NeuroImage 106 (2015) 451–463

Contents lists available at ScienceDirect

NeuroImage

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

# An algorithm for optimal fusion of atlases with different labeling protocols

Juan Eugenio Iglesias [a,*], Mert Rory Sabuncu [b,d], Iman Aganj [b], Priyanka Bhatt [c], Christen Casillas [c], David Salat [b], Adam Boxer [c], Bruce Fischl [d,b], Koen Van Leemput [b,e,f,g]

[a] Basque Center on Cognition, Brain and Language (BCBL), Spain
[b] Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School/Massachusetts General Hospital, Charlestown, MA, USA
[c] Memory and Aging Center, University of California, San Francisco, USA
[d] MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), USA
[e] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[f] Department of Information and Computer Science, Aalto University, Finland
[g] Department of Biomedical Engineering and Computational Science, Aalto University, Finland

ARTICLE INFO

*Article history:*
Accepted 14 November 2014
Available online 22 November 2014

*Keywords:*
Segmentation
Label fusion

ABSTRACT

In this paper we present a novel label fusion algorithm suited for scenarios in which different manual delineation protocols with potentially disparate structures have been used to annotate the training scans (hereafter referred to as "atlases"). Such scenarios arise when atlases have missing structures, when they have been labeled with different levels of detail, or when they have been taken from different heterogeneous databases. The proposed algorithm can be used to automatically label a novel scan with any of the protocols from the training data. Further, it enables us to generate new labels that are not present in any delineation protocol by defining intersections on the underling labels. We first use probabilistic models of label fusion to generalize three popular label fusion techniques to the multi-protocol setting: majority voting, semi-locally weighted voting and STAPLE. Then, we identify some shortcomings of the generalized methods, namely the inability to produce meaningful posterior probabilities for the different labels (majority voting, semi-locally weighted voting) and to exploit the similarities between the atlases (all three methods). Finally, we propose a novel generative label fusion model that can overcome these drawbacks. We use the proposed method to combine four brain MRI datasets labeled with different protocols (with a total of 102 unique labeled structures) to produce segmentations of 148 brain regions. Using cross-validation, we show that the proposed algorithm outperforms the generalizations of majority voting, semi-locally weighted voting and STAPLE (mean Dice score 83%, vs. 77%, 80% and 79%, respectively). We also evaluated the proposed algorithm in an aging study, successfully reproducing some well-known results in cortical and subcortical structures.



## Introduction

Automatic segmentation of brain structures from MRI data makes it possible to carry out neuroimaging studies at larger scales than manual tracings would, since the latter are very time consuming to make. Moreover, automatic segmentation methods are also more repeatable and reliable than their manual counterparts. Brain MRI segmentation has been used in a number of applications, such as tractography (Yendiki et al., 2011), surgical planning (Cline et al., 1990) and studies of aging (Walhovd et al., 2005), brain development (Knickmeyer et al., 2008) and pathologies like Alzheimer's disease (De Jong et al., 2008).

One family of supervised segmentation techniques that has become popular in brain MRI is multi-atlas segmentation (Rohlfing et al., 2004). In conventional atlas-based segmentation, the grayscale image of the atlas is nonlinearly registered to the space of the test scan, and the resulting transform is then used to warp the corresponding labels, which provide an estimate of the segmentation. Since a single atlas is not sufficient to cover the whole spectrum of variability within a population, multi-atlas segmentation has emerged as a natural extension. Using multiple atlases, this family of techniques produces more accurate segmentations (Awate and Whitaker, 2014) by: (1) independently registering several atlases to the test scan; (2) using the resulting transforms to deform the corresponding label images; and (3) combining the registered label maps into a single estimate of the segmentation with a label fusion algorithm. Multi-atlas segmentation is becoming widespread for three reasons. First, the maturity of registration algorithms (e.g., ANTs/SyN (Avants et al., 2008), Elastix (Klein et al., 2010)) enables multi-atlas techniques to achieve very high performance. Second, the public availability of such methods makes multi-atlas segmentation

* Corresponding author.
*E-mail address:* e.iglesias@bcbl.eu (J.E. Iglesias).

http://dx.doi.org/10.1016/j.neuroimage.2014.11.031
1053-8119/

easy to implement. And third, the relative computational cost associated with nonlinearly registering the atlases is quickly diminishing thanks to the rapid increase in processing power of computers.

The choice of label fusion method is critical for the performance of multi-atlas segmentation. Early algorithms include *best atlas selection* (Rohlfing et al., 2004) and *majority voting* (Heckemann et al., 2006). The former estimates the segmentation as the labels of the atlas that is most similar to the test scan after registration. In this context, similarity can be measured with the same metrics that are typically used in image registration, such as cross-correlation, mutual information or sum of squared differences. Majority voting, on the other hand, operates at the voxel level by locally assigning the most frequent deformed atlas label at each spatial location – without considering the image intensity information. The performance of majority voting can be increased by an atlas selection process, in which only the deformed atlases that are most similar to the target scan are considered in the fusion (Aljabar et al., 2009; Duc et al., 2013).

Later fusion methods compute the segmentation as a weighted combination of the labels of the registered atlases such that higher weights are given to more similar atlases. The weights can be global (Artaechevarria et al., 2008) or local (Isgum et al., 2009; Coupé et al., 2011; Wang et al., 2013; Sabuncu et al., 2010). *Sabuncu* et al. (Sabuncu et al., 2010) have shown that many of these multi-atlas methods can be written within a unified generative model. Another popular label fusion approach is STAPLE (Warfield et al., 2004) and its variants (Asman and Landman, 2012, 2013; Cardoso et al., 2013; Akhondi-Asl and Warfield, 2013); while this method was originally developed to combine multiple manual segmentations from different human raters, it is increasingly being applied in the context of multi-atlas label fusion.

All the aforementioned label fusion algorithms assume that all structures of interest are labeled in all atlases, which is a rather limiting constraint. Eliminating this requirement would have several practical implications:

- It would enable us to combine training scans from different datasets even if they have different sets of annotated structures. In turn, this would make it possible to take advantage of the increasing amount of heterogeneously labeled MRI data that are publicly available.
- It would also enable us to segment structures that are not included in any of the datasets, but defined as the intersection of labels. For instance, the intersection of the lateral postcentral region and the cerebral gray matter would define the primary somatosensory cortex.
- It would allow for the fusion of segmentations from different modalities with different field of views and resolution. For instance, it would be possible to combine standard resolution brain MRI (1 mm resolution) with high-resolution MRI with limited field of view or even histology or optical coherence tomography data.
- It would be useful if one were to manually relabel a subset of atlases to include finer structures in the annotations. For example, in a large dataset with the hippocampi already labeled, an expert rater can additionally delineate the hippocampal subfields – which is extremely difficult and time consuming – in just a few cases. Traditional label fusion methods would only be able to use these few scans in the segmentation, having to disregard the information in all the scans in which the subfields are not labeled.

Despite the practical implications that a label fusion algorithm which allows for heterogeneously labeled atlases would have, this direction remains largely unexplored in the literature. To the best of our knowledge, only a particular case of label fusion with heterogeneous labels has been considered so far: the situation in which some of the labels are missing in some of the atlases. To tackle this problem, *Landman* et al. (Landman et al., 2009, 2010, 2012) propose an ad-hoc solution by modifying the STAPLE framework such that unlabeled voxels are ignored and the confusion matrix entries corresponding to the missing structures are fixed. *Commowick* et al. (Commowick et al., 2012) propose ameliorating the effect of missing labels by adding a prior on the confusion matrices to the STAPLE algorithm that, when a label is missing, encourages higher a transition probability from that label to the background. However, such an approach treats as background all the voxels that have not been labeled with one of the foreground labels.

In this study, we present a family of probabilistic models for label fusion that make it possible to use atlases that have been annotated with different protocols. In our models, the atlases are assumed to have a hidden “fine” segmentation with all the structures present in the training data – including those defined by intersections of labels. The actual observed labels are assumed to have been obtained by collapsing groups of hidden fine labels into more general, coarse labels.

The contribution of this study is twofold:

i. We use probabilistic models of label fusion to extend three popular methods (majority voting, semi-locally weighted fusion and STAPLE) to the scenario of heterogeneously labeled atlases.
ii. We propose a new generative model for label fusion that can overcome the limitations of these generalizations – the inability to produce meaningful posteriors and to exploit the similarities between the atlases – and show that it outperforms the generalizations in experiments with four datasets.

The rest of this paper is organized as follows. In the Methods section, we describe the general framework for label fusion with heterogeneously labeled atlases, propose the generalizations of the different methods, identify their disadvantages, and present a new fusion algorithm to address their shortcomings. In the Experiments and results section, we assess the performance of the different algorithms with experiments on four different datasets. Finally, the Conclusion and discussion section closes the paper.

## Methods

In this section, we first introduce the general framework and define the variables that we will use throughout the paper. Then, we present the generalizations of majority voting, semi-locally weighted voting and STAPLE and identify their weaknesses. Finally, we introduce a label fusion method that addresses these shortcomings.

### *General framework*

Throughout the remainder of this paper, we will assume that a test scan consisting of $J$ voxels is to be segmented. We will use $\mathbf{y} = \{y_j, j = 1, \ldots, J\}$ to refer to the image intensities, and $\mathbf{s} = \{s_j, j = 1, \ldots, J\}$ to refer to its hidden, underlying segmentation. Let us also assume that a set of $N$ atlases has been pre-registered to a test scan with a non-linear algorithm. Let $\{\mathbf{i}_n\}$ (where $\mathbf{i}_n = \{i_{nj}, j = 1, \ldots, J\}$) be the observed image intensities of the $N$ registered atlases, and let $\{\mathbf{l}_n\}$ (where $\mathbf{l}_n = \{l_{nj}, j = 1, \ldots, J\}$) be the corresponding discrete labels, defined at the finest detail level. Their values range from 1 to $L$, the total number of fine labels.

These deformed labels $\{\mathbf{l}_n\}$ are not directly observed; instead, we have access to a different set of coarse labels $\{\mathbf{c}_n\}$ (where $\mathbf{c}_n = \{c_{nj}, j = 1, \ldots, J\}$), which correspond to the actual manual delineations. The coarse labels $\{\mathbf{c}_n\}$ are obtained by collapsing the fine labels $\{\mathbf{l}_n\}$ into different groups of labels by means of a set of $N$ deterministic, protocol-specific functions: $c_{nj} = f_n(l_{nj})$. A protocol function could, for instance, collapse the hippocampal subfields into a single hippocampal label. Having a separate $f_n$ for each atlas enables us to combine different labeling protocols. Different protocol functions can collapse the same fine label into different coarse labels; for instance, orbital cortex could be collapsed into the cerebral cortex by one protocol and into the frontal lobe by another.