Head and Neck Auto-segmentation Challenge: Segmentation of the Parotid Glands

Vladimir Pekar¹, Stéphane Allaire², ⁴, Arish A. Qazi², ⁴, John J. Kim², ³ and David A. Jaffray², ⁴

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

Abstract. This paper presents the results of the Head and Neck Auto-segmentation Challenge, a part of the workshop, “Medical Image Analysis in the Clinic: A Grand Challenge”, held in conjunction with the 13th International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), Beijing, China, in September 2010. The aim of the challenge was to evaluate the performance of fully automated algorithms in segmenting the parotid glands 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.

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 [1]. 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, where 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. On the other hand, 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 purpose of the Head and Neck Auto-segmentation Challenge is evaluating the performance of state-of-the-art fully automatic segmentation algorithms in CT image data used for radiotherapy planning. It is organized as a part of “The Grand Challenge” series of workshops series [2] held in conjunction with the Medical Image Computing and Computer Assisted Interventions (MICCAI). 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 first Head and Neck Auto-segmentation Challenge took place in London, UK, in September 2009 [3]. The present paper describes the offsite evaluation results for the contributions submitted to the second Head and Neck Auto-segmentation Challenge held in Beijing, China, in September 2010.

The organization of this paper follows the structure of the paper summarizing the on-site results of the previous event [3]: 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 offsite 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 quite challenging.

In addition 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, thus minimizing the dose delivered to the critical structures and leading to reduced 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.

For the first Head and Neck Auto-segmentation Challenge, which took place
in 2009, two prominent organs at risk were selected: i) mandibular bone and
ii) brainstem, which is a soft tissue structure. Both organs are always contoured
in a head and neck treatment plan and their excessive irradiation can lead to
significant morbidity for the patient. The submitted segmentation results demon-
strated that it was potentially possible to reduce the contouring time by applying
fully automatic algorithms to segment the mandible, where the segmentation ac-
ccuracy of approximately two thirds of all slices could be deemed acceptable. In
contrast, the user would need to correct about half of the slices of the brainstem
after applying auto-segmentation [3].

In this year’s challenge, the complexity of the task has been increased, and
the participants were invited to develop fully automatic solutions to segment the
parotid glands. These bilateral organs are responsible for the salivary function
and represent important structures at risk contoured in most of the head and
neck treatment plans [4]. The shape of the parotid glands is rather irregular with
concavities, and their gray-value appearance in CT data spans the intensity range
between fat and muscle but can be inhomogeneous and highly affected by dental
artifacts. All these factors make their reliable fully automatic segmentation very
challenging.

2.2 Image data

The datasets used were the planning CT images of 25 anonymized patients ac-
quired at the Princess Margaret Hospital in Toronto, Canada, and were identical
to the 2009 challenge. The reconstruction matrix for all datasets was 512×512
pixels with the pixel size of approximately 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 both left and right parotid glands were generated by
an expert radiation oncologist and stored as a set of axial contours for visual-
ization purposes and as binary masks for the quantitative evaluation.

Analogously to the 2009 challenge, the datasets were organized in 3 groups:
10 datasets could be used by the participants for training purposes, for which
the manual ground truth segmentations were provided; 8 datasets were used for
the offsite testing; 7 datasets were left for the online contest.

2.3 Evaluation metrics

The evaluation metrics used in the challenge have been selected to reflect dif-
f erent aspects of segmentation quality assessment for the clinical application in
focus. The contours produced by any auto-segmentation algorithm in radiother-
apy planning must be reviewed and approved by a clinician in order to be used
instead 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 an automatic segmentation
algorithm is to estimate the amount of manual interactions that may be needed
before accepting the segmentation. Another criterion used in the challenge mea-
sures the volumetric discrepancies between the automatic and manual ground
truth segmentations.

A formal representation of the above criteria can be done by using the fol-
lowing evaluation metrics (see Fig. 1):

**Slice-wise 2-D symmetric 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 clin-
icians, 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 extracted boundary
of the binary masks. 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 auto-
matic 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 slice-wise Hausdorff distance, the Dice coefficient was also evaluated only

---
in axial slices where the manual delineations were present, however, the total volume overlap was also computed in the whole image.

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
- Number of slices with 2-D Hausdorff distance greater than 3 mm
- Average volume overlap per slice
- Median volume overlap per slice
- Total volume overlap

2.4 Participating groups

After the data had been made available to the public, 19 groups expressed their interest in participating in the challenge, 7 groups managed to submit the offsite results according to the announced deadline and 6 groups submitted papers for the workshop. The groups who returned the results and submitted the papers describing the underlying algorithmic solution are the following:

1. CMS Software / Elekta, St. Louis, MO – CMS
2. Technical University of Denmark, Copenhagen, Denmark – DTU
3. Federal Polytechnic University of Lausanne, Switzerland – EPFL
5. MD Anderson Cancer Center, Houston, TX – MDACC
6. TomoTherapy, Madison, WI

3 Results

A common methodological background shared between all submissions consists of non-rigid registration between the testing dataset and a pre-segmented atlas, transferring segmentations from the atlas, and applying refinement. This includes application of STAPLE approach [5] to merge individual segmentations derived by registering the dataset to multiple atlases (CMS, MDACC, EPFL), registration of multiple segmentations with the test dataset through an intermediate averaged atlas and using intensity-based voting system to combine the segmentations into the final result (INRIA), using level set speed function to constrain a deformable registration framework and several optimal atlas selection strategies (EPFL), and applying level set refinement of registration results (DTU). TomoTherapy has proposed an original approach based on a combination of global and local priors in a pre-classified region of interest obtained by registration of a probabilistic atlas and slice-wise propagation of segmentation. A detailed description of each method can be found in the respective workshop paper.
The variability between automatic segmentations yielded by each individual method is shown in Fig. 2. The quantitative results are summarized in Tables 1-5, where the best value is shown in boldface. Dataset No. 13 was acquired with a truncated field of view and different neck flexion, and not all methods were able to cope with this issue. More details of the offsite evaluation can be found in the respective paper in these proceedings.

To compute the final rank for each group, the following approach has been applied. Depending on the result per dataset, 2 ranks from 1 to 6 were assigned for each of the 2 parotid glands. The two criteria used to assign the ranks were: i) median 2-D Hausdorff distance and ii) total volume overlap (Dice metric). The final ranks were computed by averaging the results for both left and right parotid.
In general, it can be noted that the quantitative results for the left and right parotid gland correlate well, e.g. the best volume overlap value for each dataset is about 0.83 which can be considered as a good result in terms of overall agreement with the ground truth. On the other hand, as reported by the workshop papers, 2-D Hausdorff distance exceeds 3 mm in more than 75% of slices even in the best case, which is strong indication that the majority of axial slices would still require manual corrections.

### 4 Conclusion

We have presented the evaluation framework and quantitative results of applying fully automated algorithms to segment the parotid glands in head and neck CT image data in the context of the Head and Neck Auto-Segmentation Challenge organized as part of the MICCAI conference in Beijing, China in September 2010.
### Table 3. Median 2-D Hausdorff distance in mm (minimum in boldface) for the right parotid.

<table>
<thead>
<tr>
<th>Dataset #</th>
<th>CMS</th>
<th>DTU</th>
<th>EPFL</th>
<th>INRIA</th>
<th>MDACC</th>
<th>TomoTherapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>6.25</td>
<td>8.05</td>
<td><strong>5.69</strong></td>
<td>6.69</td>
<td>6.69</td>
<td>8.73</td>
</tr>
<tr>
<td>12</td>
<td>5.69</td>
<td>16.00</td>
<td>9.62</td>
<td><strong>5.61</strong></td>
<td>6.87</td>
<td>12.82</td>
</tr>
<tr>
<td>13</td>
<td><strong>5.78</strong></td>
<td>10.74</td>
<td>7.01</td>
<td>40.06</td>
<td>6.40</td>
<td>12.54</td>
</tr>
<tr>
<td>14</td>
<td>5.25</td>
<td>12.70</td>
<td>9.81</td>
<td><strong>4.63</strong></td>
<td>5.69</td>
<td>10.38</td>
</tr>
<tr>
<td>15</td>
<td><strong>5.27</strong></td>
<td>9.67</td>
<td><strong>5.27</strong></td>
<td>7.03</td>
<td>6.18</td>
<td>13.36</td>
</tr>
<tr>
<td>16</td>
<td>3.09</td>
<td>8.37</td>
<td>6.02</td>
<td>4.88</td>
<td>6.91</td>
<td>13.81</td>
</tr>
<tr>
<td>17</td>
<td>3.91</td>
<td>19.53</td>
<td>10.28</td>
<td>6.91</td>
<td>8.30</td>
<td>8.73</td>
</tr>
<tr>
<td>18</td>
<td><strong>3.73</strong></td>
<td>28.44</td>
<td>5.90</td>
<td>5.90</td>
<td>4.17</td>
<td>6.52</td>
</tr>
</tbody>
</table>

### Table 4. 3-D volume overlap (maximum in boldface) for the right parotid.

<table>
<thead>
<tr>
<th>Dataset #</th>
<th>CMS</th>
<th>DTU</th>
<th>EPFL</th>
<th>INRIA</th>
<th>MDACC</th>
<th>TomoTherapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td><strong>0.881</strong></td>
<td>0.809</td>
<td>0.817</td>
<td>0.857</td>
<td>0.852</td>
<td>0.774</td>
</tr>
<tr>
<td>12</td>
<td><strong>0.849</strong></td>
<td>0.585</td>
<td>0.717</td>
<td>0.835</td>
<td>0.823</td>
<td>0.649</td>
</tr>
<tr>
<td>13</td>
<td>0.828</td>
<td>0.721</td>
<td>0.779</td>
<td>0.040</td>
<td><strong>0.837</strong></td>
<td>0.719</td>
</tr>
<tr>
<td>14</td>
<td>0.853</td>
<td>0.601</td>
<td>0.681</td>
<td><strong>0.876</strong></td>
<td>0.870</td>
<td>0.723</td>
</tr>
<tr>
<td>15</td>
<td>0.881</td>
<td>0.817</td>
<td>0.844</td>
<td><strong>0.883</strong></td>
<td>0.871</td>
<td>0.741</td>
</tr>
<tr>
<td>16</td>
<td><strong>0.868</strong></td>
<td>0.667</td>
<td>0.738</td>
<td>0.826</td>
<td>0.813</td>
<td>0.507</td>
</tr>
<tr>
<td>17</td>
<td><strong>0.838</strong></td>
<td>0.573</td>
<td>0.675</td>
<td>0.829</td>
<td>0.818</td>
<td>0.644</td>
</tr>
<tr>
<td>18</td>
<td><strong>0.882</strong></td>
<td>0.202</td>
<td>0.802</td>
<td>0.845</td>
<td>0.860</td>
<td>0.772</td>
</tr>
</tbody>
</table>

### Table 5. Final average ranks.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS</td>
<td><strong>1.38</strong></td>
</tr>
<tr>
<td>INRIA</td>
<td>2.66</td>
</tr>
<tr>
<td>MDACC</td>
<td>2.66</td>
</tr>
<tr>
<td>EPFL</td>
<td>3.75</td>
</tr>
<tr>
<td>TomoTherapy</td>
<td>5.09</td>
</tr>
<tr>
<td>DTU</td>
<td>5.34</td>
</tr>
</tbody>
</table>

### References

2. [http://www.grand-challenge.org](http://www.grand-challenge.org)