Automatic tissue segmentation of head and neck MR images for hyperthermia treatment planning

Citation for published version (APA):

Fortunati, V., Verhaart, R. F., Niessen, W. J., Veenland, J. F., Paulides, M. M., & van Walsum, T. (2015). Automatic tissue segmentation of head and neck MR images for hyperthermia treatment planning. *Physics in Medicine and Biology*, 60(16), 6547-6562. <https://doi.org/10.1088/0031-9155/60/16/6547>

Document license:
TAVERNE

DOI:
[10.1088/0031-9155/60/16/6547](https://doi.org/10.1088/0031-9155/60/16/6547)

Document status and date:
Published: 21/08/2015

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

PAPER

Automatic tissue segmentation of head and neck MR images for hyperthermia treatment planning

To cite this article: Valerio Fortunati *et al* 2015 *Phys. Med. Biol.* **60** 6547

View the [article online](#) for updates and enhancements.

Related content

- [Regional deep hyperthermia: impact of observer variability in CT-based manual tissue segmentation on simulated temperature distribution](#)
Bassim Aklon, Josefin Hartmann, Diana Zink *et al.*
- [Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT](#)
A Chen, K J Niermann, M A Deeley *et al.*
- [Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region](#)
J P Kieselmann, C P Kamerling, N Burgos *et al.*

Recent citations

- [Hyperthermia treatment planning guided applicator selection for sub-superficial head and neck tumors heating](#)
Tomas Drizdal *et al*
- [Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks](#)
Bulat Ibragimov and Lei Xing

Uncompromised.

See clearly during treatment to attack the tumor and protect the patient.

Two worlds, one future.

Captured on Elekta high-field MR-linac during 2018 imaging studies.

Elekta

Elekta MR-linac is pending FDA (k) premarket clearance and not available for commercial distribution or sale in the U.S.

Automatic tissue segmentation of head and neck MR images for hyperthermia treatment planning

Valerio Fortunati¹, René F Verhaart², Wiro J Niessen^{1,3},
Jifke F Veenland¹, Margarethus M Paulides² and
Theo van Walsum¹

¹ Departments of Medical Informatics and Radiology, Biomedical Imaging Group Rotterdam, Erasmus MC University Medical Center, 3015 CE Rotterdam, The Netherlands

² Department of Radiation Oncology, Hyperthermia Unit, Erasmus MC University Medical Center—Cancer Institute, 3015 CE Rotterdam, The Netherlands

³ Faculty of Applied Sciences, Delft University of Technology, 2628 CC Delft, The Netherlands

E-mail: vfortunati83@gmail.com

Received 17 January 2015, revised 17 May 2015

Accepted for publication 24 June 2015

Published 12 August 2015

CrossMark

Abstract

A hyperthermia treatment requires accurate, patient-specific treatment planning. This planning is based on 3D anatomical models which are generally derived from computed tomography. Because of its superior soft tissue contrast, magnetic resonance imaging (MRI) information can be introduced to improve the quality of these 3D patient models and therefore the treatment planning itself. Thus, we present here an automatic atlas-based segmentation algorithm for MR images of the head and neck.

Our method combines multiatlas local weighting fusion with intensity modelling. The accuracy of the method was evaluated using a leave-one-out cross validation experiment over a set of 11 patients for which manual delineation were available.

The accuracy of the proposed method was high both in terms of the Dice similarity coefficient (DSC) and the 95th percentile Hausdorff surface distance (HSD) with median DSC higher than 0.8 for all tissues except sclera. For all tissues, except the spine tissues, the accuracy was approaching the interobserver agreement/variability both in terms of DSC and HSD. The positive effect of adding the intensity modelling to the multiatlas fusion decreased when a more accurate atlas fusion method was used.

Using the proposed approach we improved the performance of the approach previously presented for H&N hyperthermia treatment planning, making the method suitable for clinical application.

Keywords: atlas-based method, MRI, treatment planning, image segmentation, hyperthermia, graph cut

(Some figures may appear in colour only in the online journal)

1. Introduction

Local hyperthermia treatment (HT) consists of elevating the temperature of tumours between 40 °C to 44 °C. The effectiveness of radiotherapy (RT) and chemotherapy is significantly enhanced when combined with HT (van der Zee 2002, Franckena *et al* 2009, Issels *et al* 2010). Phase III clinical studies show that hyperthermia is also effective in the head and neck (H&N) region (Valdagni 1994, Huilgol *et al* 2010, Hua *et al* 2011). In response to these positive results, a site-specific hyperthermia applicator was developed for precise HT of tumours in the entire H&N region using electromagnetic fields. This device, called the HYPERcollar, enables deep heating (>4 cm under the skin) of H&N cancers (Paulides *et al* 2007, Bakker *et al* 2010, Paulides *et al* 2010). HT of H&N tumors requires accurate treatment planning (HTP), based on 3D patient models that are derived from segmented 3D images (Verhaart *et al* 2014a). These segmentations are generally derived by manual delineation of computed tomography (CT) images. However, it is well-known that manual delineation is tedious, labour intense and prone to inter and intraobserver variability. Thus automation of the segmentation process is highly desired. It has been shown that, due to their superior soft tissue contrast, MR images provide better delineation accuracy of target and normal tissues than CT images (Toonkel *et al* 1988, Gordon *et al* 2004, Ahmed *et al* 2010, Rasch *et al* 2010, Verhaart *et al* 2014b). Hence, in this paper we investigate the accuracy of MRI-based automatic segmentation for some of the tissues needed for HTP. MRI-based segmentation need to be registered to the CT coordinate frame in order to obtain a geometrically accurate full 3D patient model (Fortunati *et al* 2014, Verhaart *et al* 2014b).

Several methods have been developed in recent years to address the problem of automatic segmentation of H&N CT images for generating anatomical models for radiotherapy (Han *et al* 2008, Qazi *et al* 2011, Fritscher *et al* 2014) treatment planning. The common aspect of all these methods is that they are based on multiatlas segmentation approaches. In the field of HT planning, we recently published a method to accurately segment all relevant tissue in CT images (Fortunati *et al* 2013) using atlases and intensity modelling. We also showed that the use of this method leads to a HTP quality comparable to the quality obtained using manually generated delineations (Verhaart *et al* 2014a).

The problem of automatic segmentation using atlases was extensively addressed in the field of brain structure segmentation in MR images (Artaechevarria *et al* 2009, Sabuncu *et al* 2010, van der Lijn *et al* 2012, Wang *et al* 2012, Akhondi-Asl and Warfield 2013). Whereas majority voting, being a simple and straightforward way, has been frequently used to combine atlas registration results, more recently multiple authors have shown that local weighting atlas fusion techniques (Artaechevarria *et al* 2009, Sabuncu *et al* 2010, Wang *et al* 2012) gives improve segmentation results. In these approaches, segmentations from multiple atlases are combined using local image similarity metrics to locally weigh the contribution of each segmentation. Among these methods, the approach developed by Wang *et al* (2012) ranked first in a recent MICCAI grand challenge (Landman and Warfield 2012). Due to the effectiveness

of this method we decided to combine local weighting multiatlas fusion with the approach we previously used for CT images (Fortunati *et al* 2013).

In the present study, we developed and evaluated an automatic segmentation algorithm for MR images of the head and neck; we used a graph cut method which combines multi-atlas-based segmentation, using atlas fusion, and intensity modeling. The method combines the approach described earlier (Fortunati *et al* 2013) with local weighting atlas fusion (Artaechevarria *et al* 2009, Wang *et al* 2012). The contribution of our work is twofold: (1) we developed an automatic segmentation method for head and neck MR images that integrates intensity modelling into state-of-the art multi atlas fusion methods; (2) we evaluated various versions of our method on imaging data of 11 patients, and compared the results to the inter-observer variability of manual delineations.

2. Methods and materials

2.1. Image data and atlas

The data used in this study was collected from 11 patients with tumours in the H&N region, which were representative for the patient-group eligible for HT. The location of the tumours was the nasopharynx (3 patients) and the oropharynx (8 patients). The data comprised axial, MR T2 weighted (T2w) images acquired on a 1.5 T scanner (Optima MR450w, GE Healthcare, Waukesha, WI) with the following acquisition parameters: Fast Recovery Fast Spin Echo sequence with TE/TR 106.8/7060 ms; in plane resolution of 0.49 mm (5 patients) and 0.68 mm (6 patients); slice thickness of 3 mm; out of plane field of view of 20 cm centred at the tumour location, yielding a slightly different field of view for each patient. For 8 patients, images were acquired in treatment position using a 6 channel surface coil while for the other 3 patients a 24 channel H&N coil was used. The T2w images were corrected for intensity inhomogeneity: for the images acquired with surface coils, a built-in algorithm (Surface Coil Intensity Correction) of the MR scanner was used at the time of acquisition; for the images acquired with H&N coils, the ITK-N4 method (Tustison *et al* 2010) was used as a post-processing procedure. In order to compensate for the differences in intensities, all the images were normalized (Udupa 1999). This normalization involved matching all the intensity histograms to the histogram of one randomly selected image. The code for this normalization is in the ITK library (www.itk.org/itkHistogramMatchingImageFilter.h). We used 256 bins for the histogram computation matching 128 histogram landmarks.

Three trained observers (O1, O2, O3) delineated the tissues in all images. We included the tissues for which the dielectric and thermal properties are different from those of the surrounding tissues. Additionally, tissue delineation had to be feasible on T2w images. With these criteria, the tissues of interest were:

- Encephalon: cerebrum (CE); cerebellum (CB); brainstem (BS).
- Spinal cord: myelum (MY) and cerebro spinal fluid (CSF).
- Eye tissues: vitreous humor (EH); sclera (SC); lens (LE).

The optical nerve was excluded from this study even though it is relevant for HTP. This was done because the optical nerve could not be consistently delineated on T2w images.

2.2. Segmentation method

Our segmentation method combines locally weighted multiatlas fusion with voxel-wise classification based on intensity values. Atlas based segmentation consists of the registration of a

target image, with an unknown segmentation, to an atlas for which the segmentation is known. The registration is used to estimate a segmentation by spatially mapping the known segmentation to the target image. In multiatlas segmentation this process is repeated for more than one atlas and an atlas fusion method is used to combine the information of all atlases into a single optimal segmentation (Rohlfing *et al* 2005).

In our method, an atlas fusion method was used to determine a spatial probability map for each tissue. The atlases, in the original coordinate frame, were further used to estimate the intensity distribution of the tissues and their corresponding background regions. This information was used in a graph cut segmentation algorithm: the probability map was used as a spatial prior and the intensity distributions as an intensity model. A regularization term was also used to ensure a smooth segmentation.

For details on how the intensity model was combined with the spatial probability map we refer the reader to our previous work (Fortunati *et al* 2013). In the following we focus on the local weighting atlas fusion technique and its integration into our segmentation framework.

2.3. Spatial probability map construction

To construct a spatial prior, first a set of J atlas images was non-rigidly registered to the target image u . Using the registration results for each atlas j and each label l we obtained the segmentation $S_j^l(m)$ and the registered images $I_j(m)$. Fusion was performed with a local weighted voting technique (Artaechevarria *et al* 2009, Wang *et al* 2012). In this method the label at voxel location m was selected as the one with the highest vote $\hat{S}^l(m)$, where $\hat{S}^l(m)$ is computed according to the rule:

$$\hat{S}^l(m) = \sum_{j=0}^J \omega_j(m) S_j^l(m), \quad (1)$$

where $\omega_j(m)$ is a weighting factor for the location m , which reflects the accuracy of each single atlas segmentation, in the proximity of m . Thus the final vote was calculated by weighting the contribution of each single atlas segmentation with respect to their local registration quality.

The weights $\omega_j(m)$ for each location m and atlas j can be estimated using the local image similarity between a patch of the target image $\mathcal{I}_u(\mathcal{N}(m))$ and a patch of atlas j registered to the coordinate frame of the target image u and interpolated at voxel location m , $\mathcal{I}_j(\mathcal{N}(m))$. The patch $\mathcal{N}(m)$ was defined as a 3D cubical region centred at m with width (in voxels) of $2r_p + 1$. We evaluated different similarity metrics (SIM) for local fusion: sum of squared differences (SSD), normalized cross-correlation (NCC) and normalized mutual information (NMI). The weights $\omega_j(m)$ were defined following the work of Artaechevarria *et al* (2009):

$$\omega_j(m) = \text{SIM}(\mathcal{I}_u(\mathcal{N}(m)), \mathcal{I}_j(\mathcal{N}(m)))^\beta, \quad (2)$$

where β is a factor which was positive for NCC and NMI, and negative for SSD in order to make ω a measure of similarity. β is a factor used to scale the similarity measure as the gain exponent in Artaechevarria *et al* (2009).

Two different advanced weighting techniques have been proposed to improve this generic local weighting framework: (1) using a local search technique to decrease the effects of small registration error; (2) computing the final weight of each atlas considering the joint image similarity between the target image and all atlases. Both approaches were integrated into our method; we discuss them in the next two subsections.

2.3.1. Local search technique. The anatomical variability between the atlas and target image makes the registration challenging and prone to errors. We used a free form deformation based non-rigid registration (Rueckert *et al* 1999) as described at the end of this section and in appendix A. The registration error can be accounted for during atlas fusion (Sabuncu *et al* 2010, Wang *et al* 2012) by locally searching for the atlas image patch that has the best similarity with the patch of the target image u . Next, the segmentation corresponding to the most similar atlas image patch can be used in the modified voting rule:

$$\hat{S}^l(m) = \sum_{j=0}^J \omega_j(\eta_j(m)) S_j^l(\eta_j(m)), \quad (3)$$

where $\eta_j(m)$ is the location in the atlas j corresponding to the patch $\mathcal{N}$ that gives the best match between atlas and target image in the local search region $\mathcal{N}_s$ according to the local search:

$$\eta_j(m) = \underset{m' \in \mathcal{N}_s}{\operatorname{argmin}} \|\mathcal{I}_j(\mathcal{N}(m')) - \mathcal{I}_u(\mathcal{N}(m))\|^2, \quad (4)$$

$\mathcal{I}_j(\mathcal{N}(m'))$ is the vector of intensities of the j -th atlas within the patch $\mathcal{N}$ centred around m' and $\mathcal{I}_u(\mathcal{N}(m))$ is the vector of intensities of the target image u within the patch $\mathcal{N}$ centred around m . The patch $\mathcal{N}_s(m)$ is defined as a 3D cubical region centred at m with width (in voxels) of $2r_s + 1$.

2.3.2. Joint weighted voting. Wang *et al* (2012) additionally introduced a weighting technique to estimate of the final segmentation taking a possibly correlated error between atlas segmentations into account. In this method the optimal weights are computed by minimizing the total expectation of segmentation error. This expectation is expressed as a matrix containing the joint probabilities of two atlases making a segmentation error at a certain location. Similarly to the general local weighting method this probability is approximated using an intensity similarity between a pair of atlases and the target image in the neighbourhood of each voxel. Let $\omega_m = (\omega_0(m), \dots, \omega_J(m))$ denote the vector of weights. Following Wang *et al* (2012) the optimal weights are calculated as:

$$\omega_m^* = \frac{(M_m + \gamma I)^{-1} \mathbf{1}_n}{\mathbf{1}_n^T (M_m + \gamma I)^{-1} \mathbf{1}_n}, \quad (5)$$

where γ is a regularization term used to ensure a unique solution for the minimization problem, $\mathbf{1}_n = [1; 1; \dots; 1]$ and I is the identity matrix. The weights in equation (5) are normalized, i.e. $\sum_j \omega_m(i) = 1$. The matrix M contains the expected pairwise joint label differences at each location m , which can be estimated using local image similarity, i.e. SSD, NCC and NMI. NCC and the NMI need to be inverted in order to represent an error measure. In case of SSD we defined the matrix M_m according to Wang *et al* (2012). For the NCC based joint fusion the matrix was defined as:

$$M_m^{\text{NCC}}(i, j) = \left(\sqrt{1 - \text{NCC}(\mathcal{I}_u(\mathcal{N}(m)), \mathcal{I}_i(\mathcal{N}(m)))} \sqrt{1 - \text{NCC}(\mathcal{I}_u(\mathcal{N}(m)), \mathcal{I}_j(\mathcal{N}(m)))} \right)^\beta, \quad (6)$$

where i and j indicate the atlases, β is a positive scaling factor and the value one is used to invert the NCC since $\text{NCC} \in [-1, 1]$. Similarly for the NMI based joint fusion we have:

$$M_m^{\text{NMI}}(i, j) = \left(\sqrt{3 - \text{NMI}(\mathcal{I}_u(\mathcal{N}(m)), \mathcal{I}_i(\mathcal{N}(m)))} \sqrt{3 - \text{NMI}(\mathcal{I}_u(\mathcal{N}(m)), \mathcal{I}_j(\mathcal{N}(m)))} \right)^\beta, \quad (7)$$

with the same notation as (6). In this case we use the value three to invert the metric since $\text{NMI} \in [1, 2]$. Following Wang *et al* (2012) for all the three metrics, the value of γ was set to approximatively 2% of the maximum value of the similarity metric i.e. 0.1 for SSD and 0.05 for NMI and NCC.

2.4. Graph cut segmentation

The multiatlas fusion methods provides a probability map $p_s^l(f_m)$ for each tissue:

$$p_s^l(f_m) = \left(\frac{1}{\sum_j \omega_j(\eta_j(m))} \sum_j S_j^l(\eta_j(m)) \omega_j(\eta_j(m)) \right)^{\lambda_2}, \quad (8)$$

where λ_2 is a weighting parameter. This probability map is used as spatial prior in the data terms of the graph cut segmentation algorithm (Boykov *et al* 2001). The other data term used is the intensity model of the tissue which consisted of the intensity distributions of the tissue and its corresponding background region. The intensity distributions is built using the unregistered atlases by a Parzen-window histogram estimation (Fortunati *et al* 2013). The intensity model aims to correct for residual registration errors or systematic error in the output segmentation. Two user-defined parameters were used: the weight parameter λ_1 to weigh the contribution of the data terms with respect to the regularization term and the parameter λ_2 to weigh the contribution of the spatial prior term with respect to the intensity model term.

For each tissue l , the graph cut segmentation depends on a vector of parameters $\theta_l = (\alpha, \lambda_1, \lambda_2)$, i.e. the data term weight λ_1 , the spatial model weight λ_2 and the foreground threshold α (Fortunati *et al* 2013). This last parameter is introduced to correct for underinclusion. For each tissue, the optimal value of the parameters $\hat{\theta}_l$ depends on the spatial and intensity model; these parameters are estimated in a cross-validation experiment on the available atlases.

A max-flow algorithm is used to obtain the optimal cut (Boykov and Kolmogorov 2004) of the graph. The graph cut segmentation is computed separately for each tissue and next all results are combined to create a single multilabel image as described in Fortunati *et al* (2013).

Atlas and target images were registered using the open source registration software elastix (Klein *et al* 2010). An affine registration followed by a free form deformation (Rueckert *et al* 1999) were performed in a multiresolution fashion using mutual information as similarity metric (Thévenaz and Unser 2000), and adaptive stochastic gradient descent optimization (Klein *et al* 2008). Details on the registration and intensity model parameters are given in appendix A.

3. Experiments

Three experiments were performed to optimize the method parameters and evaluate its performance. The purpose of the *first experiment* was to evaluate which strategy lead to the most accurate spatial priors. Six strategies were compared: local weighting voting (LW) using (1) SSD, (2) NCC and (3) NMI as image similarity metric; joint weighting voting (JW) using (4) SSD, (5) NCC and (6) NMI as image similarity metric. For each strategy the local similarity patch radius r_p , local search radius r_s , and scaling exponent β were optimized in a leave-one-out cross validation experiment. A pre-defined set of parameters was used for all methods: $r_p \in \{1, 2, 3, 4\}$ voxels; $r_s \in \{0, 1, 2, 3, 4\}$ voxels; $\beta \in \{0.5, 1, 2, \dots, 10\}$. The highest average generalized dice similarity coefficient (Dice 1945) over all images calculated using all tissues

except cerebrum and cerebellum (DSC_{Gen}) was used to select the best parameter set. These tissues were excluded since their volumes are much larger than the volume of the other tissues, and thus would dominate the DSC_{Gen} . The DSC measures the normalized overlap between two volumes and the DSC_{Gen} extends the measure to two sets of volumes. For each segmentation the DSC_{Gen} was defined as:

$$DSC_{Gen} = \frac{2 \sum_{l \in L^*} \hat{S}^l \cap G^l}{\sum_{l \in L^*} \hat{S}^l \cup G^l}, \quad (9)$$

where G^l is the gold standard segmentation of tissue l , $\hat{S}^l$ is the automatic segmentation and L^* is the set of all tissue labels excluding cerebrum and cerebellum. Using the optimal parameters, we evaluated the accuracy of the segmentation of each tissue by DSC. For both tuning and evaluation the manual delineations from O1 were used. The best method (using the best metric) was then selected as the one giving the best overall accuracy in DSC. This method was then used as spatial prior in the graph cut segmentation.

Given the optimal method for computing a spatial prior, the purpose of the *second experiment* was to optimize the parameters λ_1 , λ_2 , and α for the graph cut segmentation. As in Fortunati *et al* (2013), this optimization was done in a leave-one-out cross validation experiment consisting on an exhaustive search over a pre-defined set of parameters: $\lambda_1, \lambda_2 \in \{0.125, 0.25, 0.5, 1, 2, 4, 8\}$ and $\alpha \in \{0.4, 0.5\}$. The delineations of observer O1 were used in this experiment. Additionally, the optimal parameters using the MV to compute the spatial prior were calculated.

The purpose of the *third experiment* was to evaluate the overall performance of the combined methods. Specifically we verified the effectiveness of adding the intensity model for the problem at hand, the influence of the spatial prior on the method accuracy, and the accuracy of each method compared to the human observer ability to consistently segment a certain tissue. For this, the optimal version of the methods resulting from the previous two experiments was employed. The following methods were thus compared:

- multiatlas based majority voting (Hartmann *et al* 1999).
- multiatlas based majority voting combined with intensity model (MV + IM) (Fortunati *et al* 2013).
- Best Local/Joint weighting atlas voting (LW/JW).
- Best Local/Joint weighting atlas voting combined with intensity model (LWIM/JWIM).
- Interobserver agreement/variability (IOA/IOV).

For the evaluation the manually-generated label images from the three observers were used as gold standard. The graph cut method was applied three times using the atlases from each of the observers to build the intensity models and the spatial priors, in a leave-one-out cross-validation scenario. Then the three results were averaged for each patient, to estimate the overall method accuracy. To have a complete quantitative overview of the segmentation accuracy, two evaluation measures were used to compare the segmented volumes: the DSC and the 95th percentile Hausdorff Surface Distance (HSD). This measures the classic Hausdorff distance (Huttenlocher *et al* 1993) but taking the 95th percentile of the distances instead of the maximum.

In the first and third experiment we performed pair-wise comparison between the distribution of accuracies per patients. To assess whether a distribution was higher or lower than another, significance was tested using a one-sided Wilcoxon signed rank test. Significant difference was tested using a two-sided Wilcoxon signed rank test ($p < 0.01$).

Table 1. Optimal leave-one-out parameters for LW and JW.

Method	Metric	Patient	r_p	r_s	β
LW	SSD	5/6	4	2	-2/-3
LW	NCC, NMI	All	4	2	10
JW	SSD/NCC	All	4	2	2/1
JW	NMI	9/2	4	2	1/0.5

4. Results

In the first experiment the parameters of LW and JW were optimized for each of the three similarity metrics. Table 1 reports the leave-one-out optimal parameters for LW and JW using all metrics.

The use of JW slightly improved the DSC of the results when compared to LW. In addition, for both LW and JW, the SSD similarity metric provided better segmentation accuracy than NCC and NMI (see appendix table B1 for details). Therefore JW using SSD, which will be referred simply as JW in the following, was considered the most accurate method and thus it was used for the graph cut segmentation.

The graph cut method was optimized two times: first using the spatial prior that results from MV technique (MVIM) and second using the spatial prior that results from JW based on SSD (JWIM). Table 1 lists the set of optimal parameters $\hat{\theta}_i$ for MVIM and JWIM, considering all patients (no leave-one-out). This parameter set is suggested for an independent test image.

Figures 1 and 2 show the results of experiment 3. Figure 1 shows the distribution of DSC over all patients for all tissues. Figure 2 show the HSD over all patients, for all tissues. The distributions of DSC and HSD over the set of patients between different methods were compared pairwise. The results of statistical analysis are shown in figures 1 (top part) and 2 (bottom part). In addition, we reported for the accuracy of each method whether there is no statistically significant difference with IOA/IOV.

Figure 3 shows examples of segmentation using all methods compared in the third experiments for brain and eye tissues. The first two rows exemplifies the added value in accuracy of using intensity modelling. In the third row, it is shown a case where the appearance of the sclera is not homogeneous leading to a segmentation error when using the intensity model. In the fourth row, the intensity inhomogeneity hampers the accuracy of all methods, however joint fusion is more robust to this artefact.

We calculated the computation times on a 2.3 GHz quad-core Intel processor, with 12 GB of RAM, running a 64 bit Windows 7 operating system running multi-threads implementations for registration and atlas fusion. The complete registration procedure comprising rigid and non-rigid steps took on average 4 min per atlas. The average MV computation time was 45 s/patient, and the average LW and JW computation time was 86 min/patient. The average computation time for the graph cut optimization was 10 min/patient for all tissues. Thus the average segmentation time of a complete dataset considering registrations, JW (or LW), and the graph cut was 136 min/patient.

5. Discussion

In this study we developed and evaluated a multiatlas segmentation method for H&N MR images which combines local weighting multiatlas fusion with intensity modelling using graph cuts. We evaluated which local weighting technique provides the best spatial prior and

Figure 1. Comparison between the accuracy of the four methods and the interobserver agreement (IOA) using DSC. The boxes of the boxplot report first and third quartiles and median values. The whiskers report confidence intervals defined as the most extreme data points (± 1.5 times the inter-quartile range around the box) not considered outliers ($\circ$). Above the boxplot, the results of the statistical analysis are shown. To the left side of this table the two methods under comparison are specified. The symbols Δ and ∇ indicate respectively that the distribution of the second method is statistically significantly ($p < 0.01$) higher and lower than of the first method. The symbols $\square$ indicate that there is no significant difference ($p < 0.01$) between the corresponding distribution and the IOA.

this technique was combined with intensity modelling using graph cut. Next, accuracies of the various automatic methods were compared between each other and with the interobserver agreement/variability of manual delineation.

The accuracy of the proposed method was high both in terms of DSC and of HSD with a median DSC higher than 0.8 for all tissue except sclera, and a median HSD in the order of the slice thickness of 3 mm for all tissues except the myelum, CSF and sclera. For all tissues excluding myelum and CSF, the accuracy of the proposed method is approaching the interobserver agreement/variability. The best local weighting technique was the joint weighting approach using SSD as similarity metric and local patch search to account for registration errors. When a more accurate spatial prior was used, adding the intensity model did not substantially improve or decrease the results. However, using the JWIM method the segmentation

Figure 2. Comparison between the accuracy of the four methods and the interobserver variability (IOV) using HSD. Below the boxplot, the results of the statistical analysis are shown. All symbols have the same meaning as in figure 1 with the difference that we are now reporting the error and not the accuracy, thus lower values are better in this case.

are smoother than with JW and the computation costs are only slightly increased. Thus the use of JWIM is suggested.

In the first experiment we evaluated which multiatlas fusion approach gave the best spatial prior. We found that using JW a small but consistent improvement was obtained. In addition the use of local patch search ($r_s = 2$ in all cases) consistently led to the optimal fusion parameter. Our results are in line with the original publication proposing this approach and they also demonstrate the validity of the method for other similarity metrics, such as NCC and NMI. Although the improvement was small there was no difference in computation time between the regular and joint fusion approach, as observed by Wang *et al* (2012). Thus the use of JW gave only advantages. While the optimal patch radii and local search radii were stable over all methods, the optimal β selection showed that for LW a more non-linear scaling of the similarity metric than for JW was needed for obtaining the best accuracy (note that for LW using SSD β was negative). In LW a more non-linear scaling is equivalent to a more selective weighting. In JW the weights are not directly dependent on the (scaled) similarity metric, as the relationship between the (scaled) similarity metric of different atlases is also considered

Figure 3. Examples of segmentations. Every row shows a comparisons between MV (green), MVIM (red), JW (cyan), JWIM (magenta) and manual delineations (yellow) for different tissues. JW and JWIM shows higher accuracy than the other methods. In (b) and (d) the effect of regularization can be seen when comparing to (a) and (c). The arrow in (l) shows an example of a location where the appearance of the sclera is not homogeneous. (a) Encephalon, MV. (b) Encephalon, MVIM. (c) Encephalon, JW. (d) Encephalon, JWIM. (e) Right eye, MV. (f) Right eye, AB+IM. (g) Right eye, AB. (h) Right eye, AB+IM. (i) Left sclera, MV. (j) Left sclera, MVIM. (k) Left sclera, JW. (l) Left sclera, JWIM. (m) Left eye, MV. (n) Left eye, MVIM. (o) Left eye, JW. (p) Left eye, JWIM.

in the calculation. Therefore for JW a selective weighting could be obtained even with a less non-linear scaling of the similarity metrics.

In the second experiment we optimized the parameters of the graph cut segmentation method using both MV and JW, with the optimal parameters from the first experiment, to

Table 2. Parameters optimization experiment.

Par	CE	CB	BS	CSF	MY	SC_R	EH_R	LE_R	SC_L	EH_L	LE_L
MVIM											
λ_1	0.25	0.25	0.125	0.25	1.00	8.00	4.00	8.00	8.00	4.00	1.00
λ_2	2.00	1.00	2.00	2.00	2.00	1.00	0.50	1.00	0.50	0.125	2.00
α	0.40	0.50	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.50	0.40
JWIM											
λ_1	0.25	0.25	0.125	0.50	0.25	8.00	4.00	8.00	8.00	4.00	4.00
λ_2	4.00	4.00	4.00	4.00	8.00	1.00	1.00	4.00	2.0	0.50	4.00
α	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40

Note: Optimal parameters considering all patients. Note that these parameters were not used for the evaluation but should be used for new patients not included in the atlas set. Abbreviations: Par = parameter; _L = Left; _R = Right.

compute the spatial priors. The leave-one-out parameters selection was stable over patients for all tissues. In addition we found that a better spatial prior model (from the JW) consistently led to a higher value (most often twice as high) of λ_2 , which could be expected since λ_2 is the relative weight between spatial prior model and intensity model.

In the third experiment, we evaluated four different approaches comparing them to the manual reference and to the interobserver agreement/variability. The positive effect of adding the intensity modelling to the multiatlas fusion decreased when a more accurate atlas fusion method was used. Using the intensity model in combination with a spatial prior built using MV led to a significant improvement in DSC for all tissues and a slight decrease of accuracy (see HSD) only for the cerebrum, myelum and left sclera whereas the combination of intensity model (JWIM) and JW significantly improved the accuracy only for the right lens in DSC. For some patients outliers can be observed for the eye tissues both in DSC and HSD. These outliers are caused by signal inhomogeneity due to the use of surface coils (see the last row of figure 3). An improved inhomogeneity correction would prevent these inaccuracies. The HSD shows that for the cerebrum and myelum the use of the intensity model led to a small but statistically significant decrease of the segmentation accuracy. However, for the cerebrum this decrease was within the range of the IOV. Considering all methods, the worse accuracy for both measures was found at the neck level for the myelum and CSF. This is probably due to limitations of the acquisition coils, i.e. lower signal quality was found at the bottom region of the images which led to intensity inhomogeneity and hampered the accuracy of both spatial prior and intensity model. A better acquisition would prevent this effect. The proposed method, both with and without using the intensity model, provided a better accuracy than the MVIM which was previously proposed for a similar application (Fortunati *et al* 2013) but different modality. Visual inspection of the segmentation results confirmed that whereas the JWMI did not improve the accuracy with respect to JW it provided smoother segmentations, which are generally preferred (e.g. as shown in the first row of figure 3).

The effect of segmentation accuracy on HTP quality was previously examined by our group for CT (Verhaart *et al* 2014a) and MRI (Verhaart *et al* 2014b) images. In the first study it was shown that the effect on treatment quality of using CT-based automatic segmentation was comparable with using manual delineations, thus the method is being used in clinical practice. However, manual corrections are currently required for some of the eye tissues. In the second study (Verhaart *et al* 2014b) the HTP quality obtained using a patient model based on CT and MRI was compared to the HTP quality of patient models based on CT only. It was shown that there was no difference between the quality of MRI-based, and CT-based treatment

Table B1. Accuracy of LW and JW evaluated using DSC mean (standard deviation).

ID	LW—SSD	JW—SSD	LW—NCC	JW—NCC	LW—NMI	JW—NMI
Cerebrum	0.971(0.009) ^{a,b}	0.971(0.010) ^b	0.967(0.013)	0.969(0.011)^b	0.968(0.012)	0.968(0.012)
Cerebellum	0.945(0.013) ^{a,b}	0.945(0.014)	0.942(0.015)	0.944(0.014)^b	0.942(0.014)	0.942(0.013)
Brainstem	0.933(0.011)	0.934(0.011)^a	0.933(0.013)	0.933(0.012)	0.934(0.012)	0.934(0.012)
Myelum	0.858(0.096)	0.863(0.091)	0.854(0.111)	0.860(0.096)	0.854(0.113)	0.852(0.116)
CSF	0.818(0.075) ^{a,b}	0.820(0.072) ^{a,b}	0.803(0.076)	0.817(0.072)^b	0.806(0.076) ^a	0.811(0.075)
Sclera_R	0.693(0.047) ^{a,b}	0.687(0.059) ^b	0.660(0.077)	0.686(0.064)^b	0.665(0.075) ^a	0.672(0.073)
Lens_R	0.882(0.047) ^{a,b}	0.876(0.050)	0.813(0.189)	0.865(0.071)	0.841(0.112) ^a	0.853(0.090)
humour_R	0.960(0.014) ^{a,b}	0.960(0.012) ^b	0.953(0.021)	0.960(0.014)^b	0.955(0.017)	0.955(0.015)
Sclera_L	0.681(0.041) ^a	0.680(0.046)	0.653(0.062)	0.680(0.047)	0.661(0.057) ^a	0.670(0.056)
Lens_L	0.859(0.052)	0.857(0.060)	0.767(0.262)	0.842(0.086)	0.787(0.219)	0.827(0.124)
humour_L	0.960(0.008)	0.959(0.011)	0.948(0.026)	0.959(0.011)^b	0.951(0.020)	0.952(0.018)
Gen	0.884(0.027) ^{a,b}	0.886(0.026) ^a	0.878(0.028)	0.885(0.027)^b	0.880(0.028) ^a	0.881(0.028)

Note: Abbreviations: L = left; R = right; Gen = generalized DSC without cerebrum and cerebellum; the statistically significant higher ($p < 0.01$) distribution, between LW and JW, were printed in bold.

^a Indicates whether a distribution was statistically significantly higher than the correspondent, LW or JW, NCC-based method distribution.

^b Indicates whether a distribution was statistically significantly better than the correspondent, LW or JW, NMI-based method distribution.

planning using manual delineations. Moreover in this study we reached an accuracy very close to the inter-observer agreement/variability such that manual correction might be not required. Therefore, we conclude that the proposed method is ready to be used in clinic.

In conclusion we developed and evaluated a segmentation method to accurately segment tissue in MR images for HTP. We compared and evaluated local weighted voting techniques and combination of the best of these techniques with intensity modelling using graph cut. The proposed method provides a better accuracy than the method currently in use for HTP.

Acknowledgments

The present research has been funded by the technology foundation STW (www.stw.nl) with grant number 10846.

Appendix A. Method parameters

The affine registration was performed using four resolution levels and 500 iterations per level. The free form deformation was performed using three resolution levels and a variable number of iterations from the coarsest to the finest resolution level (500, 1000 and 2000), in order to speed up the registration. These parameters were obtained after training on an independent set of 11 T2w images acquired with the same protocol as the segmentation dataset. Twenty two anatomical landmarks (Fortunati *et al* 2014) were manually placed in the images to assess the registration accuracy and therefore enable the selection of the best parameters.

The intensity model was built by random sampling from the set of training images. Five thousands foreground samples were selected randomly from the entire training set for all tissues, except for the sclera (1500) and for the lenses (700) due to the small volume of these tissues. In all cases, the same number of samples was selected for background and foreground where the background region was defined as a band with a width of 5 mm around the foreground.

Appendix B. Atlas fusion strategies

For LW, SSD provided a statistically significantly ($p < 0.01$) better accuracy than both NCC and NMI for all tissue except the brainstem, myelum, left lens and humour. NMI provided a statistically significantly ($p < 0.01$) better accuracy than NCC for myelum, right sclera and lens, left sclera and Gen. For all the other tissues SSD never decreased the accuracy.

For JW, SSD provided a statistically significantly ($p < 0.01$) better accuracy than NCC for the brainstem, CSF, and Gen. SSD also provided more accurate results than NMI: a statistically significantly ($p < 0.01$) better accuracy was found for cerebrum, CSF, right sclera and humour, left humour and Gen. NCC provided a statistically significantly ($p < 0.01$) better accuracy than NMI for cerebrum, cerebellum, CSF, right sclera and humour, left humour and Gen. For all the other tissues the SSD never decreased the accuracy. Table B1 shows mean value and standard deviation for all methods and all tissues.

References

- Ahmed M *et al* 2010 The value of magnetic resonance imaging in target volume delineation of base of tongue tumours—a study using flexible surface coils *Radiother. Oncol.* **94** 161–7
- Akhondi-Asl A and Warfield S K 2013 Simultaneous truth and performance level estimation through fusion of probabilistic segmentations *IEEE Trans. Med. Imaging* **32** 1840–52

- Artechevarria X, Munoz-Barrutia A and Ortiz-de Solorzano C 2009 Combination strategies in multi-atlas image segmentation: application to brain MR data *IEEE Trans. Med. Imaging* **28** 1266–77
- Bakker J F, Paulides M M, Westra A H, Schippers H and Van Rhooon G C 2010 Design and test of a 434 MHz multi-channel amplifier system for targeted hyperthermia applicators *Int. J. Hyperth.* **26** 158–70
- Boykov Y and Kolmogorov V 2004 An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1124–37
- Boykov Y, Veksler O and Zabih R 2001 Fast approximate energy minimization via graph cuts *IEEE Trans. Pattern Anal. Mach. Intell.* **23** 1222–39
- Dice L R 1945 Measures of the amount of ecologic association between species *Ecology* **26** pp 297–302
- Fortunati V, Verhaart R, van der Lijn F, Niessen W J, Veenland J, Paulides M and van Walsum T 2013 Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling *Med. Phys.* **40** 071905
- Fortunati V, Verhaart R, van der Lijn F, Niessen W J, Veenland J, Paulides M and van Walsum T 2014 Feasibility of multi-modal deformable registration for head and neck tumor treatment planning *Int. J. Radiat. Oncol. Biol. Phys.* **90** 85–93
- Franckena M, Fatehi D, de Bruijne M, Canters R A M, van Norden Y, Mens J W, van Rhooon G C and van der Zee J 2009 Hyperthermia dose-effect relationship in 420 patients with cervical cancer treated with combined radiotherapy and hyperthermia *Eur. J. Cancer* **45** 1969–78
- Fritscher K D, Peroni M, Zaffino P, Spadea M F, Schubert R and Sharp G 2014 Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours *Med. Phys.* **41** 051910
- Gordon A et al 2004 Intraobserver variability in the MR determination of tumor volume in squamous cell carcinoma of the pharynx *Am. J. Neuroradiol.* **25** 1092–8
- Han X, Hoogeman M S, Levendag P C, Hibbard L S, Teguh D N, Voet P, Cowen A C and Wolf T K 2008 Atlas-based auto-segmentation of head and neck CT images *MICCAI Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* vol 11 pp 434–41
- Hartmann S L, Parks M H, Martin P R and Dawant B M 1999 Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations: part II, validation on severely atrophied brains *IEEE Trans. Med. Imaging* **18** 917–26
- Hua Y, Ma S, Fu Z, Hu Q, Wang L and Piao Y 2011 Intracavity hyperthermia in nasopharyngeal cancer: a phase III clinical study *Int. J. Hyperth.* **27** 180–6
- Huilgol N G, Gupta S and Dixit R 2010 Chemoradiation with hyperthermia in the treatment of head and neck cancer *Int. J. Hyperth.* **26** 21–5
- Huttenlocher D, Rucklidge W and Klanderman G 1993 Comparing images using the hausdorff distance *IEEE Trans. Pattern Anal. Mach. Intell.* **15** 850–63
- Issels R D et al 2010 Neo-adjuvant chemotherapy alone or with regional hyperthermia for localised high-risk soft-tissue sarcoma: a randomised phase 3 multicentre study *Lancet Oncol.* **11** 561–70
- Klein S, Pluim J P W, Staring M and Viergever M A 2008 Adaptive stochastic gradient descent optimisation for image registration *Int. J. Comput. Vis.* **81** 227–39
- Klein S, Staring M, Murphy K, Viergever M and Pluim J 2010 Elastix: a toolbox for intensity-based medical image registration *IEEE Trans. Med. Imaging* **29** 196–205
- Landman B A and Warfield S K 2012 MICCAI 2012 workshop on multi-atlas labeling *MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling Challenge Results* pp 33–42
- Paulides M M, Bakker J F, Neufeld E, van der Zee J, Jansen P P, Levendag P C and van Rhooon G C 2007 Winner of the new investigator award at the european society of hyperthermia oncology meeting 2007. The HYPERcollar: a novel applicator for hyperthermia in the head and neck *Int. J. Hyperth.* **23** 567–76
- Paulides M M, Bakker J F, Linthorst M, van der Zee J, Rijnen Z, Neufeld E, Pattynama P M T, Jansen P P, Levendag P C and van Rhooon G C 2010 The clinical feasibility of deep hyperthermia treatment in the head and neck: new challenges for positioning and temperature measurement *Phys. Med. Biol.* **55** 2465–80
- Qazi A A, Pekar V, Kim J, Xie J, Breen S L and Jaffray D A 2011 Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach *Med. Phys.* **38** 6160–70
- Rasch C, Steenbakkers R, Fitton I, Duppen J, Nowak P J, Pameijer F, Eisbruch A, Kaanders J, Paulsen F and van Herk M 2010 Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer *Radiat. Oncol.* **5** 5–21

- Rohlfing T, Brandt R, Menzel R, Russakoff D B and Maurer C R Jr 2005 Quo vadis, atlas-based segmentation? *The Handbook of Medical Image Analysis—Volume III: Registration Models* vol 3, ed J Suri et al (New York: Plenum) chapter 11 pp 435–86
- Rueckert D, Sonoda L, Hayes C, Hill D, Leach M and Hawkes D 1999 Nonrigid registration using free-form deformations: application to breast MR images *IEEE Trans. Med. Imaging* **18** 712–21
- Sabuncu M R, Yeo B T T, Van Leemput K, Fischl B and Golland P 2010 A generative model for image segmentation based on label fusion *IEEE Trans. Med. Imaging* **29** 1714–29
- Thévenaz P and Unser M 2000 Optimization of mutual information for multiresolution image registration *IEEE Trans. Image Process.* **9** 2083–99
- Toonkel L, Soila K, Gilbert D and Sheldon J 1988 MRI assisted treatment planning for radiation therapy of the head and neck *Magn. Reson. Imaging* **6** 315–9
- Tustison N J, Cook P A and Gee J C 2010 N4ITK: improved N3 bias correction *IEEE Trans. Med. Imaging* **29** 1310–20
- Udupa J K 1999 On standardizing the MR image intensity scale *Magn. Res. Med.* **42** 1072–81
- Valdagni R A M 1994 Report of long-term follow-up in a randomized trial comparing radiation therapy and radiation therapy plus hyperthermia to metastatic lymph nodes in stage IV head and neck patients *Int. J. Radiat. Oncol. Biol. Phys.* **28** 163–9.
- van der Lijn F, de Bruijne M, Klein S, den Heijer T, Hoogendam Y Y, van der Lugt A, Breteler M M B and Niessen W J 2012 Automated brain structure segmentation based on atlas registration and appearance models *IEEE Trans. Med. Imaging* **31** 276–86
- van der Zee J 2002 Heating the patient: a promising approach? *Ann. Oncol.* **13** 1173–84
- Verhaart R F, Fortunati V, Verduijn G M, van Walsum T, Veenland J F and Paulides M M 2014a CT-based patient modeling for head and neck hyperthermia treatment planning: manual versus automatic normal-tissue-segmentation *Radiother. Oncol.* **111** 158–63
- Verhaart R F, Fortunati V, Verduijn G M, van der Lugt A, van Walsum T, Veenland J F and Paulides M M 2014b The relevance of MRI for patient modeling in head and neck hyperthermia treatment planning: a comparison of CT and CT-MRI based tissue segmentation on simulated temperature *Med. Phys.* **41** 123302
- Wang H, Suh J and Das S 2012 Multi-atlas segmentation with joint label fusion *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 611–29