3.1 Publishable summary

3.1.1 Project Context and Objectives

Diagnostic decision making (using images and other clinical data) is still very much an art for many physicians in their practices today due to a limited availability of quantitative tools and measurements. Traditionally, decision making has involved using evidence provided by the patient's data coupled with a physician's a priori knowledge and experience of a limited number of similar cases. With advances in electronic patient record systems, a large number of pre-diagnosed patient data sets are now becoming available. These data sets are often multimodal consisting of images (x-ray, CT, MRI), videos and other time series, and textual data (free text diagnostic reports and structured clinical data). Analyzing these multimodal sources for disease-specific information across patients can reveal important similarities between patients and hence their underlying diseases and potential treatments. Researchers are now beginning to use techniques of content-based retrieval to search for disease-specific information in visual data to find supporting evidence for a disease or to automatically learn associations of symptoms and visual abnormalities with diseases.

With ever more sophisticated imaging techniques, the number of images acquired per day is exploding. There are however many challenges in processing 3D (MRI, CT) and 4D (MRI with a time component) radiology images that have currently not been solved, especially in the areas of anatomy identification that can be a first step to all further analysis and pathology identification that requires very localized data analysis as regions of interest are most often very small. Semi-supervised learning based on radiology reports is also of interest as fully manual annotation is extremely expensive and does not scale well if the goal is to deal with realistically sized data collections. Most current research projects work on small collections and often, a large part of the budget of research projects is spent on creating local data collections to validate algorithms and afterwards these collections cannot be shared with other researchers.

Creating local small collections means the same effort in terms of ethics approval and similar paperwork, so creating much larger collections that can be shared is more efficient. The same is true for annotation, as manual annotation is expensive and creating this for a larger number of researchers is also much more efficient.

VISCERAL is defining and executing a targeted benchmark framework to speed up progress towards the objective of automated anatomy identification and pathology identification in 3D (MRI, CT) and 4D (MRI with a time component) radiology images. The following objectives are being met in carrying out the project:

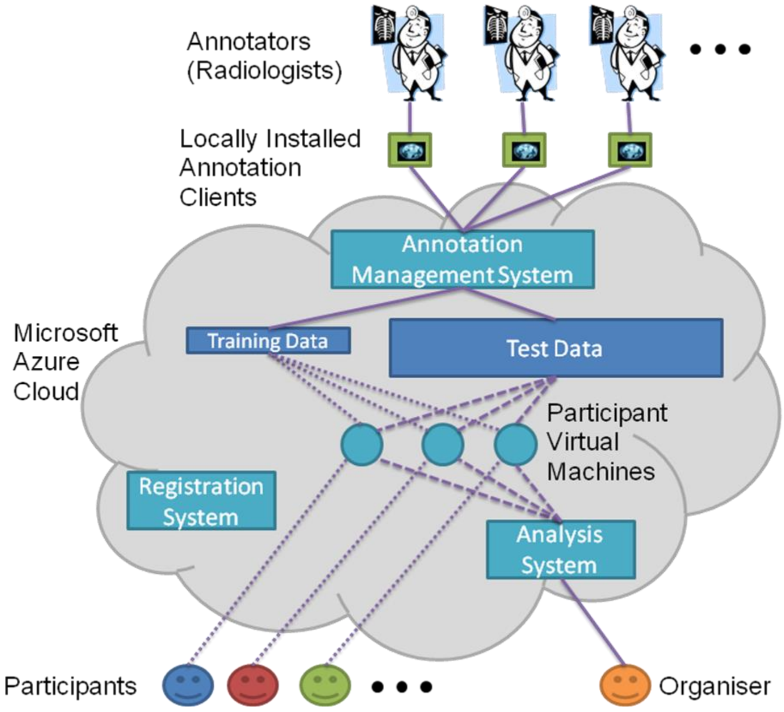
1. Create an evaluation infrastructure and software based partly on existing tools to allow the benchmarks to be carried out efficiently and effectively, but that also allows continuous evaluation to take place beyond the benchmarks;
2. Innovate through the use of a cloud infrastructure for the evaluation of algorithms on huge amounts of data;

3. Run two benchmarks and two workshops at which competition results will be discussed;
4. Fuse a large number of automated entries to create a very large silver corpus;
5. Create a small but sufficiently large gold corpus by manual annotation of the radiology images (used to evaluate the quality of the evaluation by the silver corpus);
6. Release the radiology data and the silver and gold corpora as research collections at the end of the project for a continuing impact.

3.1.2 Work Performed and Main Results

Benchmark and Annotation Framework

A benchmarking and image annotation framework was created, running on the Microsoft Azure Cloud, shown in the diagram below:



The training data is placed on the cloud. During the training phase of the benchmark (dotted lines above), participants register using the Registration System, and get assigned a Virtual Machine (VM) in the cloud with access to the training data. By the end of the training phase, participants must have installed software completing the provided segmentation task in their VM. During the testing phase (dashed lines above), the organisers take over the VMs and, using the Analysis System, run the installed software on the test data in order to evaluate how well the participant programs perform. In the second year of the project, the testing phase was automated to a large extent. It is now possible that when a participant submits a VM, metrics are calculated without intervention of the organisers and shown to the participant. Participants can decide which results should be published on a publicly visible Leaderboard.

The Annotation Management System manages the manual annotation of the radiology images. The aim of the system is to assign the manual annotation resources, given the assumption that there is a large amount of data, and that it is possible to only annotate part of it with given annotation resources. Therefore the system aims at using these resources optimally, by ranking the volumes and organs to identify those for which annotation would mean the maximum information gain, while at the same time performing quality checks of the annotations. The system generates annotation “tickets” for specific annotators that specify the volume and anatomical structures to be annotated, and accepts the upload of finished annotations.

Evaluation Metrics

In order to evaluate the segmentations generated by the participants, it is necessary to compare them objectively to the manually created ground truth. There are many ways in which the similarity between two segmentations can be measured, and at least 22 metrics have each been used in more than one paper in the medical segmentation literature. We implemented these 22 metrics in the EvaluateSegmentation software, which is available as open source on GitHub.[4] The software is specifically optimised to be efficient and scalable, and hence can be used to compare segmentations on full body volumes.

Benchmarks

The VISCERAL project has organised a series of Benchmarks described below:

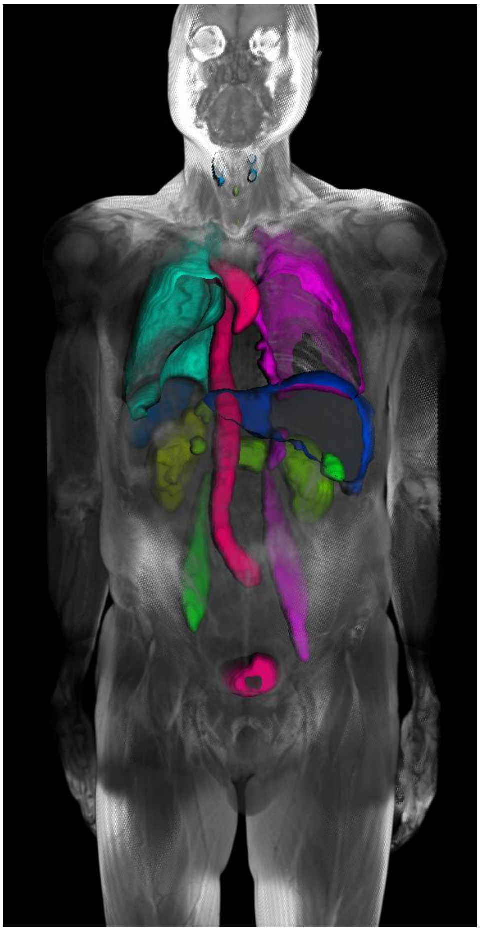
Anatomy Benchmarks: A set of medical imaging data in which organs are manually annotated is provided to the participants. The data contains segmentation of several different anatomical structures in different image modalities, e.g. CT and MRI. Participants in the Benchmarks have the task of submitting software that automatically segments the organs for which manual segmentations are provided, and this software is tested on images which the participants have not seen. Three rounds of the Anatomy Benchmark have been organised.

Detection Benchmark: A set of medical imaging data that contains various lesions manually annotated in anatomical regions such as the bones, liver, brain, lung, or lymph nodes is distributed to participants. Participants have the task of submitting software that will automatically detect these lesions. The software is tested on detecting lesions on images that the participants have not seen.

Retrieval Benchmark: One of the challenges of medical information retrieval is similar case retrieval in the medical domain based on multimodal data, where cases refer to data about specific patients (used in an anonymised form), such as medical records, radiology images and radiology reports or cases described in the literature or teaching files. The Retrieval Benchmark simulates the following scenario: a medical professional is assessing a query case in a clinical setting, e.g., a CT volume, and is searching for cases that are relevant in this assessment. The participants in the Benchmark have the task of developing software that finds clinically-relevant (related or useful for differential diagnosis) cases given a query case (imaging data only or imaging and text data), but not necessarily the final diagnosis.

[4] https://github.com/codalab/EvaluateSegmentation

Gold Corpus



The VISCERAL project is producing a large corpus of manually annotated radiology images, called the gold corpus. An innovative manual annotation coordination system has been created, based on the idea of tickets, to ensure that the manual annotation is carried out as efficiently as possible. The gold corpus is subjected to an extensive quality control process, and is therefore small but of high quality. Annotation in VISCERAL serves as the basis for all three Benchmarks. For each Benchmark, training data is distributed to the participants and testing data is kept for the evaluation.

For the Anatomy challenge series, volumes from 120 patients will be manually segmented by the end of VISCERAL by radiologists, where the radiologists trace out the extent of each organ. The following organs are manually segmented, as shown in the image on the right: 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, and left/right adrenal gland. The radiologists also manually mark landmarks in the volumes, where the landmarks include: lateral end of clavicula, crista iliaca, symphysis below, trochanter major, trochanter minor, tip of aortic arch, trachea bifurcation, aortic bifurcation, and crista iliaca.

For the detection challenge overall 300 lesions have been annotated in 50 volumes of two different modalities, in 5 different anatomical regions selected by radiologists: brain, lung, liver, bones, and lymph nodes.

For the retrieval we collected more than 10.000 medical image volumes, and selected about 2000 for the challenge, and extracted terms describing pathologies and anatomical regions from the corresponding radiology reports.

The use of these images for the purpose of automated segmentation and landmark detection has been cleared by the relevant ethics boards. The capability to provide the images once on the cloud instead of needing to distribute the images to the participants was a key aspect in getting the approvals of the ethics boards.

Silver Corpus

The silver corpus of the Anatomy Benchmark is created by fusing the submitted automated segmentations from multiple participants. For this reason the silver corpus is large and diverse, but not of the same quality as the gold corpus. The final silver corpus of VISCERAL Anatomy Benchmark contains 264 volumes of four modalities (CT, CTce, MRT1, MRT1cefs), their computed silver corpus segmentations.

3.1.3 Expected Final Results and Impact

At the end of the project, the following will be available as open source:

- The software for running the evaluation framework
- The EvaluateSegmentation metric calculation software
- The software of the active annotation framework for coordinating manual annotation work
- The software for creating silver corpora from multiple automated segmentations

Furthermore, the annotated data will continue to be available for research use, as will the results of all of the benchmarks and the evaluation methodologies used in running the benchmarks.

3.1.4 Project Website

http://visceral.eu