Fusion of Multi-Atlas Segmentations with Spatial Distribution Modeling

Subrahmanyam Gorthi[1], Meritxell Bach Cuadra[2,1], Ulrike Schick[3], Pierre-Alain Tercier[4], Abdelkarim S. Allal[4], and Jean-Philippe Thiran[1]

[1] Signal Processing Laboratory(LTS5), Ecole Polytechnique Fédérale de Lausanne(EPFL), Switzerland, subrahmanyam.gorthi@epfl.ch,
[2] Department of Radiology, University Hospital Center and University of Lausanne,
[3] Department of Radiation Oncology, University Hospital of Geneva, Switzerland,
[4] Service de Radio-oncologie, Hôpital Fribourgeois, Fribourg, Switzerland. *

Abstract. In recent years, multi-atlas fusion methods have gained significant attention in medical image segmentation. In this paper, we propose a general Markov Random Field (MRF) based framework that can perform edge-preserving smoothing of the labels at the time of fusing the labels itself. More specifically, we formulate the label fusion problem with MRF-based neighborhood priors, as an energy minimization problem containing a unary data term and a pairwise smoothness term. We present how the existing fusion methods like majority voting, global weighted voting and local weighted voting methods can be reframed to profit from the proposed framework, for generating more accurate segmentations as well as more contiguous segmentations by getting rid of holes and islands. The proposed framework is evaluated for segmenting lymph nodes in 3D head and neck CT images. A comparison of various fusion algorithms is also presented.

1 Introduction

Atlas-based image segmentation is a key area of research in medical imaging [1]. It has been shown in many recent works [2,3,4] that automated segmentations obtained by merging results from multiple atlases are more reliable and accurate than the results obtained from a single atlas. However, one of the main problems with many of the existing fusion strategies is, although the segmentations obtained from each individual atlas are contiguous, the merged segmentations can be fragmented ones with undesirable discontinuities including holes and islands [3,5].

To deal with the above mentioned problems, sometimes, the segmentation results are post-processed. For instance, in [3], the segmentation results for each structure are independently considered as binary masks, and are post-processed

* This work is supported in part by the Swiss National Science Foundation under Grant 205321-124797 and by CIBM of the Geneva–Lausanne Universities and EPFL, as well as the foundations Leenaards and Louis-Jeantet.

by first smoothing them with a Gaussian kernel, and then thresholded at 0.5; after that, they further perform connected component analysis and retain only the largest component. In [6], the segmentation results are post-processed by morphological closing, and then followed by the extraction of the largest component.

However, such postprocessing approaches have many disadvantages. First, they do not preserve edges. Second, such simple Gaussian smoothing of labels results in conflicting regions across the boundaries of adjacent structures, depending on the order in which those labels are smoothed. To avoid this conflict between regions, one could probably introduce approaches like, a more complex iterative coupled Gaussian smoothing of multiple labels, but that still will not solve the first problem of preserving the edges. Finally, it is not elegant to handle "fusion" and "smoothing" as two different, independent problems.

To address the above mentioned issues, we propose here a general MRF-based framework that simultaneously performs both fusion and edge-preserving smoothing of multiple labels. The rest of the paper is organized as follows. In the next section, we propose a general framework and reframe some of the existing fusion methods to fit into our framework. In section 3, we present an evaluation of various fusion methods that can be derived from the proposed framework. The discussion is presented in section 4, followed by the conclusions in section 5.

2 MRF-based Fusion Model

Let V be the number of voxels in the image. Let Y_p denote the label assigned to the p^{th} voxel in the output image. Let Y be the set containing labels assigned to each voxel in the output image, i.e., $Y = \{Y_1, \cdots Y_V\}$. Then, we will be formulating the atlas fusion as a general energy minimization problem of the form:

$$Y^* = \arg\min_{Y} \left\{ E_{\text{data}}(Y) + \lambda\, E_{\text{smooth}}(Y) \right\}, \tag{1}$$

where the first term is a data term (unary term), and it should be defined in such a way that it reaches to a minimum value when the chosen fusion criteria has been met; the second term is a smoothness term (pairwise term), and in the current context, it should penalize for irregular distribution of labels while allowing for the edge-discontinuities. λ is a weighting parameter between the data term and smoothness term. Energy equation of the form (1) is ubiquitous in many computer vision problems, and there exists various efficient MRF optimization methods for solving them [7]. In this paper, we use the graph cuts expansion method [8] as it guarantees convergence to a global optimum for the current model.

Note that while the above energy formulation is commonly encountered in other computer vision problems like image segmentation, denoising and stereo vision [7], such model has not been used so far, for the atlas fusion problem. We now reformulate some of the existing fusion methods so that they fit into the data term (E_{data}) of the above energy minimization problem.

(a) Majority Voting (MV) [2,4]: This is the most simple fusion method. It assigns for each voxel a label that maximum number of atlases agree on. Let N be the number of atlas images. Let X^j represent j^{th} input labeled image (corresponding to j^{th} atlas) after applying the transformation that maps the j^{th} atlas to the output intensity image. Let X_p^j be the label assigned to the p^{th} voxel of X^j. Now, it is easy to note that the original maximum-energy formulation of MV in [4] can be reframed into the following equivalent minimum-energy formulation:

$$E_{\text{data}}(Y) = \frac{1}{N} \sum_{p=1}^{V} \sum_{j=1}^{N} \left(1 - \delta\left(X_p^j, Y_p\right)\right),$$

where δ is a Kronecker delta function. The reason to include a factor of $1/N$ in the above data term is to make its magnitude independent of number of atlases when smoothness term is also used with it.

(b) Global Weighted Voting (GWV) [4]: Unlike MV, GWV attaches a weight to each atlas while counting its vote. The weight for each atlas is determined *globally*, based on its similarity to the image to be segmented: more the similarity, higher the weight, and vice versa. We again reframe the maximum-energy formulation of GWV in [4] to the following equivalent minimum energy formulation:

$$E_{\text{data}}(Y) = \frac{1}{N} \sum_{p=1}^{V} \sum_{j=1}^{N} \hat{w}^j \left(1 - \delta\left(X_p^j, Y_p\right)\right),$$

where $\hat{w}^j$ the normalized weight assigned to j^{th} atlas image. In this work, we use inverse of the mean square difference (MSD) between the atlas and output intensity images as the similarity metric for computing weights. However, extension to other similarity metrics (like mutual information or normalized cross-correlation) is straight forward.

(c) Local Weighted Voting (LWV) [3,4]: LWV is similar to GWV except that, not a single global weight is assigned to the entire atlas; rather, for each voxel, an individual weight is assigned based on *local* similarity. Let $\hat{w}_p^j$ be the normalized weight assigned to p^{th} voxel, in the j^{th} atlas. Then, the minimum-energy formulation for LWV is given by:

$$E_{\text{data}}(Y) = \frac{1}{N} \sum_{p=1}^{V} \sum_{j=1}^{N} \hat{w}_p^j \left(1 - \delta\left(X_p^j, Y_p\right)\right).$$

In the current application, $\hat{w}_p^j$ at each voxel is computed based on MSD metric computed over a local neighborhood of that voxel. It is interesting to note that the above formulations of the data term have an equivalent representations in a probabilistic framework as well [9]. Similar to [9], by introducing an additional parameter, the above three energy terms can also be seen as special cases of a more general data term.

Regarding the smoothness term, we use here the widely used edge-preserving Potts model [7]. However, one could even use models that are specific to a given

application, that incorporate prior knowledge about the spatial distribution of the labels. Let $\aleph_p$ be the set of all voxels in the predefined neighborhood of p^{th} voxel. Then the Potts model-based smoothness term is given by:

$$E_{\text{smooth}}(Y) = \sum_{p=1}^{V} \sum_{\forall q \in \aleph_p} w_{pq} \ (1 - \delta(Y_p, Y_q)) \, .$$

3 Results

3.1 Dataset and Methods

The evaluation of the proposed fusion framework is performed on a dataset of 3D head and neck (H&N) CT images, for segmenting lymph nodes. Lymph nodes are constructed volumes in the H&N region and they do not have a clear contrast with the neighboring structures, thus, making their segmentation a challenging task [10]. In clinical practice, accurate delineation of lymph nodes is essential for precise radiotherapy treatment of H&N cancer. In this work, we consider 10 lymph node volumes for automated segmentation, and are: (i) IB-Left, (ii) IB-Right, (iii) IIA-Left, (iv) IIA-Right, (v) IIB-Left, (vi) IIB-Right, (vii) III-Left, (viii) III-Right, (ix) IV-Left, (x) IV-Right. These lymph node volumes for one of the patients are shown in Fig. 1

The current dataset contains 12 atlas images and 8 patients' images to be segmented. These images typically have a resolution of $1mm \times 1mm \times 1mm$ in x, y and z directions respectively. An expert oncologist has manually delineated lymph nodes on all the images, and they are considered as the ground truth segmentations.

Regarding the registration, all the 12 atlases are registered to each patient to be segmented. An initial affine registration is performed followed by a two-level hierarchical nonrigid registration. In the first level, a region-based registration, driven by selected structures (bones, trachea and external-contour of H&N) having clear boundaries, is performed; this is followed by a second level of pixel-based nonrigid registration. Since the main focus of this work is fusion of multi-atlas segmentation results, which is independent of the registration algorithm, we skip more details on registration and refer the readers to [10].

We mainly evaluate atlas fusion results from 8 methods. 4 of them are the state-of-the-art methods that do not contain any smoothness term (i.e., $\lambda = 0$), and are: (i) majority voting (MV), (ii) global weighted voting (GWV), (iii) local weighted voting (LWV) with the weights computed over a neighborhood of $3 \times 3 \times 3$ voxels (LWV_1), (iv) LWV with the weights computed over $9 \times 9 \times 9$ neighborhood. The reason to consider above two versions of LWV is also to study the effect of neighborhood size on the segmentation of H&N lymph nodes. The other 4 algorithms are based on the MRF-based atlas fusion framework proposed here, and are referred as: (v) MV+MRF, (vi) GWV+MRF, (vii) LWV_1+MRF, and (viii) LWV_2+MRF.

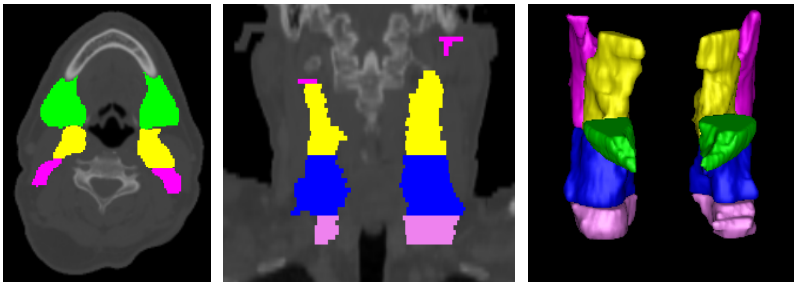


Fig. 1: Ground truth segmentations of H&N lymph nodes for one of the images: Segmentations in an axial and coronal slice are respectively shown in the first two images; the lymph node volumes are shown in the last image. Lymph nodes IB, IIA, IIB, III and IV are respectively shown in green, yellow, magenta, blue and pink.

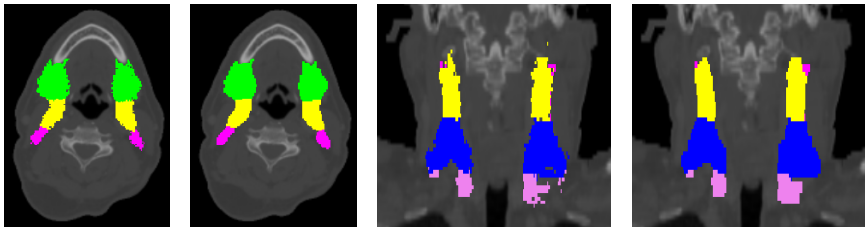
Another widely used atlas fusion algorithm is STAPLE (simultaneous truth and performance level estimation) [11], which was originally proposed for combining manual segmentations done by multiple experts. We have not yet explored reformulating the STAPLE as an energy minimization problem that fits into the current framework; however, we present here some statistical results obtained using STAPLE, in order to notice it's relative performance in the current context. In particular, we use here the multi-label implementation of the STAPLE proposed in [12].

In all the experiments, λ value is set empirically to 0.5, and has not been optimized anymore; however, this is suffice to demonstrate the advantages of the proposed framework. Finally, in the smoothness term, $\aleph_p$ is set to the standard 3D grid of 6 neighbors, and w_{pq} is set to $1/\text{card}(\aleph_p)$.

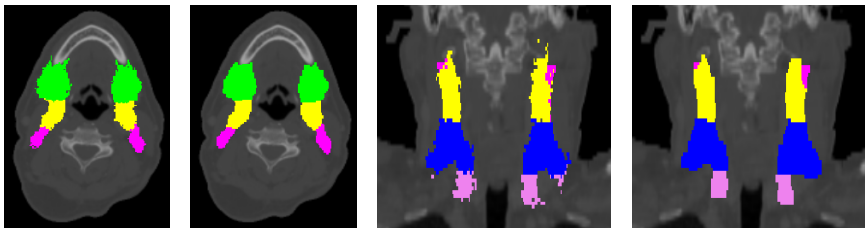
3.2 Qualitative and Quantitative Results

We present here qualitative results for one of the images. Fig. 1 shows ground truth segmentations of lymph nodes. It shows the segmentations in an axial and coronal slice, as well as the 3D volumes of the lymph nodes. Fig. 2 shows the automated segmentation results of the same, obtained from each method; the odd numbered columns of this figure show results for the 4 existing methods, while the even columns show results for their counterparts that include an MRF-based smoothing term in their fusion. The lymph node volumes obtained from these methods are shown in Fig. 3. By visual comparison of these results with the ground truth in Fig. 1, it can be noted that the proposed MRF-based smoothing can provide more accurate segmentations, and also results in getting rid of unwanted holes and islands in the output segmentations.

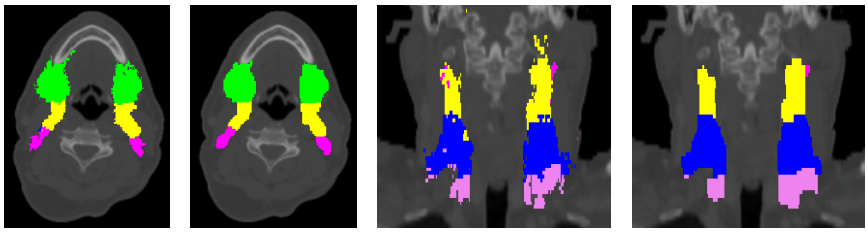
The quantitative evaluation is performed over the entire dataset, using two metrics; (i) *Dice similarity metric (DSM):* This is a commonly used metric that computes the measure of overlap (in %) between the ground truth segmentations



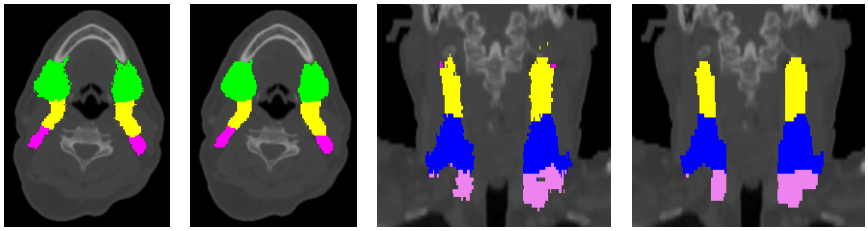
(a) MV (b) MV+MRF (c) MV (d) MV+MRF



(e) GWV (f) GWV+MRF (g) GWV (h) GWV+MRF

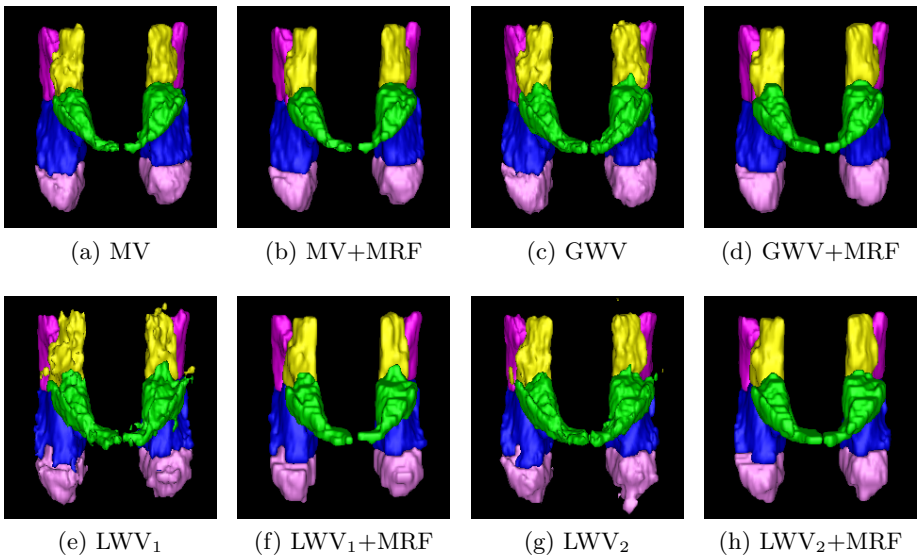


(i) LWV_1 (j) LWV_1+MRF (k) LWV_1 (l) LWV_1+MRF



(m) LWV_2 (n) LWV_2+MRF (o) LWV_2 (p) LWV_2+MRF

Fig. 2: Qualitative comparison of lymph nodes segmentation results for one of the images: Results in an axial and a coronal slice are presented in the first and the last two columns respectively. The ground truth segmentations for the above slices are shown in Fig. 1.



(a) MV (b) MV+MRF (c) GWV (d) GWV+MRF

(e) LWV_1 (f) LWV_1+MRF (g) LWV_2 (h) LWV_2+MRF

Fig. 3: Qualitative comparison of lymph node volumes obtained from different methods. The ground truth volume for the above is shown in Fig. 1.

and automated segmentations. (ii) *Number of connected regions per label:* The output segmentations of each lymph node (label) should ideally contain a single contiguous region. Hence, we evaluate the fusion algorithms also based on the number of connected regions it has created per each label (ideal value = 1); we take into account both islands and holes for computing the number of connected regions.

Table 1 presents the mean and standard deviation values of the DSM obtained from all methods for each lymph node. As mentioned earlier, just to note the relative performance of multi-label STAPLE, we presented the mean DSM values obtained from it, in the same Table. We notice that other fusion methods performed better than STAPLE in the current context; however, we did not perform any further quantitative analysis of STAPLE, as that is not in the current scope of the paper. The corresponding box plots of DSM are presented in Fig. 4. For the box plots, because of the space limitations, statistics for the left and right structures of the lymph nodes of the same number are combined together as they are approximately symmetric. Finally, Table 2 summarizes for each method, the mean and standard deviations of number of connected regions per label.

We further evaluated the statistical significance of the improvements in the segmentation results with the inclusion of MRF-based smoothness term. Wilcoxon signed-rank test [13] is used for this purpose. Notice that Wilcoxon signed-rank test can be seen as non-parametric alternative to paired-student test since it

does not make any assumptions regarding the distributions of the data population. We performed this test on DSM statistics, for each pair of methods (i.e., without and with the inclusion of MRF-based smoothness term), with the alternative hypothesis being: "Segmentation results with the inclusion of MRF-based smoothness term are statistically better (greater) than the original methods that do not use this term." From these experiments, it is found (at 0.05 significance level) that in all cases, the improvements in the segmentation results due to the inclusion of MRF-based smoothness term are statistically significant compared to the original methods.

Table 1: Mean and standard deviation values of the dice similarity metric (DSM) in %, for the H&N lymph nodes segmentation, obtained from different fusion methods. The best mean DSM values for each lymph node are shown in bold.

	Fusion Method								
	STAPLE	MV	MV +MRF	GWV	GWV +MRF	LWV_1	LWV_1 +MRF	LWV_2	LWV_2 +MRF
IB-L	61.21± 11.9	64.83± 9.2	65.10± 9.9	64.20± 7.3	65.33± 8.6	63.34± 7.4	64.93± 8.8	64.74± 7.9	**65.34± 9.1**
IB-R	58.77± 9.6	64.42± 8.2	65.15± 8.4	66.56± 6.3	67.50± 6.1	65.02± 4.9	66.58± 5.6	67.01± 5.1	**67.74± 5.2**
IIA-L	57.35± 14.4	61.50± 9.3	62.25± 9.3	61.81± 7.5	62.93± 7.8	64.36± 7.8	**67.23± 8.7**	65.66± 7.4	67.11± 7.3
IIA-R	62.91± 7.1	66.11± 6.4	67.17± 6.3	66.88± 4.4	68.23± 4.9	67.59± 4.3	69.84± 4.7	68.74± 4.4	**69.92± 5.0**
IIB-L	49.96± 14.5	55.79± 13.5	56.56± 14.4	58.57± 13.2	58.77± 13.3	65.02± 9.3	61.13± 18.4	**65.82± 7.2**	64.86± 7.7
IIB-R	52.16± 12.8	60.01± 14.0	61.07± 14.1	63.42± 12.5	63.86± 12.5	67.21± 9.2	67.09± 10.9	67.28± 10.2	**67.48± 10.8**
III-L	66.55± 10.3	69.74± 7.1	70.64± 7.0	70.10± 6.5	71.30± 6.6	69.96± 5.5	72.29± 6.8	72.10± 5.7	**73.39± 6.0**
III-R	65.90± 4.8	67.46± 4.4	67.73± 5.3	68.51± 5.0	69.16± 5.2	69.03± 3.9	71.19± 4.6	70.12± 4.2	**71.30± 4.3**
IV-L	56.86± 8.2	62.41± 6.1	63.23± 5.9	61.54± 6.1	62.31± 6.8	61.84± 4.8	64.15± 4.9	63.05± 5.0	**64.44± 5.9**
IV-R	57.06± 8.4	61.15± 5.9	61.12± 6.5	62.27± 6.6	63.41± 6.2	61.36± 6.3	63.92± 6.6	62.53± 6.3	**64.14± 6.6**
Avg.	58.87± 10.2	63.34± 8.4	64.00± 8.7	64.39± 7.5	65.28± 7.8	65.47± 6.3	66.84± 8.0	66.70± 6.3	**67.57± 6.8**

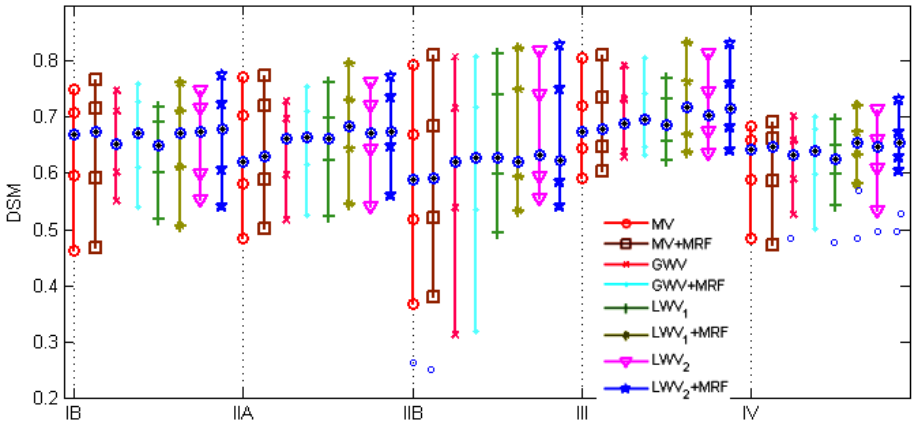


Fig. 4: Box plots of dice similarity metric (DSM) for the lymph node segmentation results obtained from all the 8 methods. Statistics for the left and right structures of the same lymph node are combined together (for example, IB-L & IB-R are combined to IB).

Table 2: Mean and standard deviations of average number of connected regions per label (considering both islands and holes), obtained from each fusion method.

MV	MV +MRF	GWV	GWV +MRF	LWV_1	LWV_1 +MRF	LWV_2	LWV_2 +MRF
16.89±3.8	1.06±0.1	21.38±3.3	1.01±0.0	46.71±8.8	1.05±0.1	24.66±4.5	1.00±0.0

4 Discussion

The following observations can be made from the above results. Methods with the proposed MRF-based edge-preserving smoothness priors in their fusion model resulted in more accurate segmentations than their counterparts that do not incorporate any smoothing priors. MRF-based atlas fusion methods, unlike their counterparts, resulted in segmentation with contiguous regions. In terms of DSM, local weight based methods are the best ones followed by global weight based and majority voting schemes. Among the two versions of the local weight based methods, the second one with the larger neighborhood (9×9×9) gave better results. From the point of view of number of connected regions, all the MRF-based methods have successfully generated segmentations with contiguous regions; on the other hand, among their counterparts, LWV_1 has produced a highly fragmented output followed by LWV_2, GWV and MV respectively.

It can be noted from Table 1 that mean values of DSM obtained from methods that use MRF-based smoothness term are clearly better than their counterparts without smoothness priors. Although a slight increase in the standard deviations is observed with the inclusion of MRF-based smoothness priors, notice that the

relative magnitude of the the improvements achieved in the mean DSM values are more than the increase in the standard deviations; it implies that the overall segmentation results are better with the MRF-based methods. That slight increase in the standard deviations can be due to the following reason: The improvements contributed by MRF-based model will be more when the segmentations obtained with the original methods (i.e., without MRF-based smoothness) are not contiguous; but for certain images, when the segmentation results with the original methods are more contiguous than for other images, the added improvements due to MRF-models are not as much as for those other images, and thereby slightly increasing the standard deviations. But, as we mentioned before, methods with MRF-based smoothness term always resulted in better segmentations, and significance of these improvements is also ascertained by the statistical test results.

Regarding the weighting parameter (λ) in eq. 1, in the current study, we have empirically set its value to 0.5. The approaches for the automatic selection of λ for this atlas fusion problem could be very similar to those used with other MRF-based energy minimization problems encountered in computer vision [7]. We would like to additionally mention here an observation specific to the current context: For a chosen value of λ, looking at the resulting number of connected regions per label could provide some insights about the λ value; when the number of regions is more than the expected value (based on the prior knowledge about that structure), it could potentially indicate to choose a larger value of λ and vice versa.

5 Conclusions

In this paper, we have proposed a general MRF-based edge-preserving fusion framework for merging segmentation results obtained from multiple atlases. Many of the existing atlas fusion methods can be reframed to profit from the proposed framework. We have demonstrated how the majority voting, global weighted voting and local weighted voting fusion methods can be fitted into this framework. We have compared the segmentation results obtained from the above methods versus the methods that additionally use the MRF-based edge-preserving smoothness term at the time of fusion. The results from the MRF-based models are found to be more accurate besides providing segmentations with contiguous regions. We have also performed a comparison of 8 fusion methods that can be derived from this framework, for segmenting 10 lymph node structures in the H&N CT images. Among all the methods, local weighted voting with MRF-based smoothness term has provided the best segmentation results. We would like to further extend this framework to other fusion methods like shape-based averaging [5], and also evaluate this framework on other important applications in medical imaging.

References

1. Rohlfing, T., Brandt, R., Menzel, R., Russakoff, D.B., Maurer, Jr., C.R.: Quo vadis, atlas-based segmentation? In: The Handbook of Medical Image Analysis – Volume III. Kluwer Academic (2005) 435–486
2. 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**(4) (2004) 1428–1442
3. Ĭsgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., van Ginneken, B.: Multi-atlas-based segmentation with local decision fusion - application to cardiac and aortic segmentation in CT scans. IEEE Transactions on Medical Imaging **28**(7) (2009) 1000–1010
4. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C.: Combination strategies in multi-atlas image segmentation: Application to brain mr data. IEEE Transactions on Medical Imaging **28**(8) (2009) 1266–1277
5. Rohlfing, T., Maurer, C.: Shape-based averaging. IEEE Transactions on Image Processing **16**(1) (2007) 153 –161
6. Ramus, L., Malandain, G.: Multi-atlas based segmentation: Application to the head and neck region for radiotherapy planning. In: MICCAI Workshop Medical Image Analysis for the Clinic - A Grand Challenge. (2010)
7. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for markov random fields with smoothness-based priors. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(6) (2008) 1068–1080
8. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11) (2001) 1222–1239
9. Sabuncu, M., Yeo, B., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging **29**(10) (2010) 1714–1729
10. Gorthi, S., Duay, V., Houhou, N., Bach Cuadra, M., Schick, U., Becker, M., Allal, A.S., Thiran, J.P.: Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. IEEE Journal on Selected Topics in Signal Processing **3**(1) (2009) 135–147
11. 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**(7) (2004) 903–921
12. Rohlfing, T., Russakoff, D.B., Maurer, Jr., C.R.: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Transactions on Medical Imaging **23**(8) (2004) 983–994
13. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6) (1945) 80–83