Figure 8C) support the quantitative improvement exhibited by Hierarchical STAPLE. Here, it is evident that each of the proposed hierarchical label representations results in: (1) substantial improvement over traditional STAPLE and (2) segmentation estimates that are qualitatively closer to the upper bound provided by the “ideal” segmentation estimate. Again, Hierarchical STAPLE using “Hierarchy 3” results in the largest improvement in mean DSC across the considered labels with an improvement of 0.0957, while “Hierarchy 1” and “Hierarchy 2” result in smaller, yet substantial improvement: 0.0752 and 0.0525, respectively.

Discussion

Herein, we propose a novel statistical fusion framework using a reformulated hierarchical performance model. Given an *a priori* model of the hierarchical label relationships for a given segmentation task, the proposed generative model of rater 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 provided a theoretical advancement to the statistical fusion framework that enables the simultaneous estimation of multiple (hierarchical) confusion matrices for each rater., (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 demonstrated statistically significant improvement in overall segmentation accuracy on both simulated and empirical data

Specifically, through a motivating simulation we have demonstrated the substantial impact that hierarchical label representations have on segmentation accuracy (Figure 2). For a 133 label whole-brain multi-atlas segmentation task, we have shown substantial and significant accuracy improvement in terms of overall accuracy (Figure 3), non-cortical segmentation (Figures 4 and 5), and cerebral cortex segmentation (Figure 6). These accuracy improvements are supported by qualitative inspection (Figure 7). Finally, using a multi-atlas segmentation framework for the orbital region on CT, we evaluated the accuracy of Hierarchical STAPLE using 3 different logical hierarchical representations of the orbital anatomy (Figure 8). Additionally, using the “ideal” performance parameters as an upper bound, the *empirical* and *theoretical* benefits of the hierarchical performance estimation framework is highlighted.

Despite the promise of the proposed framework, there are several potential advancements that require future exploration. For example, 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. With that said, using the model described in Eq. 2, there are certain qualities of optimal hierarchies that can be inferred from mathematical intuition. For instance, when estimating

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the performance parameters for a given observed label, it is evident that the hierarchical parents that led to this observed label have a substantial impact on the final performance parameter representation. Thus, the quality of the rater (or atlas) at the hierarchical parent labels should be strongly indicative of the performance of the leaf label (i.e., the final label representation). This type of framework is largely supported by the results in Figures 1 and 8. For example, in Figure 8, the worst performing "logical" hierarchy is the representation that groups the optic nerves together. This is not particularly surprising because the left and right optic nerves are spatially separated from one another, and, thus, the quality of an individual atlas observation will be more strongly tied to the intra-orbital labels (see "Hierarchy 3") as opposed to the inter-orbital labels (see "Hierarchy 2").

Additionally, 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. With that said, these multi-protocol methods would inherently rely on an anatomical representation that can be accurately modeled through a hierarchy. For instance, while hierarchies provide a natural framework for capturing many intuitive anatomical relationships, they are limited by their very nature to relationships that can be represented as a tree. Unfortunately, many physical structures have heterogeneous composition (in terms of both tissue and function) that simply cannot be represented in this form.

In the end, we have presented a powerful theoretical advancement to the statistical fusion context for leveraging the complex inter-structure relationships. 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. Most importantly, we have demonstrated the proposed hierarchical performance model is completely amenable to many of the state-of-the-art statistical fusion algorithms. As a result, the primary contribution of this manuscript is a foundational level improvement to statistical fusion theory that is applicable to the entire gamut of fusion-based applications.

Acknowledgements

We would like to thank Andrew Worth and Neuromorphometrics, Inc. for the whole-brain data, and Alexander Dagley for his assistance in labeling the orbital data and constructing the hierarchical representations. This work was supported in part by NIH R21EY024036, 1R21NS064534, 2R01EB006136, R01EB006193, 1R03EB012461. This research 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

Akhondi-Asl A, Warfield S. Simultaneous Truth and Performance Level Estimation Through Fusion of Probabilistic Segmentations. IEEE transactions on medical imaging. 2013

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Aljabar P, Heckemann R, Hammers A, Hajnal J, Rueckert D. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. Neuroimage. 2009; 46:726–738. [PubMed: 19245840]

Artaechevarria X, Muñoz-Barrutia A, Ortiz-de-Solorzano C. Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE Trans. Med. Imaging. 2009; 28:1266–1277. [PubMed: 19228554]

Asman, A.; Landman, B. Information Processing in Medical Imaging (IPMI). Vol. 6801. Springer; 2011. Characterizing spatially varying performance to improve multi-atlas multi-label segmentation; p. 85-96.

Asman, AJ.; Dagley, AS.; Landman, BA. Statistical label fusion with hierarchical performance models. SPIE Medical Imaging. San Diego, CA.: 2014.

Asman AJ, DeLisi MP, Mawn LA, Galloway RL, Landman BA. Robust non-local multi-atlas segmentation of the optic nerve, SPIE Medical Imaging. International Society for Optics and Photonics. 2013a:86691L–86691L-86698.

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

Asman, AJ.; Landman, BA. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; Nice, France: 2012b. Non-Local STAPLE: An Intensity-Driven Multi-Atlas Rater Model; p. 417-424.

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

Asman, AJ.; Smith, SA.; Reich, DS.; Landman, BA. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; Nagoya, Japan: 2013b. Robust GM/WM Segmentation of the Spinal Cord with Iterative Non-Local Statistical Fusion; p. 759-767.

Avants B, Epstein C, Grossman M, Gee J. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis. 2008; 12:26–41. [PubMed: 17659998]

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; 54:2033–2044. [PubMed: 20851191]

Bai W, Shi W, O'Regan DP, Tong T, Wang H, Jamil-Copley S, Peters NS, Rueckert D. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. IEEE transactions on medical imaging. 2013; 32:1302–1315. [PubMed: 23568495]

Beucher, S. Mathematical morphology and its applications to image processing. Springer; 1994. Watershed, hierarchical segmentation and waterfall algorithm; p. 69-76.

Cao Y, Yuan Y, Li X, Turkbey B, Choyke PL, Yan P. Segmenting images by combining selected atlases on manifold. Med Image Comput Comput Assist Interv. 2011; 14:272–279. [PubMed: 22003709]

Cardoso MJ, Leung K, Modat M, Barnes J, Ourselin S. Locally Ranked STAPLE for template based segmentation propagation. MICCAI Workshop on Multi-Atlas Labeling and Statistical Fusion. 2011

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; 17:671–684. [PubMed: 23510558]

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]

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; 31:1593–1606. [PubMed: 22562727]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

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.

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; 54:940–954. [PubMed: 20851199]

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.

Depa, M.; Sabuncu, MR.; Holmvang, G.; Nezafat, R.; Schmidt, EJ.; Golland, P. Statistical Atlases and Computational Models of the Heart. Springer; 2010. Robust atlas-based segmentation of highly variable anatomy: Left atrium segmentation; p. 85-94.

Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945; 26:297–302.

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

Iglesias JE, Sabuncu MR, Van Leemput K. A unified framework for cross-modality multi-atlas segmentation of brain MRI. Medical Image Analysis. 2013; 17:1181–1191. [PubMed: 24001931]

Isgum I, Staring M, Rutten A, Prokop M, Viergever MA, van Ginneken B. Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging. 2009; 28:1000–1010. [PubMed: 19131298]

Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, Christensen GE, Collins DL, Gee J, Hellier P. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage. 2009; 46:786–802. [PubMed: 19195496]

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.

Klein A, Hirsch J. Mindboggle: a scatterbrained approach to automate brain labeling. Neuroimage. 2005; 24:261. [PubMed: 15627570]

Langerak TR, van der Heide UA, Kotte ANTJ, Viergever MA, van Vulpen M, Pluim JPW. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans. Med. Imaging. 2010; 29:2000–2008. [PubMed: 20667809]

Lucas BC, Bogovic JA, Carass A, Bazin P-L, Prince JL, Pham DL, Landman BA. The Java Image Science Toolkit (JIST) for rapid prototyping and publishing of neuroimaging software. Neuroinformatics. 2010; 8:5–17. [PubMed: 20077162]

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; 19:1498–1507. [PubMed: 17714011]

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

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

Rohlfing T, Brandt R, Menzel R, Maurer CR. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. Neuroimage. 2004a; 21:1428–1442. [PubMed: 15050568]

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. 2004b; 23:983–994. [PubMed: 15338732]

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. 2004c; 23:983–994. [PubMed: 15338732]

Author Manuscript

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; 29:1714–1729. [PubMed: 20562040]

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; 35:611–623.

Wang H, Yushkevich PA. Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. Frontiers in neuroinformatics. 2013; 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; 23:903–921. [PubMed: 15250643]

Weisenfeld N, Warfield S. Learning likelihoods for labeling (L3): a general multi-classifier segmentation algorithm, Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer. 2011; 6893:322–329.

Wilcoxon F. Individual comparisons by ranking methods. Biometrics bulletin. 1945; 1:80–83.

Wolz R, Aljabar P, Hajnal JV, Hammers A, Rueckert D. LEAP: Learning embeddings for atlas propagation. Neuroimage. 2010; 49:1316–1325. [PubMed: 19815080]

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- We provide a theoretical advancement to the statistical fusion framework to enable hierarchical performance estimation.
- Advancements are highly amenable to many of the state-of-the-art advancements to the statistical fusion framework.
- We demonstrate statistically significant improvement on both simulated and empirical data.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

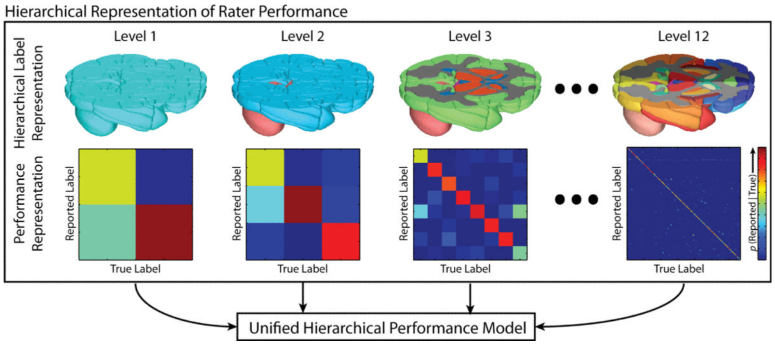

Hierarchical Representation of Rater Performance
Hierarchical Label Representation
Performance Representation
Level 1
Level 2
Level 3
Level 12
Reported Label
True Label
p (Reported | True)
Unified Hierarchical Performance Model

Figure 1.

Hierarchical representation of rater performance. Volumetric renderings of the brain anatomy at the various levels are shown. At each level, the rater performance is quantified using a representative confusion matrix. Each level is then unified through a complete hierarchical performance model.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

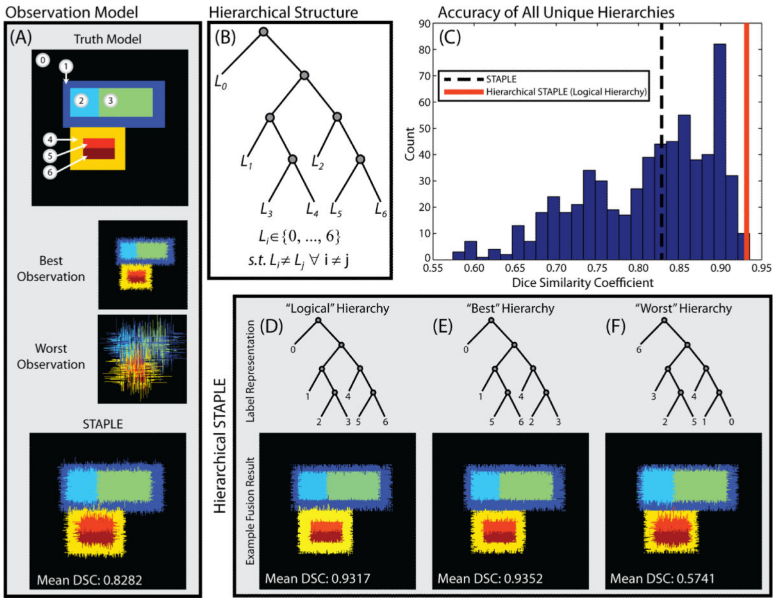


Figure 2.

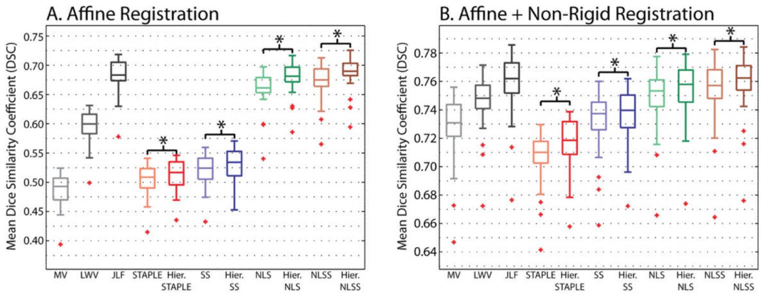
Motivating simulation data and results. A simple 2D simulated dataset was constructed with observations using a boundary error model (A). Given a pre-defined hierarchical structure (B), the accuracy of all possible unique hierarchies via label permutation was quantified (C). Representative estimates using the “logical” (D), “best” (E), and “worst” (F) hierarchies are also presented.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



Note: * indicates statistical difference using a Wilcoxon signed-rank test (p < 0.01).

Figure 3.

Mean accuracy of the various benchmarks and their corresponding hierarchical implementations for both the affine and the non-rigid registration frameworks. The accuracy of a majority vote (MV), locally-weighted vote (LWV), and joint label fusion (JLF) are presented to provide a reference baseline. The hierarchical implementations for STAPLE, Spatial STAPLE (SS), Non-Local STAPLE (NLS), and Non-Local Spatial STAPLE (NLSS) provide consistent and statistically significant improvement over their non-hierarchical counterparts.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

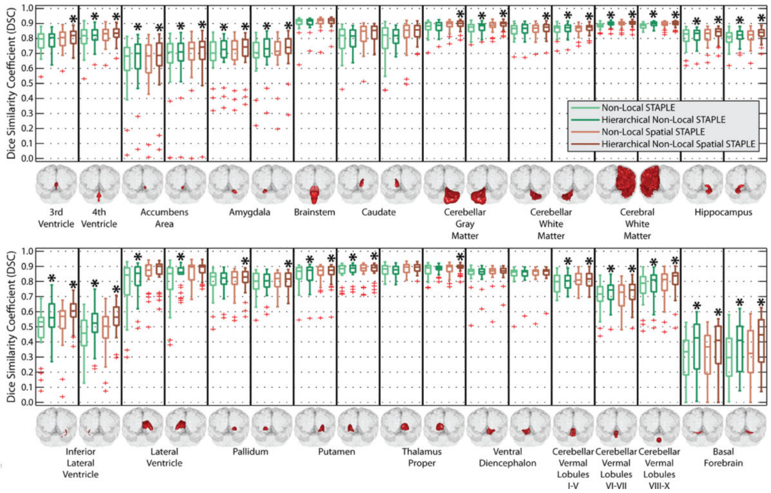


Figure 4.

Per-label accuracy for non-cortical labels for hierarchical implementations of NLS and NLSS using the affine registration framework. The hierarchical reformulations provide substantial and significant improvement for many of the considered labels. A “*” over the hierarchical NLS or NLSS results indicate statistically significant improvement over the non-hierarchical implementation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

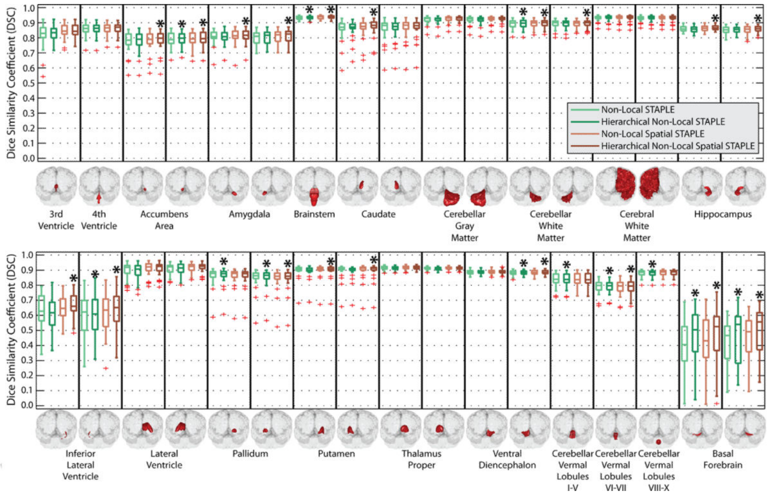


Figure 5.

Per-label accuracy for non-cortical labels for hierarchical implementations of NLS and NLSS using the non-rigid registration framework. As with the affine-only registration framework (Figure 4), the hierarchical implementations provide substantial and significant improvement for many of the considered labels. A “*” over the hierarchical NLS or NLSS results indicate statistically significant improvement over the non-hierarchical implementation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

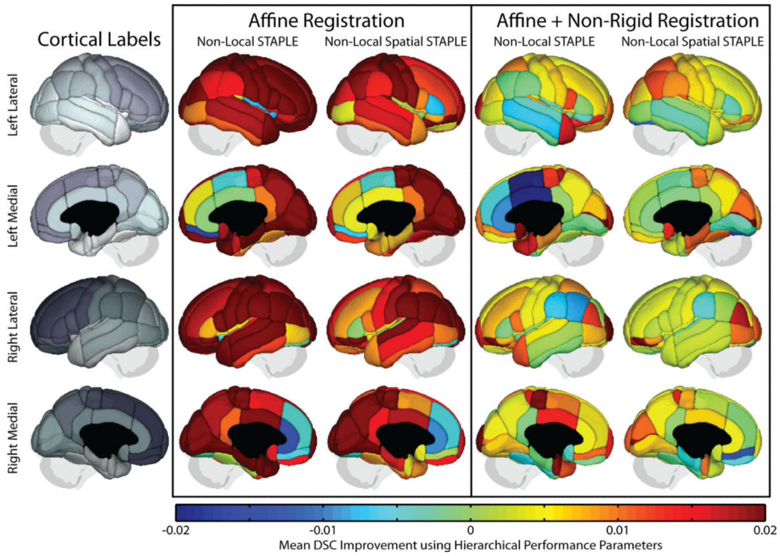


Figure 6.

Mean per-label accuracy improvement for cortical labels using the hierarchical implementations of NLS and NLSS for the both of the considered registration frameworks. Particularly for the affine registration framework, the hierarchical reformulations provide substantial improvement in mean DSC accuracy for many of the cortical labels.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

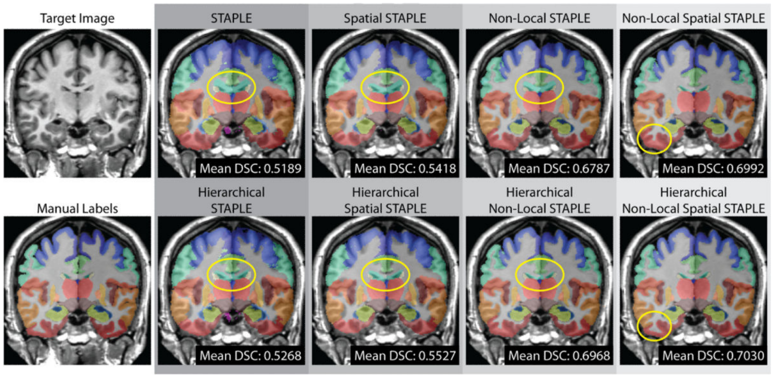


Figure 7.

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. The ellipses highlight regions exhibiting particular qualitative improvement.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

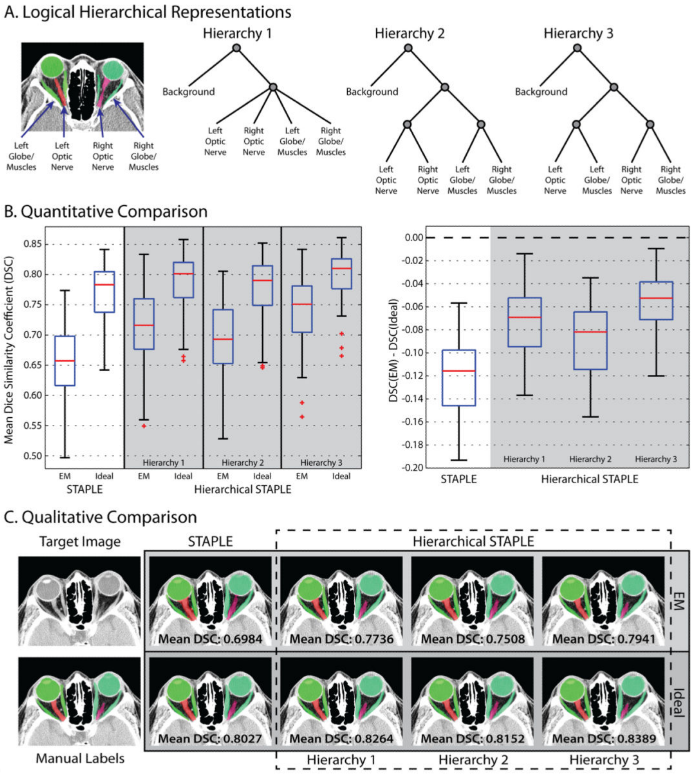


Figure 8.

Empirical evaluation of the impact of the various logical hierarchical representations on STAPLE applied to multi-atlas segmentation of orbital anatomy on CT. The three considered logical hierarchical representations are shown in (A). The quantitative comparison (B) demonstrates that Hierarchical STAPLE provides significant improvement using both the “ideal” performance parameters, and the parameters estimated via EM. The quantitative accuracy benefits support the qualitative improvement shown in (C).