EVALUATION OF ATLAS FUSION STRATEGIES FOR SEGMENTATION OF HEAD AND NECK LYMPH NODES FOR RADIOTHERAPY PLANNING

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

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

ABSTRACT

Accurate segmentation of lymph nodes in head and neck (H&N) CT images is essential for the radiotherapy planning of the H&N cancer. Atlas-based segmentation methods are widely used for the automated segmentation of such structures. Multi-atlas approaches are proven to be more accurate and robust than using a single atlas. We have recently proposed 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. There are three main contributions of this paper: First, we reformulate the "shape based averaging" (SBA) fusion method to fit into the general MRF-based fusion framework. Second, we evaluate the following fusion algorithms for the segmentation of H&N lymph nodes: (i) STAPLE, (ii) SBA, (iii) SBA+MRF, (iv) majority voting (MV), (v) MV+MRF, (vi) global weighted voting (GWV), (vii) GWV+MRF, (viii) local weighted voting (LWV) and (ix) LWV+MRF. Finally, we also study the effect varying the number of atlases on the performance of the above algorithms.

Index Terms— Atlas-based segmentation, label fusion, lymph nodes, radiotherapy, MRF, IMRT.

1. INTRODUCTION

Intensity-modulated radiotherapy (IMRT) is a high precision technique used for radiation treatment of different tumor locations of the patients. This requires accurate delineation of various structures on 3-D CT images. One of the main challenges for the widespread implementation of IMRT for head and neck (H&N) cancer is, obtaining automated and accurate segmentation of lymph nodes.

Lymph nodes are constructed volumes in H&N CT images; they have poor contrast in the CT images, and do not posses distinctly visible boundaries with the surrounding structures [1, 2]. Hence, automated segmentation of lymph nodes is a challenging task and, incorporation of some prior knowledge is required for accurately delineating them. Atlas-based segmentation methods are widely used for exploiting the prior anatomical knowledge [3].

It is well known that the segmentations obtained by the appropriate fusion of results obtained from multiple atlases are more accurate and robust than the results from a single atlas [4, 5, 6, 7]. The widely used fusion methods include STAPLE [8, 9], majority voting (MV) [4, 5], global weighted voting (GWV) [5, 6], local weighted voting (LWV) [5, 6, 7], and shape based averaging (SBA) [10]. One of the issues with many of the fusion methods is that, although the segmentation results from each individual atlas are contiguous, the merged segmentation results can contain unwanted discontinuities like holes and islands.

In order to deal with such discontinuities in the segmentations, the final results are generally post-processed using different approaches like Gaussian smoothing [7], and morphological operations [11]. However, such post-processing approaches have many disadvantages. In particular, they handle the fusion and smoothing as two different, independent problems, and do not preserve the edges of the labels. To address this issue, a Markov random field (MRF)-based framework has been proposed recently in [12]. It combines the tasks of fusion and smoothing, and performs edge-preserving smoothing at the time of fusing the labels itself. It is also shown in [12] that, MV, GWV, LWV methods can be reformulated to profit from this MRF-based framework.

In this paper, we further adapt the MRF-based general framework to SBA method. We evaluate all the above mentioned methods in the specific context of H&N lymph nodes segmentation. We also study the effect of varying number of atlases on these methods. The rest of the paper is organized as follows. In the next Section, we present the general MRF-based framework and the reformulation of SBA method. A detailed evaluation for lymph nodes segmentation is presented in Section 3. The discussion and conclusions are presented in Section 4.

2. FUSION METHODS

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, the atlas fusion is formulated 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 will 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.

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.

The reformulation of MV, GWV and LWV methods to fit into the above MRF-based framework has been presented in [12]; hence, we skip those details here, and refer the readers to [12]. We now show how the SBA method can be reformulated, and, this is one of the contributions of this paper.

Shape Based Averaging (SBA) [10]: When compared to other approaches like MV, GWV, LWV methods, SBA looks at the fusion problem from a different perspective. For each voxel in the output image, SBA assigns a label that results in minimum "signed Euclidean distance" (SED) when summed up over all the input label images; note that the signed Euclidean distances are computed with respect to each label. Please refer to [10] for a nice intuitive illustration of this approach, and we now present its mathematical formulation.

Let N be the number of atlas images. Let X^j represent the 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 $u_p^j(l)$ represent the SED for label:l, at p^{th} voxel in X^j. Now, for each voxel p, SBA assigns independently, that label which results in the minimum value of the following summation:

$$Y_p = \arg\min_l \frac{1}{N}\sum_{j=1}^{N} u_p^j(l).$$

Unlike other fusion methods, the label selection in SBA involves computation of distance metric, and these distances can be negative value also. But, for the convergence of the MRF-based model to a global optimum, all the energy terms should be nonnegative. For this purpose, we modify the original $u_p^j(l)$ to the following:

$$\hat{u}_p^j(l) = \begin{cases} u_p^j(l) + u_{th}, & \text{if } -u_{th} \le u_p^j \le u_{th};\\ 0, & \text{if } u_p^j < -u_{th};\\ 2u_{th}, & \text{if } u_p^j > u_{th}, \end{cases}$$

where u_{th} (>0) is a threshold applied to $u_p^j(l)$. Notice that, the above reformulation is equivalent to first thresholding $u_p^j(l)$ to the range: $[-u_{th}, u_{th}]$, and then, adding an offset value of u_{th} and thereby, modifying the range to: $[0, 2u_{th}]$. The thresholding is done for making the algorithm sensitive to even small changes in the SED, and the offset value is added to make all SED values nonnegative. Adding offset value alone without thresholding makes the algorithm insensitive to fine-changes in the SED. Note that, when u_{th} is large enough compared to the SEDs for adjacent labels, minimizing the above equation results in exactly the same labeling as the original SBA in [10].

Regarding the smoothness term, we use here the widely used edge-preserving Potts model [13]. 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. Let δ represent a Kronecker delta function. Then, with the above mentioned data model, the Potts model-based smoothness term, the energy equation (1) can be rewritten as follows:

$$\arg\min_Y \frac{1}{N}\sum_{p=1}^{V}\sum_{j=1}^{N}\hat{u}_p^j(Y_p) + \lambda\sum_{p=1}^{V}\sum_{\forall q\in\aleph_p} w_{pq}\,\left(1-\delta(Y_p,Y_q)\right).$$

Energy equation of the form (1) is ubiquitous in other computer vision problems like image denoising, segmentation and stereo matching; there are various efficient MRF optimization methods for solving them [13]. In this paper, we use the graph cuts expansion method [14] since it guarantees convergence to a global optimum for the current model.

3. RESULTS

3.1. Data

The data set contains 12 atlas images and 8 patients' images to be segmented; these CT images acquired at Divisions of Radiotherapy, Geneva University Hospital (HUG), during routine clinical practice. they typically have a resolution of $1mm{\times}1mm{\times}1mm$. We considered 10 lymph node volumes for automated segmentations, 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 structures have been manually delineated by an expert oncologist, according to the guidelines given in [2], and these manual segmentations are considered as ground truth segmentations.

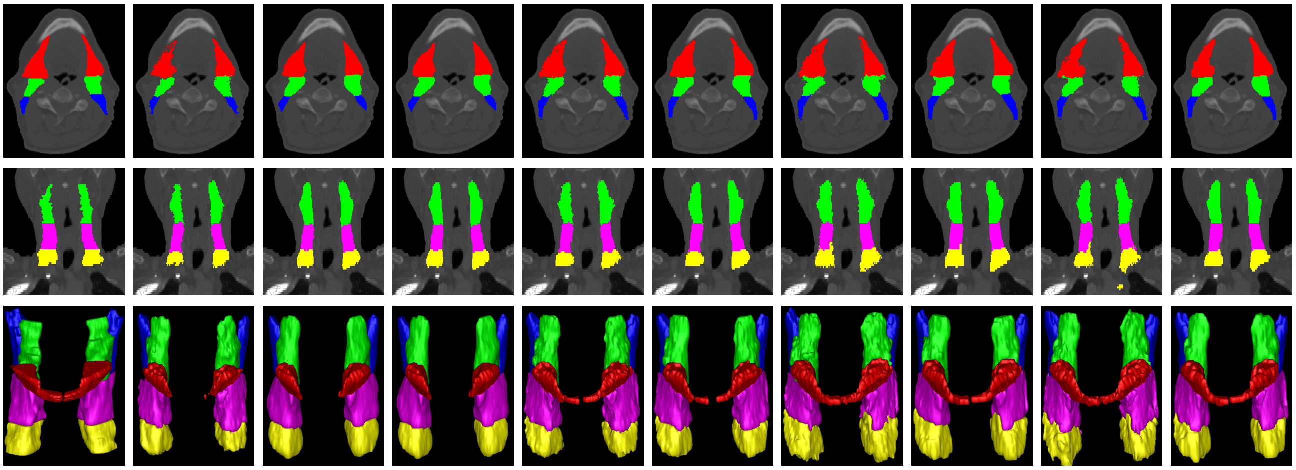
Regarding the registration, all the atlases are registered to each patient to be segmented. An initial affine registration is performed followed by a region-based registration, and a final pixel-based non-rigid registration. We skip here details about registration methods since the main focus of this work is label fusion, and this is independent of the registration algorithms used. we refer the readers to [15, 12] for more details on registration methods and parameters.

Fusion methods evaluated in this paper are: STAPLE, SBA, MV, GWV and LWV. All these methods (except STAPLE) are evaluated for both, 'with' and 'without' the MRF-based smoothness term. Those methods with the smoothness term are denoted with a suffix: "+MRF" to the name of the method. The performance of these methods has been evaluated for varying number of atlases also. For each patients' image, the atlases are ordered based on the overall dice similarity measure between the ground truth and the segmentation results obtained from single atlases.

For MV+MRF, GWV+MRF and LWV+MRF methods, λ is empirically set to 0.5. Note that, while the data term for the above three methods is a function of number of votes, for SBA, the data term is a function of distances; hence, the same smoothness term, when used with these two different classes of data terms, could require different scalings. For SBA+MRF method, λ is empirically set to 3. We did not optimize λ values; however, for obtaining more accurate segmentations, λ value may be further fine-tuned iteratively, based on the prior knowledge about the output structures (for instance, based on the prior knowledge about the "number of connected regions" in the structures of interest). The weights for LWV methods are computed over 9×9×9 neighborhood. For STAPLE, we use its multi-label version proposed in [5].

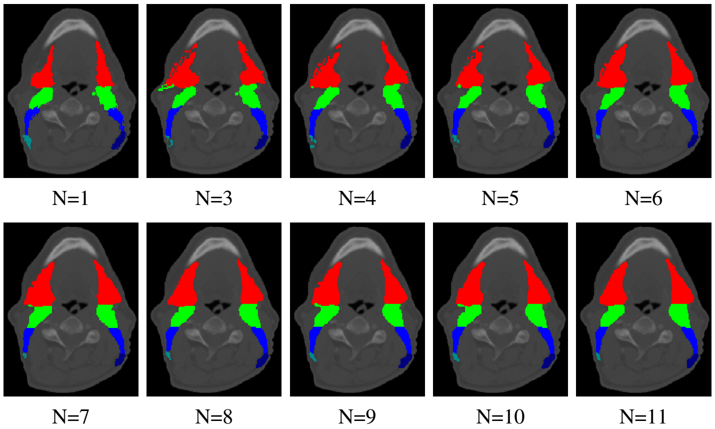
3.2. Evaluation

Figure 1 shows the qualitative results obtained from different methods, for one of the patients' images, for a fixed number of atlases (N=12). The following observations can be made from the visual inspection: MRF-based methods in general, provided more contiguous segmentations without any holes and islands. Among the methods that do not use MRF-based smoothness term, SBA provided relatively more contiguous structures. We can notice the presence of thin regions for structures IB-Left and IB-Right (shown in red color); we observe that, compared STAPLE and SBA-based methods, the remaining voting-based methods are more successful in retaining such thin regions. Figure 2 shows segmentation results obtained from LWV+MRF, for varying number of atlases. Because of the space limitations, we could not present here qualitative results for varying number of atlases for the remaining methods. It can be noted from Figure 2 that for LWV+MRF method, the segmentation results are quite stable for $N \geq 6$.



Ground truth (1)STAPLE (2)SBA (3)SBA+MRF (4)MV (5)MV+MRF (6)GWV (7)GWV+MRF (8)LWV (9)LWV+MRF

Fig. 1: Screenshots of segmentation results obtained from different methods (for N=12). The first two rows respectively show the results in axial and coronal slices, and the last row shows the resultant lymph node volumes. Each column presents results from different methods.



N=1 N=3 N=4 N=5 N=6

N=7 N=8 N=9 N=10 N=11

Fig. 2: Qualitative results from LWV+MRF method, for varying number of atlases, in one of the axial slices, for one of the patients' image.

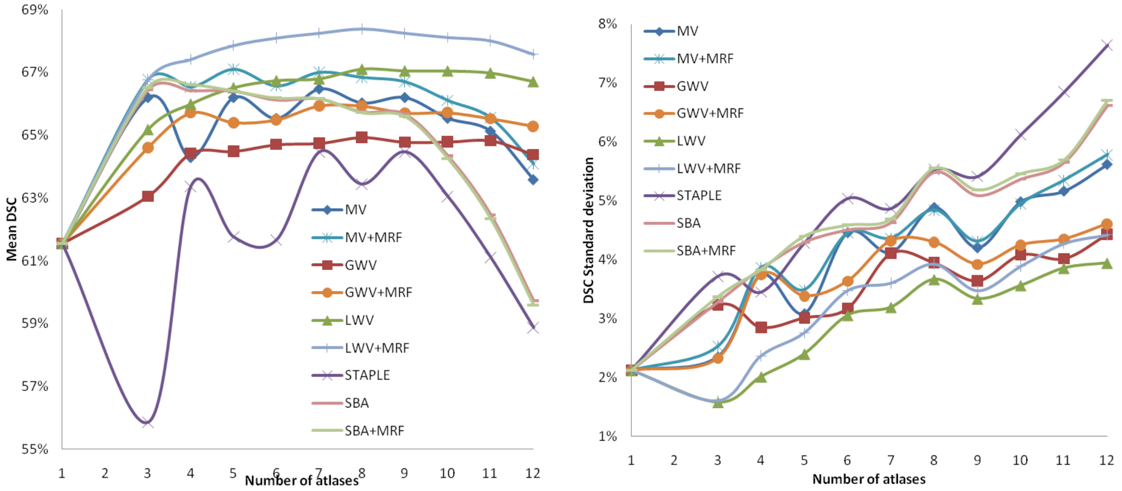


Fig. 3: Mean and standard deviations of Dice Similarity Coefficient (DSC) for different atlas fusion methods, for varying number of atlases.

Table 1: Mean and standard deviations of average number of connected regions per label, obtained from each fusion method (for N=12).

STAPLE	SBA	SBA +MRF	MV	MV +MRF	GWV	GWV +MRF	LWV	LWV +MRF
12.36±4.0	1.29±0.3	1.0±0.0	16.89±3.8	1.06±0.1	21.38±3.3	1.01±0.0	24.66±4.5	1.00±0.0

The quantitative evaluation is performed based on (i) Dice Similarity Coefficient (DSC), which is a widely used measure of overlap between the ground truth and automated segmentations, and (ii) Number of connected regions per label; since the output segmentations of each lymph node should ideally contain a single contiguous region, we are also evaluating the fusion algorithms based on the number of connected regions they create per each label; we take into account both islands and holes for computing the number of regions.

Figure 3 presents mean and standard deviations of DSC for different atlas fusion methods, and for varying number of atlases. LWV+MRF provided the best DSC results, followed by LWV, GWV+MRF, GWV, SBA-based methods, and STAPLE respectively. There is no significant improvement from SBA to SBA+MRF since lymph node segmentations from SBA itself are already contiguous (this is evident from the number-of-connected-regions). For $N \geq 6$, The behavior of mean DSC curve for voting-based methods is found to be more stable compared to STAPLE and SBA-based methods.

We evaluated the statistical significance of improvements in DSC with the inclusion of MRF-based term for MV, GWV, and LWV methods, using the Wilcoxon signed-rank test. It is found (at 0.05 significance level) that in all cases, the improvement in the segmentation results due to the inclusion of MRF-based smoothness term are statistically significant compared to their original methods.

Finally, Table 1 summarizes, for each method, the mean and standard deviations of number of connected regions per label. It can be noted that, SBA method, even without MRF-based term, provided contiguous lymph node segmentations. Among the methods that do not use MRF-based term, from the perspective of contiguous regions, SBA performed the best followed by STAPLE, MV, GWV and LWV respectively. Inclusion of MRF-based term clearly resulted in contiguous labels with no holes and islands.

4. CONCLUSIONS

In this paper, we have evaluated various atlas fusion strategies in the context of lymph nodes segmentation in the head and neck CT images. We have also reformulated the shape based averaging algorithm to fit into the general MRF-based framework that simultaneously performs the tasks of fusion and smoothing. Among all the methods, local weighted voting combined with MRF-based edge-preserving smoothing provided the best results, both in terms of overlap measure and contiguous regions.

5. REFERENCES

[1] S. Gorthi *et al.,* "Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 135–147, 2009.

[2] V. Grégoire *et al.,* "CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines," *Radiotherapy and Oncology*, vol. 69, no. 3, pp. 227–236, 2003.

[3] T. Rohlfing *et al.,* "Quo vadis, atlas-based segmentation?," in *The Handbook of Medical Image Analysis – Volume III*, chapter 11, pp. 435–486. Kluwer Academic, 2005.

[4] T. Rohlfing *et al.,* "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, 2004.

[5] X. Artaechevarria *et al.,* "Combination strategies in multi-atlas image segmentation: Application to brain mr data," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1266–1277, 2009.

[6] M.R. Sabuncu *et al.,* "A generative model for image segmentation based on label fusion," *IEEE Transactions on Medical Imaging*, vol. 29, no. 10, pp. 1714–1729, 2010.

[7] I. Ǐsgum *et al.,* "Multi-atlas-based segmentation with local decision fusion - application to cardiac and aortic segmentation in CT scans," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1000–1010, 2009.

[8] S.K. Warfield *et al.,* "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[9] T. Rohlfing *et al.,* "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 8, pp. 983–994, 2004.

[10] T. Rohlfing and C.R. Maurer, "Shape-based averaging," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 153 – 161, 2007.

[11] L. Ramus and G. Malandain, "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.

[12] S. Gorthi *et al.,* "Fusion of multi-atlas segmentations with spatial distribution modeling," in *MICCAI Worskshop on Multi-Atlas Labeling and Statistical Fusion*, 2011.

[13] R. Szeliski *et al.,* "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, 2008.

[14] Y. Boykov *et al.,* "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[15] S. Gorthi *et al.,* "Active Deformation Fields: Dense Deformation Field Estimation for Atlas-based Segmentation using the Active Contour Framework," *Medical Image Analysis*, vol. 15, no. 6, pp. 787–800, 2011.