HHS Public Access

Author manuscript

Proc SPIE Int Soc Opt Eng. Author manuscript; available in PMC 2010 December 06.

Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2010 March 1; 7623: . doi:10.1117/12.844214.

Statistical Fusion of Surface Labels Provided by Multiple Raters

John A. Bogovic[a], **Bennett A. Landman**[b,d], **Pierre-Louis Bazin**[c], and **Jerry L. Prince**[a,b]

[a]Electrical, Johns Hopkins University, Baltimore, MD, USA

[b]Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

[c]Radiology, Johns Hopkins University, Baltimore, MD, USA

[d]Electrical Engineering, Vanderbilt University, Nashville, TN, USA

Abstract

Studies of the size and morphology of anatomical structures rely on accurate and reproducible delineation of the structures, obtained either by human raters or automatic segmentation algorithms. Measures of reproducibility and variability are vital aspects of such studies and are usually acquired using repeated scans and repeated delineations (in the case of human raters). Methods exist for simultaneously estimating the true structure and rater performance parameters from multiple segmentations and have been demonstrated on volumetric images. In this work, we extend the application of previous methods onto two-dimensional surfaces parameterized as triangle meshes. Label homogeneity is enforced using a Markov random field formulated with an energy that addresses the challenges introduced by the surface parameterization. The method was explored using both simulated raters and surface labels obtained from an atlas registration. Simulated raters are computed using a global error as well as a novel and more realistic boundary error model. We study the impact of raters and their accuracy based on both models, and show how effectively this method estimates the true segmentation on simulated and real surfaces.

Keywords

Rater evaluation; labeling; label fusion; surface mesh; statistics; STAPLE

1. INTRODUCTION

Assessment of structural and morphological anatomical characteristics plays an essential role in the application of medical imaging in clinical research. Such assessments depend upon the ability to accurately and precisely label structures in multidimensional images. The labeling process may be carried out by human raters or by automated segmentation algorithms. Statistically motivated approaches1,2 have been developed (e.g., Simultaneous Truth and Performance Level Estimation, STAPLE) to fuse or combine labels from multiple individuals or algorithms into an estimate of a truth label set that can be made more reliable than any of the individual, underlying labelings. These approaches have been applied to volumetric studies, and additionally provide measures of rater or algorithm performance. Although typically acquired in two or three dimensions, medical images may be processed and/or analyzed to form representations on manifolds (e.g., cortical surfaces). In this paper, we extend the STAPLE algorithm1 to two-dimensional surfaces represented

by triangle meshes in the STAPLE Surface (STAPLES) method. To characterize this approach, we develop a new, realistic model of "simulated rater" behavior and explore application of STAPLES in an atlas-based approach for determining gyral labels on brain cortical surfaces.

2. METHODS

STAPLE1 simultaneously estimates a true segmentation and a reliability characterization for each rater. It was originally presented on a voxel-wise basis for multiple-label, volumetric images, and included a Markov Random Field (MRF) to model spatially correlated structures. Here we modify the original approach to operate on surface labels. First, the update equations of the STAPLE algorithm are applied vertex-wise to the triangle mesh (Section 2.1). Second, a new mesh-based MRF accounts for spatial consistency with multiple labels (Section 2.2).

2.1 STAPLE on Triangle Meshes

In this formulation, labels exist on a surface parameterized as a triangle mesh with vertices $\mathbf{v}_i \in \Re^3$, $i \in \{1 \ldots N\}$ and labels edges $e_j \in (i, i')$, $j \in \{1 \ldots M\}$. A label is to be assigned each vertex, $T_i \in \{0, 1, \ldots L,\}$ given the labels from R raters $D_i = D_{ir}$, $r \in \{1 \ldots R\}$. As well, we seek to estimate rater performances, modeled as a set of confusion matrices $\Theta = \theta_{rs's}$, $s, s' \in \{1 \ldots L\}$.For any vertex v_i, $\theta_{rs's}$ denotes the probability that rater r assigns label s when s' is the true label.

The algorithm is initialized by setting the performance parameters equal to a matrix close to the identity for all raters as suggested in Warfield et al.1 At iteration k, the segmentation given the performance parameters is then estimated using the update equation

$$p\left(T_i=s|\boldsymbol{D}_i,\boldsymbol{\Theta}^{(k)}\right)=W_{si}^{(k)}=\frac{f\left(T_i=s\right)\prod_{r:D_{ir}=s'}\theta_{rs's}^{(k)}}{\sum_s f\left(T_i=s\right)\prod_{r:D_{ir}=s'}\theta_{rs's}^{(k)}}. \quad (1)$$

Next, the rater performance parameters for the $(k + 1)^{th}$ iteration are estimated with

$$\theta_{rs's}^{(k+1)}=\frac{\sum_{i:D_{ir}=s'}W_{si}^{(k)}}{\sum_i W_{si}^{(k)}}. \quad (2)$$

These equations are identical to the conventional STAPLE algorithm since at this level the specifics of the locations of the points on which labels are defined are irrelevant. The segmentation and performance estimation are iterated until a convergence criterion is met.

2.2 Markov Random Field

As with the original approach, we utilize a Markov Random Field (MRF) to model the spatial homogeneity of labels. However, a reformulation of the MRF model was required due to two major challenges of the mesh parameterization not present in the conventional

pixel or voxel image representation. First, vertices need not be equally spaced nor regularly connected as they are in the volumetric framework. A Gaussian kernel was applied to account for this, as the relative contribution to the local conditional probability from vertices within the clique are weighted by the distance of the clique vertex to the vertex itself. Second, cliques need not be of the same order (i.e. vertices are connected by different numbers of edges with different angular separation). Here we have addressed via simple normalization, but more sophisticated approaches will be explored in the future. Multiple labels are addressed by penalizing any "wrong" labels equally. The necessary changes were incorporated by defining the following MRF potential function

$$-E(\boldsymbol{T}) = \sum_{i \in V} W_i + \sum_i \sum_{\substack{i' \neq i \\ T_{i'} \neq T_i}} \beta_{ii'} e^{-\frac{d_{ij}^2}{d_0}} \quad (3)$$

where T is the full label configuration, d_{ij} is the distance between vertices i and j, d_0 is fixed and controls the size of the kernel. The interaction weight, $\beta_{ii'} = \beta/\mathrm{M}_i$ if i and i' are neighbors where β is fixed and M_i denotes the number of neighbors of vertex i. If vertices i and i' are not neighbors, $\beta_{ii'} = 0$. Optimization was achieved using an iterative conditional modes (ICM) scheme,3 which successfully converged for the presented problems.

3. DATA

3.1 Simulated Raters

Simulated rater errors were modeled in two ways. First, we used confusion matrix in which the i, j^{th} element indicates the probability that the rater will assign the j^{th} label when the i^{th} label is correct at any particular location. In this case, the identity matrix describes the "perfect" rater. In these simulations, confusion matrices were constructed such that each rater had equal expected performance for all labels, and errors were uniformly distributed among the remaining labels. Modeled errors are equally likely to occur throughout the image domain, i.e., every vertex is equally likely to be incorrectly labeled. An example of such a rater is shown in Figure 1(b).

Since in practice, raters are more likely to make errors near the object boundaries, we introduce a second rater error model that models this kind of behavior. Three vector parameters describe a rater's performance: r, l, and b. The scalar r is the rater's global true positive fraction. The vector l encodes the probability, given an error occurred, that it was at the i^{th} boundary. Finally the vector b describes the error bias at every boundary. The unbiased rater has $b_i = 0.5\ \forall i$, while values of 0.0 or 1.0 indicate that one label or the other is always favored when an error occurs. A labeling produced by this type of simulated rater is shown in Figure 1(c). This is visually similar to the types of errors commonly observed in automatic gyral labeling approaches.

4. RESULTS

4.1 Simulations

We first examined how the quality of the segmentation estimate produced with STAPLES varies with changes in the number of raters and their performance parameters. This was done by measuring the average label overlap (measured by the Dice Coefficient, defined as $DC = 2|A \cap B|/(|A| + |B|)$), between the true label configuration and estimated the true configuration. Figure 2 shows the results of 50 Monte Carlo iterations for the confusion matrix raters. We observe that the quality of the estimated true segmentation improves as the number of raters increases, and as the rater performance increases. A more detailed analysis will be provided in the full manuscript.

4.2 Multi-Atlas Parcellation

We also examined the applicability of STAPLES as part of a multi-atlas parcellation scheme in the spirit of Rohlfing et al.,2 using a leave-one-out experiment. Four subjects were drawn from the OASIS data7 (http://www.oasis-brains.org) and gyral labels were obtained using Freesurfer.5 These labels were transferred to cortical surfaces that were obtained using CRUISE.4 Three labeled surfaces were used as an atlas, and the remaining brain surface were considered as a target for the atlas-based parcellation. The parcellation was done by first partially inflating8 both the target and source surfaces. Next, the iterative closest point algorithm9 was used to register the target to the source using an affine tranformation. Labels were then transfered to vertices on the source surface from the nearest vertex on the target surface. This method was repeated to obtain a labelling of the source surface for each of the three atlases. These were combined using STAPLES to provide an estimate of the true labeling of the cortical gyri on the source surface, the results of which are shown in Figure 3. We see that a reasonable parcellation is obtained, despite the variability in the atlases and the simplicity of the registration and label transfer. Future work will include the incorporation of surface features such as curvature in this labelling scheme. Figure3(d) shows a quantitative, fair comparison of STAPLES against average individual atlas performance in our four test cases. Though not conclusive, these results suggest that STAPLES may improve the segmentation obtained relative to single-atlas methods. More thorough results with larger data pools will be provided in the full manuscript.

5. CONCLUSION

This work introduced STAPLES, an extension of STAPLE that enables statistical label recombination on 2D surfaces parameterized as a triangle meshes. This enables the improvement of labels on surfaces obtained from biomedical images, including surface representations of the cortex, cerebellum, and subcortical structures, for example. It also suggests a further generalization to arbitrary higher dimensional manifolds such as time series, tensor spaces and the like. This type of generalization is easily accomplished for a point-wise STAPLE approach, as demonstrated in this work. Generalizing the MRF potential for any manifold, however, presents a significant challenge, but would be extremely beneficial to research involving these types of high-dimensional data.

6. ACKNOWLEDGEMENTS

This work was supported in part by NIH/NINDS grants R01NS56307 and R01NS37747. The OASIS project is supported in part by grants P50AG05681, P01AG03991, R01AG021910, P50MH071616, U24RR021382, and R01MH56584.

REFERENCES

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

[2]. Rohlfing T, Russakoff DB, Maurer CR. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE TMI. 2004; 23(8):983–994.

[3]. Besag J. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society. 1986; 48:259–302.

[4]. Han X, Pham DL, Tosun D, Rettman ME, Xu C, Prince JL. CRUISE: Cortical Reconstruction Using Implicit Surface Evolution. Neuroimage. 2004; 23(3):997–1012. [PubMed: 15528100]

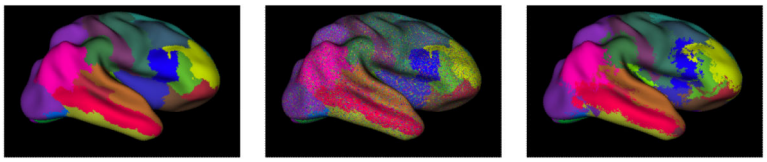
[5]. Desikan RS, Sgonne F, Fischl B, Albert MS, Killiany RJ, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage. 2006; 31(3):968–980. [PubMed: 16530430]

[6]. Makris M, Schlerf JE, Hodge SM, Schmahmann D, et al. MRI-based surface-assisted parcellation of human cerebellar cortex: An anatomically specified method with estimate of reliability. Neuroimage. 2005; 25(4):1146–1160. [PubMed: 15850732]

[7]. Marcus DS, Wang TH, Buckner RL, et al. Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience. 2004; 19:1498–1507. [PubMed: 17714011]

[8]. Tosun D, Rettmann M, Prince JL. Mapping techniques for aligning sulci across multiple brains. Medical Image Analysis. 2004; 8(3):295–309. [PubMed: 15450224]

[9]. Besl PP, McKay NN. A Method for Registration of 3-D Shapes. IEEE Trans. PAMI. 1992; 14:239–256.



(a) (b) (c)

Figure 1.

Examples of simulated cortical surface raters. (a) The true surface segmentation. Sample segmentations by a rater with true positive rates of 0.7 parameterized by (b) a confusion matrix and (c) boundary error rate.

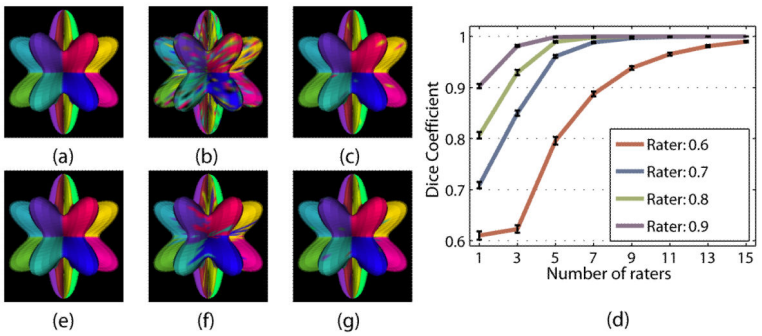


Figure 2.

Results for simulated surfaces: (a) The ground truth phantom parcellation, (b) simulated confusion matrix rater (true positive fraction = 0.7), (c) STAPLES result using 7 such raters without spatial correlation, (d) the results of 50 Monte Carlo iterations varying the expected rater performance and the number of raters, and (e) the STAPLES result using a MRF. Also shown are (f) a sample simulated boundary error rater (true positive fraction = 0.7), and (g) the STAPLES result using 7 such raters without spatial correlation.

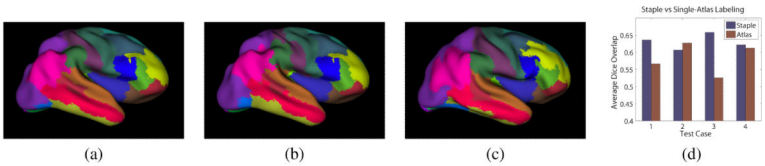


Figure 3.

An example of our method applied to a cortical gyral labeling task. Shown are (a) the true cortical parcellation, (b) the parcellation estimated from this multi-atlas / STAPLES approach, (c) one of the atlas subject surfaces, and (d) a comparison of STAPLES and single-atlas parcellation schemes.