Chapter 7
VISCERAL Anatomy Benchmarks for Organ Segmentation and Landmark Localization: Tasks and Results

Orcun Goksel and Antonio Foncubierta-Rodríguez

Abstract While a growing number of benchmark studies compare the performance of algorithms for automated organ segmentation or lesion detection in images with restricted fields of view, few efforts have been made so far towards benchmarking these and related routines for the automated identification of bones, inner organs and relevant substructures visible in an image volume of the abdomen, the trunk or the whole body. The VISCERAL project has organized a series of benchmark editions designed for segmentation and landmark localization in medical images of multiple modalities, resolutions and fields of view acquired during daily clinical routine work. Participating groups are provided with data and computing resources on a cloud-based framework, where they can develop and test their algorithms, the submitted executables of which are then run and evaluated on unseen test data by the VISCERAL organizers.

7.1 Introduction

While a growing number of benchmark studies compare the performance of algorithms for automated organ segmentation or lesion detection in images with restricted fields of view, few efforts have been made so far towards benchmarking these and related routines for the automated identification of bones, inner organs and relevant substructures visible in an image volume of the abdomen, the trunk or even the whole body. The VISual Concept Extraction challenge in RAdioLogy (VISCERAL[1]) project established a cloud-based infrastructure for the evaluation of medical image analysis techniques in computed tomography (CT) and magnetic resonance (MR) imaging. The aim of VISCERAL was to create a single, large and multipurpose

[1] http://www.visceral.eu.

O. Goksel (✉) · A. Foncubierta-Rodríguez
Computer Vision Laboratory, Swiss Federal Institute of Technology (ETH) Zurich,
Sternwartstrasse 7, 8092 Zurich, Switzerland
e-mail: ogoksel@ethz.ch

A. Hanbury et al. (eds.), *Cloud-Based Benchmarking of Medical Image Analysis*, DOI 10.1007/978-3-319-49644-3_7

medical image dataset and infrastructure, on which research groups can test their specific applications and solutions. The Anatomy Benchmark of the VISCERAL project with its two tasks, landmark localization and segmentation of bones, inner organs and other relevant structures, has a series of cycles. Anatomy1 and Anatomy2 (where the latter includes an ISBI challenge as an early teaser) Benchmarks have been completed, and the last Benchmark Anatomy3 is an ongoing open benchmark, to which any research group can still submit new methods for their evaluation to be included in the online leader board. In this chapter, the Anatomy Benchmark tasks and results are described.

7.2 Data and Data Format

This section gives a brief overview of the data used in the Anatomy Benchmarks, as well as a discussion of the choice of data format for these Benchmarks.

7.2.1 Data

The datasets used for the Benchmarks have been acquired during daily clinical routine work. Whole-body MRI and CT scans or examinations of the whole trunk are used. Furthermore, imaging of the abdomen in MRI and contrast-enhanced CT for oncological staging purposes are also included in the benchmark dataset, since there is a higher resolution for segmentation especially of smaller inner organs, such as the adrenal glands. Accordingly, these four image-anatomy combinations are available:

1. Abdomen/thorax contrast-enhanced CT (ThAb/CTce)
2. Whole-body CT (Wb/CT)
3. Whole-body MR T1 (Wb/MRT1)
4. Abdomen contrast-enhanced fat-saturated MR T1 (Ab/MRT1cefs).

We call the image data together with its manual annotations as the *Gold Corpus*; this is in contrast to *Silver Corpus* that was generated by the VISCERAL consortium by fusing the results of several automatic methods to (approximately and automatically) annotate a large set of images. The Gold Corpus is the reference annotation to train and evaluate the algorithms for segmenting and localizing anatomical structures. The Anatomy Benchmarks focus on labelling large-field-of-view 3D medical imaging data. For the Gold Corpus, manual annotations were performed and the quality was checked by trained and experienced radiologists. The Gold Corpus was built up during the cycle of Anatomy Benchmarks, as described below. The final Gold Corpus is described in detail in Chap. 5.

7.2.2 Gold Corpus: Training Set

The training Gold Corpus comprises 28 fully annotated volumes in Anatomy1 (segmentations of organs/structures and landmarks). Although the MR annotations were only manually performed in one MR sequence (T1-weighted), the T2-weighted MR volumes from the same patients were also made available to the participants in the training set. In total, 42 volumes were available to the participants during the Anatomy1 benchmark. For Anatomy2, 80 volumes were fully annotated and 120 volumes were in total distributed to the participants. The total volumes included the corresponding 40 MR T2-weighted volumes not annotated for each annotated MR T1-weighted volume. For the ISBI VISCERAL Challenge that took place during the Anatomy2 Benchmark, a subset of the Anatomy2 training set was available to participants (60 annotated volumes, 90 volumes distributed in total). Once the ISBI Challenge concluded, the test set used for this challenge was added to the Anatomy2 training set. Table 7.1 provides a summary of the volumes annotated for each of the Benchmarks from the different modalities and regions.

Since not all structures are visible in all images, the total number of annotations are not a simple multiple of images and structures; e.g. for Anatomy2-ISBI, for 6 volumes, there are only 946 annotated segmentations (instead of $60 \times 20 = 1200$). As an example, Fig. 7.1 shows a breakdown of structures and landmarks segmented for the Anatomy2-ISBI challenge. Similarly, Fig. 7.2 shows the breakdown of segmented structures for Anatomy3.

Table 7.1 Summary of the training Gold Corpus volumes annotated for each of the Benchmarks

Benchmark	Vol.	Wb/CT	ThAb/CTce	Ab/MRT1cefs	Wb/MRT1	Segmentations	Landmarks
Anatomy1	42	7	7	7	7	491	42 volumes
Anatomy2 ISBI	90	15	15	15	15	946	60 volumes
Anatomy2 Main	120	20	20	20	20	1295	80 volumes
Anatomy3	120	20	20	20	20	1295	N/A

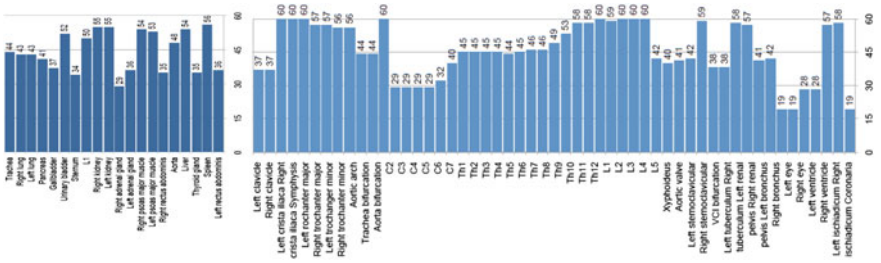

Fig. 7.1 Number of segmented structures (*left*) and annotated landmarks (*right*) for Anatomy2-ISBI

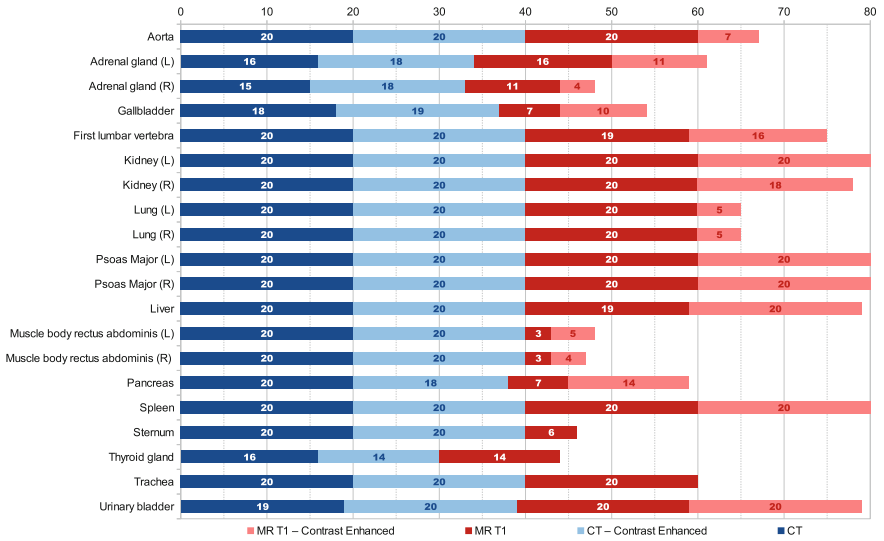

Fig. 7.2 Number of segmented structures per image modality for Anatomy3

Table 7.2 Summary of test Gold Corpus volumes annotated for each of the Benchmarks

Benchmark	Vol.	Wb/CT	ThAb/Ctce	Ab/MRT1ce	Wb/MRT1	Structures	Landmarks
Anatomy1	48	12	12	12	12	761	48 volumes
Anatomy2 ISBI	20	5	5	5	5	305	20 volumes
Anatomy2 Main	40	10	10	10	10	643	40 volumes
Anatomy3	40	10	10	10	10	643	N/A

7.2.3 Gold Corpus: Test Set

Overall, 48 volumes were included in the Gold Corpus test set for Anatomy1 (12 CT whole-body datasets, 12 CT contrast-enhanced Thorax/Abdomen datasets, 12 MRT1 whole body, 12 MRT1 contrast-enhanced Abdomen). For Anatomy2 and Anatomy3, 40 volumes were evaluated in the Gold Corpus test set, as summarized in Table 7.2.

7.2.4 Data Format

Clinical medical imaging is dominated by the Digital Imaging and Communications in Medicine (DICOM) file format. It is ubiquitous in hospital image management systems such as picture archiving and communication systems (PACS), and its standard

has facilitated clinical integration and widespread deployment of medical informatics frameworks substantially. Notably, the DICOM standard was developed in a time of significantly different information technology environments than we typically face today. One example is the slower data transfer times that made the splitting of large amounts of data sensible, which is no more required considering current data storage and transfer capabilities.

In the VISCERAL project, we revisited the choice between image format alternatives and decided for the Neuroimaging Informatics Technology Initiative (NIfTI) format. The NIfTI format was established by the NIfTI Data Format Working Group (NIfTI-DFWG) as part of an effort to enhance and disseminate neuroimaging informatics tools. NIfTI-1 was adapted from the ANALYZE 7.5 format, and NIfTI-2 was updated to support 64 bits. Our reasons for choosing NIfTI were as follows:

1. NIfTI files are easier to handle and to exchange, since each imaging volume (or volume+time information) is stored as a single self-contained file (in contrast to DICOM format), together with the header information for dimensions and coordinate transformations that establish the link between image and physical spaces.
2. In computer science research scenarios, data are typically managed by individuals and not by central image management systems such as PACS in hospitals. Dealing with a single file (instead of hundreds of files) facilitates file management considerably, since file naming allows for a straightforward identification of files—in contrast to DICOM directory information.
3. Transferring and storing of these compact large files (which also support additional ZIP compression) is typically more efficient in newer file systems.
4. Read and write functionality for NIfTI files exists for most of the popular computing frameworks, such as MATLAB, Python and R.
5. Despite the relative ease of reading DICOM files, *writing* them for annotations is significantly complicated and prone to compatibility errors, and it is a major limitation for the development environments that can be used.

Feedback from benchmark participants also corroborated these points; data transfer was reported to be swift and easy to manage, and no complaints were raised on the choice of data format.

7.3 Tasks

There were two tasks in the Anatomy Benchmarks:

1. Segmentation of anatomical structures (lung, liver, kidney, …) in the given image modalities, where participants could choose which organs to segment, and
2. Localization of anatomical landmarks.

Considering semi-automatic algorithms that can segment organs accurately only once they are localized (e.g. given a seed point), we also established a third challenge

category, the participants of which were provided with initialization information as organ centroids (computed from the manual segmentations of the test set). We call this the *half-run segmentation* segmentation task, as opposed to the *full-run* segmentation task, where no initialization is provided. No groups have participated in the half-run segmentation task.

During the Training Phase (Fig. 7.3), the training image data together with annotations for the benchmark tasks above were made available to all participants. Participants then developed algorithms on the provided virtual machines (VM) and submitted their executables tailored for our predefined input–output convention. In the Test Phase, we took over the VM to run the participant algorithms, where the *algorithms* (not the participants) were given access to the test data (Fig. 7.4). This is fundamentally different from typical benchmark set-ups, where the participants

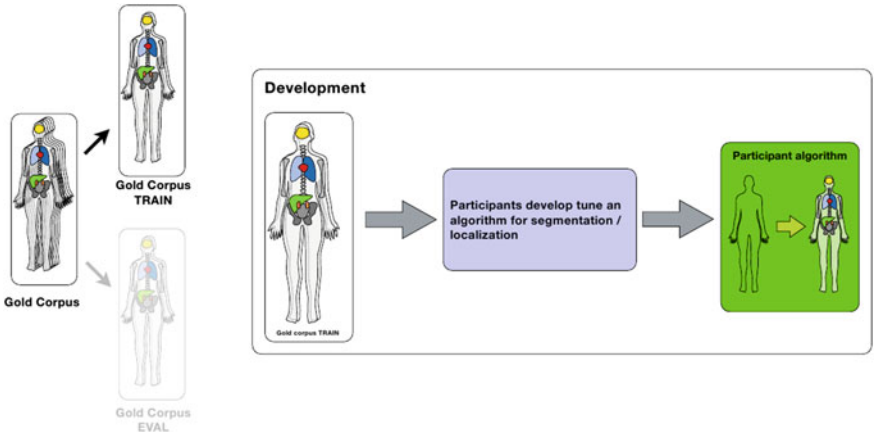

Fig. 7.3 During the development phase, annotated data are available to the participants

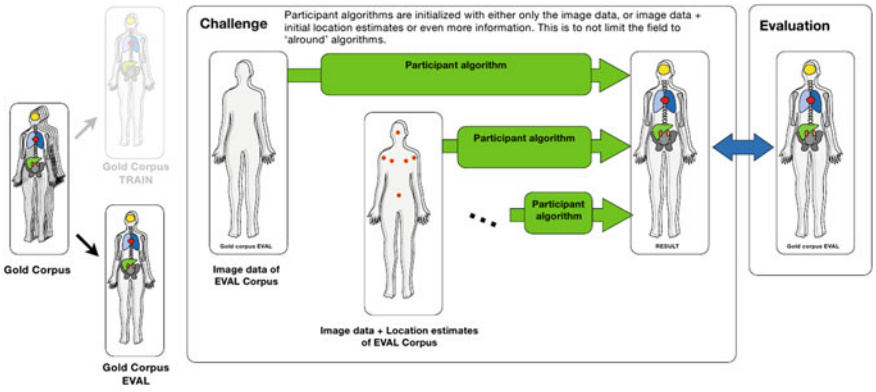

Fig. 7.4 During the evaluation phase, participant algorithms perform localization and/or segmentation tasks and are evaluated against Gold Corpus test set that is never released publicly

themselves are given the test images, where it becomes infeasible to control how much manual participant input is provided. Such release of test data also limits its repeatable use in further benchmarks or for evaluating future participants.

7.4 Results

This section presents the results of the Anatomy1, Anatomy2 (intermediate and final) and Anatomy3 Benchmarks.

7.4.1 Anatomy1

For the first Anatomy Benchmark, the following seven participants submitted algorithms, with their scores shown in Tables 7.3 and 7.4:

Dabbah et al. ($P1_{A1}$) use a voxel-level trained solution based on classification forests for landmark detection. Datasets are first aligned and downsampled to an isotropic resolution of 4 mm per voxel. Features are the Hounsfield units at chosen random offsets from each landmark.

Gass et al. ($P2_{A1}$) use multiatlas-based techniques for both segmentation and landmark detection, focusing on modality- and anatomy-independent techniques to be applied in a wide range of image modalities, in contrast to methods customized to a specific anatomy or modality. For segmentation, label propagation from several atlases to a target image is proposed. For landmark localization, consensus-based fusion of location estimates from several atlases identified by a customized template-matching approach is used.

Huang et al. ($P3_{A1}$) propose an automatic and robust coarse-to-fine liver image segmentation method. The workflow can be divided into four steps: liver localization, shape model fitting, appearance profile fitting and free-form deformation.

Jiménez del Toro et al. ($P4_{A1}$) use a multiatlas-based segmentation approach. Multiple atlases identify the location of one or more structures in the patient volume. The label volumes of the atlases are transformed using the image registrations of each atlas to the target volume. A stochastic gradient descent optimization is performed for the desired metric during the process.

Kechichian et al. ($P5_{A1}$) present an automatic multiple organ segmentation method based on a multilabel graph cuts using prior information of organ spatial relationships and shape. The former is derived from shortest-path pairwise constraints defined on a graph model of structure adjacency relations, and the latter is represented by probabilistic organ atlases learned from a training dataset.

Spanier et al. ($P6_{A1}$) describe a new generic method for the automatic rule-based segmentation of multiple organs from 3D CT scans. The rules determine the order

Table 7.3 Anatomy1 segmentation DICE scores of participants on Wb/CT, ThAb/CTce, Wb/MRT1 and Ab/MRT1cefs images

DICE	CT			CTce						MR	MRce
	$P2_{A1}$	$P3_{A1}$	$P7_{A1}$	$P2_{A1}$	$P3_{A1}$	$P4_{A1}$	$P5_{A1}$	$P6_{A1}$	$P7_{A1}$	$P2_{A1}$	$P2_{A1}$
Left kidney	0.805	-	0.820	0.903	-	0.921	0.747	0.631	0.820	0.730	0.782
Right kidney	0.754	-	0.802	0.877	-	0.913	0.632	0.663	0.872	0.733	0.787
Spleen	0.688	-	0.868	0.802	-	0.852	0.768	0.690	0.891	0.668	0.689
Liver	0.830	0.892	0.910	0.899	0.892	0.918	0.806	0.747	0.914	0.822	0.847
Left lung	0.952	-	0.961	0.961	-	0.955	0.853	0.848	0.965	0.533	0.650
Right lung	0.960	-	0.963	0.968	-	0.965	0.892	0.975	0.969	0.900	0.664
Urinary bladder	0.640	-	0.732	0.676	-	0.700	0.718	-	0.805	0.656	0.280
Left rectus abdominis	-	-	-	-	-	-	0.130	-	-		
Right rectus abdominis	-	-	-	-	-	-	0.171	-	-		
Lumbar vertebra 1	0.350	-	-	0.604	-	0.522	0.447	-	-	0.396	0.060
Thyroid	0.469	-	-	0.469	-	-	0.004	-	-	0.367	
Pancreas	0.438	-	-	0.465	-	-	0.155	-	-		0.356
Left psoas major	0.772	-	0.764	0.811	-	-	0.706	-	0.792	0.801	0.644
Right psoas major	0.787	-	0.771	0.787	-	-	0.633	-	0.811		
Gallbladder	0.102	-	-	0.334	-	0.566	0.281	-	-	0.023	0.035
Sternum	0.648	-	0.683	0.648	-	-	0.454	-	0.713	0.358	
Aorta	0.723	-	-	0.785	-	-	0.495	-	-	0.744	0.616
Trachea	0.822	-	-	0.847	-	0.836	0.696	0.785	-	0.736	
Left adrenal gland	0.165	-	-	0.204	-	-	0.000	-	-	0.109	0.000
Right adrenal gland	0.138	-	-	0.164	-	-	0.000	-	-	0.215	0.107

Table 7.4 Anatomy1 landmark localization scores as average Euclidean distances in ThAb/CTce and Ab/MRT1cefs images

Avg Error [mm]	CTce		MRce
	$P1_{A1}$	$P2_{A1}$	$P2_{A1}$
Aorta bifurcation	16.34	33.65	48.65
Aortic arch	9.70	16.05	-
Left clavicle	18.50	8.21	-
Right clavicle	20.65	9.36	-
Left crista iliaca	11.19	9.50	74.6
Right crista iliaca	7.80	9.35	55.92
Symphysis	7.13	9.38	52.25
Trachea bifurcation	3.90	4.51	-
Left trochanter major	7.44	4.74	66.69
Right trochanter major	7.03	4.17	77.79
Left trochanter minor	9.88	6.32	98.11
Right trochanter major	8.88	5.41	39.63

in which the organs are isolated and detected from simple to difficult. Following the isolation of the body, first respiratory structures are segmented, the trachea and the left/right lungs. Next, the organs with high blood content are segmented: the spleen, the liver and the left/right kidneys.

Wang et al. ($P7_{A1}$) propose multiorgan segmentation using fast model-based level set method and hierarchical shape priors. Segmentation starts with stripping the body of skin and subcutaneous fat using threshold-based level set methods. After registering the image to be processed against a standard subject picked from the training datasets, a series of model-based level set segmentation operations are carried out guided by hierarchical shape priors.

7.4.2 Anatomy2: Intermediate Results at the ISBI Challenge

Participants in Anatomy2 were given the opportunity to submit intermediate results for the Anatomy Challenge co-located with the IEEE International Symposium in Biomedical Imaging (ISBI) 2014. Five participants submitted their algorithms, with their scores shown in Tables 7.5 and 7.6. Methods used by participating groups are described in these references:

Gass et al. ($P1_{a2}$) *Segmentation and Landmark Localization Based on Multiple Atlases* [3].

Huang et al. ($P2_{a2}$) *Automatic Liver Segmentation using Multiple Prior Knowledge Models and Free-Form Deformation* [6].

Jiménez del Toro et al. ($P3_{a2}$) *Hierarchical Multistructure Segmentation Guided by Anatomical Correlations* [8].

Spanier et al. ($P4_{a2}$) *Rule-based ventral cavity multiorgan automatic segmentation in CT scans* [14].

Wang et al. ($P5_{a2}$) *Automatic multiorgan segmentation using fast model-based level set method and hierarchical shape priors* [16].

7.4.3 Anatomy2: Main Benchmark

Eight groups submitted algorithms for the final Anatomy2 Benchmark, with scores reported in Tables 7.7 and 7.8. Approaches used are described in the following references:

Gass et al. ($P1_{A2}$) submitted a multiatlas-based segmentation and landmark localisation method in images with large field of view [2].

Jiménez del Toro et al. ($P2_{A2}$) submitted an algorithm based on hierarchical multiatlas-based segmentation for anatomical structures [7].

Kéchichian et al. ($P3_{A2}$) submitted a generic multilabel graph cut method, which uses location likelihood and spatial relationships between organs [12].

Li et al. ($P4_{A2}$) submitted an automatic and robust coarse-to-fine liver image segmentation method [13].

Mai et al. ($P5_{A2}$) submitted an approach for landmark detection in volumetric images based on the popular Histograms of Oriented Gradients Descriptor (HOG) and linear support vector machines (SVM).

Spanier et al. ($P6_{A2}$) submitted a rule-based algorithm [14, 15].

Vincent et al. ($P7_{A2}$) submitted a specific, automatic model-based framework for segmenting the aorta, kidneys, liver, lungs and the psoas major muscles in Wb/CT and ThAb/CTce images.

Wang et al. ($P8_{A2}$) submitted the method described in [16].

7.4.4 Anatomy3

Five participants submitted algorithms to the Anatomy3 Benchmark before an initial kick-off deadline, with their scores reported in Table 7.9. Results from subsequent and more recent submissions can be found in the online leaderboard.[2] The approaches submitted are described in the following references:

[2] http://visceral.eu:8080/register/Leaderboard.xhtml.

Table 7.5 Anatomy2-ISBI challenge segmentation DICE scores for Wb/CT, ThAb/CTce, Wb/MRT1 and Ab/MRT1cefs images

DICE	CT				CTce					MR	MRce
	$P1_{a2}$	$P2_{a2}$	$P3_{a2}$	$P5_{a2}$	$P1_{a2}$	$P2_{a2}$	$P3_{a2}$	$P4_{a2}$	$P5_{a2}$	$P1_{a2}$	$P1_{a2}$
Left kidney	0.756	-	0.678	0.729	0.885	-	0.923	0.902	0.896	0.548	0.888
Right kidney	0.679	-	0.649	0.777	0.827	-	0.905	-	0.890	0.589	0.732
Spleen	0.684	-	0.677	0.887	0.803	-	0.859	0.934	0.842	0.646	0.785
Liver	0.798	0.911	0.823	0.904	0.882	0.922	0.908	-	0.887	0.817	0.861
Left lung	0.955	-	0.969	0.971	0.960	-	0.952	0.970	0.956	0.486	-
Right lung	0.965	-	0.967	0.972	0.966	-	0.963	0.979	0.942	0.909	-
Urinary bladder	0.636	-	0.616	0.806	0.657	-	0.680	-	0.738	0.577	0.334
Left rectus abdominis	-	-	-	-	-	-	-	-	-	-	-
Right rectus abdominis	-	-	-	-	-	-	-	-	-	-	-
Lumbar vertebra 1	0.333	-	0.44	-	0.548	-	0.472	-	-	0.623	0.084
Thyroid	0.439	-	-	-	0.315	-		-	-	0.488	-
Pancreas	0.466	-	-	-	0.442	-	-	-	-	-	0.356
Left psoas major	0.773	-	-	0.722	0.797	-	-	-	0.737	0.765	0.654
Right psoas major	0.78	-	-	0.764	-	-	-	-	0.752	-	-
Gallbladder	0.078	-	0.271	-	0.212	-	0.400	-	-	0.044	0.000
Sternum	0.63	-	-	0.712	0.612	-	-	-	0.590	0.359	-
Aorta	0.724	-	-	-	0.787	-	-	-	-	0.783	-
Trachea	0.837	-	0.855	-	0.839	-	0.830	0.856	-	0.747	-
Left adrenal gland	0.282	-	-	-	0.099	-	-	-	-	0.144	-
Right adrenal gland	0.133	-	-	-	0.019	-	-	-	-	0.268	-

Table 7.6 Anatomy2-ISBI challenge landmark localization scores as average Euclidean distances in Wb/CT, ThAb/CTce, Wb/MRT1 and Ab/MRT1cefs images

Avg Error [mm]	CT	CTce	MR	MRce
	$P1_{a2}$	$P1_{a2}$	$P1_{a2}$	$P1_{a2}$
Aorta bifurcation	19.05	36.22	252.49	61.28
Aortic arch	17.68	16.18	43.67	-
Left clavicle	9.27	16.26	13.05	-
Right clavicle	5.69	32.35	23.31	-
Left crista iliaca	7.7	13.93	23.29	88.92
Right crista iliaca	6.12	10.38	19.21	57.65
Symphysis	8.01	15.59	122.45	50.86
Trachea bifurcation	3.99	3.35	61.2	-
Left trochanter major	34.37	37.84	29.57	30.49
Right trochanter major	36.18	38.31	44.4	59.81
Left trochanter minor	5.16	11.22	18.51	28.54
Right trochanter major	4.06	12.64	62.4	34.84

Dicente Cid et al. ($P1_{A3}$) participated with a fully automatic method for the segmentation of the lung volumes in CT [1].

He et al. ($P2_{A3}$) submitted an automatic multiorgan segmentation based on multiboost learning and statistical shape model search [4].

Heinrich et al. ($P3_{A3}$) submitted a discrete medical image registration framework to multiorgan segmentation in different modalities [5].

Jiménez del Toro et al. ($P4_{A3}$) contributed a hierarchical multiatlas multi structure segmentation approach guided by anatomical correlations (AnatSeg-Gspac) [9].

Kahl et al. ($P5_{A3}$) proposed a method for multiorgan segmentation in whole-body CT images based on a multiatlas approach [11].

7.4.5 Discussion

Participation in the various editions of the Anatomy Benchmarks allows us to answer questions regarding popularity of tasks and image modalities, potentially also relating to the (perceived) difficulty of each task/modality. Specifically, the popular modality in Anatomy1 and Anatomy2 editions was contrast-enhanced CT, followed by standard CT. Magnetic resonance imaging did not attract more than a single participant for the segmentation tasks, and only in the Anatomy2 landmark localization task, was able to attract two participants, potentially due to the relative difficulty of automatic analysis using this modality. Some algorithms were organ or modality specific, so were only submitted for that anatomy, whereas other methods were

Table 7.7 Anatomy2 Benchmark segmentation DICE scores for Wb/CT, ThAb/CTce, Wb/MRT1 and Ab/MRT1cefs images

DICE	CT					CTce							MR	MRce
	$P1_{A2}$	$P2_{A2}$	$P4_{A2}$	$P7_{A2}$	$P8_{A2}$	$P1_{A2}$	$P2_{A2}$	$P3_{A2}$	$P4_{A2}$	$P6_{A2}$	$P7_{A2}$	$P8_{A2}$	$P1_{A2}$	$P1_{A2}$
Left kidney	0.778	0.784	-	0.925	0.873	0.913	0.910	0.855	-	0.829	0.943	0.927	0.808	0.845
Right kidney	0.748	0.790	-	0.866	0.871	0.914	0.889	0.805	-	0.870	0.927	0.923	0.812	0.880
Spleen	0.671	0.703	-	-	0.914	0.781	0.721	0.812	-	0.822	-	0.867	0.684	0.659
Liver	0.831	0.866	0.831	0.934	0.934	0.908	0.882	0.925	0.937	-	0.942	0.930	0.827	0.834
Left lung	0.952	0.972	-	0.970	0.960	0.961	0.959	0.955	-	0.970	0.969	0.965	0.567	0.528
Right lung	0.960	0.974	-	0.970	0.962	0.965	0.962	0.953	-	0.968	0.974	0.866	0.903	0.725
Urinary bladder	0.666	0.698	-	-	0.713	0.683	0.674	0.774	-	-	-	-	0.709	0.205
L rectus abd	-	0.551	-	-	-	-	0.444	0.111	-	-	-	-	-	-
R rectus abd	-	0.498	-	-	-	-	0.453	0.211	-	-	-	-	-	-
L1	0.412	0.718	-	-	-	0.624	0.523	0.486	-	-	-	-	0.415	0.077
Thyroid	0.450	0.549	-	-	-	0.184	0.410	0.037	-	-	-	-	0.306	-
Pancreas	0.415	0.408	-	-	-	0.460	0.406	0.544	-	-	-	-	0.196	0.372
L psoas major	0.777	0.806	-	0.858	0.833	0.813	0.794	0.775	-	-	0.864	0.820	0.820	0.640
R psoas major	0.747	0.787	-	0.848	0.828	-	0.799	0.693	-	-	0.874	0.847	-	-
Gallbladder	0.191	0.276	-	-	-	0.381	0.484	-	-	-	-	-	0.000	0.043
Sternum	0.633	0.742	-	-	-	0.635	0.714	0.573	-	-	-	0.773	0.006	-
Aorta	0.741	0.748	-	0.823	0.660	0.785	0.758	0.535	-	-	0.838	-	0.750	0.525
Trachea	0.840	0.888	-	-	-	0.847	0.849	0.592	-	0.851	-	-	0.731	-
L adr gland	0.067	0.353	-	-	-	0.250	0.331	0.000	-	-	-	-	0.151	0.048
R adr gland	0.186	0.355	-	-	-	0.213	0.341	0.000	-	-	-	-	0.077	0.020

Table 7.8 Anatomy2 Benchmark landmark localization scores as average Euclidean distances in Wb/CT, ThAb/CTce, Wb/MRT1 and Ab/MRT1cefs images

Avg Error [mm]	CT		CTce		MR		MRce	
	$P1_{A2}$	$P5_{A2}$	$P1_{A2}$	$P5_{A2}$	$P1_{A2}$	$P5_{A2}$	$P1_{A2}$	$P5_{A2}$
Aorta bifurcation	35.48	79.44	-	5.83	91.83	429.13	56	17.06
Aortic arch	14.67	8.55	-	10.98	37.12	10.78	-	-
Aortic valve	54.25	7.73	-	6.48	189.02	117.64	192.35	-
Left bronchus	6.98	2.81	-	6.12	74.45	850.85	-	-
Right bronchus	16.85	3.34	-	3.87	95.08	116.19	-	-
Cervical vertebra 2	36.43	9.21	-	-	16.54	14.11	-	-
Cervical vertebra 3	17.82	12.41	-	-	127.65	11.21	-	-
Cervical vertebra 4	21.29	8.36	-	-	282.72	15.15	-	-
Cervical vertebra 5	11.33	11.04	-	-	127.35	15.32	-	-
Cervical vertebra 6	7.63	11.94	-	-	125.01	11.74	-	-
Cervical vertebra 7	9.56	15.77	-	16.7	328.86	14.63	-	-
Left clavicle	5.86	5.09	-	5.53	9.81	12.53	-	-
Right clavicle	11.09	11.27	-	8.25	17.56	19.07	-	-
Coronaria	20.33	10.34	-	8.16	-	-	-	-
Left crista iliaca	10.63	13.27	-	13.77	59.92	63.94	68.54	64.85
Right crista iliaca	10.72	11.31	-	14.84	19.28	13.44	37.35	38.16
Left eye	81.68	3.31	-	-	193.16	12.01	-	-
Right eye	75.66	2.82	-	-	192.99	1.99	-	-
Left ischiadicum	10	3.31	-	14.18	46.87	11.24	60.01	35.89
Right ischiadicum	10.08	3.89	-	13.7	40.57	9.52	70.15	35.59
Lumbar vertebra 1	33.9	24.3	-	14.62	40.67	20.28	49.57	16.38
Lumbar vertebra 2	21.34	120.4	-	6.16	55.68	9.03	43.27	11.85
Lumbar vertebra 3	28.47	23.75	-	16.4	95.44	28.07	62.16	11.75
Lumbar vertebra 4	22.14	15.48	-	16.17	89.66	23.02	56.83	20.01
Lumbar vertebra 5	23.2	11.92	-	18.2	35.43	11.94	45.07	29.68
Left renal pelvis	58.57	56.18	-	6.77	48.75	51.95	72.45	22.3
Right renal pelvis	71.83	85.01	-	20.55	53.31	50.99	45.46	43.96
Left sternoclavicular joint	11.51	3.36	-	3.34	118.18	204.31	-	-
Right sternoclavicular joint	4.89	2.52	-	3.77	143.15	122.02	-	-
Symphysis	10.73	7.23	-	4.41	191.88	13.19	48.19	53.87
Thoracic vertebra 1	14.04	14.15	-	11.1	216.69	12.81	-	-
Thoracic vertebra 2	19.86	14.62	-	9.86	36.27	60.55	-	-
Thoracic vertebra 3	12.29	11.76	-	7.07	46.46	13.84	-	-
Thoracic vertebra 4	24.66	9.51	-	9.17	81.69	14.1	-	-
Thoracic vertebra 5	21.08	39.36	-	13.21	165.64	75.8	-	-
Thoracic vertebra 6	33.18	4.82	-	15.77	137.01	9.8	178.79	-
Thoracic vertebra 7	38.44	7.04	-	17.27	145.01	55.59	156.32	-

(continued)

Table 7.8 (continued)

Avg Error [mm]	CT		CTce		MR		MRce	
Thoracic vertebra 8	55.84	12.35	-	11.85	184.15	13.35	187.26	309.45
Thoracic vertebra 9	55.86	12.44	-	19.19	139.07	20.19	168.7	163.17
Thoracic vertebra 10	66.8	12.58	-	25.32	188.43	67.44	84.32	36.62
Thoracic vertebra 11	38.77	26.55	-	22.96	140.11	15.57	85.63	18.85
Thoracic vertebra 12	32.68	20.75	-	26.6	51.38	18.93	61.47	8.04
Trachea bifurcation	4.68	2.6	-	4.94	17	9.94	-	-
Left trochanter major	4.44	4.58	-	6.27	37.06	38.84	127.11	85.97
Right trochanter major	4.77	6.19	-	3.7	64.89	97.45	68.21	71.75
Left trochanter minor	8.53	4.97	-	2.82	55.54	7.47	125.94	131.36
Right trochanter minor	6.57	4.49	-	2.67	157.91	9.13	30.6	41.91
Left tuberculum	8.45	120.91	-	12.68	17.5	53.16	-	-
Right tuberculum	11.59	7.69	-	83.16	17.6	20.11	-	-
Inferior vena cava bifurcation	16.14	10.19	-	14.14	88.35	239.12	80.31	19.99
Left ventricle	6.32	4.72	-	-	129.68	803.14	-	-
Right ventricle	7.14	5.28	-	-	116.43	1076.85	-	-
Xyphoid process	28.76	122.47	-	14.32	217.86	154.09	210.03	39.69

more general. Some participants with such generic methods seemingly pre-tested their methods on different inputs and only submitted them for the organs/modalities where these methods could actually provide a value (i.e. satisfactory results), whereas other participants simply submitted their method for all organs/modalities, whether they generalized successfully or not.

Regarding the tasks, segmentation gathered a vast majority of the submissions. Most popular organs attempted in these benchmarks were liver, lungs, spleen, kidneys and urinary bladder. Some structures were segmented by very few methods, e.g. rectus abdominis muscles.

In terms of segmentation results, the organs that obtained the highest DICE coefficient values for each modality were the lungs and the liver in CT and the kidneys and the liver in MRI. Other structures that achieved relatively accurate segmentation across different Anatomy benchmarks include trachea, aorta, urinary bladder, psoas major muscles and spleen, with DICE coefficients ranging between 0.80 and 0.95. On the other hand, thyroid, adrenal glands, rectus abdominis muscles and gall bladder have been shown to be the most difficult structures for segmentation, with DICE coefficients below 0.5.

The landmark localization tasks have shown a large variation in performance even for the same method, but accurate results with average localization errors below 3

Table 7.9 Anatomy3 Segmentation DICE coefficient on CT volumes

DICE	CT				CTce			MRce
	$P1_{A3}$	$P2_{A3}$	$P4_{A3}$	$P5_{A3}$	$P1_{A3}$	$P2_{A3}$	$P4_{A3}$	$P3_{A3}$
Left kidney	-	-	0.784	0.934	-	0.91	0.91	0.862
Right kidney	-	-	0.79	0.915	-	0.922	0.889	0.855
Spleen	-	0.874	0.703	0.87	-	0.896	0.73	0.724
Liver	-	0.923	0.866	0.921	-	0.933	0.887	0.837
Left lung	0.972	0.952	0.972	0.972	0.974	0.966	0.959	-
Right lung	0.974	0.957	0.975	0.975	0.973	0.966	0.963	-
Urinary bladder	-	-	0.698	0.763	-	-	0.679	0.494
Left rectus abdominis	-	-	0.551	0.746	-	-	0.474	-
Right rectus abdominis	-	-	0.519	0.679	-	-	0.453	-
Lumbar vertebra 1	-	-	0.718	0.775	-	-	0.523	-
Thyroid	-	-	0.549	0.424	-	-	0.410	-
Pancreas	-	-	0.408	0.383	-	-	0.423	-
Left psoas major	-	-	0.806	0.861	-	-	0.794	0.801
Right psoas major	-	-	0.787	0.847	-	-	0.799	0.772
Gallbladder	-	-	0.276	0.19	-	-	0.484	-
Sternum	-	-	0.753	0.775	-	-	0.762	-
Aorta	-	-	0.761	0.847	-	-	0.721	-
Trachea	-	-	0.92	0.931	-	-	0.855	-
Left adrenal gland	-	-	0.373	0.282	-	-	0.331	-
Right adrenal gland	-	-	0.355	0.22	-	-	0.342	-

voxels could be achieved, e.g. for the eyes and the trachea bifurcation. Modality also had a strong impact, with some structures being much easier to localize in CT (for instance, sternoclavicular joints), whereas others in MRI (e.g. aorta bifurcation and the coronaria).

Additional discussion and further information on the organization and the results of the Anatomy benchmarks can be found in [10].

7.5 Conclusion

During the VISCERAL Anatomy Benchmarks, segmentation and landmark localization methods on large medical image datasets have been evaluated. Organization of these benchmarks led to the creation of large amounts of annotated medical imaging data, which continue to be available beyond the end of the VISCERAL project (see Chap. 5). The use of a cloud-based evaluation not only represents an opportunity for larger datasets, but also impacts the number of participants. However, the series has shown that yearly cycles of evaluation can attract larger numbers of participants, when sufficient data are provided for training and testing.

Acknowledgements The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 318068 (VISCERAL).

References

1. Dicente Cid Y, Jiménez del Toro OA, Depeursinge A, Müller H (2015) Efficient and fully automatic segmentation of the lungs in CT volumes. In: Goksel O, Jiménez del Toro OA, Foncubierta-Rodríguez A, Müller H (eds) CEUR workshop proceedings of the VISCERAL anatomy3 organ segmentation challenge at ISBI, New York, USA
2. Gass T, Szekely G, Goksel O (2014) Multi-atlas segmentation and landmark localization in images with large field of view. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai WT, Metaxas D (eds) MCV 2014. LNCS, vol 8848. Springer, Cham, pp 171–180. doi:10.1007/978-3-319-13972-2_16
3. Goksel O, Gass T, Szekely G (2014) Segmentation and landmark localization based on multiple atlases. In: Goksel O (ed) CEUR workshop proceedings of the VISCERAL challenge at ISBI. Beijing, China, pp 37–43
4. He B, Huang C, Jia F (2015) Fully automatic multi-organ segmentation based on multi-boost learning and statistical shape model search. In: Goksel O, Jiménez del Toro OA, Foncubierta-Rodríguez A, Müller H (eds) CEUR workshop proceedings of the VISCERAL anatomy3 organ segmentation challenge at ISBI, New York, USA
5. Heinrich MP, Maier O, Handels H (2015) Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. In: Goksel O, Jiménez del Toro OA, Foncubierta-Rodríguez A, Müller H (eds) CEUR workshop proceedings of the VISCERAL anatomy3 organ segmentation challenge at ISBI, New York, USA
6. Huang C, Li X, Jia F (2014) Automatic liver segmentation using multiple prior knowledge models and free-form deformation. In: Goksel O (ed) CEUR workshop proceedings of the VISCERAL challenge at ISBI. Beijing, China, pp 22–24

7. Jiménez del Toro OA, Müller H (2014) Hierarchic multi–atlas based segmentation for anatomical structures: evaluation in the VISCERAL anatomy benchmarks. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai WT, Metaxas D (eds) MCV 2014. LNCS, vol 8848. Springer, Cham, pp 189–200. doi:10.1007/978-3-319-13972-2_18
8. Jiménez del Toro OA, Müller H (2014) Hierarchical multi-structure segmentation guided by anatomical correlations. In: Goksel O (ed) CEUR workshop proceedings of the VISCERAL challenge at ISBI. Beijing, China, pp 32–36
9. Jiménez del Toro OA, Dicente Cid Y, Depeursinge A, Müller H (2015) Hierarchic anatomical structure segmentation guided by spatial correlations (anatseg-gspac): VISCERAL anatomy3. In: Goksel O, Jiménez del Toro OA, Foncubierta-Rodríguez A, Müller H (eds) CEUR workshop proceedings of the VISCERAL anatomy3 organ segmentation challenge at ISBI, New York, USA
10. Jiménez del Toro OA, Müller H, Krenn M, Gruenberg K, Taha AA, Winterstein M, Eggel I, FoncubiertaRodríguez A, Goksel O, Jakab A, Kontokotsios G, Langs G, Menze B, Fernandez TS, Schaer R, Walley A, Weber M, Cid YD, Gass T, Heinrich M, Jia F, Kahl F, Kechichian R, Mai D, Spanier AB, Vincent G, Wang C, Wyeth D, Hanbury A (2016) Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. IEEE Trans Med Imaging 35(11):2459–2475
11. Kahl F, Alvén J, Enqvist O, Fejne F, Ulén J, Fredriksson J, Landgren M, Larsson V (2015) Good features for reliable registration in multi-atlas segmentation. In: Goksel O, Jiménez del Toro OA, Foncubierta-Rodríguez A, Müller H (eds) CEUR workshop proceedings of the VISCERAL anatomy3 organ segmentation challenge at ISBI, New York, USA
12. Kéchichian R, Valette S, Sdika M, Desvignes M (2014) Automatic 3D multiorgan segmentation via clustering and graph cut using spatial relations and hierarchically-registered atlases. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai WT, Metaxas D (eds) MCV 2014. LNCS, vol 8848. Springer, Cham, pp 201–209. doi:10.1007/978-3-319-13972-2_19
13. Li X, Huang C, Jia F, Li Z, Fang C, Fan Y (2014) Automatic liver segmentation using statistical prior models and free-form deformation. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai WT, Metaxas D (eds) MCV 2014. LNCS, vol 8848. Springer, Cham, pp 181–188. doi:10.1007/978-3-319-13972-2_17
14. Spanier AB, Joskowicz L (2014) Rule-based ventral cavity multi-organ automatic segmentation in CT scans. In: Goksel O (ed) CEUR workshop proceedings of the VISCERAL challenge at ISBI. Beijing, China, pp 16–21
15. Spanier AB, Joskowicz L (2014) Rule-based ventral cavity multi-organ automatic segmentation in CT scans. In: Menze B, Langs G, Montillo A, Kelm M, Müller H, Zhang S, Cai WT, Metaxas D (eds) MCV 2014. LNCS, vol 8848. Springer, Cham, pp 163–170. doi:10.1007/978-3-319-13972-2_15
16. Wang C, Smedby O (2014) Automatic multi-organ segmentation using fast model based level set method and hierarchical shape priors. In: Goksel O (ed) CEUR workshop proceedings of the VISCERAL challenge at ISBI. Beijing, China, pp 25–31