ASSESSING SELECTION METHODS
IN THE CONTEXT OF MULTI-ATLAS BASED SEGMENTATION

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

[1] INRIA - Asclepios Team, Sophia Antipolis, France
[2] DOSIsoft S.A., Cachan, France
{Liliane.Ramus,Gregoire.Malandain}@sophia.inria.fr

ABSTRACT

In atlas-based segmentation, using one single atlas for segmenting all patients introduces a bias. Multi-atlas techniques overcome this drawback by selecting and fusing the most appropriate atlases among a database for a given patient. Globally assessing different multi-atlas strategies provides a biased evaluation of the atlas selection methods. To address this problem, we propose to evaluate atlas selection methods independently from the number of atlases selected and from the atlas fusion step. Briefly, we first cluster the selection methods on the basis of rank correlation and then assess each sub-group of methods with respect to a sub-group of reference selection methods. We apply our method to 105 images of the head and neck region.

Index Terms— Medical imaging, patient-specific atlas, multi-atlas segmentation, atlas selection.

1. INTRODUCTION

Radiotherapy treatment planning requires the delineation of both the target tumor and the organs at risk (OARs). These delineations are traditionally made manually by experts but this task is tedious and not reproducible. Atlas-based segmentation has proved to be an efficient procedure to automatically get these delineations for the head and neck region [1, 2].

The principle of atlas-based segmentation is to non-linearly deform an already segmented anatomy, called atlas, on the patient image. Non-linear registration requires a trade-off between accuracy and smoothness and performs better when the atlas image is similar to the patient image.

Under this assumption, using one single atlas for any patient may be satisfactory if the atlas image is a good representative of the population. When the atlas used is a particular segmented patient image, it may be very similar to some patients, but very different from others. An average atlas built from a database of several segmented images provides better results [1] , but it still has difficulties to cope with very high anatomical variabilities as those in the head and neck region.

To overcome this drawback, methods have been introduced to design patient-specific atlases. A way to address this problem is to automatically select among a set of potential atlases the most appropriate one for each new image to segment. The potential atlases can be average atlases precomputed from homogeneous sub-groups of the database that can possibly be obtained by *atlas stratification* [3]. Alternatively, each image of the database can be considered as a potential atlas [4, 5]. By extension, multi-atlas based segmentation consists in selecting the most appropriate images among the database (*atlas selection* step) and fusing their segmentations, either globally [6, 7, 8] or locally [9] (*atlas fusion* step).

In all these approaches, atlas selection is a crucial step. The selection can be done on the basis of meta-information such as the age [6] or any clinical information. However, this is not always possible nor relevant. Thus, several image-based selection methods have been proposed. The selection can be based on similarity measures between the patient image and each potential atlas [6, 7, 8, 4] or based on local deformations [8, 5, 9]. Therefore, the comparison of these different methods for a given application is of interest. Usually, this is done by assessing the resulting segmentations with respect to the manual segmentation, but this does not enable to evaluate atlas selection independently neither from the method of atlas fusion used nor from the number of atlases selected.

Our objective is to compare different atlas selection methods independently from the atlas fusion step. In order to determine whether some selection methods are equivalent and to assess them, we consider the issue of *atlas selection* in terms of *atlas ranking*, and we propose to use pairwise rank correlations values to cluster the ranking methods. We present various ranking methods and our evaluation framework in sections 2 and 3 respectively. Results obtained using 105 CT images of the head and neck region are shown in section 4.

2. ATLAS RANKING METHODS

Let N be the number of delineated images in the database. For each patient P of the database, we consider the $N - 1$ remaining images as potential atlases, and we rank them from the most similar to least similar to the patient P using various ranking methods. We describe automatic ranking methods in 2.1 and 2.2, and a reference ranking method in 2.3.

2.1. Intensity-based ranking methods

Intensity-based ranking methods consist in computing similarity measures between the patient image and the potential atlases warped into the same referential. Any intensity-based ranking method can then be defined by:

1. the similarity measure used: it can be Sum of Squared Differences (SSD), Correlation Coefficient (CC), Mutual Information (MI), Normalized Mutual Information (NMI);
2. the registration algorithm used to put the patient image and the potential atlases in the same referential: it can be either affine or non-linear registration;
3. the mask on which the similarity measure is computed: it can be the whole image or a mask of the Region Of Interest (ROI) corresponding to the structures and their neighborhood.

In the affine case, we register the potential atlases on the patient image, and we compute the similarity measures in the referential of the patient. In the non-linear case, we chose to deform the potential atlases and the patient image onto the average image pre-computed from the $N-1$ remaining patients using [1], and to compute the similarity measures in this referential. This is computationally interesting because the non-linear deformation between each potential atlas and the average image has already been computed during the construction of the average image. Thus, the only non-linear registration to perform is the one between the patient and the average image.

2.2. Deformation-based ranking methods

Alternatively, atlas ranking can be based on local deformations [8, 5, 9]. In this case, the most similar atlas is the one that requires the smallest local deformation to be matched on the patient. As in [5], we estimate these local deformations through the intermediate referential of the average image.

2.3. Reference ranking method

To assess automatic ranking methods, a reference has to be defined. Since our purpose is the delineation of the OARs, the potential atlases will be ranked according to their ability to yield accurate segmentation. Each of them is non-linearly registered onto the patient image P, and its efficiency is quantified using various measures (such as the Dice index, the sensitivity or the Hausdorff distance between the deformed OARs and the patient's ones) which represent so many reference criteria to rank the potential atlases. The reference ranking method thus defined does not rely on any fusion method of several atlases, but solely on the non-linear registration.

3. EVALUATION FRAMEWORK

In section 2, we described automatic and reference ranking methods. Our evaluation framework has two objectives. The first one (detailed in sections 3.1 and 3.2) is to identify subgroups of equivalent ranking methods. To this end, we propose to cluster the ranking methods (both automatic and reference methods) with the affinity propagation algorithm using pairwise rank correlation. The second objective (detailed in section 3.3) is to analyze the average correlation between the rankings provided by each sub-group of automatic ranking methods and the reference rankings presented in section 2.3.

3.1. Rank correlation analysis

For each patient P of the database, the rank correlation between two atlas ranking methods M_i and M_j can be quantified using Spearman's rank correlation coefficient between the corresponding rankings of the $N-1$ remaining images, called $\rho(P, M_i, M_j)$. Let M be the number of atlas ranking methods tested. Thus, for each patient, we have a $M \times M$ matrix of Spearman's coefficients. Averaging Spearman's coefficients over the N patients of the database provides a $M \times M$ matrix of average Spearman's values $\bar{\rho}(M_i, M_j)$.

3.2. Clustering of the atlas ranking methods

In this step, we use the average Spearman's values $\bar{\rho}(M_i, M_j)$ as pairwise similarity measures between the different ranking methods to apply the affinity propagation clustering algorithm proposed by Frey et al. [10]. Briefly, given a set of data points (the ranking methods here) and the pairwise similarities between these data points (the average Spearman's values here), this algorithm iteratively identifies the underlying clusters and also finds for each cluster the exemplar that best represents this cluster.[1] The optimal number of clusters is automatically estimated according to the input self-similarities that quantify the suitability for each data point to be an exemplar. Since none of the methods is more suitable to be an exemplar than the others, we chose to give the same self-similarity value to each method. The higher the shared self-similarity value is, the higher the number of clusters is.

3.3. Cluster assessment

We assume here that the reference ranking methods group together in one or several sub-groups, called reference clusters. In practice, this was always the case in our experiments. Under this assumption, we assess the clusters of automatic methods by computing their average correlation with the cluster(s) of reference methods. Given the input pairwise similarities, we define the inter-cluster correlation between two clusters c_{auto} and c_{ref} as the average similarity value computed over all pairs of methods $(M_a \in c_{auto}, M_r \in c_{ref})$. Comparing the inter-cluster correlation between each cluster c_{auto} and the cluster(s) c_{ref} enables to determine whether one sub-group of automatic methods performs better than the others.

[1] Basically, the algorithm is based on an iterative message-passing procedure where each data point is iteratively reassessed as a potential exemplar by exchanging messages with other data points and taking into account the input similarities as well. After convergence, the clusters and exemplars are estimated from the final messages exchanged. For further details, see [10].

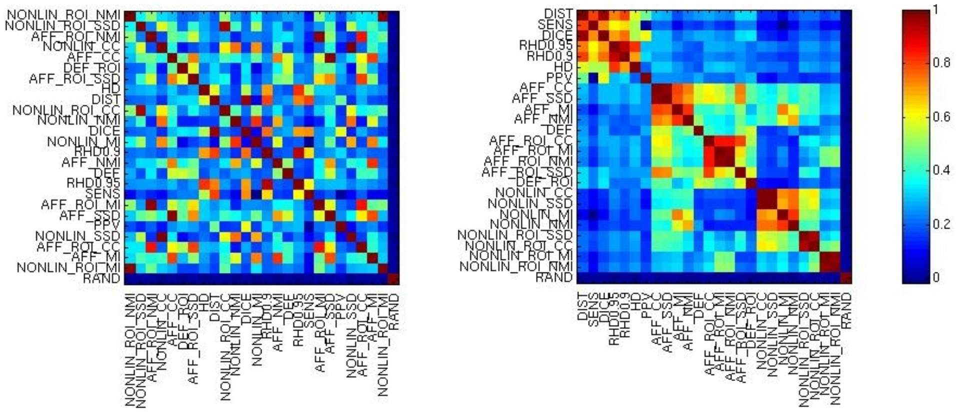

Fig. 1. Spearman's coefficients for each pair of ranking methods (both automatic and reference methods) before (left) and after (right) clustering with a self-similarity set to the median value of all input similarities (0.2628). See text for abbreviations.

4. RESULTS

We applied our methodology on a database of 105 CT images of the head and neck region that were delineated by experts following the guidelines of [11]. The structures involved are organs at risk (parotids, sub-mandibular glands, mandible, brainstem, spinal cord) and the lymph node levels II, III, IV.

In the remainder of the article, deformation-based ranking methods are abbreviated DEF (when computed on the whole image) and DEF_ROI (when computed on the ROI). Intensity-based ranking methods are abbreviated either AFF_SIM, NONLIN_SIM (when computed on the whole image) or AFF_ROI_SIM, NONLIN_ROI_SIM (when computed on the ROI) where SIM is CC, SSD, MI or NMI, and where AFF and NONLIN refer to the registration used for the normalization. As to the reference ranking methods, we considered the Dice index (DICE), the sensitivity (SENS), the distance to the best achievable measure (sensitivity=1;specificity=1) (DIST), the Positive Predictive Value (PPV), the Hausdorff distance (HD) and the Robust Hausdorff Distance that exclude the worst 5% or 10% cases (RHD0.95 and RHD0.9). By way of example, we also included a random ranking method (RAND).

4.1. Clustering results

First, we applied the framework described in section 3 with a self-similarity value set to the median value of all the input similarities as recommended in [10]. In this case, the ranking methods are divided into 6 sub-groups (see the red dotted line in Figure 2 for details). The first sub-group gathers all reference ranking methods. Then, the automatic ranking methods split into 4 sub-groups, and the random ranking method is alone in its own cluster. For this value of self-similarity, this configuration of sub-groups is the one that maximizes the intra-cluster similarities (on the diagonal blocks) and minimizes the inter-cluster similarities, as illustrated in Figure 1.

Secondly, we were interested in the influence of the self-similarity value on the resulting clusters configuration, as shown in Figure 2. When the self-similarity is high (0.95 for instance), each method tends to be its own exemplar and to create its own cluster. Conversely, when the self-similarity is low (-3 for instance), all methods are in a unique cluster. In between these two extrema, intensity-based methods computed after non-linear normalization tend to group together, whereas deformation-based methods tend to group together with intensity-based methods computed after affine normalization. This might be explained by the fact that deformation-based methods and intensity-based methods computed after affine normalization are two classes of methods that both depend on the residue after affine registration, either encoded as intensity or as deformation. On the contrary, intensity-based methods computed after non-linear normalization depend on the residue after non-linear registration, which is less discriminant and possibly increases the noise in the measures. As to the reference methods, they split between overlap measures (Dice, sensitivity, distance, PPV) and Hausdorff distance measures, which was expected.

4.2. Assessment of the clusters of automatic methods

For a self-similarity set to the median value of the input similarities, we computed the inter-cluster correlation values between the cluster of reference methods (cluster of exemplar DIST) and the clusters of automatic methods as described in section 3.3 (see the red dotted line in Figure 2 for details on the clusters). First, the random ranking method shows a very low correlation with the cluster of reference methods (0.0054). The 2 clusters gathering the intensity-based methods after affine normalization and the deformation-based methods (clusters of exemplars AFF_ROI_CC and AFF_CC) provide higher inter-cluster correlation with the reference cluster (respectively 0.2495 and 0.2331) than the 2 clusters of intensity-based methods after non-linear normalization (0.1745 and 0.1880). Therefore, the intensity-based methods after affine registration and the deformation-based methods seem more appropriate for our application.

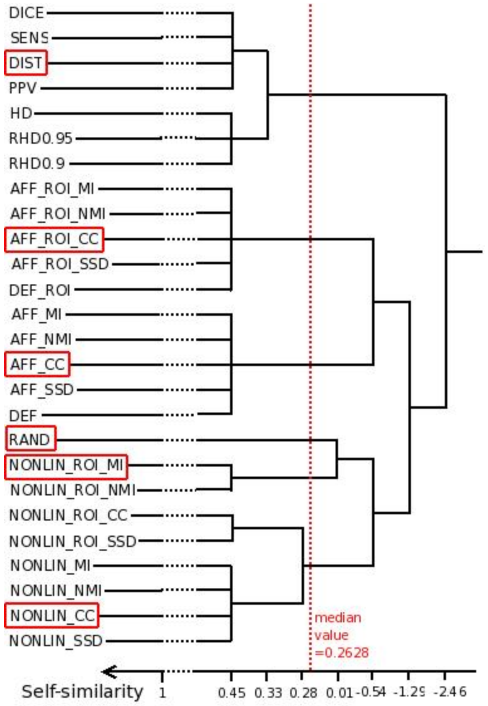

Fig. 2. Resulting clusters according to the input self-similarity. The boxes correspond to the exemplars of the configuration obtained for a self-similarity set to the median value of the input similarities (0.2628). See text for abbreviations.

5. CONCLUSION

In this article we presented an evaluation framework to compare atlas ranking methods in the context of multi-atlas based segmentation. Briefly, we first cluster the ranking methods according to the pairwise Spearman's rank correlation values to identify sub-groups of equivalent methods. Secondly, we assess each cluster of automatic ranking methods by computing its average correlation value with a cluster of reference ranking methods.

The first advantage of this framework is that it enables to compare the atlas selection methods independently from the number of atlases selected and from the atlas fusion step. This is of interest because even if the majority vote rule is often used for fusing the segmentations of the selected atlases, more sophisticated atlas fusion methods have also been presented. For instance, Isgum et al. proposed to locally take into account the quality of the non-linear registration of each selected atlas on the patient [12]. Besides, the number of atlases selected and fused also plays an important role in the quality of the resulting segmentation as shown in [6].

The second advantage of our methodology is that it can be used to reduce the number of ranking methods to consider for a potential in-depth evaluation or a visual inspection, or more generally for future work. Finally, this framework can be applied for any other application dealing with ranking methods.

6. ACKNOWLEDGMENTS

This work was undertaken in the framework of the MAESTRO project (IP CE503564) funded by the European Commission, and was also partially funded by ANRT. The authors gratefully acknowledge Pr. V. Grégoire for the manually delineated database and for his expertise.

7. REFERENCES

[1] O. Commowick, V. Grégoire, and G. Malandain, "Atlas-based delineation of lymph node levels in head and neck computed tomography images," *Radiotherapy Oncology*, vol. 87, no. 2, pp. 281–289, 2008.

[2] X. Han, L.S. Hibbard, N. O'Connell, and V. Willcut, "Automatic segmentation of head and neck CT images by GPU-accelerated multi-atlas fusion," in *MICCAI'09 Workshop on 3D Segmentation Challenge for Clinical Applications*, 2009.

[3] D.J. Blezek and J.V. Miller, "Atlas stratification," *Med Image Anal*, vol. 11, no. 5, pp. 443–57, October 2007.

[4] M. Wu, C. Rosano, P. Lopez-Garcia, et al., "Optimum template selection for atlas-based segmentation," *Neuroimage*, vol. 34, no. 4, pp. 1612–8, February 2007.

[5] O. Commowick and G. Malandain, "Efficient selection of the most similar image in a database for critical structures segmentation," in *Proc. MICCAI'07, Part II*, 2007, vol. 4792 of *LNCS*, pp. 203–210.

[6] P. Aljabar, R.A. Heckemann, A. Hammers, et al., "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy," *Neuroimage*, vol. 46, no. 3, pp. 726–38, July 2009.

[7] P. Aljabar, R. Heckemann, A. Hammers, et al., "Classifier selection strategies for label fusion using large atlas databases," in *Proc. MICCAI'07, Part I*, 2007, vol. 4791 of *LNCS*, pp. 523–31.

[8] T. Rohlfing, R. Brandt, R. Menzel, 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. 1428–42, April 2004.

[9] O. Commowick, S.K. Warfield, and G. Malandain, "Using Frankenstein's creature paradigm to build a patient specific atlas," in *Proc. MICCAI'09, Part II*, 2009, vol. 5762 of *LNCS*, pp. 993–1000.

[10] B.J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–6, February 2007.

[11] V. Grégoire, P. Levendag, K.K. Ang, 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 Oncology*, vol. 69, no. 3, pp. 227–36, Dec. 2003.

[12] I. Isgum, M. Staring, A .Rutten, et al., "Multi-atlas-based segmentation with local decision fusion–application to cardiac and aortic segmentation in CT scans," *IEEE Trans Med Imaging*, vol. 28, no. 7, pp. 1000–10, July 2009.