Head and Neck Auto-segmentation Challenge

Vladimir Pekar[1], Stéphane Allaire[2,4], John Kim[2,3] and David A. Jaffray[2,4]

[1]Philips Research North America, Markham, ON, Canada
[2]Princess Margaret Hospital, Radiation Medicine Program, Toronto, ON, Canada
[3]Dept. of Radiation Oncology, University of Toronto, Toronto, ON, Canada
[4]Dept. of Medical Biophysics, University of Toronto, Toronto, ON, Canada

Abstract

This paper presents the results of the Head and Neck Auto-segmentation Challenge, which was part of the workshop "3D Segmentation in the Clinic: A Grand Challenge". This workshop took place in London, UK, in September 2009, in conjunction with the 12th conference on Medical Image Computing and Computer Assisted Interventions (MICCAI). The aim of the challenge was to evaluate the performance of fully automated algorithms in segmenting the mandible and the brainstem in head and neck CT image data used in radiotherapy planning. We describe the motivation behind the clinical application selected for the challenge, the image data used, and the metrics applied for the quantitative assessment of the segmentation accuracy with respect to the ground truth segmentations provided by a clinical expert. The quantitative evaluation results of the auto-segmentations submitted by the workshop participants are included.

Contents

1 Introduction

Radiation therapy is one of the three principal treatments for cancer besides surgery and chemotherapy. It is based on the principle of damaging DNA of the malignant cells by applying ionizing radiation. External beam radiation treatment planning is a process of setting up the treatment protocol including dose computation and beam placement and is typically done using 3-D computed tomography (CT) image data. Accurate segmentation of the target volumes and risk organs in the patient's image is a crucially important part of the planning procedure. Although some commercial software products allowing for semi- and fully automated segmentation of risk organs have recently become available, their application is limited for many anatomical structures, and the common clinical practice is still 2-D manual contouring in axial slices using standard drawing tools.

The planning of head and neck cancer radiation therapy is especially labor-intensive due to the complexity of the underlying anatomy and the large number of contours that need to be generated. Manual contouring can often require several hours to be spent on a single plan. At the same time, automated segmentation of many organs at risk in the head and neck area is challenging due to poor soft tissue discrimination in CT, artifacts from dental fillings and large variability of patient's anatomy.

The aim of the Head and Neck Auto-segmentation Challenge [1] was to evaluate the performance of state-of-the-art fully automatic segmentation algorithms in CT image data used for radiotherapy planning. It was organized as a part of the "3D Segmentation in the Clinic: The Grand Challenge" workshops series [2, 3, 4] held in conjunction with the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference and was held in London, UK, in September 2009. These workshops have been attracting considerable attention from the scientific community as they provide an excellent testground for systematic and unbiased evaluation of segmentation algorithms with a focus on important clinical applications. The other challenges organized at this year's workshop were devoted to the segmentation of the left ventricle in MR image data [5], carotid lumen segmentation and stenosis grading in CT images [6], and prostate segmentation in MR image data [7].

This paper is organized as follows. Section 2 describes the challenge objectives, presents the image data used for the contest, and introduces the evaluation metrics used for the quantitative assessment of segmentation accuracy. In Section 3, the results of the evaluation study for the data submitted by the participants are discussed. Section 4 concludes the paper.

2 The challenge

2.1 Clinical background

Anatomical structures that need to be contoured in the planning routine include the treatment target volumes and a set of structures at risk. Among the target volumes, a differentiation is made between the gross tumor volumes (GTV) encompassing the visible extents of the disease and regional lymph nodes, the clinical target volumes (CTV) accounting for possible microscopic infiltration into the surrounding tissue, and the planning target volumes (PTV), which add margins to the CTV due to various geometric uncertainties at treatment time, such as patient setup differences, changes in the tumor volume, etc. The definition of the target volumes is usually highly patient specific. In most cases it cannot rely on image information alone and is based on the clinical case, hospital common practice, physician's experience, and other subjective factors. Due to these reasons, automatic contouring of the target volumes is difficult and was not considered for the challenge.

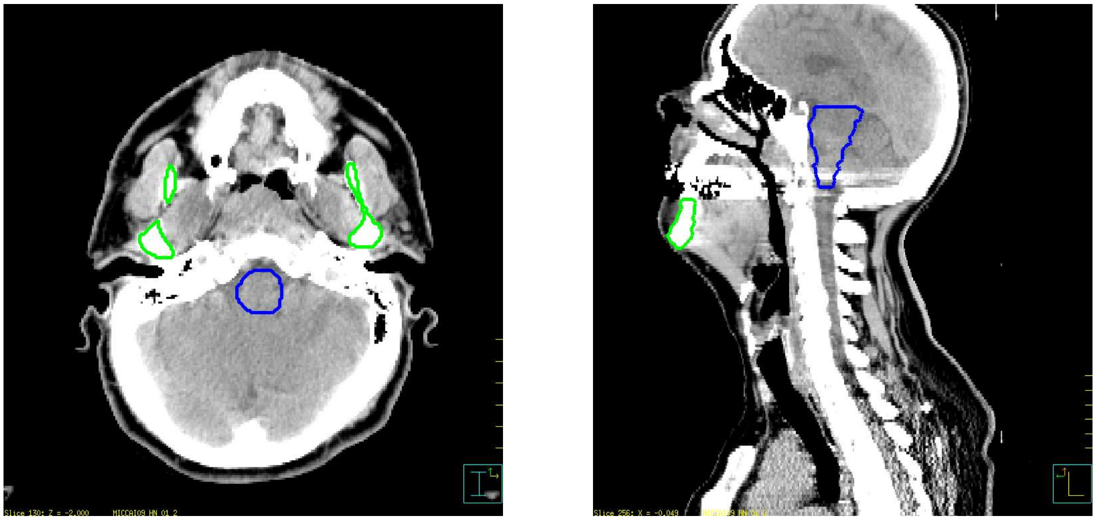

Figure 1: Axial and sagittal slices of a CT dataset used in the challenge with expert delineations of the mandible (green) and brainstem (blue).

Additionally to the target volumes, a set of critical structures at risk must also be contoured. The goal is to incorporate this information into the treatment plan to minimize the dose delivered to the critical structures and in this way minimize radiation induced toxicity. Due to the complexity of the head and neck anatomy, a large number of organs needs to be contoured. Their size and appearance in the image is highly variable across patients. Based on their appearance in CT, anatomical structures can be divided in two groups: high-contrast bones and low-contrast soft tissues. One clinically important representative from each group was selected for the challenge: i) mandibular bone and ii) brainstem. Both structures are always contoured in a head and neck treatment plan and their excessive irradiation can lead to significant morbidity for the patient.

2.2 Image data

All 25 CT datasets used in the challenge were acquired at the Princess Margaret Hospital in Toronto, Canada. The reconstruction matrix for all datasets was 512×512 pixels with the pixel size of approx. 0.98×0.98 mm. The number of slices was in the range of 100-200 slices with the slice thickness of 2 mm.

Manual delineations of the mandible and brainstem were generated by an expert radiation oncologist and stored as a set of contours for visualization purposes and as binary masks for the quantitative evaluation. Fig. 1 shows an axial and a sagittal slice of an exemplary dataset and manual expert delineations of the mandible and the brainstem.

The datasets were organized in 3 groups: 10 datasets could be used by the participants for the training purposes, for which the manual ground truth segmentations were provided; 8 datasets were used for the off-site testing; 7 datasets were used for the online contest.

2.3 Evaluation metrics

The evaluation metrics used in the challenge have been selected to reflect different aspects of segmentation quality assessment for the clinical application in focus. The contours produced by any auto-segmentation algorithm in radiotherapy planning must be reviewed and approved by a clinician in order to be used instead

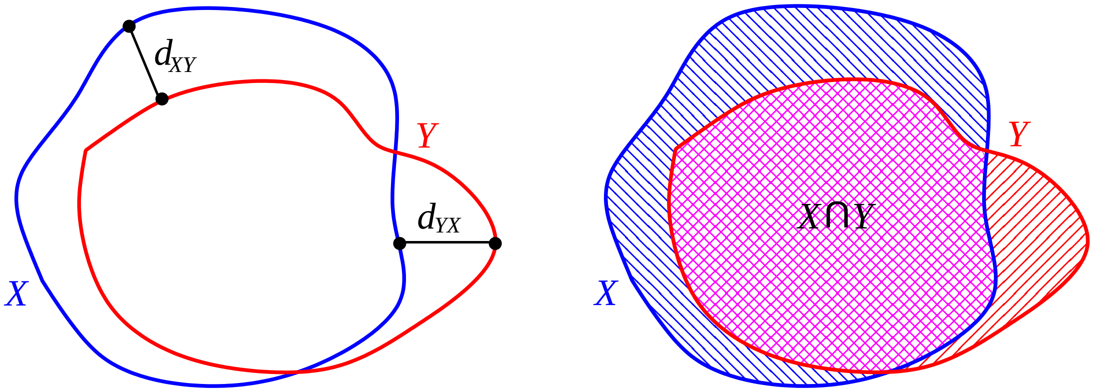

Figure 2: Illustration of the evaluation metrics: Hausdorff distance (left) and the parameters of the Dice metric (right).

of manual delineations, and interactive corrections are often required for the problematic areas. The review and correction process is typically performed, analogously to manual contouring, by inspecting axial slices of the dataset. Thus, one relevant criterion to evaluate the performance of a segmentation algorithm is to estimate the amount of manual interactions that may be needed before accepting the automatically generated segmentation. Another criterion used in the challenge measures the volumetric discrepancies between the automatic and manual ground truth segmentations.

A formal representation of the above criteria can be done by using the following evaluation metrics (see Fig. 2):

2-D Hausdorff distance.

The Hausdorff metric measures the maximum distance of a point in a set to the nearest point in the other set:

$$d_H(X,Y) = \max\{d_{XY}, d_{YX}\} = \max\{\max_{x \in X} \min_{y \in Y} d(x,y), \max_{y \in Y} \min_{x \in X} d(x,y)\}.$$

In the segmentation challenge context, this distance was only computed in the axial slices where the expert manual delineations were present. A large value indicates that the automated segmentation was not accurate in that particular slice. Since deviations below 3 mm are often considered acceptable by the clinicians, the number of slices per dataset with the Hausdorff distance exceeding 3 mm is directly related to the amount of manual corrections required.

The technical implementation of the Hausdorff metric for the challenge was done by computing a Euclidean distance map around the binary masks, where their interior was considered to have the distance value of 0. A particular slice was given a symbolic Hausdorff distance value of -1 when it contained no automatically generated delineation whereas a manual expert delineation existed, thus definitely requiring manual interaction.

Volume overlap (Dice similarity coefficient).

This criterion was used to measure the volumetric overlap between the automatic and manual segmentations represented by binary masks. It is valued from 0 to 1, and is computed as:

$$\kappa = 2 \times \frac{|X \cap Y|}{|X| + |Y|},$$

where $|\cdot|$ is the number of pixels/voxels contained in a region. Analogously to the Hausdorff distance, the

Dice coefficient was also evaluated in axial slices where the manual delineations were present, however the total volume overlap has also been computed.

In summary, the following evaluation results were sent to the groups participating in the challenge:

- Mean 2-D Hausdorff distance
- Median 2-D Hausdorff distance
- Percentage of slices with 2-D Hausdorff distance greater than 3 mm
- Average volume overlap per slice
- Median volume overlap per slice
- Total volume overlap
- 2-D Hausdorff distance and volume overlap for each axial slice

2.4 Participating groups

After the data had been made available to the public, twelve groups expressed their interest in participating in the challenge, five groups managed to submit the off-site results according to the announced deadline and four groups were able to participate in the on-site contest in London, representing three academic institutions and one company. The groups who returned the results and submitted the papers describing the underlying algorithmic solution are the following:

1. University of Manchester [8]
2. Federal Polytechnic University of Lausanne [9]
3. CMS Software / Elekta [10]
4. Konrad Zuse Institute Berlin [11]
5. Institute of Automation, Chinese Academy of Science [12]

The top four groups were able to take part in the on-site contest.

3 Results

The results were evaluated separately for the mandible and brainstem. A common feature of the segmentation methods applied by the participants included active appearance models [8], as well as various atlas registration and refinement strategies [9, 10, 12]. The group from the Zuse Institute Berlin was using a dedicated model-based bone segmentation algorithm [11] and was only able to participate in the mandible segmentation contest. The quantitative results for the off-site evaluation can be found in the respective above referenced publications.

The quantitative results for the four groups are summarized in Tables 1-8, where the best value is shown in boldface. Depending on the result per dataset, a rank from 1 to 4 was assigned. The three criteria used to assign the ranks were: i) median 2-D Hausdorff distance, ii) percentage of slices with 2-D Hausdorff distance above 3 mm, and iii) total volume overlap. The final ranks were computed by summing up all the ranks for the three criteria and dividing by the number of datasets multiplied by 3. Note that although the resulting values for the first two ranks in the mandible segmentation contest are different, the quantitative values for the 2-D Hausdorff distance and volume overlap are very close and would obviously fall in the range of inter-user variability.

It can be seen from the results that segmentation of soft tissue organs, such a brainstem, is still challenging in CT data due to poor contrast. Even for the best performing algorithm, the user would still need to correct about half of the slices. Auto-segmentation of bones is, on the other hand, more feasible, and the segmentation accuracy of approximately two thirds of all slices can be deemed acceptable.

4 Conclusion

We have presented the evaluation framework and quantitative results of applying fully automated algorithms to segment the mandible and the brainstem in head and neck CT image data used for radiotherapy planning. The presented results have been obtained in the online setting at the Grand Challenge Workshop organized as part of the MICCAI 2009 conference in London, UK.

Dataset #	U Manchester	EPF Lausanne	CMS	ZIB
1	7.81	13.88	**2.18**	2.76
2	8.79	13.57	**2.18**	**2.18**
3	6.25	7.44	**2.18**	**2.18**
4	6.18	3.23	**2.18**	**2.18**
5	4.88	3.09	**1.95**	2.18
6	4.88	11.13	**2.07**	2.18
7	6.84	10.71	**2.18**	2.47

Table 1: Median 2-D Hausdorff distance (mm) for the mandible segmentation.

Dataset #	U Manchester	EPF Lausanne	CMS	ZIB
1	100%	100%	40%	**33%**
2	92%	100%	39%	**32%**
3	100%	100%	**26%**	36%
4	100%	50%	38%	**30%**
5	90%	52%	**21%**	24%
6	95%	100%	35%	**28%**
7	100%	100%	**32%**	**32%**

Table 2: Percentage of slices with 2-D Hausdorff distance above 3 mm for the mandible segmentation.

Dataset #	U Manchester	EPF Lausanne	CMS	ZIB
1	0.83	0.68	**0.93**	**0.93**
2	0.84	0.71	0.92	**0.93**
3	0.80	0.75	**0.93**	0.91
4	0.80	0.90	**0.93**	**0.93**
5	0.83	0.85	**0.94**	0.93
6	0.81	0.71	**0.93**	**0.93**
7	0.84	0.69	**0.94**	0.93

Table 3: Total volume overlap for the mandible segmentation.

U Manchester	EPF Lausanne	CMS	ZIB
3.29	3.57	**1.24**	1.43

Table 4: Final ranks for the mandible segmentation.

Dataset #	U Manchester	EPF Lausanne	CMS	ZIB
1	5.86	6.18	**2.18**	-
2	6.69	5.86	**3.30**	-
3	9.42	5.52	**2.76**	-
4	4.14	5.69	**2.93**	-
5	7.81	4.08	**1.95**	-
6	3.09	4.98	**3.01**	-
7	8.79	6.54	**3.30**	-

Table 5: Median 2-D Hausdorff distance (mm) for the brainstem segmentation.

Dataset #	U Manchester	EPF Lausanne	CMS	ZIB
1	100%	97%	**43%**	-
2	100%	96%	**59%**	-
3	93%	100%	**29%**	-
4	81%	97%	**48%**	-
5	100%	85%	**15%**	-
6	62%	77%	**50%**	-
7	100%	93%	**60%**	-

Table 6: Percentage of slices with 2-D Hausdorff distance above 3 mm for the brainstem segmentation.

Dataset #	U Manchester	EPF Lausanne	CMS	ZIB
1	0.72	0.80	**0.83**	-
2	0.64	0.73	**0.85**	-
3	0.67	0.77	**0.90**	-
4	0.88	0.76	**0.90**	-
5	0.65	0.83	**0.92**	-
6	0.87	0.80	**0.89**	-
7	0.59	0.69	**0.88**	-

Table 7: Total volume overlap for the brainstem segmentation.

U Manchester	EPF Lausanne	CMS	ZIB
2.67	2.33	**1.00**	-

Table 8: Final ranks for the brainstem segmentation.

References

[1] http://www.grand-challenge2009.ca. 1

[2] http://mbi.dkfz-heidelberg.de/grand-challenge2007. 1

[3] http://grand-challenge2008.bigr.nl. 1

[4] http://grand-challenge2009.bigr.nl. 1

[5] http://smial.sri.utoronto.ca/LV_Challenge/Home.html. 1

[6] http://cls2009.bigr.nl. 1

[7] http://wiki.na-mic.org/Wiki/index.php/2009_prostate_segmentation_challenge_MICCAI. 1

[8] K. Babalola and T. Cootes. AAM segmentation of the mandible and brainstem. *The MIDAS Journal*, 2009. http://hdl.handle.net/10380/3097. 1, 3

[9] S. Gorthi, V. Duay, M. Bach Cuadra, P. Tercier, A.S. Allal, and J. Thiran. Active contour-based segmentation of head and neck with adaptive atlas selection. *The MIDAS Journal*, 2009. http://hdl.handle.net/10380/3092. 2, 3

[10] X. Han, L. Hibbard, N. O'Connel, and V. Willcut. Automatic segmentation of head and neck CT images by GPU- accelerated multi-atlas fusion. *The MIDAS Journal*, 2009. http://hdl.handle.net/10380/3111. 3, 3

[11] D. Kainmueller, H. Lamecker, H. Seim, and S. Zachow. Multi-object segmentation of head bones. *The MIDAS Journal*, 2009. http://hdl.handle.net/10380/3099. 4, 3

[12] X. Zhang, J. Tian, Y. Wu, J. Zheng, and K. Deng. Segmentation of head and neck CT scans using atlas-based level set method. *The MIDAS Journal*, 2009. http://hdl.handle.net/10380/3094. 5, 3