Overview of the VISCERAL Challenge at ISBI 2015

Orcun Goksel[1], Antonio Foncubierta-Rodríguez[1], Oscar Alfonso Jiménez del Toro[2], Henning Müller[2], Georg Langs[3], Marc-André Weber[4], Bjoern Menze[5], Ivan Eggel[2], Katharina Gruenberg[4], Marianne Winterstein[4], Markus Holzer[3], Markus Krenn[3], Georgios Kontokotsios[6], Sokratis Metallidis[2], Roger Schaer[2], Abdel Aziz Taha[6], András Jakab[3], Tomàs Salas Fernandez[7], Allan Hanbury[6]

ETH Zürich, Switzerland[1]; HES-SO Valais, Sierre, Switzerland[2]; MUW, Vienna, Austria[3]; University of Heidelberg, Germany[4]; TUM, Munich, Germany[5]; TUWien, Vienna, Austria[6]; AQuAS, Barcelona, Spain[7]

Abstract

This is an overview paper describing the data and evaluation scheme of the VISCERAL Segmentation Challenge at ISBI 2015. The challenge was organized on a cloud-based virtual-machine environment, where each participant could develop and submit their algorithms. The dataset contains up to 20 anatomical structures annotated in a training and a test set consisting of CT and MR images with and without contrast enhancement. The test-set is not accessible to participants, and the organizers run the virtual-machines with submitted segmentation methods on the test data. The results of the evaluation are then presented to the participant, who can opt to make it public on the challenge leaderboard displaying 20 segmentation quality metrics per-organ and per-modality. Dice coefficient and mean-surface distance are presented herein as representative quality metrics. As a continuous evaluation platform, our segmentation challenge leaderboard will be open beyond the duration of the VISCERAL project.

1 Introduction

In this challenge, a set of annotated medical imaging data was provided to the participants, along with a powerful complimentary cloud-computing instance (8-core CPU with 16GB RAM) where participant algorithms can be developed and evaluated. The available data contains segmentations of several different anatomical structures in different image modalities, e.g.ĊT and MRI. Annotated

In: O. Goksel (ed.): Proceedings of the VISCERAL Anatomy Grand Challenge
at the 2015 IEEE International Symposium on Biomedical Imaging (ISBI), New York, NY, Apr 16th, 2015
published at http://ceur-ws.org

structures in the training and testing data corpus included the segmentations of left/right kidney, spleen, liver, left/right lung, urinary bladder, rectus abdominis muscle, 1st lumbar vertebra, pancreas, left/right psoas major muscle, gallbladder, sternum, aorta, trachea, left/right adrenal gland.

As training, 20 volumes each were provided for four different image modalities and field-of-views, with and without contrast enhancement, which add up to 80 volumes in total. In each volume, up to 20 structures were segmented. The missing annotations are due to poor visibility of the structures in certain image modalities or due to such structures being outside the field-of-view. Accordingly, in all 80 volumes, a total of 1295 structures are segmented. A breakdown of annotations per anatomy can be seen in Figure 1.

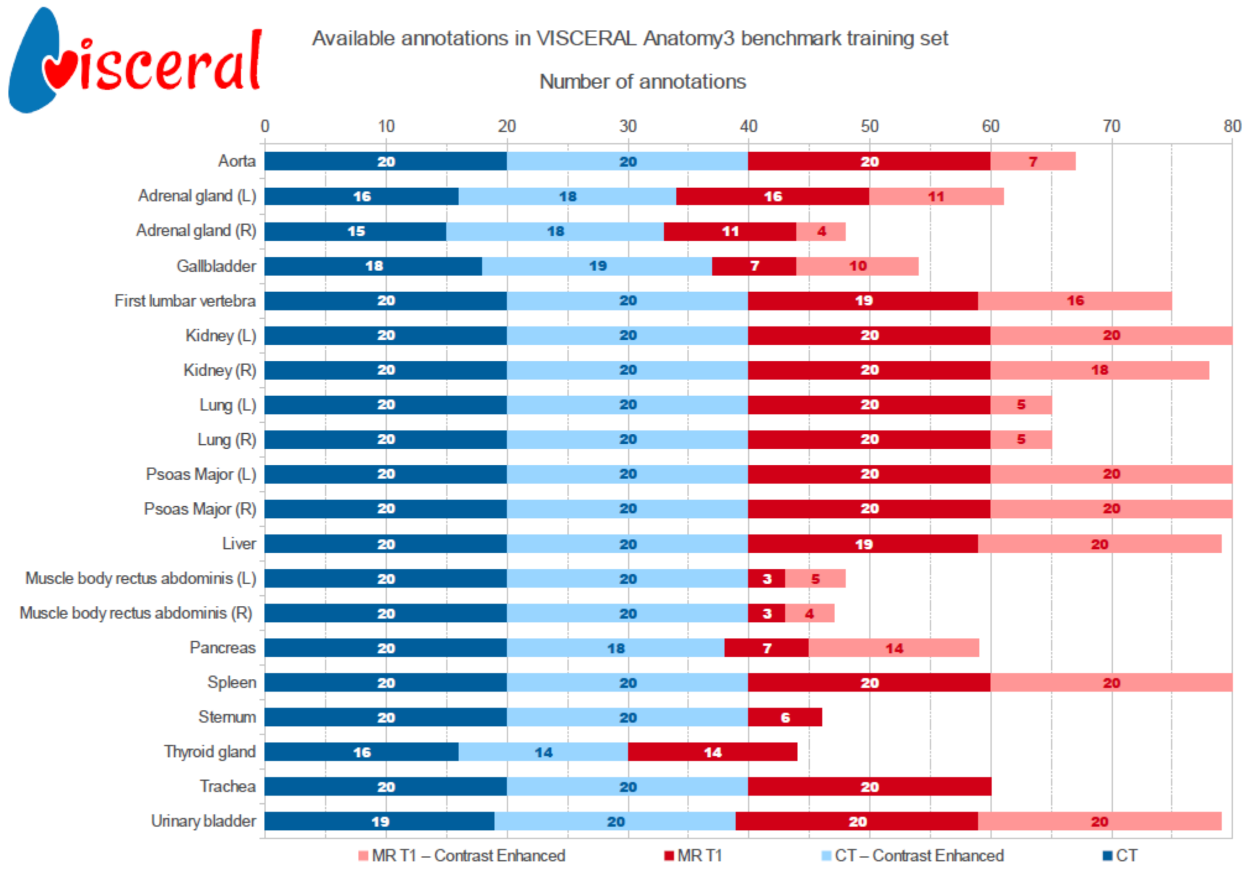


Figure 1: Number of annotations in the Anatomy3 training set classified by modality and organ.

Participants did not need to segment all the structures involved in such data, but rather they could attempt any single anatomical structure or a combination thereof. For instance, an algorithm that could segment only some organs in some of the modalities was evaluated only in those categories for which it outputted any results. Accordingly, our evaluation results were presented in a per-anatomy, per-modality fashion depending on the attempted segmentation task/s by each participating algorithm. This is, indeed, in line with the VISCERAL vision of creating a single, large, and multi-purpose medical image dataset, on which different research groups can test their specific applications and solutions.

Participants first registered for a benchmark account at the VISCERAL registration website. Among the options during the registration, they could request their choice of operating system (Linux, Windows, etc) for the virtual machine (VM), in order to get access to the VM and the data. Having signed the data usage agreement and uploaded it to the participant dashboard, they could then access the VM for algorithm development and also use the training data accessible therein. Participants could additionally download the training dataset via FTP for offline training.

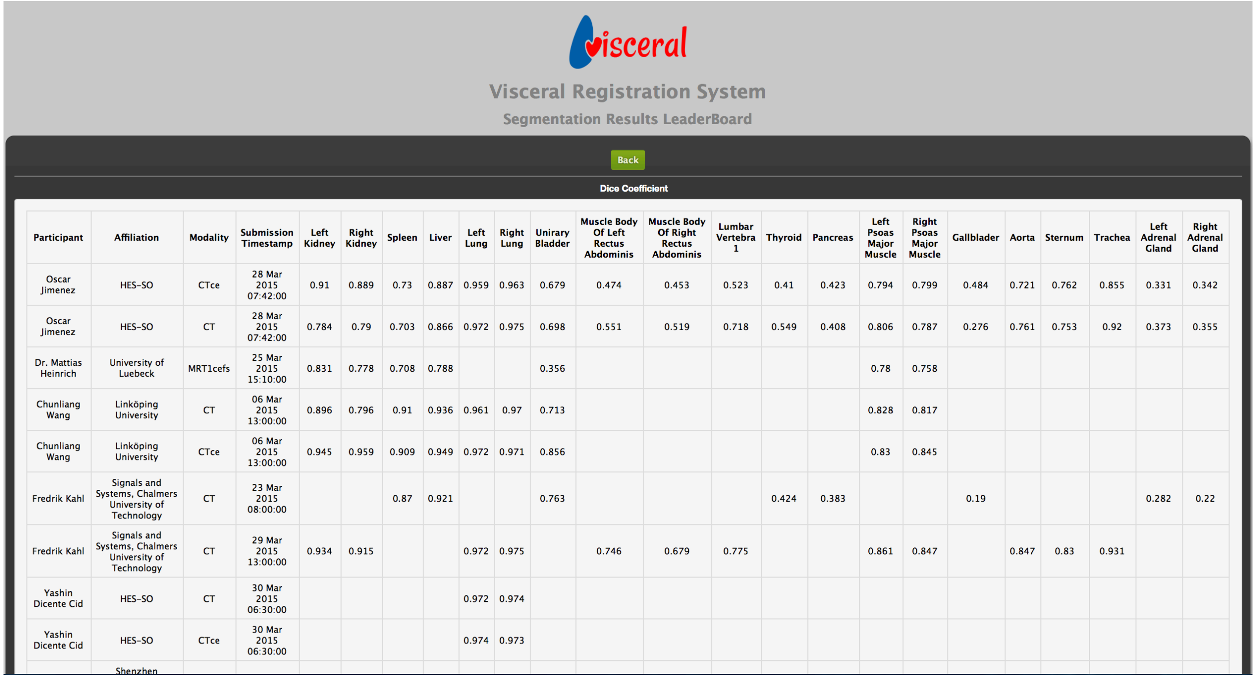
Participants accordingly developed and installed their algorithms in the VM, while adapting and testing them on the training data. They then prepared their executable on the VM according to the input/output specifications announced by us earlier in the Anatomy3 Guidelines for Participation, and submitted their VMs (through "Submit VM" button in the online participant dashboard) for evaluation on the test data. We subsequently ran their VM (and hence their algorithm) on the test data, and computed the relevant metrics. This evaluation process could be performed several times during the training phase, nevertheless, we limited submissions to once per week, in order to prevent the participants "training on the test data". The participants received feedback from their evaluations in a private leaderboard and had the option to make their results publicly available on the online public leaderboard, which included the results considered in our benchmark results.

2 Evaluation

For the Anatomy3 benchmark, a different evaluation approach was implemented compared to the previous Anatomy benchmarks [LMMH13, JdTGM$^+$14]. For this benchmark, participants had the opportunity to submit their algorithms several times, giving them the opportunity to improve their algorithms prior to the final evaluation analysis during ISBI 2015. They could also choose to make any of their results from the test-set public at any time. To allow a continuous workflow with this evaluation approach, the steps during the evaluation phase were automated to a large extent.

The continuous evaluation approach included the following steps:

1. The participant registers for the challenge; fills, signs, and uploads the participant agreement.
2. The organizers provide the participant with a virtual machine (VM) from the VISCERAL cloud infrastructure.
3. The participant implements a segmentation algorithm in the VM according to the benchmark specifications.
4. The VM is submitted by the participant using the participant dashboard.
5. The organizers isolate the VM to prevent the participant from accessing it during the evaluation phase.
6. The participant executable is run in a batch-script to test if its output files correspond with those expected by the evaluation routines.
7. If the previous step is successful, the evaluation proceeds for all the volumes in the test set.
8. Each generated output segmentation file is uploaded by the batch script to the cloud storage reserved for that participant.
9. Once all the images in the test-set are processed by the participant executable, the output segmentations are cleared from the VM, which is in turn returned to the participant.
10. The output segmentations uploaded in the cloud storage are then evaluated against the ground-truth (manual annotations) and the results are presented in the participant dashboard.
11. The participant can then analyze and interpret the results of their submission, and choose to make them public or not on the public leaderboard.
12. The participant is allowed to submit again for testing only after a minimum of one week from their latest submission.



Visceral Registration System

Segmentation Results LeaderBoard

Dice Coefficient

Participant	Affiliation	Modality	Submission Timestamp	Left Kidney	Right Kidney	Spleen	Liver	Left Lung	Right Lung	Unirary Bladder	Muscle Body Of Left Rectus Abdominis	Muscle Body Of Right Rectus Abdominis	Lumbar Vertebra 1	Thyroid	Pancreas	Left Psoas Major Muscle	Right Psoas Major Muscle	Gallblader	Aorta	Sternum	Trachea	Left Adrenal Gland	Right Adrenal Gland
Oscar Jimenez	HES-SO	CTce	28 Mar 2015 07:42:00	0.91	0.889	0.73	0.887	0.959	0.963	0.679	0.474	0.453	0.523	0.41	0.423	0.794	0.799	0.484	0.721	0.762	0.855	0.331	0.342
Oscar Jimenez	HES-SO	CT	28 Mar 2015 07:42:00	0.784	0.79	0.703	0.866	0.972	0.975	0.698	0.551	0.519	0.718	0.549	0.408	0.806	0.787	0.276	0.761	0.753	0.92	0.373	0.355
Dr. Mattias Heinrich	University of Luebeck	MRT1cefs	25 Mar 2015 15:10:00	0.831	0.778	0.708	0.788			0.356						0.78	0.758						
Chunliang Wang	Linköping University	CT	06 Mar 2015 13:00:00	0.896	0.796	0.91	0.936	0.961	0.97	0.713						0.828	0.817						
Chunliang Wang	Linköping University	CTce	06 Mar 2015 13:00:00	0.945	0.959	0.909	0.949	0.972	0.971	0.856						0.83	0.845						
Fredrik Kahl	Signals and Systems, Chalmers University of Technology	CT	23 Mar 2015 08:00:00			0.87	0.921			0.763				0.424	0.383			0.19				0.282	0.22
Fredrik Kahl	Signals and Systems, Chalmers University of Technology	CT	29 Mar 2015 13:00:00	0.934	0.915			0.972	0.975		0.746	0.679	0.775			0.861	0.847		0.847	0.83	0.931		
Yashin Dicente Cid	HES-SO	CT	30 Mar 2015 06:30:00					0.972	0.974														
Yashin Dicente Cid	HES-SO	CTce	30 Mar 2015 06:30:00					0.974	0.973														
	Shenzhen																						

Figure 2: A snapshot of VISCERAL Anatomy3 public leaderboard at the time of ISBI 2015 challenge.

3 Benchmark results

Detailed results from 20 metrics can be seen in the online leaderboard[1], a snapshot of which at the time of ISBI 2015 Anatomy3 challenge is shown in Gig. 2. Participant evaluation results are summarized in tables 1 and 2, respectively for Dice coefficient and mean surface distance, as commonly-used segmentation evaluation metrics. The former is an overlap metric, describing how well an algorithm estimates target anatomical region. The latter is a surface distance metric, summarizing the overall surface estimation error by a given algorithm. The participant row in the tables contains the citation for the publication contribution within this Anatomy3 proceedings Part II.

In the Dice results table, the highest ranking methods per-modality per-organ are marked in bold. Any other method within 0.01 (1%) Dice of this are also considered a winner (or a tie) due to the insignificance of the difference. Dice values below a threshold are considered unsuccessful segmentations, and thus are not declared as a winner – even though the reader should note that depending on particular clinical application such results can potentially still be useful. This threshold was selected as 0.6 Dice, coinciding with a gap in the reported participant results.

The results corresponding to the same bold values in the Dice table are also marked in the mean surface distance table, in order to facilitate comparison of the segmentation surface errors for the best methods in terms of the Dice metric. For successfully segmented organs (defined by the empirical 0.6 Dice cutoff), both metrics agree on the results for all structures and modalities, except for the first lumbar vertebra in CT. The reader should note that the mean surface distances are presented in voxels, therefore the values between modalities (e.g. MR-ce and CT) are not directly comparable in the latter table.

According to these tables, there are different algorithms performing well for different anatomy. In contrast-enhanced MR modality, we had only a single participant, Heinrich *et al.*, potentially

[1] The leaderboard is accessible at `http://visceral.eu:8080/register/Leaderboard.xhtml`

MODALITY	**MR ce**	**CT contrast-enhanced (ce)**			**CT**			
PARTICIPANT	**Heinrich** *et al.*	**Jiménez** *et al.*	**He** *et al.*	**Cid** *et al.*	**Kahl** *et al.*	**Jiménez** *et al.*	**He** *et al.*	**Cid** *et al.*
Left kidney	**0.862**	**0.91**	**0.91**	-	**0.934**	0.784	-	-
Right kidney	**0.855**	0.889	**0.922**	-	**0.915**	0.79	-	-
Spleen	**0.724**	0.73	**0.896**	-	**0.87**	0.703	**0.874**	-
Liver	**0.837**	0.887	**0.933**	-	**0.921**	0.866	**0.923**	-
Left lung	-	0.959	**0.966**	**0.974**	**0.972**	**0.972**	0.952	**0.972**
Right lung	-	**0.963**	**0.966**	**0.973**	**0.975**	**0.975**	0.957	**0.974**
Bladder	0.494	**0.679**	-	-	**0.763**	0.698	-	-
Pancreas	-	0.423	-	-	0.383	0.408	-	-
Gallbladder	-	0.484	-	-	0.19	0.276	-	-
Thyroid	-	0.41	-	-	0.424	0.549	-	-
Aorta	-	**0.721**	-	-	**0.847**	0.761	-	-
Trachea	-	**0.855**	-	-	**0.931**	0.92	-	-
Sternum	-	**0.762**	-	-	**0.83**	0.753	-	-
1^{st} lumbar vertebra	-	0.523	-	-	**0.775**	0.718	-	-
Left adrenal gland	-	0.331	-	-	0.282	0.373	-	-
Right adrenal gland	-	0.342	-	-	0.22	0.355	-	-
Left psoas major	**0.801**	**0.794**	-	-	**0.861**	0.806	-	-
Right psoas major	**0.772**	**0.799**	-	-	**0.847**	0.787	-	-
Left rectus abdominis	-	0.474	-	-	**0.746**	0.551	-	-
Right rectus abdominis	-	0.453	-	-	**0.679**	0.519	-	-

Table 1: Segmentation results in terms of DICE coefficient classified by modality and organ.

MODALITY	**MR ce**	**CT contrast-enhanced (ce)**			**CT**			
PARTICIPANT	**Heinrich** *et al.*	**Jiménez** *et al.*	**He** *et al.*	**Cid** *et al.*	**Kahl** *et al.*	**Jiménez** *et al.*	**He** *et al.*	**Cid** *et al.*
Left kidney	**0.251**	**0.172**	**0.171**	-	**0.147**	1.209	-	-
Right kidney	**0.3**	0.243	**0.131**	-	**0.229**	1.307	-	-
Spleen	**1.138**	2.005	**0.385**	-	**0.534**	1.974	**0.36**	-
Liver	**0.935**	0.514	**0.203**	-	**0.299**	0.78	**0.239**	-
Left lung	-	0.071	**0.069**	**0.05**	**0.045**	**0.043**	0.101	**0.05**
Right lung	-	**0.065**	**0.078**	**0.052**	**0.043**	**0.038**	0.094	**0.046**
Bladder	2.632	**1.879**	-	-	**1.057**	1.457	-	-
Pancreas	-	3.804	-	-	4.478	5.521	-	-
Gallbladder	-	3.603	-	-	9.617	5.938	-	-
Thyroid	-	3.337	-	-	2.163	1.466	-	-
Aorta	-	**0.899**	-	-	**0.542**	0.938	-	-
Trachea	-	**0.223**	-	-	**0.083**	0.103	-	-
Sternum	-	**1.094**	-	-	**0.798**	1.193	-	-
1^{st} lumbar vertebra	-	4.504	-	-	**2.424**	1.953	-	-
Left adrenal gland	-	3.115	-	-	3.298	2.672	-	-
Right adrenal gland	-	2.66	-	-	7.046	3.445	-	-
Left psoas major	**0.493**	**0.742**	-	-	**0.443**	0.595	-	-
Right psoas major	**0.569**	**0.757**	-	-	**0.55**	0.775	-	-
Left rectus abdominis	-	6.068	-	-	**1.614**	0.355	-	-
Right rectus abdominis	-	6.6	-	-	**1.922**	4.032	-	-

Table 2: Segmentation results in terms of mean surface distance in pixels (which may have different physical meanings based on the resolution of a particular modality).

due to the difficulty of automatic segmentations in this modality. This group thus became the unchallenged winner of MRce for the structures they participated in. Note that the surface error results are reported in voxels, where MRce has a significantly lower resolution than the other modalities. In CTce, He *et al.* performed the best for the 6 structures they participated in, with some ties with Jimenez *et al.* The latter group segmented all the given structures in CTce, some of them with satisfactory accuracy, while for the others with potentially unusable results. We had the most participants for the CT modality, in which the lungs –a relatively easier segmentation problem– were segmented successfully by most participants; potentially close to the accuracy of inter-subject annotations. For most other structures for which successful segmentations were achieved in CT, Kahl *et al.*were the winner of the challenge. Nevertheless, for structures where lower fidelity segmentations (below the 0.6 Dice cutoff) were attained, Jimenez *et al.*are seen to provide better segmentations estimations; likely due to their segmentation approach being atlas-based. It is also observed that, despite the relatively good contrast of CT, several structures (prominently the pancreas, gallbladder, thyroid, and adrenal glands) are still quite challenging to segment from CT — potentially due to the lower sensitivity of CT to those structures also complicated by the difficult-to-generalize shapes of these anatomies.

4 Conclusions

The VISCERAL Anatomy3 Challenge had a total of 23 virtual machines allocated for participants at a time, although not all participants ultimately submitted results for the challenge. Most participants relied on atlas-based segmentation methods, although there were also techniques that use anatomy-based reasoning and locational relations. By using an online leaderboard evaluation method, more participants are expected to submit results for our Anatomy3 challenge in the future.

5 Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2014) under grant agreement n° 318068 VISCERAL.

References

[JdTGM$^+$14] Oscar Alfonso Jiménez del Toro, Orcun Goksel, Bjoern Menze, Henning Müller, Georg Langs, Marc-André Weber, Ivan Eggel, Katharina Gruenberg, Markus Holzer, Georgios Kotsios-Kontokotsios, Markus Krenn, Roger Schaer, Abdel Aziz Taha, Marianne Winterstein, and Allan Hanbury. VISCERAL – VISual Concept Extraction challenge in RAdioLogy: ISBI 2014 challenge organization. In Orcun Goksel, editor, *Proceedings of the VISCERAL Challenge at ISBI*, number 1194 in CEUR Workshop Proceedings, pages 6–15, Beijing, China, May 2014.

[LMMH13] Georg Langs, Henning Müller, Bjoern H. Menze, and Allan Hanbury. Visceral: Towards large data in medical imaging – challenges and directions. *Lecture Notes in Computer Science*, 7723:92–98, 2013.