Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure

Olivier Commowick, Christian Barillot

To cite this version:


HAL Id: inserm-01895603
https://www.hal.inserm.fr/inserm-01895603
Submitted on 15 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
MSSEG Miccai 2016 Challenge:
Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure

Olivier Commowick, Christian Barillot and FLI / OFSEP

Workshop FLI-SFR – October 11, 2018


Background: multiple sclerosis

- Highly variable evolution
- Clinical classification in 4 types
- Two main stages
  - Early: variable evolution
  - Later: parallel evolution

Lesion segmentation in MS

- Lesion load and lesion count crucial in MS
  - Part of diagnosis (McDonald criteria)
  - Evaluation of drug effectiveness

- Delineation of lesion tedious
  - Manual → time consuming
  - Subject to intra- / inter-individual variability

⇒ Automatic segmentation is key

Why a segmentation challenge?

• A huge number of automatic segmentation methods
  • Tissue classification & outlier detection
  • Machine learning (random forests, deep, etc.)
  • Many others

• Large variety of modalities used
  • T1, T2, FLAIR, PD…

• Large variety of implementations
  • GPU, Matlab, Python, C++ …

5 surveys in the last 5 years involving 50+ methods
Why a segmentation challenge?

- Evaluation complicated
  - Each method evaluated on a specific set
  - No comparison possible

- The challenge concept
  - Have all methods evaluated on a common dataset

- Main drawbacks
  - Possibility to adapt parameters to each patient
  - Ground truth not well defined

Styner et al., 2008. 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. Insight journal.
An OFSEP and FLI challenge @ MICCAI

• Evaluation objectives
  • Evaluate algorithms developed in the community
  • In a well defined computational framework (FLI)
    • Same set of parameters for all images
    • With respect to a solid ground truth

• Additional objectives (OFSEP)
  • Evaluate lesion segmentation algorithms for MS
  • Fully automatic, on standardized images
    • Standardized but different centers

http://www.ofsep.org
MICCAI challenge: The Data

- Challenge data
  - 53 patients from 4 different scanners
  - Modalities: 3DFLAIR, T2/DP, 3DT1, 3DT1-Gado
    - OFSEP consensus
  - 7 manual segmentations for each patient

- Two datasets drawn
  - Training (open): challengers tune their algorithms
  - Testing (closed): evaluation database

<table>
<thead>
<tr>
<th>Center / #exams</th>
<th>Training set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 - Siemens Verio 3T (Rennes)</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>03 - GE Discovery 3T (Bordeaux)</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>07 - Siemens Aera 1.5T (Lyon)</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>08 - Philips Ingenia 3T (Lyon)</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>15</strong></td>
<td><strong>38</strong></td>
</tr>
</tbody>
</table>
Dataset examples (with experts consensus)

FLAIR from center 01

FLAIR from center 03
Not in the Training

FLAIR from center 07

FLAIR from center 08
A well defined execution and evaluation framework

- Pipelines provided by the challengers
  - Black box (docker) including their optimal parameters
  - Parameters chosen or optimized on training set

- Pipelines started automatically on testing set
  - On France Life Imaging (FLI-IAM) computing platform
  - By FLI-IAM project engineers
  - Ensures a uniform set of parameters on the whole testing database

https://portal.fli-iam.irisa.fr/msseg-challenge/overview
France Life Imaging computing platform
Challenge participations

- Thirteen pipelines including a variety of algorithms
  - Machine learning:
    - Random forests
    - Deep learning
  - Model Inference (Bayes, Markov, …):
    - Tissue classification approaches
- Training phase: 2 months \((at \ home)\)
- Integration phase: 3 to 4 months \(on \ FLI-IAM \ system\)
  - Docker packaging and integration help by FLI
- Evaluation (independent from challengers): 2 months
Which evaluation? Metric categories

- Evaluation of MS lesions segmentation: tough topic
  - Which ground truth? → LOP STAPLE consensus
  - What is of interest to the clinician?

- Two metric categories:
  - Detection: are the lesions detected, independently of the precision of their contours? → $F_1$ score
  - Segmentation: are the lesions contours exact?
    - Overlap → $Dice$ score
    - Surface-based measures → $Mean$ surface distance

https://portal.fli-iam.irisa.fr/msseg-challenge/evaluation
### No lesion case results

<table>
<thead>
<tr>
<th>Evaluated method</th>
<th>Lesion volume (cm$^3$)</th>
<th>Number of lesions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team 1</td>
<td>8.25</td>
<td>18</td>
</tr>
<tr>
<td>Team 2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Team 3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Team 4</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Team 5</td>
<td>28.44</td>
<td>522</td>
</tr>
<tr>
<td>Team 6</td>
<td>0.47</td>
<td>7</td>
</tr>
<tr>
<td>Team 7</td>
<td>5.99</td>
<td>168</td>
</tr>
<tr>
<td>Team 8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Team 9</td>
<td>2.55</td>
<td>33</td>
</tr>
<tr>
<td>Team 10</td>
<td>11.09</td>
<td>31</td>
</tr>
<tr>
<td>Team 11</td>
<td>3.44</td>
<td>42</td>
</tr>
<tr>
<td>Team 12</td>
<td>0.06</td>
<td>1</td>
</tr>
<tr>
<td>Team 13</td>
<td>0.07</td>
<td>4</td>
</tr>
</tbody>
</table>
Visual results for center 01

Consensus

Team 7

Team 8

Team 9

Team 10

Team 11

Team 12

Team 13
Visual results for center 03 *(not in the training phase)*
Groups of methods: Comparison to Experts

The diagram shows a scatter plot comparing experts' performance to that of automatic methods. The x-axis represents the Dice score, while the y-axis represents the F1 score. There are three groups represented:

- **Experts**: The experts' data points are highlighted, showing a range from a Dice score of 0.6 to 0.7 and an F1 score of 0.5 to 0.8.
- **Consensus of Automatic**: This group shows a consensus of automatic methods, with a Dice score ranging from 0.4 to 0.6 and an F1 score from 0.4 to 0.7.
- **Automatic #1**: This group includes a smaller set of automatic methods, with a Dice score from 0.3 to 0.5 and an F1 score from 0.3 to 0.5.
- **Automatic #2**: This group has a wider range, with a Dice score from 0.2 to 0.4 and an F1 score from 0.2 to 0.5.

The graph also divides the data into three groups: Group 1, Group 2, and Group 3, based on their performance metrics.
Segmentation performance vs lesion load

Average Dice as a function of total lesion load

$R^2 = 0.82197$
Take home messages from the challenge

• Standardized acquisitions necessary for MS
  • Yet differences remain
  • Need for large database with many expert delineations (i.e. big issue in medical imaging)

• Automatic computing platform
  • Great tool for
    • challenges organization
    • Open Science
    • Certification of algorithms (e.g. industrial solutions)
  • Fair comparison → no parameter tuning during test
  • No work from challengers after pipeline integration

• Main results
  • Individual algorithms still trailing behind experts
  • Unknown images lead to more failures