Medical Image Analysis 18 (2014) 1070–1081

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Hierarchical performance estimation in the statistical label fusion framework

Andrew J. Asman [a,*], Bennett A. Landman [a,b]

[a] Electrical Engineering, Vanderbilt University, Nashville, TN 37235, USA
[b] Institute of Imaging Science, Vanderbilt University, Nashville, TN 37235, USA

ARTICLE INFO

Article history:
Received 7 January 2014
Received in revised form 17 April 2014
Accepted 16 June 2014
Available online 4 July 2014

Keywords:
Label fusion
Multi-atlas segmentation
STAPLE
Hierarchical segmentation
Rater performance models

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.

1. 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, 2012b; 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, 2012b; 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

* Corresponding author. Address: Vanderbilt University EECS, 2301 Vanderbilt Pl., PO Box 351679 Station B, Nashville, TN 37235-1679, USA. Tel.: +1 (615) 322 2338.
E-mail address: andrew.j.asman@vanderbilt.edu (A.J. Asman).

http://dx.doi.org/10.1016/j.media.2014.06.005
1361-8415/

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., 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 (Fig. 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-of-the-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.

2. Theory

2.1. Problem Definition

Let $T \in L^{N\times 1}$ be the latent representation of the true target segmentation, where $\boldsymbol{L} = \{0, \ldots, L-1\}$ 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 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 M levels. At each level of the hierarchy, let $\mathcal{S}_m \in \boldsymbol{\mathcal{S}} = \{\mathcal{S}_0, \ldots, \mathcal{S}_{M-1}\}$ be a mapping vector that maps a label in the original collection of labels, $s \in L$, to the corresponding label at the mth level of the hierarchy, $\mathcal{S}_{ms} \in L^m$, where $\boldsymbol{L}^m = \{0, \ldots, L^m - 1\}$ is the collection labels at the mth 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, $\theta^m_{j\mathcal{S}_{ms'}\mathcal{S}_{ms}}$ is the probability that rater j observes label s' given that the true label is s at the mth level of the hierarchy. Additionally, let $\boldsymbol{\beta} \in \mathbb{R}^{R\times L}$ be a collection of exponential normalization values that ensure that the generative model is properly normalized. Thus, the generative model is described by

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

which can be directly interpreted as the probability that rater j observes label s' given the true label, hierarchical model, and the

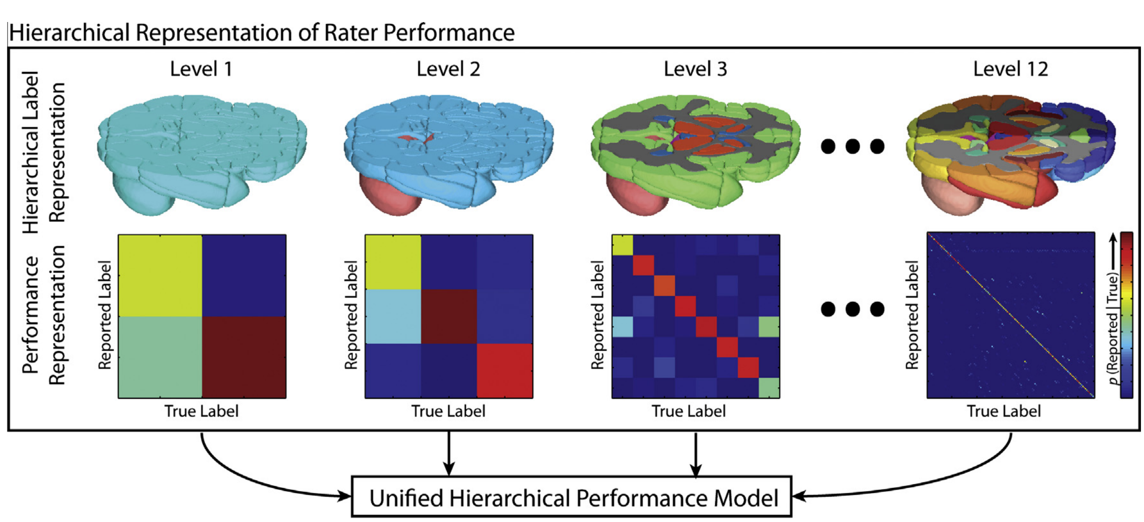

Fig. 1. Hierarchical representation of rater performance. Volumetric renderings of the brain anatomy at the various levels are shown. At each level, the rater performance is quantified using a representative confusion matrix. Each level is then unified through a complete hierarchical performance model.