SIMPLE Is a Good Idea (and Better with Context Learning)

Zhoubing Xu[1], Andrew J. Asman[1],
Peter L. Shanahan[2], Richard G. Abramson[2], and Bennett A. Landman[1, 2]

[1] Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235
{zhoubing.xu,andrew.j.asman,bennett.landman}@vanderbilt.edu
[2] Radiology and Radiological Science, Vanderbilt University, Nashville, TN, USA 37235
{peter.l.shanahan,richard.abramson}@vanderbilt.edu

Abstract. Selective and iterative method for performance level estimation (SIMPLE) is a multi-atlas segmentation technique that integrates atlas selection and label fusion that has proven effective for radiotherapy planning. Herein, we revisit atlas selection and fusion techniques in the context of segmenting the spleen in metastatic liver cancer patients with possible splenomegaly using clinically acquired computed tomography (CT). We re-derive the SIMPLE algorithm in the context of the statistical literature, and show that the atlas selection criteria rest on newly presented principled likelihood models. We show that SIMPLE performance can be improved by accounting for exogenous information through Bayesian priors (so called context learning). These innovations are integrated with the joint label fusion approach to reduce the impact of correlated errors among selected atlases. In a study of 65 subjects, the spleen was segmented with median Dice similarity coefficient of 0.93 and a mean surface distance error of 2.2 mm.

Keywords: Selective and Iterative Method for Performance Level Estimation (SIMPLE), Context Learning, Multi-Atlas Segmentation, Abdomen.

1 Introduction

Multi-atlas segmentation is a technique for transferring information from canonical atlases to target images via registration. While this family of techniques has proven effective in neuroimaging [2], the importance of atlas selection has become increasingly clear for variable anatomical targets (e.g., the prostate [3]).

Here, we revisit atlas selection and fusion techniques in the context of segmenting the spleen in metastatic liver cancer patients with possible splenomegaly using clinically acquired computed tomography (CT). Abdominal anatomy is variable both between individuals (e.g., weight, stature, age, disease status) and within individuals (e.g., pose, respiratory cycle, clothing). Splenomegaly exacerbates inter-individual spleen variability (and can result in a ≈10 fold increase in spleen volume over normal individuals) and complicates inter-subject registration (Figure 1). Note that in this situation a majority of atlases tend to be poorly registered, but a subset (shown for

P. Golland et al. (Eds.): MICCAI 2014, Part I, LNCS 8673, pp. 364–371, 2014.

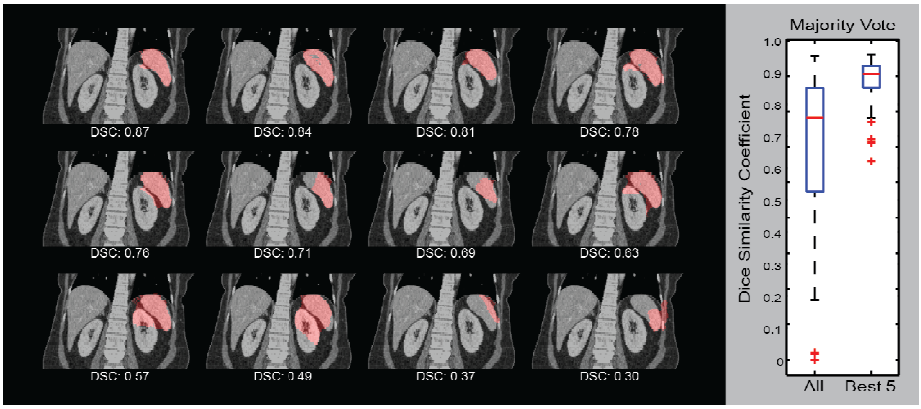

Fig. 1. (left) 12 registered (using [1]) spleen labels are overlaid on a slice of a target image. DSC values are computed on a volumetric basis as examples. (right) Using leave-one-out validation, multi-atlas segmentation is tested on 65 subjects for spleen segmentation. For each subject, majority vote is used to fuse either all registered atlases or the best five atlases in terms of DSC.

five best matches) could result in excellent segmentation *if* they could be identified without using the true labels to compute Dice similarity coefficient (DSC) [4] (as in the Figure 1 illustration).

Two broad families of techniques have emerged to address these issues: voting (well synthesized in [2]) and statistical fusion (largely following [5, 6]). The selective and iterative method for performance level estimation (SIMPLE) [3] algorithm identifies and combines a set of registered images by selecting self-consistent sets of atlases based on DSC. SIMPLE addressed extensive variation in prostate anatomy and was presented from the perspective of voting using sensible, but *ad hoc,* criteria to reduce the impact of outlier atlases. Recently, joint label fusion (JLF) [7] has provided a technique to reduce the impact of correlated errors among atlases. SIMPLE and JLF effectively address different aspects of the fusion problem: atlas selection and label determination. Hitherto, they have not been combined into a common framework.

The primary contributions of this manuscript are as follows (Figure 2). (1) We present SIMPLE in the context of the statistical literature and show that "simple" atlas selection criteria rest on principled likelihood models. (2) We generalized the SIMPLE theoretical framework to account for exogenous information (e.g., from separate models of tissue likelihood) – referred to as *context learning*. (3) We integrate context learning and SIMPLE with JLF. (4) Finally, we combine these contributions to segment the spleen on metastatic liver cancer patients.

2 Theory

First, we consider SIMPLE from the perspective of Expectation-Maximization (EM) while focusing on the atlas selection step. Consider the hidden true segmentation as a vector, $\boldsymbol{T} \in \boldsymbol{L}^{N\times 1}$, where $\boldsymbol{L} = \{0, \ldots, L-1\}$ is the set of possible label types.

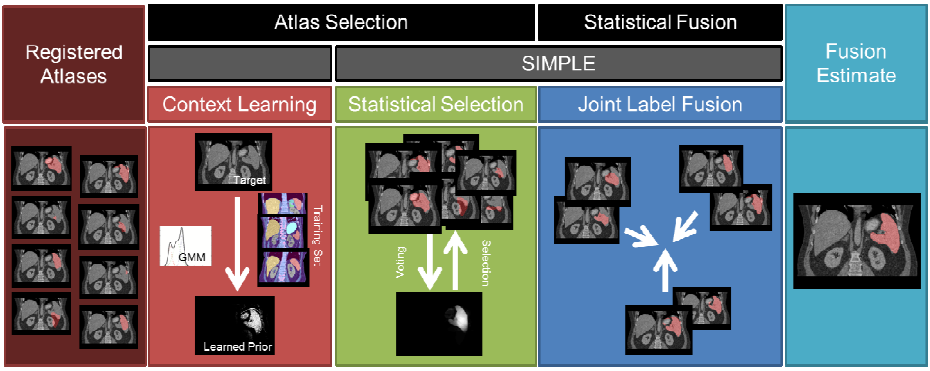

Fig. 2. Flowchart of the proposed method. Given registered atlases with variable qualities, atlas selection and statistical fusion are considered as two necessary steps to obtain a reasonable fusion estimate of the target segmentation. The SIMPLE algorithm implicitly combines these two steps to fusion selected atlases; however, more information can be incorporated to improve the atlas segmentation, and a more advanced fusion technique can be used after the atlases are selected. We propose to (1) extract a probabilistic prior of the target segmentation by context learning to regularize the atlas selection in SIMPLE and (2) use Joint Label Fusion to obtain the final segmentation while characterizing the correlated errors from among atlases.

Consider a collection of R registered atlases with label decisions, $\boldsymbol{D} \in \boldsymbol{L}^{N\times R}$. Let $\boldsymbol{c} \in \boldsymbol{S}^{R}$ represent the atlas selection decisions, where $\boldsymbol{S} = \{0,1\}$ indicates that the associated atlas is ignored or selected, respectively. We propose a non-linear rater model, $\theta \in \mathbb{R}^{R\times 2\times L\times L}$, where each elemen $\theta_{jns's}$ represents the probability that the registered atlas j observes label s' given the true label is s and the atlas selection decision is n. Let the ignored atlases be no better than random chance, and the selected atlases be slightly inaccurate with error factors $\boldsymbol{\epsilon} \in \mathbf{E}^{\mathrm{R}\times 1}$, where $\mathbf{E} \in (0, \frac{L-1}{L})$. Thus

$$\theta_{j0s's} = \frac{1}{L}, \quad \forall s' \,;\, \theta_{j1s's} = \begin{cases} 1-\epsilon_j, & s' = s \\ \dfrac{\epsilon_j}{L-1}, & s' \neq s \end{cases} \tag{1}$$

For a registered atlas j at each voxel i, the generative model is

$$f\left(D_{ij} = s' \middle| T_i = s, c_j = n, \epsilon_j\right) \equiv \theta_{jns's} \tag{2}$$

Following [5, 6], let $\boldsymbol{W} \in \mathbb{R}^{L\times N}$, where W_{si} represents the probability that the true label associated with voxel i is label s. Using Bayesian expansion and conditional independence between the atlases, the k^{th} iteration of $\boldsymbol{W}$ is

$$W_{si}^{(k)} = \frac{f(T_i = s)\prod_j f\left(D_{ij} \middle| T_i = s, c_j^{(k)} = n, \epsilon_j^{(k)}\right)}{\sum_{s'} f(T_i = s')\prod_j f\left(D_{ij} \middle| T_i = s', c_j^{(k)} = n, \epsilon_j^{(k)}\right)} \tag{3}$$

where $f(T_i = s)$ is a voxel-wise *a priori* distribution of the underlying segmentation. Note that the ignored atlases do not affect $\boldsymbol{W}$, and the selected atlases contribute to $\boldsymbol{W}$ in a similar way as majority vote (MV) given the symmetric form of $\theta_{1s's}$.

The estimate of the parameters (M-Step) is obtained by maximizing the expected value of the conditional log likelihood function found in Eq. 3. For the error factor,

$$\begin{aligned}\epsilon_j^{(k+1)} &= \arg\max_{\epsilon_j} \sum_i E\left[\ln f\left(D_{ij}\middle|T_i, c_j^{(k)}, \epsilon_j\right)|\boldsymbol{D}, c_j{}^{(k)}, \epsilon_j^{(k)}\right] \\ &= \arg\max_{\epsilon_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_s W_{si}^{(k)} \ln \theta_{jc_j^{(k)}s's} \equiv L_{\epsilon_j}\end{aligned} \tag{4}$$

Consider the binary segmentation for simplicity, let $M_{TP} = \sum_{i:D_{ij}=1} W_{1i}^{(k)}$, $M_{FP} = \sum_{i:D_{ij}=1} W_{0i}^{(k)}$, $M_{FN} = \sum_{i:D_{ij}=0} W_{1i}^{(k)}$, $M_{TN} = \sum_{i:D_{ij}=0} W_{0i}^{(k)}$, and $\mathrm{M}_T = M_{TP} + M_{TN}$, $M_F = M_{FP} + M_{FN}$. After taking partial derivative of L_{ϵ_j},

$$\epsilon_j^{(k+1)} = \frac{M_F}{M_T + M_F}, i.e., 1 - \epsilon_j^{(k+1)} = \frac{M_T}{M_T + M_F} \tag{5}$$

Then for the atlas selection decision

$$\begin{aligned}c_j^{(k+1)} &= \arg\max_{c_j} \sum_i E\left[\ln f\left(D_{ij}\middle|T_i, c_j, \epsilon_j^{(k+1)}\right)|\boldsymbol{D}, c_j{}^{(k)}, \epsilon_j^{(k+1)}\right] \\ &= \arg\max_{c_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_s W_{si}^{(k)} \ln \theta_{jc_j s's}\,.\end{aligned} \tag{6}$$

In general, this is a combinatoric problem; however, assuming known true labels, it can be maximized separately for each 0/1 atlas selection decision. Noting the behavior of selecting/ignoring atlases in Eq. 6 is parameterized with the error factor ϵ_j, and thus, as in Eq. 5, affected by the four summed values of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).

In SIMPLE, atlases are selected based on DSC with the previous majority vote estimate. Above, Eq. 3 reduces to a majority vote of atlases with $c_j^{(k)} = 1$ and the relative weight of atlases is scaled by Eq. 5, which differs from DSC in that DSC does not factor the impacts of TN. Typical practice for a fusion approach might use the prior probability, $f(T_i = s)$, to weight by expected volume of structure. With outlier atlases, one could reasonably expect a much larger region of confusion (i.e., non "consensus"[8]) than true anatomical volume. Hence, an informed prior would greatly deemphasize the TN and yield a metric similar to DSC. Therefore, we argue that SIMPLE is legitimately viewed as a statistical fusion algorithm that is approximately optimal for the non-linear rater model proposed in Eq. 1.

2.1 Context Learning (CL)

Different classes of tissues in CT images can be characterized with multi-dimensional Gaussian mixture models using intensity and spatial "context" features. On a voxel-wise basis, let $\boldsymbol{v} \in \mathbb{R}^{d\times 1}$represent a d dimensional feature vector, $m \in \boldsymbol{M}$ indicate the tissue membership, where $\boldsymbol{M} = \{1, \ldots, M\}$ is the set of possible tissues. The probability of the observed features given the tissue is m can be represented with the mixture of N_G Gaussian distributions,

$$f(\boldsymbol{v}|m=t)=\sum_{k=1}^{N_G}\frac{\alpha_{kt}}{(2\pi)^{\frac{d}{2}}|\boldsymbol{C}_{kt}|^{\frac{1}{2}}}exp\left[-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{\mu}_{kt})^T\boldsymbol{C}_{kt}^{-1}(\boldsymbol{v}-\boldsymbol{\mu}_{kt})\right] \quad (7)$$

where $\alpha_{kt}\in\mathbb{R}^{1\times1}$, $\boldsymbol{\mu}_{kt}\in\mathbb{R}^{d\times1}$, and $\boldsymbol{C}_{kt}\in\mathbb{R}^{d\times d}$ are the unknown mixture probability, mean, and covariance matrix to estimate for each Gaussian mixture component k of each tissue type t by the EM algorithm following [9].

The context model can be learned from datasets with known tissue separations, and then the tissue likelihoods on unknown dataset can be inferred by Bayesian expansion and flat tissue membership probability from extracted feature vectors.

$$f(m=t|\boldsymbol{v})=\frac{f(\boldsymbol{v}|m=t)f(m=t)}{\sum_{t'}f(\boldsymbol{v}|m=t')f(m=t')}=\frac{f(\boldsymbol{v}|m=t)}{\sum_{t'}f(\boldsymbol{v}|m=t')} \quad (8)$$

3 Methods and Results

Under an Institutional Review Board (IRB) waiver, the first-session CT abdomen scans of 65 metastatic liver cancer patients were randomly selected from an ongoing colorectal cancer chemotherapy trial. Images are with variable field of views (approx. 300 x 300 x 400 mm ~ 500 x 500 x 700 mm) and resolutions (approx. 0.5 x 0.5 x 1.5 mm ~ 1.0 x 1.0 x 7.0 mm). Spleens were manually labeled by an experienced graduate student on a volumetric basis using the MIPAV software (NIH, Bethesda, MD [10]). All images and labels are cropped along the cranio-caudal axis with a tight border without excluding liver, spleen, and kidneys before any processing (following [1]).

We used 12 of the 65 subjects as training datasets for learning context models for eight tissue types, including five manually traced organs (i.e., spleen, liver, kidneys, pancreas, and stomach) and three automatically retrieved tissues (i.e., muscle, fat, and other) using intensity clustering and excluding the traced organ regions. These 12 datasets were not considered for quantitative evaluation in the leave-one-out analyses. Six context features were extracted, including intensity, gradient, and local variance, and three spatial coordinates with respect to a single landmark, which was loosely identified as the mid-frontal point of the lung at the plane with the largest cross-sectional lung area. We specified the number of components of Gaussian mixture model, $N_G=3$. The spleen and non-spleen likelihoods on each target image were inferred, and used as a two-fold spatial prior to regularize the SIMPLE atlas selection, referred as CL+SIMPLE. We constrained the number of selected atlases as no less than five and no larger than ten. When using JLF, we specified the local search radii as $3\times3\times3$, the local patch radii as $2\times2\times2$, and set the intensity difference mapping parameter (β), and the regularization term (α) as 2 and 0.1, respectively (i.e., default parameters). We appended Markov Random Field (MRF) for smoothing the Gaussian filtered ($\sigma_G=1$) result of CL+SIMPLE+JLF with the smoothness parameter as 0.2, and the incompatibility parameter as -5.

3.1 Motivating Simulation

To demonstrate and motivate the benefits of the CL+SIMPLE approach, a simulation was constructed using a single CT slice from a representative subject with the spleen manually labeled (see Figure 3). Eighty simulated observations were estimated by

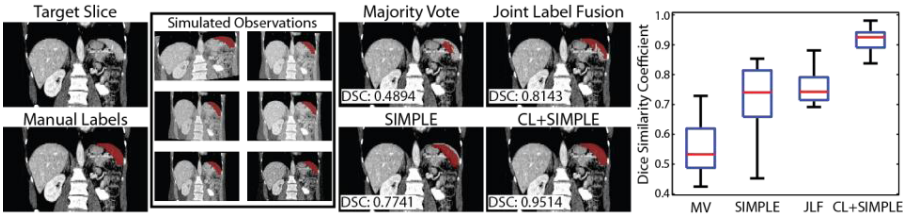

Fig. 3. The results of the motivating simulation demonstrate the benefits of the CL+SIMPLE approach. Using a model in which simulations are drawn from a randomly generated affine transformation, CL+SIMPLE more accurately estimates the location, size, and orientation of the spleen, and significantly outperforms the considered benchmarks.

applying a random five degree-of-freedom affine transformation to the target slice. Each transformation consisted of a rotational component as well as two translational and two scaling components, with the effect of each component drawn from a zero-mean Gaussian distribution with standard deviations of 2 degrees for the rotational component, 1 mm for the translational components and 0.1 mm for the scaling components. A representative fusion result is shown for MV, JLF, SIMPLE, and CL-SIMPLE, with CL+SIMPLE resulting in an estimate that substantially more accurately represents the shape, location, and orientation of the spleen. With 20 Monte Carlo iterations, the spread of DSC values demonstrate significant improvement exhibited by CL+SIMPLE, with a median DSC improvement of approximately 0.15 over SIMPLE and JLF, and approximately 0.4 over majority vote.

3.2 Volumetric Spleen Multi-atlas Segmentation

We performed leave-one-out cross validation for the multi-atlas spleen segmentation for 53 scans (excluding the 12 subjects used for context learning). For each scan, all other scans (including the training dataset) were considered as atlases (hence, 64 atlases), and aligned to the target with a multi-stage registration, in the order of rigid, affine and a multi-level non-rigid registration using free-form deformations with B-spline control point spacings of 20, 10, and 5mm [11]. We tested on seven label fusion methods (as listed in Table 1). The performances of the methods were evaluated on DSC, symmetric mean surface distance (Sym. MSD), and symmetric Hausdorff distance (Sym. HD).

Combined with CL, CL+SIMPLE and CL+SIMPLE+JLF improved mean DSC by at least 0.03 over SIMPLE and SIMPLE+JLF, respectively, while the direct integration of SIMPLE with JLF does not provide higher accuracy over JLF. CL+SIMPLE+JLF outperforms the other methods with the higher median DSC, narrower range of DSC, and lower MSD. An extra MRF step effectively removes outlier speckled structures in the segmentation (Figure 4), and thus further reduces the surface distance errors (Table 1).

4 Discussion

We reformulated SIMPLE as a statistical fusion algorithm that is approximately optimal for a newly presented non-linear rater model tailored for heterogeneous atlases. Revealed in the generalized SIMPLE theoretical framework, we find

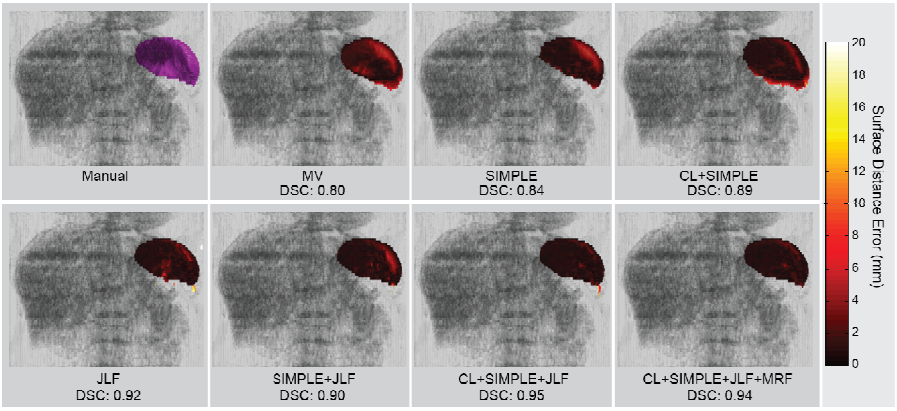

Fig. 4. Representative qualitative results. The subject above was selected to fall within 30th and 70th percentile in terms of DSC for each tested method. The 3D rendered segmentation of each method is colored in terms of the voxel-wise surface distance from the estimated segmentation to the manual segmentation.

Table 1. Quantitative metrics for tested methods on 53 of 65 subjects

Metrics / Methods	Dice Similarity Coefficient		Surface Distance (mm)	
	Median [Min, Max]	Mean ± Std	Sym. MSD	Sym. HD
MV	0.74 [0, 0.94]	0.65 ± 0.25	6.28 ± 4.81	25.89 ± 11.95
SIMPLE	*0.84 [0.24, 0.95]	0.78 ± 0.16	*4.27 ± 3.87	*21.71 ± 11.03
CL+SIMPLE	*0.87 [0.44, 0.95]	0.83 ± 0.11	*3.35 ± 2.78	*19.91 ± 9.96
JLF	*0.9 [0.4, 0.98]	0.86 ± 0.13	4.45 ± 7.67	*67.07 ± 41.75
SIMPLE+JLF	0.91 [0.33, 0.97]	0.85 ± 0.14	*4.04 ± 7.85	*28.17 ± 26.4
CL+SIMPLE+JLF	***0.93 [0.54, 0.97]**	**0.89 ± 0.1**	***2.2 ± 2.27**	**22.16 ± 9.73**
CL+SIMPLE+JLF+MRF	***0.93 [0.54, 0.97]**	**0.88 ± 0.1**	**2.17 ± 2.33**	***18.46 ± 9.85**

* indicates that the DSC, Sym. MSD, or Sym. HD values are statistically different than the corresponding value in the row immediately above (e.g, SIMPLE vs. MV) as determined by a Wilcoxon signed rank test ($p<0.05$).

adaptation of the spatial priors to be critical for contexts with large numbers outliers. Using exogenous information, these can be estimated separately with the proposed context learning. In a study of 65 liver cancer patients with potential splenomegaly using clinically acquired CT, we combined context learning and SIMPLE with JLF to address the problems of atlas selection and label determination in multi-atlas segmentation, achieving a mean DSC of 0.89 with a range of 0.54 to 0.97 (evaluated on the 53 that were not used for context learning). In a study of 150 gastric / cholecystitis / colorectal cancer patients, Wolz *et al.* applied multi-scale non-rigid registration followed by non-local fusion and graph-cut post processing to achieve a mean DSC of 0.92 with a range of 0.26 to 0.98 [1]. The body shape variations of their patient population (acquired in Japan) is expected to be relatively smaller comparing to ours (acquired in U.S.), while all their scans were acquired in portal venous phase as opposed to ours with variable phases. With larger variations on our clinically data, our method yields a slightly lower DSC, but substantially narrower range. The performance of the Wolz approach using our dataset was disappointing (mean DSC of

0.73); however, we considered the direct comparison unfair since our atlas structure, i.e., single organ on 65 datasets, did not support some innovative aspects of the Wolz approach (re-weighting atlases based on different organs). Lastly, we note that proposed generative model naturally leads to an iterative atlas selection, which differs from the STEPS approach [12] that first locally ranks atlases, and uses the top local atlases for statistical fusion. In the further study, a systematic integration between CL+SIMPLE and JLF theories could yield atlas selection and label determination.

Acknowledgements. This research was supported by NIH 1R03EB012461, NIH 2R01EB006136, NIH R01EB006193, ViSE/VICTR VR3029, NIH UL1 RR024975-01, and NIH UL1 TR000445-06. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D.: Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Transactions on Medical Imaging 32, 1723–1730 (2013)
2. Sabuncu, M.R., Yeo, B.T., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging 29, 1714–1729 (2010)
3. Langerak, T.R., van der Heide, U.A., Kotte, A.N., Viergever, M.A., van Vulpen, M., Pluim, J.P.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Transactions on Medical Imaging 29, 2000–2008 (2010)
4. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26, 297–302 (1945)
5. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23, 903–921 (2004)
6. Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr., C.R.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21, 1428–1442 (2004)
7. Wang, H., Suh, J.W., Das, S.R., Pluta, J., Craige, C., Yushkevich, P.A.: Multi-Atlas Segmentation with Joint Label Fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012)
8. Asman, A.J., Landman, B.A.: Robust statistical label fusion through COnsensus Level, Labeler Accuracy, and Truth Estimation (COLLATE). IEEE Transactions on Medical Imaging 30, 1779–1794 (2011)
9. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18, 885–896 (1999)
10. McAuliffe, M.J., Lalonde, F.M., McGarry, D., Gandler, W., Csaky, K., Trus, B.L.: Medical image processing, analysis and visualization in clinical research. In: Proceedings of the 14th IEEE Symposium on Computer-Based Medical Systems, pp. 381–386. IEEE (2001)
11. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast MR images. IEEE Trans. Med. Imaging 18, 712–721 (1999)
12. Jorge Cardoso, M., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S.: STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Med. Image Anal. 17, 671–684 (2013)