Multi-atlas Based Segmentation : Application to the Head and Neck Region for Radiotherapy Planning

Liliane Ramus[1,2] and Grégoire Malandain[1]

[1] INRIA Sophia Antipolis - Asclepios Team, France
[2] DOSIsoft S.A., France

Abstract. Radiotherapy treament planning implies to delineate on the CT image of the patient the organs at risk where the dose has to be controlled. To avoid manual contouring that is tedious and prone to inter-expert variability, algorithms able to provide these delineations automatically can be helpful for the clinicians. This paper presents a multi-atlas based segmentation procedure to segment the parotid glands in the context of the *Head And Neck Auto-Segmentation Challenge 2010*. In this procedure, the images of the training database and their manual segmentations are first resampled on the query image through the intermediate coordinate system of an average atlas. Then, a voxel-wise combination strategy based on a weighted majority vote rule is performed to estimate the segmentation of the query. For each voxel, the weight assigned to a given resampled segmentation reflects the local degree of similarity between the corresponding resampled image and the query image. We applied our method with the database provided by the workshop.

1 Introduction

During a radiotherapy treatment planning, the clinicians have to parametrize and shape the beams in order to maximize the dose received by the target tumor while controlling the dose received by the surroundings organs at risk. Manual contouring can provide these delineations but it is time consuming and prone to inter-expert variability. To overcome these drawbacks, algorithms can be used to get automatic delineations of the organs at risk. The *Head and Neck Auto-Segmentation Challenge 2010* aims at comparing several of these algorithms in the context of the delineation of the parotid glands, that are among the most critical structures in the head and neck. Indeed, these glands are the most important salivary glands and their irradiation can result in xerostomia and dental complications. Therefore, sparing them from irradiation as much as possible is essential to preserve the quality of life of the patient after the treatment.

Atlas-based approaches have been proposed to get automatic delineations of the organs at risk in the brain [1], and automatic delineations of the lymph nodes and/or organs at risk in the head and neck region [2, 3]. When using atlas-based segmentation, the choice of the atlas is crucial, and several strategies have been proposed. The first introduced strategy was to use a single atlas for

segmenting all the images. This atlas can be an artificial delineated volume [1] or a particular manually delineated image but the accuracy of the resulting segmentation may be low for patients whose anatomy is too different from the atlas. If a database of several manually delineated images is available, there are two ways to solve this problem. The first way is to build an average atlas from the database and to use it for segmenting the patients [3]. If the database is sufficient to represent the variability of the population, this solution enables defining an atlas that is in the center of the population, and therefore close enough to most of the patients. However, such an atlas can still fail to segment some particular anatomies that are not well-represented in the database (for instance corpulent patients or patients with high neck flexion). The second way to take advantage of the training database is to select, for each new query image to segment, the image of the database that is the most similar to the query image, and to use it as atlas [4, 5]. This enables being more robust to non-common anatomies. By extension, it has also been proposed to select several of the most similar images and to combine their segmentations for segmenting the query image [6]. This approach is commonly called multi-atlas segmentation.

In the multi-atlas segmentation approaches, the selection of the most similar images can also be performed regionally as in [7–10] instead of globally. We propose here to go one step further by using a voxel-wise selection. In addition, instead of selecting the most similar images and combine their segmentations with equal weights, we keep all the images and weight their segmentations according to an intensity-based weighting system. Though a bit different, our approach has some similarities with one method proposed in [11].

Finally, we applied our method with the database provided by the Princess Margaret Hospital, which is composed of 10 training images and 8 testing images.

2 Method

We denote by P a query image to segment. We assume that a training database of N manually delineated images $\{I_k\}_{k\in[1\ldots N]}$ is available.

2.1 Efficient resampling of the training images on the query image

Registering each training image I_k on the query image requires to perform N non-linear registrations. To avoid these multiple registrations, we use here the intermediate coordinate system of an average intensity image M. This average intensity image is built from the N training images using the algorithm of Guimond et al. [12]. As registration method, we perform an affine registration [13] and then a non-linear registration [14] that are both based on a block-matching framework. Guimond's algorithm not only provides us the average intensity image M, but also the transformations $T_{I_k \leftarrow M}$ enabling to resample each image I_k of the database on M. All these preliminary steps are done off-line.

Then, the query image and the average intensity image M are non-linearly registered using the same framework than the one detailed above. This provides the transformation $T_{M \leftarrow P}$. At the end, each image I_k and its segmentation can be resampled on the query image with the transformation $T_{I_k \leftarrow M} \circ T_{M \leftarrow P}$.

2.2 Intensity-weighted majority vote rule

Let us denote by $s \in [1 \dots L]$ the labels of the anatomical structures of interest, the label $s = 0$ representing the background. At this step, the images I_k and their associated segmentations S_k have been resampled onto the query image P. We call $\tilde{I}_k$ and $\tilde{S}_k$ the images and segmentations resampled onto P. Each resampled segmentation $\tilde{S}_k$ can be seen as a candidate segmentation for the query image. Combining the N candidate segmentations $\{\tilde{S}_k\}_{k \in [1 \dots N]}$ enables compensating the local errors that can be introduced by some of them, and can therefore enhance accuracy and robustness.

To estimate the probability $p(x \in s)$ of the voxel x to belong to each label $s \in [0 \dots L]$, we apply a weighted majority vote rule, as described below:

$$\forall s \in [0 \dots L] \qquad p(x \in s) = \frac{1}{\sum_{k=1}^{K} \omega_k(x)} \sum_{k=1}^{K} \omega_k(x) \; \delta\left(\tilde{S}_k(x), s\right) \tag{1}$$

where $\delta(\tilde{S}_k(x), s) = 1$ if $\tilde{S}_k(x) = s$ and 0 otherwise. In this equation, the weight $\omega_k(x)$ reflects the local degree of similarity between P and $\tilde{I}_k$ on a neighborhood $V(x)$ of the voxel x. We defined it as the inverse of the SSD (Sum of Squared Differences), so that high weights are given to the candidate segmentations $\tilde{S}_k$ for which the intensity similarity between P and $\tilde{I}_k$ is high on $V(x)$:

$$\forall k \in [1 \dots K] \qquad \omega_k(x) = \frac{1}{SSD_{V(x)}(P, \tilde{I}_k)} = \frac{1}{\frac{1}{\text{card}(V(x))} \sum_{y \in V(x)} [P(y) - \tilde{I}_k(y)]^2} \tag{2}$$

Then, we assign the voxel x to the label $\tilde{s}(x)$ having the highest probability: $\tilde{s}(x) = \arg\max_{s \in [0 \dots L]} p(x \in s)$. In case there is one or several images $\tilde{I}_k$ for which $SSD_{V(x)}(P, \tilde{I}_k) = 0$, then we compute $\tilde{s}(x)$ using a simple majority vote rule from the corresponding $\tilde{S}_k$, without taking into account the other segmentations.

Finally, we apply a morphological closing to smooth the resulting segmentation and we extract the main connected component.

3 Evaluation

The database of images was provided by the Princess Margaret Hospital in Toronto. The training database was composed of 10 images for which the parotids were manually delineated by an expert. The testing database was composed of 8 images. The voxel size was $0.976562 \times 0.976562 \times 2 \; mm^3$ for almost all images and the matrix size was 512×512. To increase the size of the training database, we symmetrized each training image and its manual delineations with respect to the mid-sagittal plane. Thus, we had a training database of $N = 20$ delineated images. We built an average intensity image from these 20 images using [12]. Then, for each of the 8 query images of the testing database, we resampled the 20 images of the training database on it (as described in 2.1), and we performed the multi-atlas based segmentation process in the coordinate system of the query image (as described in 2.2). In this last step, the local similarity measures were computed on a $3 \times 3 \times 3$ neighborhood. The figure 1 and the overlap and Hausdorff distance raw results were provided by the organizers of the workshop.

3.1 Qualitative Results

Figure 1 shows the automatic and manual segmentations for a particular patient of the testing database. The top images illustrate well the problem of dental artifacts, which are one of the main issues in segmenting organs of the head and neck on CT images. Indeed, these artifacts introduce a bias in the intensity distribution especially in the area of the parotids, which can corrupt our intensity-based similarity metric and result in local errors in the segmentation.

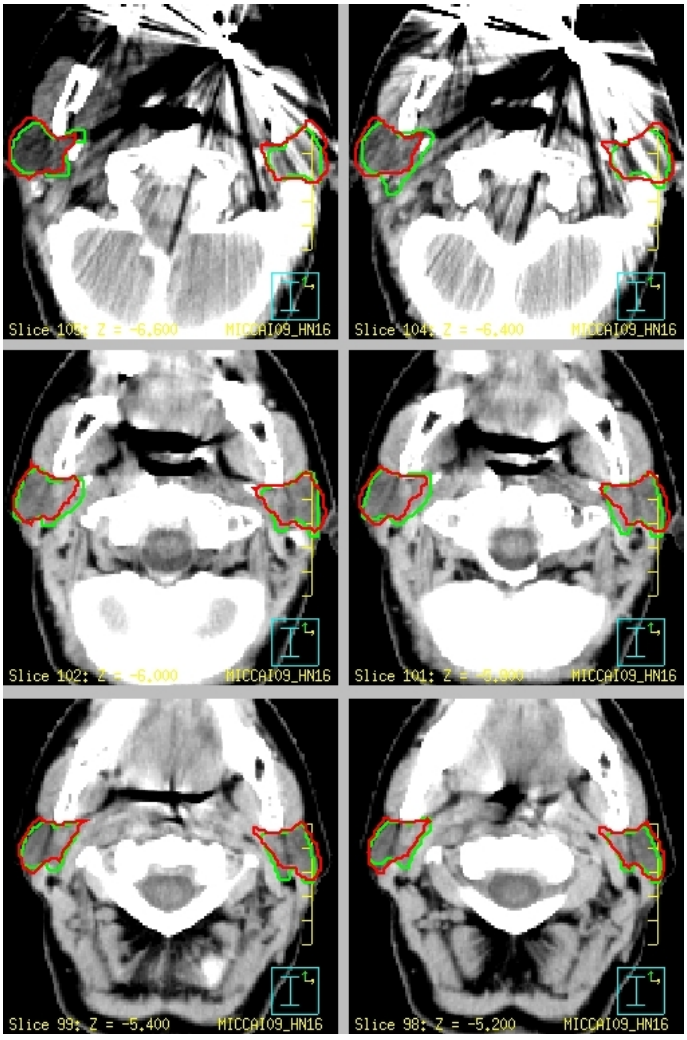

Fig. 1. Qualitative results of segmentation obtained with our method (red contours) compared to the ground truth manual segmentation performed by a single observer (green contours).

3.2 Quantitative Results

The quantitative evaluation is based on overlap measures (Dice index, presented in Table 1) and distance measures (Hausdorff distances (HD), presented in Table 2). These measures were computed slice by slice and the average/median values over all the slices are shown in the columns 2 and 3 of Tables 1 and 2. The volumetric Dice was also computed (column 4 of Table 1). Finally, the column 4 of Table 2 contains the number of slices for which the Hausdorff distance was greater than $3mm$. More details on these measures can be found in [15].

First, the quantitative measures show that our method fails to segment the dataset #13. The particularity of this dataset is that it has a truncated field of view and an important neck flexion compared to the majority of the patients in the training database. We visually inspected the intermediate results for this patient, and concluded that the failure occurred during the non-linear registration of the query image with the average atlas. The rigid and affine part of this registration work pretty well and the truncated field of view of the patient is well-taken into account in the registration. The failure appears in the non-linear part of the registration and is due to the high neck flexion of the patient. As the remaining of the algorithm (resampling of the training images/segmentations on the query and multi-atlas segmentation) directly relies on this registration, the resulting automatic segmentation is not localized in the correct area and the numerical results are very bad (Dice ≈ 0 and $HD > 40mm$).

Statistics computed on the remaining of the testing datasets (ie left and right parotids of datasets #11, #12, #14, #15, #16, #17 and #18) are shown at the bottom of Tables 1 and 2. Numerical results obtained for both parotids were considered together to compute these statistics. We chose to exclude the dataset #13 to compute statistics in order to illustrate the capacities of the algorithm when no registration failure happens. We believe that such registration failure is quite rare (this is confirmed by experiments launched on another database, with registration failure occuring for only 2 images out of 105) and that taking it into account while computing statistics over 8 datasets only would introduce a bias. All remaining datasets provided a median slice HD that ranges between $3.91mm$ and $10.34mm$, with an average of $6.42mm$. As to the median slice overlap values, they range between 75.0 % and 90.4 %, with an average of 85.3 %.

Finally, our algorithm does not perform very well with respect to the number of slices for which the Hausdorff distance was greater than $3mm$ (column 4 of Table 2). Indeed, the best case is the left parotid of dataset #16 for which 22 slices out of 29 have an HD greater than $3mm$.

3.3 Discussion

The low accuracy obtained for dataset #13 is caused by a failure in the registration between the query image and the average atlas. This failure is due to the different neck flexions in the two images to register (high neck flexion in the query image, and standard neck flexion in the average atlas). Actually, this kind of problem could potentially occur with any other particular anatomy that is marginally represented in the training database (like corpulence for instance).

Dataset No.		Average slice OV	Median slice OV	Total volume OV
Right parotid	11	81.6 %	87.1 %	85.7 %
	12	81.7 %	83.2 %	83.5 %
	13	4.0 %	0.6 %	4.0 %
	14	85.3 %	89.0 %	87.6 %
	15	83.6 %	88.4 %	88.3 %
	16	79.9 %	83.7 %	82.6 %
	17	77.8 %	83.3 %	82.9 %
	18	80.8 %	85.1 %	84.5 %
Left parotid	11	77.0 %	82.1 %	83.5 %
	12	82.8 %	87.4 %	86.2 %
	13	0.0 %	0.0 %	0.0 %
	14	85.3 %	90.4 %	88.8 %
	15	83.1 %	86.0 %	86.9 %
	16	82.4 %	85.4 %	84.5 %
	17	67.7 %	75.0 %	76.4 %
	18	80.0 %	83.5 %	83.1 %
Statistics with dataset #13 excluded:				
Min-Max		**[67.7 % - 85.3 %]**	**[75.0 % - 90.4 %]**	**[76.4 % - 88.8 %]**
Average		**80.6 %** ± 4.5 %	**85.0 %** ± 3.8 %	**84.6 %** ± 3.1 %
Median		**81.7 %**	**85.3 %**	**84.5 %**

Table 1. Overlap (OV) statistics for left and right parotid segmentation in the testing datasets.

Dataset No.		Average slice HD	Median slice HD	No. of slices (HD > 3 mm)
Right parotid	11	7.04	6.69	32 (32)
	12	5.35	5.61	30 (30)
	13	41.22	40.06	26 (26)
	14	5.12	4.63	24 (23)
	15	7.32	7.03	26 (24)
	16	4.89	4.88	31 (25)
	17	9.12	6.91	27 (27)
	18	6.38	5.90	25 (23)
Left parotid	11	6.72	6.40	34 (33)
	12	5.70	4.63	28 (26)
	13	50.32	48.65	26 (26)
	14	6.19	6.94	24 (20)
	15	8.96	7.91	23 (23)
	16	4.37	3.91	29 (22)
	17	12.37	10.34	33 (33)
	18	8.15	8.07	24 (24)
Statistics with dataset #13 excluded:				
Min-Max		**[4.37 - 12.37]**	**[3.91 - 10.34]**	-
Average		**6.98** ± 2.13	**6.42** ± 1.69	-
Median		**6.55**	**6.55**	-

Table 2. Hausdorff distance (HD) statistics for left and right parotid segmentation in the testing datasets.

A way to avoid this kind of problem is to directly register each training image with the query image instead of using the average atlas as intermediate coordinate system. Indeed, the resulting segmentation does no longer rely on a single registration but on multiple independent registrations. If the query image has a particular anatomy that is represented by a small minority of datasets in the training database, then at least the direct registrations with those particular datasets will succeed, which would be sufficient to improve significantly the accuracy of the resulting segmentation. Of course, the main drawback of this method is that it increases the computation time. However, the recent advances in parallel programming may help to solve this problem.

In the perspective of evaluating our algorithms before submission to the workshop, we also provided to the organizers the resulting segmentations obtained for the testing database when the query image was directly registered to each training image. The improvement with respect to the algorithm using the average atlas as intermediate coordinate system is significant for the dataset #13: the total volume overlap reaches 47 % instead of 4 % for the right parotid, and 15 % instead of 0 % for the left parotid. There is no significant difference between the two methods for the other datasets, indicating that the approximation $T_{I_k \leftarrow P} \approx T_{I_k \leftarrow M} \circ T_{M \leftarrow P}$ is valid for these datasets.

4 Conclusion

We presented a multi-atlas based segmentation approach and applied it with the database provided by the *Head And Neck Auto-Segmentation Challenge 2010*. Our method uses a pre-built average intensity image as intermediate coordinate system to deform the images/segmentations of the training database onto the query. Then, a multi-atlas segmentation algorithm is performed in the coordinate system of the query. This algorithm enables weighting locally the influence of each training dataset with respect to its local intensity similarity to the query.

Our results show that our approach provides segmentation with reasonable accuracy for 7 testing datasets out of 8 (average Dice of 84.6 %). Our framework fails to segment the remaining patient (dataset #13) because of its high neck flexion that causes errors in the registration with the average atlas. We demonstrated that registering directly each training image to the query enabled improving significantly the segmentation accuracy for this dataset. This approach is therefore more robust to particular anatomies (high neck flexion, corpulence) than the approach that uses the average atlas as intermediate coordinate system. However, it is computationally too expensive for the on-site live contest, which is the reason why we submitted the approach that uses the average atlas.

As to the evaluation, this study uses 10 training images and 8 testing images. Increasing the training database would enable a better representation of the anatomical variability present in the head and neck region, and it is likely to improve the results as our method consists in weighting the training images according to their similarity to the query. As to the testing database, the online

contest will enable assessing our method on 7 more unseen datasets, which will increase the statistical significance of the results.

Finally, we will study the impact of the neighborhood size on the weights and also test to what extent undersampling to 256×256 decreases the accuracy.

Acknowledgments This work was undertaken with the MAESTRO project (IP CE503564) funded by the European Commission, and was also funded by ANRT.

References

1. Bondiau, P.Y., Malandain, G., Chanalet, S., et al.: Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context. IJROBP **61**(1) (2005) 289–98
2. Han, X., Hoogeman, M.S., Levendag, P.C., et al.: Atlas-based auto-segmentation of head and neck CT images. In: Proc. MICCAI'08. Volume 5242 of LNCS. (2008) 434–441
3. Commowick, O., Grégoire, V., Malandain, G.: Atlas-based delineation of lymph node levels in head and neck computed tomography images. Rad Oncol **87**(2) (2008) 281–289
4. Wu, M., Rosano, C., Lopez-Garcia, P., et al.: Optimum template selection for atlas-based segmentation. Neuroimage **34**(4) (2007) 1612–8
5. Commowick, O., Malandain, G.: Efficient selection of the most similar image in a database for critical structures segmentation. In: Proc. MICCAI'07, Part II. Volume 4792 of LNCS. (2007) 203–210
6. Aljabar, P., Heckemann, R.A., et al.: Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage **46**(3) (2009) 726–38
7. van Rikxoort, E.M., Isgum, I., et al.: Adaptive local multi-atlas segmentation: application to heart segmentation in chest CT scans. MedIA **14**(1) (2010) 39–49
8. Commowick, O., Warfield, S.K., Malandain, G.: Using Frankenstein's creature paradigm to build a patient specific atlas. In: Proc. MICCAI'09, Part II. Volume 5762 of LNCS. (2009) 993–1000
9. Shi, F., Yap, P.T., Fan, Y., et al.: Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation. Neuroimage **51**(2) (2010) 684–93
10. Ramus, L., Commowick, O., Malandain, G.: Construction of patient specific atlases from locally most similar anatomical pieces. In: Proc. MICCAI'10. LNCS, Beijing, China (September 2010)
11. Artaechevarria, X., Munoz-Barrutia, A., de Solorzano, C.O.: Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans Med Imaging **28**(8) (August 2009) 1266–77
12. Guimond, A., Meunier, J., Thirion, J.P.: Average brain models: A convergence study. CVIU **77**(2) (2000) 192–210
13. Ourselin, S., Roche, A., Prima, S., Ayache, N.: Block matching: A general framework to improve robustness of rigid registration of medical images. In: Proc. MICCAI'00. Volume 1935 of LNCS. (2000) 557–566
14. Garcia, V., Commowick, O., Malandain, G.: A robust and efficient block matching framework for non linear registration of thoracic CT images. In: Grand Challenges in Medical Image Analysis, Beijing, China (September 2010)
15. Pekar, D.: http://www.grand-challenge2010.ca/Evaluation.pdf