www.visceral.eu

Definition of the evaluation protocol and goals for competition 2

Deliverable number	*D4.2*
Dissemination level	*Public*
Delivery date	*7 February 2014*
Status	*Final*
Author(s)	*Georg Langs, Bjoern Menze, Orcun Göksel, Abdel Aziz Taha, Marianne Winterstein, Katharina Gruenberg, Henning Müller, Allan Hanbury*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Grant Agreement Number: 318068

Executive Summary

This deliverable describes the second benchmark organized in the course of VISCERAL. The objective of the benchmark is to provide data and means of evaluation for methods that perform *retrieval and analysis on large medical imaging data sets*. The benchmark is split into tasks to allow for a heterogeneous community of research groups to take advantage of the data and take part in the benchmarks.

The basis of the benchmark is medical imaging data collected from three sites. It comprises on the order of 7000 medical image volumes together with meta information such as RadLex terms extracted from radiology reports and partial annotations of pathological entities such as lesions.

The benchmark is split into three tasks from which participants can choose.

1. **Retrieval**: Given a query image volume, algorithms have to find other similar volumes that are clinically relevant for differential diagnosis in the data set. Algorithms should generate a ranking for the cases based on how relevant they are for that query case.
2. **Lesion detection:** Given a query image and an indicated location of a lesion, the algorithms have to find similar lesions in the data set (benign vs. malignant, patients with similar disease).
3. **Exploratory identification of structure in the data:** This benchmark aims at evaluating algorithms that identify structure in large heterogeneous multimodal (images and semantic annotations) data sets. Participants have to identify a structure in the form of groupings on a training set. During evaluation we will test the algorithms on new data and evaluate if they can identify equivalent clusters on such new data, i.e. if the groupings are a characteristic of the overall data.

The deliverable describes the data that is being prepared for the benchmarks, explains the training and evaluation phases of the three tasks, and specifies the evaluation protocols to be used.

Table of Contents

List of Abbreviations

MRI	Magnetic Resonance Imaging
CT	Computed Tomography
PACS	Picture Archiving and Communication System
CBMIR	Content based medical image retrieval

1 Introduction

Medical image analysis is a highly active research field that impacts both medical research and clinical practice. The increasing prevalence of digital acquisition and storage of medical imaging data in the clinical workflow has opened the possibility to access and analyse multiple data at the same time, or even mine data bases of large amounts of cases. This has triggered two lines of development that are emerging as active and volatile fields of research:

1. **Content-based medical image retrieval (CBMIR)**: CBMIR is based on a query case for which a medical expert or a researcher wants to find relevant related cases. CBMIR aims at finding relevant cases in a large set of medical imaging data, based on the query case and potentially additional information, such as marked regions of interest or textual information provided by the user. CBMIR enables access to huge medical imaging archives in hospitals and will allow individual doctors to use and compare against the knowledge of other experts.
2. **Identifying structure in large data**: Given imaging and metadata of large populations, we are reaching a state where the available digitized data samples the variability in the patient population to a high extent. It can be sufficient for inference and for identification of patterns and relationships that allow for prediction of clinically relevant variables, or interactions among measurements.

CBMIR is viewed as a particularly promising direction and instead of providing a direct automated assessment it allows for the efficient search for comparable and relevant cases. These cases are presented to aid the physician who is performing reading and interpretation of the case at hand. The visual content of medical images has been used for information retrieval for over ten years [LAHS98], and has been shown to improve quality of diagnosis [ABW03]. Visual retrieval can extract rich information beyond the associated textual cues [MJC09], and it is a promising direction to make use of the medical imaging data bases in hospitals. Currently, image retrieval is mostly used for very simple tasks of classifying images into modality and anatomic region [TSL06]. This is only the first step for the extraction of information usable for providing matching cases and anatomical regions, or mining the data for pathologies. Methods are in transition from supervised learning and recognition towards unsupervised learning or semi-supervised learning from huge amounts of data. This direction of research aims to make the wealth of visual information available without the need for supervised training [DHB11]. Identifying structures in large amounts of imaging data is a more recent approach, and it is currently rooted mainly in the machine learning and pattern recognition communities. We foresee that substantial methodological novelty will be required in order to transition these approaches from general computer vision to medical imaging data.

The goal of the VISCERAL benchmark 2 is to provide data, annotations and tasks that allow a diverse set of research groups to develop and evaluate algorithms that perform content-based retrieval or large data analysis in medical imaging data.

2 Definition of the benchmark

The benchmark comprises three tasks. They cover retrieval, detection and localization, and the unsupervised learning of structure in imaging data. The aim of retrieval is to find cases that are relevant when assessing a query case. The aim of detection is to localize and identify similar lesions in imaging data. The aim of structure learning is to identify groupings in the data, that can be transferred and reproduced on new data. While the first two tasks have a clear clinical motivation, the latter is the most exploratory of the benchmarks. We expect a range of research groups that do not necessarily have medical imaging background to participate. In the following we outline the three tasks.

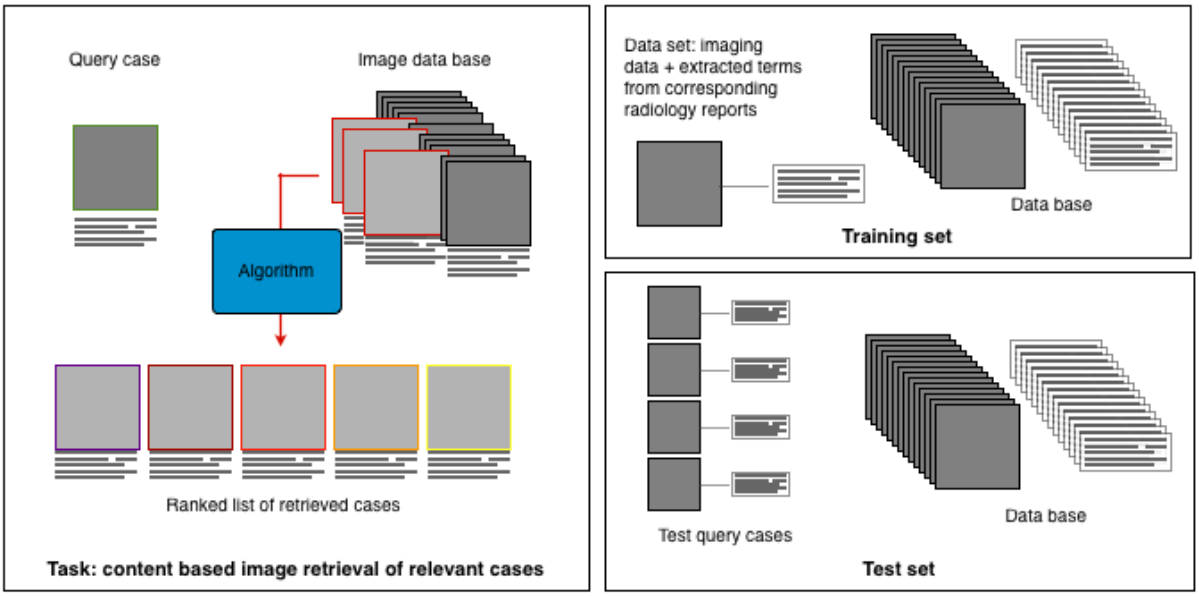

Figure 1: Task 1 - Retrieval: content-based retrieval of medical imaging data.

2.1 Task 1 - Retrieval: Content-based medical image retrieval

In this task, we evaluate the retrieval of relevant cases based on a query case. It serves the following scenario: a user 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 algorithm has to find cases that are relevant in a large database of cases. For each case there is imaging data (CT, MRI) and textual information available (RadLex terms in the radiology report and basic demographic information such as age and sex). Figure 1 provides an overview of this task.

2.1.1 Training phase

The participants are provided with imaging data and textual meta-information corresponding to that imaging data. They have to submit an algorithm that finds clinically-relevant (related) cases given a query case (imaging and text data).

2.1.2 Evaluation

Goal: We will evaluate the algorithms with a set of 10 unseen test query cases. For each query case, the algorithms should generate a ranked list of search results out of a large data base of cases (each case containing imaging data and text data). Then, experts will perform relevance assessment of the top ranked cases by each approach, to judge the quality of retrieval.

Evaluation measures: The participant algorithms will be evaluated based on the result rankings based on the 10 unseen query cases. Experts will assess the relevance of the ranked cases. The evaluation measures that will be used are the precision of the top-ranked X cases. Since we don't know a priori, how many relevant cases exist, we will evaluate precision for top-ranked 10 and 30 cases (P@10, P@30), mean uninterpolated average precision (MAP), and in cases where the establishing of the number of relevant cases in the entire data set is feasible, we will evaluate the R-precision. We will also evaluate the bpref measure [BV04].

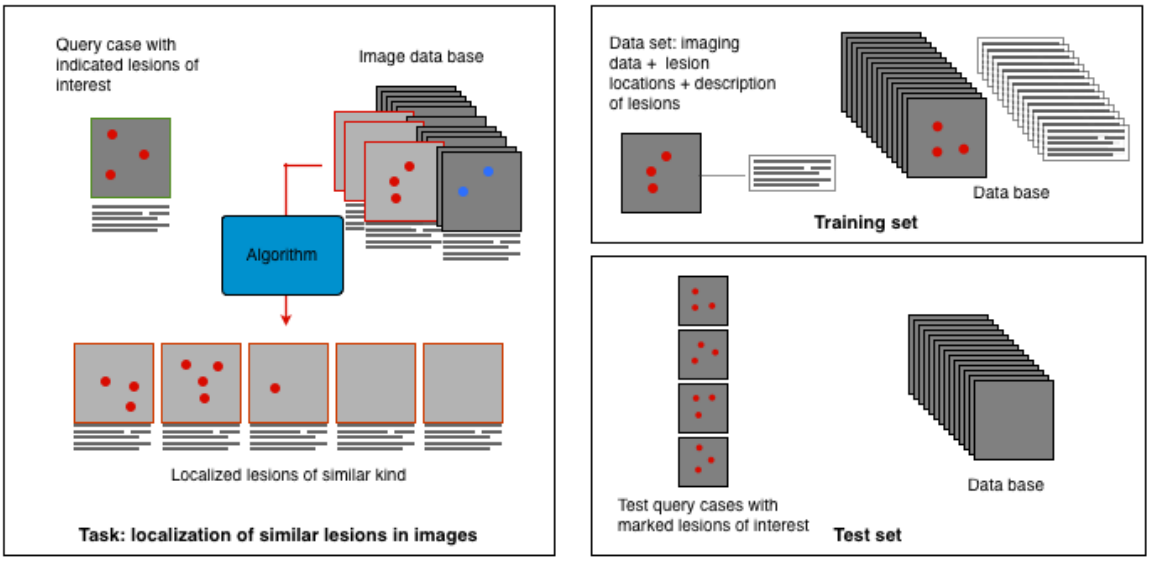

Figure 2: Task 2 – Detection: detection of similar lesions in imaging data.

2.2 Task 2 - Detection: Detection and localization of similar lesions

In this task we evaluate the detection and localization of lesions of similar type to a query lesion. The goal is to parse large amounts of imaging data and to detect instances of lesions of similar type to a query lesion. Similar type is for instance a similar imaging appearance – same signal behaviour in MRI or same densities in CT. Also, lesions that are located in the same area and have the same size fulfil the criteria of "similar type." Figure 2 provides an overview of this task. Overall categories of diseases present in the data are bone marrow neoplasms, such as multiple myeloma, malignant lymphoma and other oncological diseases, e.g. cancer of the gastrointestinal tract:

- In the multiple myeloma you find focal bone lesions (osteolysis in CT and bone marrow affection in MRI) and extra osseous lesions, e.g. soft tissue masses.
- In the malignant lymphoma, there are lesions like pathological lymph nodes or other organ affection of the lymphoma, e.g. liver, lung or spleen.
- In the other oncological diseases we expect lung, liver, bone and lymph node metastasis.

2.2.1 Training phase

During the training phase, the participants are provided with imaging data and the locations of lesions in such data. They have to submit an algorithm that, given a query case together with an indicated lesion of interest, detects and localizes similar lesions in other imaging data.

2.2.2 Evaluation

Goal: We will evaluate the algorithms with a set of 20 query lesions that are not available to the participants. The algorithms parse a set of images and detect and localize lesions of the same type.

Evaluation measures: The evaluation is performed on an image set for which we have obtained annotations of lesions in five anatomical structures (brain, lung, liver, bones, lymph nodes). Annotation is performed on two levels. For one set of images, lesion locations are annotated in the imaging data. For a second larger set of images, annotations in the form of lesion counts per anatomical structure are available (1, 2, 3, 4, 5, more than 5). Based on an image set for which we have ground truth annotations of lesions we evaluate the accuracy of the lesion localization. On the larger set of coarsely annotated lesions, we evaluate if lesions have been detected in the correct organs.

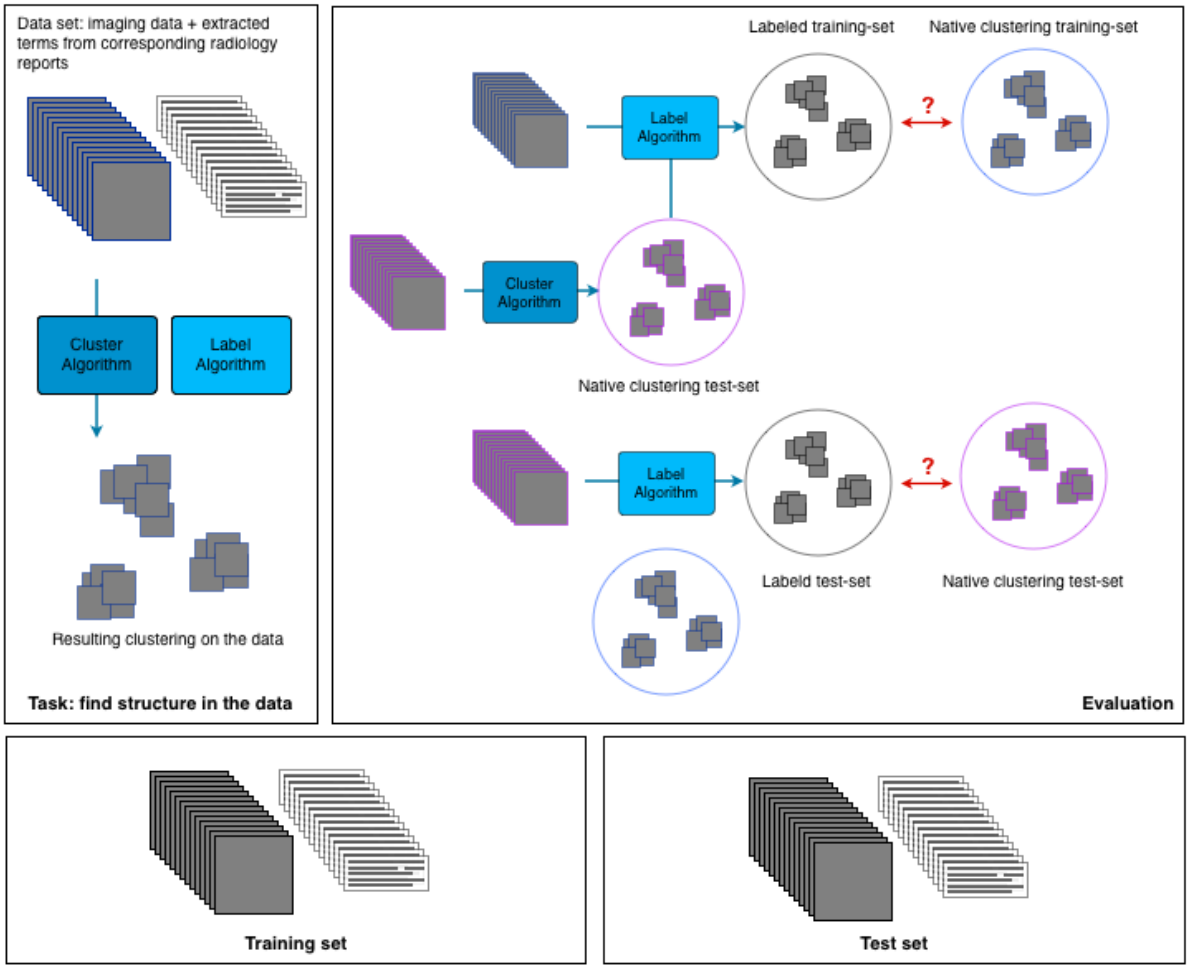

Figure 3: Task 3 – Exploration: identify structure in large data.

2.3 Task 3 - Exploration: Modelling of structure in imaging data

This tasks aims at evaluating algorithms that can identify structure in data. In this case, by structure we mean groupings of data examples to sets of similar properties. We do not restrict the kind of property, or specify which purpose this grouping should serve, except that the clustering itself is only based on the imaging data. At this phase the foremost objective is to provide medical imaging data to researchers who work on methodological approaches and who are not necessarily involved with a clinical application. The criterion that we will evaluate is the ability of the submitted algorithm to identify structure in a data set and to detect the same structure in a new data set, i.e., we will evaluate if the algorithm results in reproducible groupings that can be identified stably across different samples of the data. Figure 3 provides an overview of this task.

2.3.1 Training phase

Participants are provided with a large set of imaging data together with corresponding textual information (basic demographics, and RadLex IDs extracted from the radiology reports corresponding to the imaging data). They have to identify structure in this data. Participants are free to choose which structure they want to identify, examples might be groupings of patients with similar observations, similar disease, relationships among different variables. The objective is to be able to detect this

structure in a stable way and to be able to generalize this to new data sampled from the same overall population. This is the most flexible objective. Participants have to submit two algorithms: the first algorithm finds a grouping in the image/text data, and the second algorithm transfers this structure to new data, i.e., it assigns new examples to one of the groups in the clustering.

2.3.2 Evaluation

Goal: During evaluation, we will first apply the structure identification algorithm to the new data. Then, we transfer labels from the training set to the test data set. We compare the grouping structure, in order to evaluate if the algorithm results in groupings that are reproducible across different sub-sets of the overall population.

Evaluation measures: To evaluate the clustering structure, we will measure the overlap of the groupings found by native clustering on a set, and by transferring cluster labels from a native clustering to new data. We will use the Rand index [R71].

3 Conclusion

In this document we have specified the tasks for the second VISCERAL benchmark. The benchmark comprises three different tasks that make use of overlapping data while evaluating different methodological directions. The first task is the content based medical image retrieval, in which the algorithms have to identify related relevant cases based on a query case. The second task is detection, in which the algorithms have to detect lesions similar to a query lesion in a set of volumes. The third task aims at identifying structure in a large medical imaging data set. The benchmark is designed to provide a wide range of research groups with interesting evaluation data, and we expect groups to typically participate in only one task.

4 References

[LAHS98] Lowe HJ, Antipov I, Hersh W, Smith CA. Towards knowledge-based retrieval of medical images. The role of semantic indexing, image content representation and knowledge-based retrieval. Proc AMIA Symp, pp. 882-886, 1998.

[MJC09] Henning Müller, Jayashree Kalpathy-Cramer, Charles E. Kahn Jr., William Hatt, Steven Bedrick, William Hersh, Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task, Springer Lecture Notes in Computer Science 5706, pages 500-510, 2009.

[ABW03] Aisen AM, Broderick LS, Winer-Muram H, Brodley CE, Kak AC, Pavlopoulou C, et al. Automated Storage and Retrieval of Thin-Section CT Images to Assist Diagnosis: System Description and Preliminary Assessment. Radiology. 2003 Jul 1;228(1):265-270

[TSL06] Thies C, Schmidt-Borreda M, Seidl T, Lehmann TM, A classification framework for content-based extraction of biomedical objects from hierarchically decomposed images, Proceedings SPIE;6144:559-68, 2006.

[DHB11] Donner, R., Haas, S., Burner, A., Holzer, M., Bischof, H., and Langs, G. Evaluation of fast 2D and 3D medical image retrieval approaches based on image miniatures. In *Medical Content-Based Retrieval for Clinical Decision Support* (pp. 128-138). Springer Berlin Heidelberg, 2012.

[BV04] C. Buckley and E. M. Voorhees, Retrieval evaluation with incomplete information, In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 25-32), ACM, 2004.

[R71] W. M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association (American Statistical Association) 66 (336): 846–850, 1971.