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 will define and execute 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 will be met in carrying out the project:

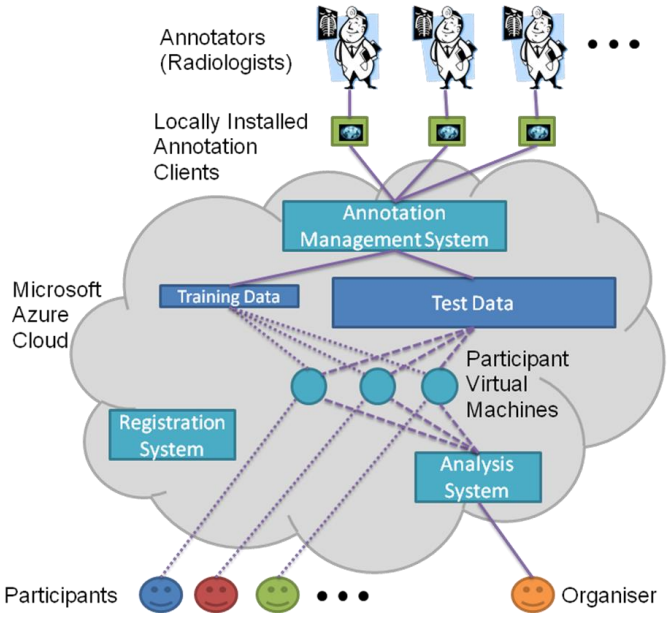
1. Create an evaluation infrastructure and software based partly on existing tools to allow the benchmarks to be carried out efficiently and effectively, but that also allows continuous evaluation to take place beyond the benchmarks;
2. Innovate through the use of a cloud infrastructure for the evaluation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

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.

Manually Annotated Dataset

By the end of the VISCERAL project, volumes from 100 patients will be manually segmented 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.

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.

First Benchmark

The first Benchmark on organ segmentation and landmark detection was organised using the framework and dataset mentioned above. 37 participants registered their interest in participating, and of these, 17 submitted a signed data use agreement form and were assigned a Virtual Machine. Finally, 7 participants submitted their Virtual Machines for their software to be evaluated on the test dataset. Evaluation is currently underway.

3.1.3 Expected Final Results and Impact

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

- The software for running the evaluation framework

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

- An updated version of 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