USING CONSENSUS MEASURES FOR ATLAS CONSTRUCTION

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

[1] INRIA Sophia Antipolis - Asclepios Team,
2004 route des Lucioles - BP 93, 06 902 Sophia Antipolis Cedex, France
[2] DOSIsoft S.A., 45 avenue Carnot, 94 230 Cachan, France
{Liliane.Ramus,Gregoire.Malandain}@sophia.inria.fr

ABSTRACT

Atlas-based segmentation has been shown to provide promising results to delineate critical structures for radiotherapy planning. However, it requires to have a reference image with its reference segmentation available. Classical methods used to build an average segmentation can lead to over-segmentation in case of high variability among the manual segmentations. We propose in this paper a consensus-based approach to construct a reference segmentation from a database of manually delineated images. We first compute local consensus measures to estimate a variability map, and then deduct from it a consensus segmentation. Finally, the proposed method is evaluated using a dataset of 64 manually delineated images of the head and neck region.

Index Terms— Medical imaging, atlas, structure averaging, kappa statistics, consensus measures.

1. INTRODUCTION

In radiotherapy planning, dose optimization consists in both maximizing the dose on the tumor and minimizing the dose on Organs At Risks (OARs). This requires to accurately delineate OARs. Usually, these delineations are manually made by a clinician, but this is time-consuming, tedious, and not reproducible. Therefore, tools providing automatic delineations of OARs are useful as they enable to save time and to give reproducible delineations. Among these tools, atlas-based segmentation has been shown to provide promising results for the brain [1] and for the head and neck region [2, 3] but it requires to have an atlas available.

When the inter-patient anatomical variability is low, it may be satisfactory to use for the atlas one particular manually delineated image. However, this method introduces a bias and is therefore not appropriate in anatomical regions with high variability. To address this problem, the standard method is to build an average image and an average segmentation from a database of manually delineated images.

The algorithm of Guimond et al. [4] can be used to construct the average image, and provides the appropriate transformations to bring all manual segmentations on the average image. Several methods exist to build the average segmentation from the manual segmentations warped onto the average image. Shape-based approaches such as [5] consist in matching corresponding points between all the meshes and then performing a geometric average to get the average shape. However, these methods require to parametrize shapes by extracting characteristic points in the manual delineations, which is not possible for all shapes.

Conversely, other approaches focus on a voxel-based estimation of the representative segmentation and therefore do not require to parametrize shapes. Among them, the STAPLE algorithm proposed by Warfield et al. [6] simultaneously estimates the true segmentation and the performance parameters for each segmentation of the database. This approach has the advantage of working for multi-label segmentations. It was already used to delineate critical structures of the head and neck region, but an over-segmentation was noticed for some critical structures [2].

To overcome this drawback, we propose in this article a new approach to build a reference segmentation. This approach is based on kappa-type statistics and is composed of two steps: (i) compute local consensus measures to construct a variability map, and (ii) use this variability map to estimate a consensus segmentation.

This article is organized as follows. We will first briefly introduce kappa-type statistics in section 2 and then describe our approach in section 3. Finally, we will present in section 4 some results and evaluations performed on a database of 64 manually delineated CT images of the head and neck region.

2. KAPPA-TYPE STATISTICS

Kappa-type statistics consist in assessing the inter-rater level of agreement with chance-corrected coefficients. They were historically introduced by Cohen [7] to evaluate the reliability of medical diagnosis.

2.1. Cohen's kappa

Let consider two raters classifying items into L mutually exclusive categories. The proportion of items classified into the

category i by the first rater and into the category j by the second rater is called $p(i,j)$. The marginal probabilities are named $p(i,.) = \sum_{j=1}^{L} p(i,j)$ and $p(.,i) = \sum_{j=1}^{L} p(j,i)$.

The proportion of observed agreement $p_{obs} = \sum_{i=1}^{L} p(i,i)$ is the proportion of items classified into the same category by the two raters. However, as agreement among raters may be due to pure chance, the previous formula over-estimates the real level of agreement. To address this problem, the chance-expected proportion of agreement $p_{ch} = \sum_{i=1}^{L} p(i,.)p(.,i)$ is estimated by taking into account the marginal probabilities. Thus, $p_{obs} - p_{ch}$ represents the excess of agreement beyond chance level.

Finally, the level of agreement among the two raters can be assessed with the normalized chance-corrected coefficient $\kappa = \frac{p_{obs} - p_{ch}}{1 - p_{ch}}$ [7].

2.2. Generalizations

Many generalizations of Cohen's kappa have been proposed. Weighted kappa-type coefficients were introduced by Cohen [8] in order to scale disagreement or to give partial credit to disagreement of low gravity for two raters.

Cohen's kappa [7] and Cohen's weighted kappa [8] are only defined for two raters. For more than two raters, there are two ways to measure the level of agreement among raters: one can measure either pairwise agreement or majority agreement, also called consensus agreement.

Concerning pairwise agreement, Schouten [9] focused on the case where every item was classified by each of the raters whereas Fleiss [10] dealt with the case where all the items were not necessarily classified by all the raters.

Landis and Kock [11] presented consensus agreement measures using weights that express the extent to which raters classified items into the same category.

3. METHODS

In this section, we describe how kappa-type statistics can be used to build a reference segmentation from a dataset of manually delineated images. As described in section 3.1, all manual segmentations are at first brought into the same referential. Then, our method is divided into two steps, which are detailed in sections 3.2 and 3.3. First, we use consensus agreement measures to quantify the local variability among the warped manual segmentations. Then, from the resulting variability map, we estimate a consensus segmentation.

3.1. Spatial normalization of manual segmentations

The variability among manual segmentations includes anatomical inter-patient variability, due to corpulence or neck flexion for instance, and inter and intra-expert variability. To overcome the anatomical variability, we build an average image with the method of Guimond et al. [4] and we apply the resulting deformation fields to bring the manual segmentations on the average image. We use here the locally affine registration method proposed by Commowick et al. [12]. However, the remaining variability after registration is high. It results from inter and intra-expert variability and remaining anatomical variability due to registration residual errors.

3.2. Quantification of the local variability

The manual segmentations warped onto the average image can be considered as raters who assigned each voxel to one of the anatomical structures or to the background. Therefore, we can apply kappa statistics considering that the raters are the manual segmentations warped onto the average image, the items classified are the voxels, and the mutually exclusive categories are the anatomical structures and the background.

As the variability among manual segmentations is high, the voxels with perfect agreement among all manual segmentations (that is to say the voxels assigned to the same anatomical structure in all manual segmentations) are very few in number. Therefore, we also take into account the voxels for which the agreement among manual segmentations is not perfect and we associate to each case of disagreement an appropriate weight reflecting the degree of local disagreement.

Using pairwise agreement measures such as [9] requires in addition to assess the degree of gravity of each case of disagreement according to its nature. In other words, it requires to decide whether the disagreement *structure A versus structure B* has a higher or a lower gravity than the disagreement *structure A versus structure C*. Since we do not have any rationale for doing so, we chose to use consensus agreement measures such as Landis and Kock [11] to quantify the variability among the manual segmentations.

Let K be the number of manual segmentations in our database and L the number of anatomical structures including the background. Each voxel of the average image can be associated to one of the L^K possible combinations $\tilde{c} = [\tilde{c}(1), ..., \tilde{c}(k), ..., \tilde{c}(K)]$ where $\tilde{c}(k) \in [1, ..., L]$ represents the index of the anatomical structure chosen for this voxel in the manual segmentation k warped onto the average image.

The number of manual segmentations that chose the structure l in the combination $\tilde{c} = [\tilde{c}(1), ..., \tilde{c}(K)]$ is given by $nbSegm(\tilde{c}, l) = \sum_{k=1}^{K} \delta(\tilde{c}(k), l)$ where $\delta(.,.)$ represents the Kronecker function.

Given these notations, a weight reflecting the level of agreement in the combination $\tilde{c} = [\tilde{c}(1), ..., \tilde{c}(K)]$ can be defined as follows:

$$\omega(\tilde{c}) = \frac{nbSegm(\tilde{c}, s_1(\tilde{c})) - nbSegm(\tilde{c}, s_2(\tilde{c}))}{K} \quad (1)$$

where $s_1(\tilde{c})$ and $s_2(\tilde{c})$ are respectively the indexes of the first and the second anatomical structure the most represented in the combination $\tilde{c}$. Thus, $\omega(\tilde{c})$ equals 1 if all the manual segmentations agreed for the same structure, that is to say if

$\tilde{c}(k) = s_1(\tilde{c})$ for each $k \in [1, ..., K]$. It equals 0 if there are as many manual segmentations that chose $s_1(\tilde{c})$ and $s_2(\tilde{c})$, which is the case of highest disagreement. It is between 0 and 1 for intermediate cases.

Furthermore, as we want to quantify the local variability, the proportion of observed weighted agreement is estimated locally for each voxel i of the average image on a local neighborhood called $\mathcal{N}(i)$ with the following formula:

$$p_{obs}(i) = \sum_{\tilde{c}} \omega(\tilde{c}) \frac{n(i, \tilde{c})}{N} \quad (2)$$

where $n(i, \tilde{c})$ is the number of voxels in $\mathcal{N}(i)$ that were associated to the combination $\tilde{c}$, and N is the total number of voxels in $\mathcal{N}(i)$. In practice, we use for $\mathcal{N}(i)$ a window of size $3 \times 3 \times 3$ voxels centered in the voxel i.

In order to enable the comparison of local consensus measures between the different voxels, the proportion of chance-expected weighted agreement is computed by considering all the combinations present in the whole image as follows:

$$p_{ch} = \sum_{\tilde{c}} \omega(\tilde{c}) \prod_{k=1}^{K} p_k(\tilde{c}(k)) \quad (3)$$

where $p_k(l)$ is the proportion of voxels assigned to the structure l in the manual segmentation k and is computed on the whole image.

Finally, we can quantify the local consensus agreement among the manual segmentations in each voxel i of the average image with the following chance-corrected coefficient:

$$\kappa(i) = \frac{p_{obs}(i) - p_{ch}}{1 - p_{ch}} \quad (4)$$

3.3. Estimation of a consensus segmentation

The first step of our method provides us a variability map representing the local variability among the warped manual segmentations. Then, we extract the regional minima of the inverse of the variability map and use them as seeds for a watershed transformation applied on the inverse of the variability map. This gives an estimation of the consensus segmentation.

4. RESULTS

The proposed method was evaluated using a database of 64 CT images of the head and neck region, which were manually delineated following the guidelines provided in [13]. The anatomical structures involved are: the lymph node levels II, III and IV, the parotids, the submandibular glands, the brainstem, the spinal cord, and the mandible.

Using the whole database, we constructed an average image, consensus delineations using our method and consensus delineations using the multi-label STAPLE algorithm [6]. In section 4.1, we first show the variability maps provided by our method and then qualitatively compare the reference delineations obtained using our method with those obtained using the STAPLE algorithm. In section 4.2, we present a quantitative comparison of the two methods.

4.1. Qualitative evaluation

Figure 1 shows the resulting variability maps zoomed in on two structures of the head and neck region. For the parotid, there are two areas with particularly high variability among the manual segmentations (areas marked with arrows in Figure 1). The first area corresponds to the beginning of the accessory lobe, which is a part of the parotid not present in all patients. The second area is the deep lobe, which is a region with low contrast and hard to delineate for clinicians.

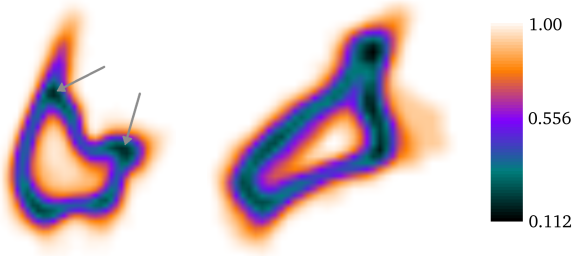

Fig. 1. Variability maps zoomed in on two anatomical structures of the head and neck region. From left to right: right parotid and right lymph node levels III.

Figure 2 compares on top of the variability maps the reference delineations obtained using our approach with those obtained using the STAPLE algorithm for two structures of the head and neck region. This figure shows that our reference delineations are smaller than those obtained with the STAPLE algorithm. Indeed, the outer areas of intermediate variability as well as the areas of high variability are not included in our contours whereas they are inside those obtained with the STAPLE algorithm. Therefore, our method enables us to reduce the over-segmentation in the head and neck region.

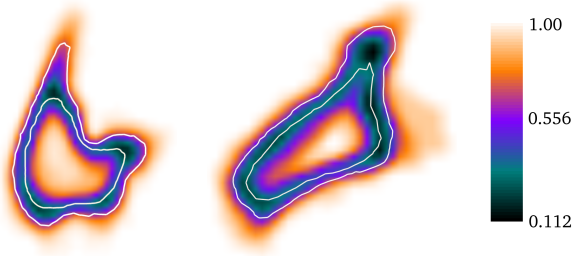

Fig. 2. Reference delineations obtained with our method (inner delineations) and with the STAPLE algorithm (outer delineations) represented on top of the variability maps for two anatomical structures of the head and neck region. From left to right: right parotid and right lymph node levels III.

4.2. Quantitative evaluation

The visual results shown above suggest that our approach enables us to reduce over-segmentation. In this section, we now compare the quantitative results obtained using our method with those obtained using the STAPLE algorithm.

For both methods, the atlas built with the 64 manually delineated images was evaluated on each of the 64 patients successively[1]. The resulting quality measures were at first averaged over all the structures for each patient, and subsequently averaged over the 64 patients. The resulting average quality measures are presented in Figure 3. This figure shows that our approach provides a significantly higher average specificity than the STAPLE algorithm (0.85 versus 0.62). At the same time, the average sensitivity is lower with our approach but the key point is that there is no significant difference between the two methods for the average Dice coefficient. Thus, our approach gives smaller contours than the STAPLE algorithm without reducing the Dice coefficient, and therefore enables us to correct efficiently the over-segmentation.

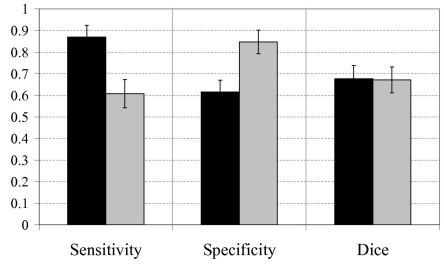

Fig. 3. Average quantitative measures computed on 64 patients and on 13 anatomical structures with our approach (in grey) and with the STAPLE algorithm (in black).

5. CONCLUSION AND PERSPECTIVES

We have presented in this article an original approach to build a consensus segmentation for atlas construction from a database of manually delineated images. The proposed method consists in first computing local consensus measures to map the local variability among the warped manual delineations and then extracting a consensus segmentation from the obtained variability map.

Using the resulting consensus segmentation for atlas construction proved to be an efficient way to reduce the over-segmentation in the head and neck region. Indeed, in comparison with classical methods as the STAPLE algorithm, the average specificity was significantly improved with our method without any decrease in the average Dice coefficient.

Our method has the advantage of working for multi-label segmentations but missing structures in the delineations are not taken into account yet. Future work will focus on the extension to the case of missing structures in the delineations.

[1]Rigorously, a Leave-One-Out evaluation should have been prefered but our experiments did not show any difference when using the whole database.

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] P.Y. Bondiau, G. Malandain, S. Chanalet, et al., "Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context," *Int J Radiat Oncol Biol Phys*, vol. 61, no. 1, pp. 289–98, Jan. 2005.

[2] 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.

[3] X. Han, M.S. Hoogeman, P.C. Levendag, et al., "Atlas-based auto-segmentation of head and neck CT images," in *Proc. MICCAI'08, Part II*, 2008, vol. 5242 of *LNCS*, pp. 434–441.

[4] A. Guimond, J. Meunier, and J.P. Thirion, "Average brain models: A convergence study," *Computer Vision and Image Understanding*, vol. 77, no. 2, pp. 192–210, 2000.

[5] R.H. Davies, C.J. Twining, P.D. Allen, et al., "Building optimal 2d statistical shape models," *Image and Vision Computing*, vol. 21, pp. 1171–1182, 2003.

[6] S.K. Warfield, K.H. Zou, and W.M. Wells, "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, July 2004.

[7] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[8] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, October 1968.

[9] H.J.A. Schouten, "Measuring pairwise interobserver agreement when all subjects are judged by the same observers," *Statistica Neerlandica*, vol. 36, no. 2, pp. 45–61, 1982.

[10] J.L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[11] J.R. Landis and G.G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, vol. 33, no. 2, pp. 363–374, June 1977.

[12] O. Commowick, V. Arsigny, A. Isambert, et al., "An efficient locally affine framework for the smooth registration of anatomical structures," *Medical Image Analysis*, vol. 12, no. 4, pp. 427–441, 2008.

[13] 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.