iSTAPLE: Improved Label Fusion for Segmentation by Combining STAPLE with Image Intensity

Xiaofeng Liu*, Albert Montillo, Ek T. Tan and John F. Schenck

GE Global Research Center, One Research Circle, Niskayuna, NY, 12309

ABSTRACT

Multi-atlas based methods have been a trend for robust and automated image segmentation. In general these methods first transfer prior manual segmentations, i.e., label maps, on a set of atlases to a given target image through image registration. These multiple label maps are then fused together to produce segmentations of the target image through voting strategy or statistical fusing, e.g., STAPLE. STAPLE simultaneously estimates the true segmentation and the label map performance level, but has been shown inaccurate for multi-atlas segmentation because it is determined completely on the propagated label maps without considering the target image intensity. We develop a new method, called iSTAPLE, that combines target image intensity into a similar maximum likelihood estimate (MLE) framework as in STAPLE to take advantage of both intensity-based segmentation and statistical label fusion based on atlas consensus and performance level. The MLE framework is then solved using a modified EM algorithm to simultaneously estimate the intensity profiles of structures of interest as well as the true segmentation and atlas performance level. Unlike other methods, iSTAPLE does not require the target image to have same image contrast and intensity range as the atlas images, which greatly extends the use of atlases. Experiments on whole brain segmentation showed that iSTAPLE performed consistently better than STAPLE.

Keywords: Label fusion, brain segmentation, STAPLE

1. INTRODUCTION

Multi-atlas based methods have been the subject of considerable research for single and multi-structure segmentation of brain[1–3] as well as other organs.[4] These methods require a set of atlases, each of which includes a structural image and a corresponding manual segmentation of structures of interest. For a target image to be segmented, these methods typically first register all atlas images to the target image, propagate the manual segmentations, or labels, by applying the computed transform, and finally generate the segmentation results on the target image by fusing the propagated labels from all atlases. Two commonly used label fusion strategies are voting-based methods[5,6] and statistical fusion, e.g., STAPLE.[7]

STAPLE was originally developed for combining labels by human raters to compensate the fact that different raters have different labeling performance levels, and was recently applied to multi-atlas segmentation. Different from most voting-based methods, STAPLE does not assume the atlases perform equally well on the target image. Instead the atlas labeling performance levels for all the structures of interest are modeled and incorporated into a probabilistic framework which is solved for the true segmentation. This makes STAPLE more robust to anatomical variation between the atlas images and the target image and has been shown advantageous over majority voting.[7] However a disadvantage of STAPLE for multi-atlas segmentation is that STAPLE (as well as voting strategy) blindly fuses the labels without considering the target image intensity information, which makes it prone to error especially at the region boundaries.

Several efforts[8,9] have been made to integrate improve STAPLE by incorporating image intensity information into its probabilistic framework. They either use atlas images to refine the label maps before applying STAPLE,[8] or locally search for better matches at each pixel by comparing the target to all atlas images while executing STAPLE.[9] However, these methods generally require the atlas images have same contrast and intensity profiles as the target image. In reality that's not the case because target images are commonly acquired with different imaging parameters, and even atlas images can have different intensity profiles. While image intensity differences may not be a problem in the registration step before label fusion, they are not properly handled by these methods.

In this work, we develop a novel label fusion method, called iSTAPLE, which extends STAPLE by incorporating the target intensity image into conventional STAPLE framework. The intuition is illustrated in Fig. 1 with a simple example,

*Send correspondence to Xiaofeng Liu. E-mail: xiaofeng.liu@ge.com. Telephone: (518)387-5028.

Medical Imaging 2013: Image Processing, edited by Sebastien Ourselin, David R. Haynor,
Proc. of SPIE Vol. 8669, 86692O · © 2013 SPIE · CCC code: 1605-7422/13/$18
doi: 10.1117/12.2006447

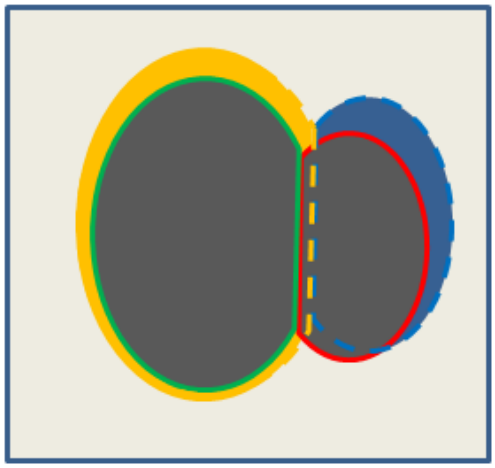


Figure 1. An example that illustrates the intuition of iSTAPLE.

where two touching structures with similar intensity (shown in gray) are present in the background. Because of imperfect registration, the transformed label maps do not completely align with the actual structures (shown in orange and blue). It is observed the mis-labeled regions around the boundaries with background have different intensity values and can be corrected based on image intensity, while the boundary between the two regions can be determined based on atlas consensus and performances using label fusion, i.e., STAPLE. Based on this observation, we develop the iSTAPLE method that takes advantage of both intensity-based segmentation and STAPLE label fusion. Moreover, iSTAPLE integrates into the MLE framework the intensity information solely based on the target image, which negates the need of using atlas images in label fusion and allows for applying multi-atlas segmentation methods to images with different modalities from the atlas images. Experiments on whole brain segmentation have shown that iSTAPLE is more robust and produces better results.

2. METHODS

2.1 Notations and Conventional STAPLE

Consider a target image $\mathbf{I}$ that is to be segmented, where I_i is the image intensity at voxel i for $i = 1, 2, ..., N$ with N being the number of voxels in $\mathbf{I}$. Let L be the number of structures or labels, and R be the number of atlases. The labels on the atlases are propagated to the target image domain after image registration and denoted as a $N \times R$ matrix $\mathbf{D}$, which describes the label decision from each atlas at each voxel. Let $\mathbf{t}_i$ be an indicator vector with a length of L representing the true segmentation at i. If i belongs to the k^{th} structure, the k^{th} element of $\mathbf{t}_i$, i.e., t_{ik}, equals to 1 and all other elements are 0. Let $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^{N}$ be the set of true labels on all voxels.

The goal of STAPLE[7] is to estimate the true segmentation $\mathbf{T}$ of $\mathbf{I}$ as well as performance parameters $\boldsymbol{\theta}$. As compared to most voting based methods,[5,6] STAPLE takes into account of the anatomical variation such that the atlases are considered to perform differently for segmenting the target image, and their performance is also different on different structures. This is modeled using performance level parameters θ_j, an $L \times L$ matrix of parameters where its elements models the probability that atlas j will assign label s' while the true label is s. Let $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_R\}$. $\boldsymbol{\theta}$ is unknown and is estimated by maximizing the likelihood function while solving for the true segmentation, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta}) \,. \tag{1}$$

This is solved iteratively using an Expectation-Maximization (EM) algorithm. Let $\boldsymbol{\theta}^{(t)}$ be the estimated performance parameters at iteration t. The expected value of the log likelihood function is given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\left[\log f(\mathbf{D}, \mathbf{T}|\boldsymbol{\theta})|\mathbf{D}, \boldsymbol{\theta}^{(t)})\right] \\ &= E\left[\log f(\mathbf{D}|\mathbf{T}, \boldsymbol{\theta}) f(\mathbf{T})|\mathbf{D}, \boldsymbol{\theta}^{(t)})\right] \\ &= \sum_{\mathbf{T}} \log f(\mathbf{D}|\mathbf{T}, \boldsymbol{\theta}) f(\mathbf{T}) f(\mathbf{T}|\mathbf{D}, \boldsymbol{\theta}^{(t)}) \end{aligned} \tag{2}$$

During the E-step, the conditional distribution of $\mathbf{T}$ at voxel i is computed and refered as the weight variable. I.e.,

$$\begin{aligned} W_{si}^{(t)} &= f(T_i = s|\mathbf{D}_i, \boldsymbol{\theta}^{(t)}) \\ &= \frac{f(T_i = s)\prod_j f(D_{ij}|T_i = s, \theta_j^{(t)})}{\sum_{s'} f(T_i = s')\prod_j f(D_{ij}|T_i = s', \theta_j^{(t)})} \end{aligned} \tag{3}$$

where $W_{si}^{(t)}$ is the probability that the true label at voxel i is s.

During the M-step, the performance parameters are estimated as

$$\theta_{js's}^{(t+1)} = \frac{\sum_{i:D_{ij}=s'} W_{si}^{(t)}}{\sum_i W_{si}^{(t)}}\,. \tag{4}$$

The two steps are then iterated until the convergence is reached. Refer to Warfield et al.[7] for detailed descriptions of the STAPLE algorithm.

2.2 The iSTAPLE Method

STAPLE fuses labels based on the propagated atlas labels without considering the target image. Therefore when the target image exhibits large anatomical variation from the atlas images, the registration step may consistently fail on certain structures and STAPLE will not work. In addition STAPLE may be less accurate along structure boundaries. Drawing on the intuition shown in Fig. 1, we want to improve the segmentation results along tissue boundaries and compensate for anatomical variation using the structure appearance information in the target image. Thus we develop the *iSTAPLE* method by extending STAPLE, which takes into account the target intensity image $\mathbf{I}$ and incorporating it into a similar probabilistic framework as STAPLE. The framework is then solved in a similar EM algorithm for multi-structure segmentation.

Assuming $\mathbf{I}$ is independent to atlas labels $\mathbf{D}$ and performance parameters $\boldsymbol{\theta}$, the log likelihood function for iSTAPLE can be expressed as

$$\log f(\mathbf{D}, \mathbf{T}, \mathbf{I}|\boldsymbol{\theta}) = \log f(\mathbf{D}|\mathbf{T}, \mathbf{I}, \boldsymbol{\theta}) f(\mathbf{I}|\mathbf{T}) f(\mathbf{T})\,. \tag{5}$$

The conditional expectation function for iSTAPLE at iteration t is then

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\left[\log f(\mathbf{D}, \mathbf{T}, \mathbf{I}|\boldsymbol{\theta})|\mathbf{D}, \mathbf{I}, \boldsymbol{\theta}^{(t)})\right] \\ &= \sum_{\mathbf{T}} \log[f(\mathbf{D}|\mathbf{T}, \mathbf{I}, \boldsymbol{\theta}) f(\mathbf{I}|\mathbf{T}) f(\mathbf{T})] f(\mathbf{T}|\mathbf{D}, \mathbf{I}, \boldsymbol{\theta}^{(t)}) \end{aligned} \tag{6}$$

Using the fact that the propagated atlas labels $\mathbf{D}$ is independent to the target image $\mathbf{I}$, the conditional probability can then be written as

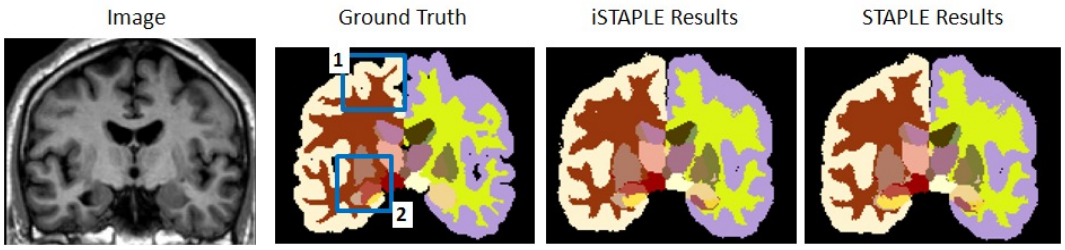
$$f(\mathbf{T}|\mathbf{D}, \mathbf{I}, \boldsymbol{\theta}^{(t)}) = \frac{f(\mathbf{D}|\mathbf{T}, \boldsymbol{\theta}^{(t)}) f(\mathbf{I}|\mathbf{T}) f(\mathbf{T})}{\sum_{\mathbf{T}'} f(\mathbf{D}|\mathbf{T}', \boldsymbol{\theta}^{(t)}) f(\mathbf{I}|\mathbf{T}') f(\mathbf{T}')} \tag{7}$$

In the E-step, the weight function at voxel i for iSTAPLE can be written as

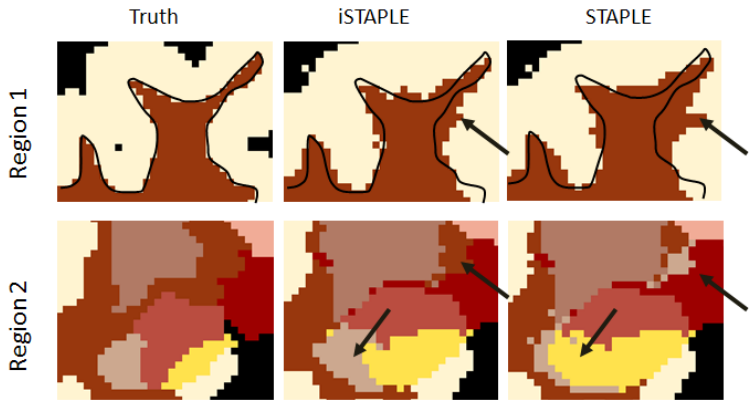
$$W_{si}^{(t)} = \frac{f(T_i = s) f(I_i|T_i = s)\prod_j f(D_{ij}|T_i = s, \theta_j^{(t)})}{\sum_{s'} f(T_i = s') f(I_i|T_i = s')\prod_j f(D_{ij}|T_i = s', \theta_j^{(t)})} \tag{8}$$

The difference between Eqn. (8) and Eqn. (3) is that Eqn. (8) includes a new term $f(I_i|T_i = s)$, which models the probability that a voxel that belongs to the s^{th} structure has an intensity of I_i. This enables iSTAPLE to take advantage of appearance differences of different structures and results in more accurate segmentation along structure boundaries and in cases of large anatomical variation. For neighboring structures with similar intensity distributions, $W_{si}^{(t)}$ is largely determined by the atlas consensus and performance parameters, and thus works similarly as in conventional STAPLE. Here the intensity distribution $f(I_i|T_i = s)$ is modeled using Gaussian functions. I.e.,

$$f(I_i|T_i = s) = \frac{1}{\sqrt{2\pi\sigma_s^{2(t)}}} e^{-\frac{(I_i - \mu_s^{(t)})^2}{2\sigma_s^{2(t)}}}\,, \tag{9}$$



(a)



(b)

Figure 2. Results on whole-brain segmentation on one IBSR data set. (a) From left to right: one coronal slice of the T1 weighted image, the ground truth, the segmentation results of iSTAPLE, and the segmentation results of STAPLE. (b) The ground truth, iSTAPLE results, and STAPLE results on two zoomed regions that are labeled in (a). For region 1, the white matter contour was drawn on the ground truth and overlayed on iSTAPLE and STAPLE results to show the differences. Improved segmentation and boundary delineation were seen in the majority of structures (arrows).

where $\mu_s^{(t)}$ and $\sigma_s^{2(t)}$ are the mean and variance of the target image intensities for structure s, respectively.

In the M-step, the parameters, $\boldsymbol{\theta}^{(t)}$, $\mu_s^{(t)}$, and $\sigma_s^{2(t)}$, are computed by maximizing the conditional expectation function shown in Eqn. (6). $\boldsymbol{\theta}^{(t)}$ is estimated in a similar way as in Eqn. (4). $\mu_s^{(t)}$ and $\sigma_s^{2(t)}$ are computed by

$$\left(\mu_s^{(t)}, \sigma_s^{2(t)}\right) = \arg\max_{\mu_s,\sigma_s} \sum_i W_{si}^{(t)} \log f(I_i|T_i = s)\,. \tag{10}$$

We find

$$\mu_s^{(t)} = \frac{\sum_i W_{si}^{(t)} I_i}{\sum_i W_{si}^{(t)}}\,, \qquad \sigma_s^{2(t)} = \frac{\sum_i W_{si}^{(t)} (I_i - \mu_s^{(t)})^2}{\sum_i W_{si}^{(t)}} \tag{11}$$

In summary, the iSTAPLE algorithm is:

1. Set $k = 0$. Initialize $\boldsymbol{\theta}^{(0)}$. Initialize $W_{si}^{(0)}$ as in (8) by assuming $f(I_i|T_i = s) = 1$.
2. Compute $\boldsymbol{\theta}^{(k+1)}$ using Eqn. (4).
3. Compute $\mu_s^{(k+1)}$ and $\sigma_s^{(k+1)}$ using Eqn. (11)
4. Compute $W_{si}^{(t)}$ using Eqn. (8)
5. Iterate steps 2-4 until the algorithm converges or reaches certain number of iterations.

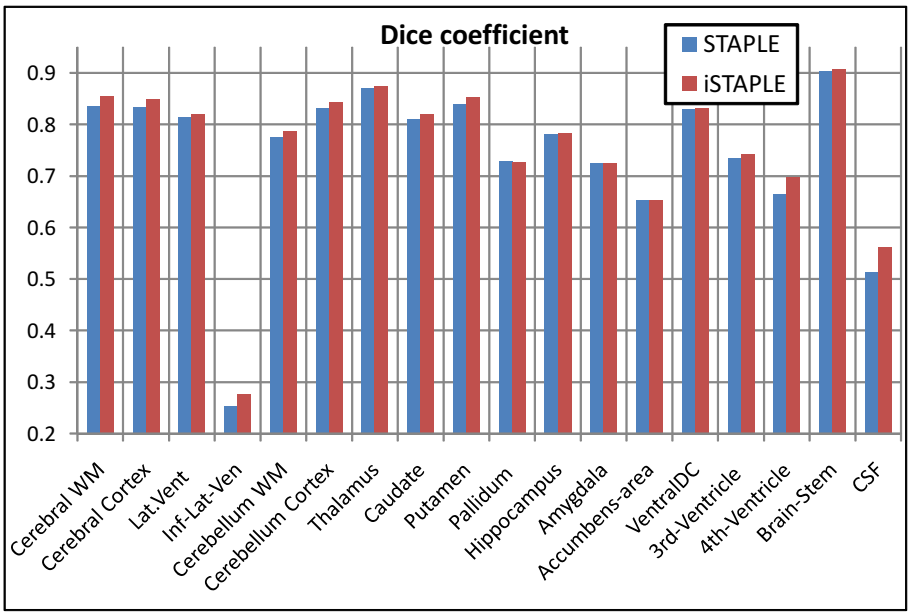


Figure 3. The mean Dice coefficients of STAPLE and iSTAPLE methods on different brain structures.

3. EXPERIMENTS AND RESULTS

We tested iSTAPLE method on whole brain segmentation using Internet Brain Segmentation Repository (IBSR)* atlases. IBSR contains 18 healthy subjects with T1 weighted images, and 32 brain structures were manually delineated on each image by experts and served as ground truth. Leave-one-out experiments were performed for cross-validation. i.e., in each experiment, one image was selected as the target image, and the rest 17 data were used as the atlases in the multi-atlas segmentation.

For each experiment, the atlas images were registered to the target image using Symmetric Diffeomorphic Normalization (SyN) method in ANTS,[10] and the labels in the atlases were propagated to the target image domain. After that the target image was segmented through label fusion with both iSTAPLE and conventional STAPLE methods. The segmentation results were then compared to the ground truth using the Dice coefficient, i.e., $D = \frac{2|X \cap Y}{X \cup Y}$ where X is the voxel set of ground truth, Y is the voxel set of the segmentation result, and $|\cdot|$ is the set cardinality.

Fig. 2 shows the segmentation results on one dataset. Visually, it was clear that iSTAPLE provided improved segmentation of the boundaries between adjacent structures as compared to conventional STAPLE. The mean Dice coefficients for the 18 experiments on the 32 brain structures are shown in Fig. 3 for quantitative comparison. Here the results on same structure at the left and right sides are shown together. Overall, iSTAPLE outperformed STAPLE methods especially on structures whose intensity distributions are different from their neighboring structures, e.g., ventricles and cortex. For subcortical structures (e.g., thalamus, caudate, puttaman, hippocampus, and amygdala) iSTAPLE performed slightly better. This is because their intensity distributions are close to their neighboring structures and thus intensity information for these structures is less effective.

4. DISCUSSIONS

In this paper we develop a new label fusion method, called iSTAPLE, that extends the STAPLE method by incorporating the target intensity image into the statistical formulation of STAPLE. By considering the different appearances of different structures as well as taking advantage of statistical label fusion based on atlas consensus and performance level, iSTAPLE improves the label fusion results especially for structures with different appearance as their neighboring structures. Experiments were performed on the brain structure segmentation on 18 IBSR data sets and the results showed that iSTAPLE consistently outperformed the STAPLE method.

*http://www.cma.mgh.harvard.edu/ibsr/

It can be seen from Fig. 2 that both iSTAPLE and STAPLE did not perform well on certain regions of the brain, e.g., white matter, mainly because image registration did not perform consistently on these regions for different atlas images, and the intensity weighting in iSTAPLE was not able to correct. In the future we are going incorporate other image cues including boundaries and textures to further improve the method. In our method the structure intensity distributions are modeled using Gaussian. Other methods, for example Parzen window,[11] may be used to model the distributions more accurately. This is our future work.

REFERENCES

1. R. Heckemann, J. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage* **33**(1), pp. 115–126, 2006.
2. P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *Neuroimage* **46**(3), pp. 726–738, 2009.
3. D. Collins and J. Pruessner, "Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting animal with a template library and label fusion," *NeuroImage* **52**(4), pp. 1355–1366, 2010.
4. I. Isgum, M. Staring, A. Rutten, M. Prokop, M. Viergever, and B. van Ginneken, "Multi-atlas-based segmentation with local decision fusion: Application to cardiac and aortic segmentation in ct scans," *Medical Imaging, IEEE Transactions on* **28**(7), pp. 1000–1010, 2009.
5. R. Heckemann, J. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage* **33**(1), pp. 115–126, 2006.
6. X. Artaechevarria, A. Munoz-Barrutia, and C. O. de Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Trans. Med. Imag.* **28**(8), pp. 1266–1277, 2009.
7. S. Warfield, K. Zou, and W. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *Medical Imaging, IEEE Transactions on* **23**(7), pp. 903–921, 2004.
8. N. Weisenfeld and S. Warfield, "Learning likelihoods for labeling (L3): A general multi-classifier segmentation algorithm," in *MICCAI 2011, Part III, LNCS 6893*, pp. 322–329, 2011.
9. A. Asman and B. Landman, "Non-local staple: An intensity-driven multi-atlas rater model," in *MICCAI 2012, Part III, LNCS7512*, pp. 139–146, 2012.
10. B. Avants, C. Epstein, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Imag. Anal.* **12**, pp. 26–41, 2008.
11. E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics* **33**, pp. 1065–1076, 1962.