Chapter 8
Retrieval of Medical Cases for Diagnostic Decisions: VISCERAL Retrieval Benchmark

Oscar Jimenez-del-Toro, Henning Müller, Antonio Foncubierta-Rodriguez, Georg Langs and Allan Hanbury

Abstract Health providers currently construct their differential diagnosis for a given medical case most often based on textbook knowledge and clinical experience. Data mining of the large amount of medical records generated daily in hospitals is only very rarely done, limiting the reusability of these cases. As part of the VISCERAL project, the Retrieval benchmark was organized to evaluate available approaches for medical case-based retrieval. Participant algorithms were required to find and rank relevant medical cases from a large multimodal dataset (including semantic RadLex terms extracted from text and visual 3D data) for common query topics. The relevance assessment of the cases was done by medical experts who selected cases that are useful for a differential diagnosis for the given query case. The approaches that integrated information from both the RadLex terms and the 3D volumes (mixed techniques) obtained the best results based on five standard evaluation metrics. The benchmark set up, dataset description and result analysis are presented.

O. Jimenez-del-Toro (✉)
University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland
e-mail: oscar.jimenez@hevs.ch

H. Müller
University and University Hospitals of Geneva, Geneva, Switzerland
e-mail: henning.mueller@hevs.ch

A. Foncubierta-Rodriguez
Swiss Federal Institue of Technology (ETH), Zurich, Switzerland
e-mail: antonio.foncubierta@vision.ee.ethz.ch

G. Langs
Medical University of Vienna, Vienna, Austria
e-mail: georg.langs@meduniwien.ac.at

A. Hanbury
TU Wien, Vienna, Austria
e-mail: allan.hanbury@tuwien.ac.at

A. Hanbury et al. (eds.), *Cloud-Based Benchmarking of Medical Image Analysis*, DOI 10.1007/978-3-319-49644-3_8

8.1 Introduction

The majority of diagnostic and treatment decisions taken by clinicians in their daily routine are based on acquired textbook knowledge and their experience [13]. Going through additional resources such as medical image repositories and interpatient radiology reports for medical case-based retrieval is currently inefficient and is not generally performed in clinical practice. Moreover, developing search and access technologies for information retrieval in the medical domain is still a challenging task for the information research community [3].

The VISual Concept Extraction challenge in RAdioLogy (VISCERAL) project was oriented towards improving medical image analysis tools through the evaluation on big datasets [11], and by running benchmarks in the cloud it aims to bring the algorithms and computation to the data [8]. The VISCERAL Retrieval Benchmark[1] was particularly designed to evaluate and promote improvements in the state of the art for this field. The benchmark provides a large dataset of multimodal clinical data (text and images) for the evaluation of medical retrieval and analysis approaches. In this chapter, the 2015 Retrieval Benchmark dataset, evaluated task and results from the submitted approaches are presented.

8.2 Dataset

The VISCERAL Retrieval dataset includes 2311 patient volumes obtained from computed tomography (CT) scans and T1- or T2-weighted magnetic resonance (MR) imaging. These volumes were selected from a pool of 2544 studies generated in two different clinical institutions. Only one volume per study was included in the dataset from a total of 10595 volumes in order to promote the inclusion of multiple independent clinical cases. For a subset of these scans, a list of anatomy-pathology RadLex terms (APterms) is also provided (1813 medical cases). These terms were extracted from German reports utilizing a natural language processing (NLP) framework described in [5] for automatic extraction of terms characterizing pathological findings and their anatomy in radiology reports. The German RadLex version is an older version than the English counterpart with fewer terms and a slightly different structure but many terms can be mapped from one to the other and are thus language independent. More details on the VISCERAL Retrieval datasets are given in Chap. 5.

[1] http://www.visceral.eu/benchmarks/retrieval-benchmark, as of 9 July 2016.

8.3 Medical Case-Based Retrieval

The general Benchmark task was to evaluate the retrieval ranking of relevant medical cases from the dataset having a query case as reference. The defined use case resembles a clinician assessing a query case in a medical practice setting, for example a CT volume, and is searching for cases that are relevant for the assessment in terms of a differential diagnosis. Ten query topics (Table 8.1) were judged by medical experts to generate the gold standard against which the algorithms were evaluated. Each topic (query case) included the following (Fig. 8.1):

- List of RadLex anatomy-pathology terms from the radiology report
- 3D patient scan (CT or MRT1/MRT2)
- Manually annotated 3D mask of the main organ affected
- Manually annotated 3D region of interest (ROI) from the radiologist's perspective

The participants then had to develop an algorithm that finds clinically relevant (related) cases given a query case (imaging and text data), but with no information about the final diagnosis of the case.

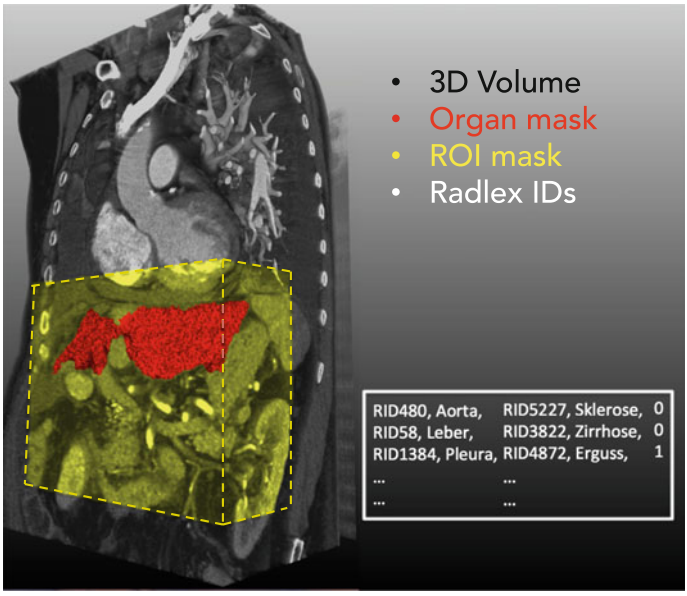


Fig. 8.1 Graphic representation of the provided data per query case. Each query topic included text information as a list of RadLex anatomy-pathology terms and a 3D volume of the patient. The manually annotated organ mask with the target diagnosis was a binary mask volume (*red*). The *yellow block* represents the region of interest (ROI) for the given case. The ROI contained either the full organ or only a region of it depending on the radiologic diagnosis

8.4 Evaluation

This section describes how the relevance judgements were obtained, as well as the metrics used for the evaluation.

8.4.1 Relevance Judgements

The submitted results by the participants were evaluated with an interface using the CrowdFlower platform.[2] This choice was made following the suggestions of [2, 4] as the interface can be used internally both without payment or with paid crowd workers. The evaluation task was divided into two parts: a task based on RadLex terms before the participant submissions and a task based on pooling after the submissions.

Relevance judgements in this benchmark needed to be performed by medical doctors, which is an expensive and time-consuming task. Therefore, a simplified preliminary task was designed in order to gather as many relevance judgements as possible before the participants submitted their runs. The task is based on the assumption that if, given a topic (diagnosis and case description), the assessors can identify a set of RadLex terms that are always relevant for this topic, then there is no need to individually evaluate all the retrieved cases that contain this term. This can produce a reduction in the number of full cases that need to be judged after the runs are submitted, when results need to be quickly computed following the benchmark. In addition, since the decision is based only on pairs of diagnosis–RadLex terms with a limited possibility to check details in the images, there is a gain also in terms of judging speed. After analysing the number of judgements received during the preliminary task, the average decision time for each pair of diagnosis–RadLex terms is 5 s.

The second task consisted in judging the relevance of the cases retrieved by the participants. A pooling strategy creates a subset of cases with the top k results of the rankings from the runs submitted by the different retrieval algorithms. The rest of the cases that are not retrieved by the participant algorithms are removed and considered as non-relevant for the corresponding run [12]. A pool with the top 100 retrieved cases by each of the submitted runs was built. The cases previously judged as non-relevant in the preliminary task were removed from the pool. In this case, each individual judgement required an average of 11–29 s depending on the topic.

The relevance criterion for the judgements was the usefulness of a case as a differential diagnosis for a given query case.

[2] http://www.crowdflower.com/, as of 9 July 2016.

Table 8.1 Query topics of the VISCERAL Retrieval benchmark. For each topic, the following features are shown as follows: imaging modality, diagnosis, main affected organ or region, size of region of interest (ROI) in voxels, number of RadLex terms in list and number of cases considered as relevant for diagnosis by medical experts during the relevance judgements

Topic	Modality	Diagnosis	Organ	ROI	RadTerms	Relevant
01	MRT1_Ab	Gall bladder sludge	Gall bladder	93 × 93 × 52	18	118
02	CT_undefined	Liver cirrhosis	Liver	258 × 351 × 284	12	428
03	CT_undefined	Liver cirrhosis	Liver	326 × 271 × 212	10	428
04	CT_Th	Lung bronchiectasis	Lung	124 × 137 × 132	14	161
05	CT_Th	Mediastinal lymphadenopathy	Mediastinum	194 × 273 × 80	8	248
06	CT_ThAb	Liver cyst	Liver	250 × 262 × 102	20	339
07	CT_Th	Pulmonary bullae	Lung	108 × 107 × 35	28	333
08	CT_ThAb	Kidney cyst	Kidney	125 × 107 × 57	16	336
09	CT_Th	Pericardial effusion	Heart	273 × 57 × 155	8	24
10	CT_Th	Rib fracture	Rib	56 × 147 × 39	26	47

8.4.2 Metrics

The standard NIST (US National Institute of Standards and Technology) evaluation procedures used in the Text Retrieval Conference (TREC) [15] were revised for selecting the Retrieval Benchmark evaluation metrics. The trec_eval tool[3] was used to compute several evaluation metrics from the results of the participant algorithms. Although multiple evaluation metrics were computed with trec_eval, the five main evaluation metrics considered for the Retrieval Benchmark were as follows:

- *Mean average precision (MAP)*: mean average fraction of retrieved cases that are relevant.
- *Geometric mean average precision (GM-MAP)*: mean average fraction of retrieved cases that are relevant, using the product of their values.
- *Binary preference (bpref)*: top number of relevant cases judged as non-relevant.
- *Precision after 10 cases retrieved (P10)*: fraction of retrieved cases that are relevant in the top 10 cases retrieved.

[3] http://trec.nist.gov/trec_eval, as of 9 July 2016.

- *Precision after 30 cases retrieved (P30)*: fraction of retrieved cases that are relevant in the top 30 cases retrieved.

8.5 Participants

There were 30 participants registered in the VISCERAL registration system. Thirteen groups had access to the data by signing the license agreement with finally four research groups submitting results for the benchmark.

Choi [1] submitted runs for text, visual and mixed (multimodal) queries. The text retrieval is based on a heuristic approach that measures case similarity with a list of conditions addressing the paired anatomy-pathology RadLex terms lists. For the image retrieval, the group used key point detection using Speeded Up Robust Features (SURF) from different sets of voxels in the images (e.g. region of interest vs. rest of the image). They then ranked the dataset images with an applied query-specific support vector machine classifier. The fusion of text and visual rankings was performed with the weighted Borda-fuse method.

Jiménez del Toro et al. [6] submitted a semi-automatic retrieval approach that generates weighting rules based on the textual and visual similarities from the query case. The main component in the final ranking is the similarity between the APterm lists of the cases, with a predefined set of rules based on clinical correlations such as same anatomy, same pathology or same imaging modalities. For the visual analysis, the images are compared using an indirect location of the region of interest from the query in a common spatial domain with the previously registered dataset. By combining 3D Riesz wavelet-based texture features with covariance descriptors, the local visual image similarity is added to the text information as an additional weight.

Spanier et al. [14] proposed a retrieval method that evaluates the similarity between cases generating an augmented RadLex graph with case-specific relations from the provided RadLex APterms lists. The sum of the link distance between term nodes from the augmented RadLex graph of each query topic is established as the similarity measure. The main organ affected is determined with the segmentation of anatomical structures in the images, and the main pathologies can be flagged by the user for the search query. This group submitted six runs using a mixed retrieval technique, differentiated by the type of imaging used in the database cases, pathologic findings, region of interest or using all these features together.

Zhang et al. [16] participated with five runs in all query types (text, visual and mixed). A co-occurrence matrix was built between the APterms and the cases for the text-only approaches. The terms were weighted by computing the term frequency–inverse document frequency (TF-IDF) or with probabilistic Latent Semantic Analysis (pLSA) to generate a probability distribution of the terms. For the visual approach, the scale-invariant feature transform (SIFT) was used to generate content descriptors for a Bag of Visual Words and was refined with relevance feedback for one of their runs. The sum combination of all text and visual retrieval results was also submitted as a mixed query method.

Table 8.2 Submitted runs of the VISCERAL Retrieval benchmark. The Type column mentions the data used in the run. A mixed type includes both text and visual data. The Input column describes how the algorithms generate a ranking of relevant cases. The Topics column shows the topics for which the runs submitted a ranking of cases

RunID	Group	Type	Input	Topics
Choi_1	SNUMedinfo	Visual	Automatic	01-10
Choi_2	SNUMedinfo	Visual	Automatic	01-10
Choi_3	SNUMedinfo	Visual	Automatic	01-10
Choi_4	SNUMedinfo	Text	Automatic	01-10
Choi_5	SNUMedinfo	Mixed	Automatic	01-10
Choi_6	SNUMedinfo	Mixed	Automatic	01-10
Choi_7	SNUMedinfo	Mixed	Automatic	01-10
Choi_8	SNUMedinfo	Mixed	Automatic	01-10
Choi_9	SNUMedinfo	Mixed	Automatic	01-10
Choi_10	SNUMedinfo	Mixed	Automatic	01-10
Jiménez_1	MedGIFT	Mixed	Semi-auto	01-10
Spanier_1	HebrewUniv	Mixed	Automatic	03-10
Spanier_2	HebrewUniv	Mixed	Automatic	03-10
Spanier_3	HebrewUniv	Mixed	Automatic	03-10
Spanier_4	HebrewUniv	Mixed	Automatic	03-10
Spanier_5	HebrewUniv	Mixed	Automatic	03-10
Spanier_6	HebrewUniv	Mixed	Automatic	03-10
Zhang_BoVW	USYD	Visual	Automatic	01-10
Zhang_fusion	USYD	Mixed	Automatic	01-10
Zhang_iter	USYD	Visual	Automatic	01-10
Zhang_plsa	USYD	Text	Automatic	01-10
Zhang_tfidf	USYD	Text	Automatic	01-10

The information that the participants provided about their techniques is summarized in Table 8.2.

8.6 Results

The results of the Retrieval Benchmark were originally presented at the *Multimodal Retrieval in the Medical Domain (MRMD) 2015* workshop, as part of the 37th European Conference on Information Retrieval (ECIR) 2015 [7]. In this chapter, a more detailed analysis of the Benchmark results is presented. Participants could submit a maximum of 10 runs and a ranked list of up to 300 cases per query topic. The 300 case threshold was defined based on experience from the previous ImageCLEF benchmarks [4], where no more than 200 results were selected as relevant in the relevance judgements. In this VISCERAL Benchmark, a few runs did have more relevant results. However, as all the participant algorithms shared this submission

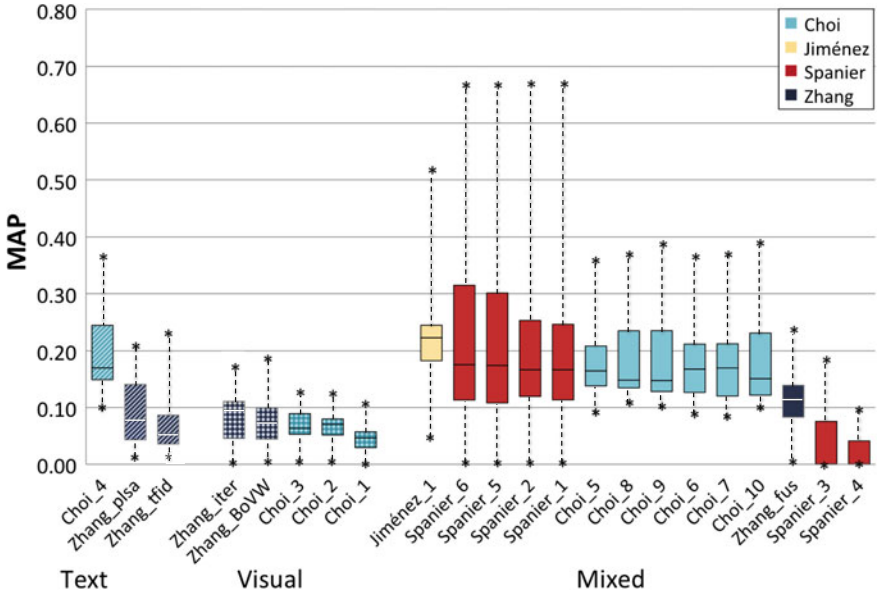


Fig. 8.2 Mean average precision (MAP) of the 22 runs in the Retrieval Benchmark. Each run is represented by a box that is extended from the first to the third quartile of the query topic MAP. The median MAP is shown as a *horizontal line* inside the *box*. The minimum and maximum MAP obtained on individual query topics are shown as *asterisks* below and above their corresponding boxes. The runs are grouped by technique (only text, only visual and mixed). The colour of the boxes is defined by the submitting group as shown in the *upper right legend*. The colour is striped in text-only runs, visual-only runs are *checkered* and mixed runs are in *solid colour*

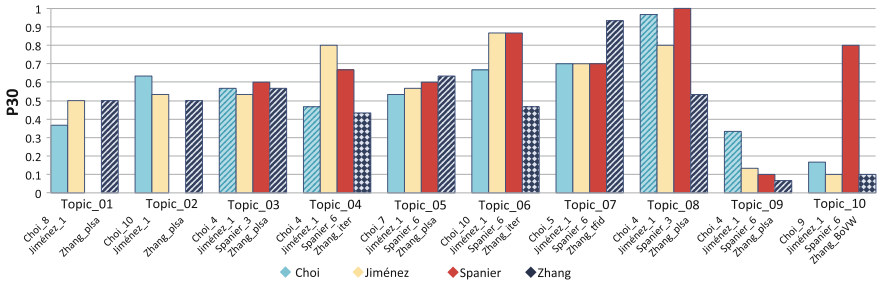
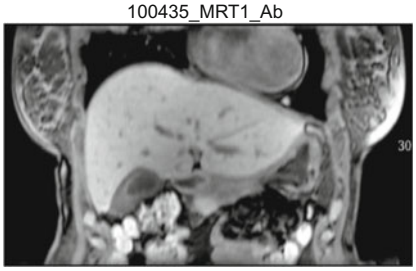


Fig. 8.3 P30 score obtained by the best run of each group, including text, visual and mixed, in the various query topics. The colour from text-only runs is *striped*, visual-only runs are *checkered* and mixed runs are in *solid colour bars*. The name of the selected runs is shown below the corresponding bar

restriction, no bias was generated towards any method. The relative performance, when algorithms are compared to other participants, was therefore the main target of the evaluation.

Topic_01

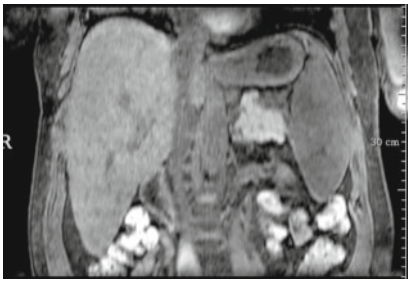
100435_MRT1_Ab



AnatRID	Anatomy	PathoRID	Pathology	Neg
RID187	Gallenblase	RID3885	Cholesterolpoly	0
RID187	Gallenblase	CIR51017	Sludge	0
RID187	Gallenblase	CIR51007	wandverdickt	0
RID205	Niere	RID3890	Zyste	0
…	…	…	…	…

Top match

101159_MRT1_Ab



AnatRID	Anatomy	PathoRID	Pathology	Neg
RID205	Niere	RID3890	Zyste	0

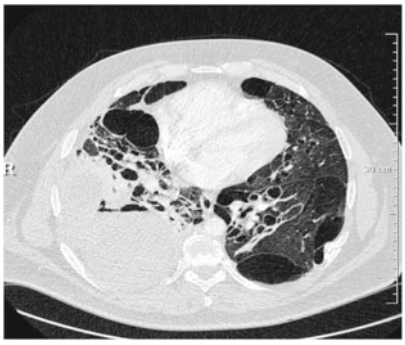
Topic_04

102758_CT_Th

AnatRID	Anatomy	PathoRID	Pathology	Neg
RID480	Aorta	RID3798	Lymphadenopt.	0
RID1301	Lunge	RID28496	Bronchiektasie	0
RID1301	Lunge	RID3820	Fibrose	0
RID1362	Pleura	RID4872	Erguss	1
…	…	…	…	…

Top match

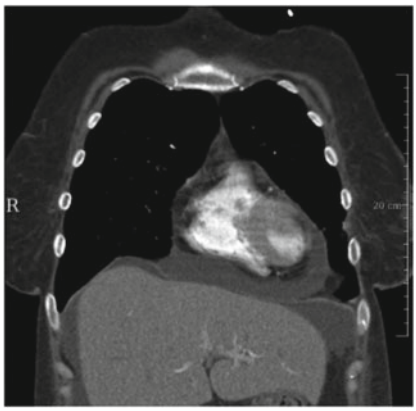
101223_CT_Th



AnatRID	Anatomy	PathoRID	Pathology	Neg
RID1301	Lunge	RID28496	Bronchiektasie	0
RID1301	Lunge	RID28502	Bulla	0
RID1301	Lunge	RID3820	Fibrose	0
RID1362	Pleura	RID4872	Erguss	0
…	…	…	…	…

Fig. 8.4 Four sample query topics (*left column*) and the corresponding top match (*right column*) obtained in the ranking of relevant cases from the algorithm with the best MAP for this topic in the Retrieval Benchmark. A sample 2D slice from the patient scan includes the affected organ together with a subset or full list of the RadLex anatomy-pathology terms

Topic_09
102423_CT_Th

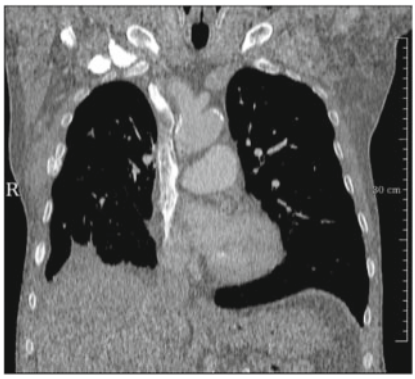


AnatRID	Anatomy	PathoRID	Pathology	Neg
RID1385	Herz	RID38588	PericardEffusio	0
RID1301	Lunge	RID39317	Infiltrat	0
RID1407	Perikard	RID4872	Erguss	0
RID1338	U.l.Lunge	RID45726	Milchglas	0

Top match
102423_CT_Th

AnatRID	Anatomy	PathoRID	Pathology	Neg
RID1407	Perikard	RID4872	Erguss	0
RID1407	Perikard	RID4705	Hämatom	0
RID1362	Pleura	RID4872	Erguss	0
RID1338	U.l.Lunge	RID4526	Milchglas	0

Topic_10
100471_CT_Th



AnatRID	Anatomy	PathoRID	Pathology	Neg
RID2471	Rippe	RID4650	Fraktur	0
RID39064	Skelett	RID5045	Degeneration	0
RID1315	U.r. Lunge	RID4799	Emphysem	1
RID1362	Pleura	RID4872	Erguss	0
...	...	...	...	...

Top match
101688_CT_Th

AnatRID	Anatomy	PathoRID	Pathology	Neg
RID1301	Lunge	RID4799	Emphysem	0
RID1362	Pleura	RID4872	Erguss	0
RID2471	Rippe	RID4650	Fraktur	0
RID39064	Skelett	RID38780	Läsion	0
...	...	...	...	...

Fig. 8.4 (continued)

A box plot chart with the MAP scores for all the individual runs is shown in Fig. 8.2, and a box plot chart with the P30 scores is presented in Fig. 8.3. Sample query topics and their corresponding top match from the algorithm with the best MAP score in the Benchmark for the corresponding topic are shown in Fig. 8.4. The runs are divided into three subtasks according to the techniques used for the query: text, visual and mixed. The scores from the individual runs for each of the subtasks are presented in Tables 8.3, 8.4 and 8.5, respectively. A table with the top participant scores for individual runs per metric per topic is shown in Table 8.6.

The four participating research groups submitted a total of 22 runs: 3 text, 5 visual and 14 mixed. Five evaluation metrics computed with the trec_eval tool are provided as the mean average score of all the topics (10 in total) for each run. Each run contained results for the 10 query topics, except for the approaches from Spanier et al. which submitted results only for 8 query topics (3–10). The results from this participant are also shown as the mean of 10 query topics just like the other participants. A score of 0 was given to the 2 missing query topics of this participant. The results computing the mean of only the 8 query topics in which Spanier et al. participated were presented in [7].

From the techniques that used only text, the run *Choi_4* with a heuristic ranking function based on the RadLex terms obtained the best scores. This algorithm had the highest AP score (0.2198) in the benchmark for topic 9–Pericardial effusion among all the techniques. This topic had the lowest number of cases (24) marked as relevant during the relevance judgements from the 10 query topics evaluated in the Retrieval Benchmark. The run by Choi, using only text data, was able to find the best features to characterize this diagnosis among the participants. Topic 10–Rib fracture had the lowest scores with only text techniques. The number of relevant cases for this topic was also low (47). Still, the results were better overall than techniques using only visual features (see Fig. 8.2).

Only visual techniques obtained the lowest scores in the benchmark. The most promising algorithm was *Zhang_iter* that reached 0.33 precision after the first 30 cases retrieved (P30, see Table 8.4). Topic 01–Gall bladder sludge obtained the highest scores from only visual techniques. This was the only topic using MR images, which suggest that differentiating between imaging modalities can already improve the retrieval of cases when only visual features are considered. On the contrary, a poor performance was achieved with only visual retrieval techniques when an uncommon disease, such as topic 09–Pericardial effusion, is present in a recurrent imaging modality (i.e. thorax CT). The challenge of successfully detecting and selecting purely visual biomarkers for general medical retrieval is still an unsolved problem in the literature [9].

There were two groups (Jiménez–del–Toro et al. and Spanier et al.) who submitted only mixed runs, using text and visual information in the same run. It is not straightforward to compare the influence of the visual or textual features based only on these results to the other algorithms (by Choi and Zhang et al.) who contributed also with results using only textual or only visual features. Nevertheless, it should be highlighted that these last two groups obtained overall higher scores using only textual features than their mixed runs. The overall highest MAP was obtained by

Table 8.3 Scores from the runs using only text retrieval techniques

Text						
RunID	*Type*	*MAP*	*GM-MAP*	*bpref*	*P10*	*P30*
Choi_4	Text	0.1942	0.1806	0.3221	0.5700	0.4967
Zhang_plsa	Text	0.0944	0.0697	0.1830	0.4100	0.3800
Zhang_tfidf	Text	0.0810	0.0582	0.1623	0.3700	0.2767

Table 8.4 Scores from the runs using only visual retrieval techniques

Visual						
RunID	*Type*	*MAP*	*GM-MAP*	*bpref*	*P10*	*P30*
Zhang_iter	Visual	0.0828	0.0541	0.1881	0.3300	0.3300
Zhang_BoVW	Visual	0.0783	0.0572	0.1900	0.0000	0.0333
Choi_3	Visual	0.0672	0.0474	0.1647	0.2700	0.3267
Choi_2	Visual	0.0661	0.0485	0.1671	0.2200	0.2633
Choi_1	Visual	0.0462	0.0188	0.1430	0.1400	0.1867

Table 8.5 Scores from the runs using mixed (text and visual) retrieval techniques

Mixed						
RunID	*Type*	*MAP*	*GM-MAP*	*bpref*	*P10*	*P30*
Jiménez_1	Mixed	0.2367	0.2016	0.3664	0.5700	0.5533
Spanier_6	Mixed	0.2295	0.2137	0.3157	0.5500	0.5100
Spanier_5	Mixed	0.2265	0.2109	0.3118	0.5500	0.5100
Spanier_2	Mixed	0.2100	0.1967	0.2976	0.5100	0.4967
Spanier_1	Mixed	0.2088	0.1954	0.2952	0.5500	0.5033
Choi_5	Mixed	0.1875	0.1722	0.3082	0.5400	0.4600
Choi_8	Mixed	0.1867	0.1721	0.3099	0.5300	0.4533
Choi_9	Mixed	0.1861	0.1700	0.3143	0.4300	0.4700
Choi_6	Mixed	0.1858	0.1697	0.3102	0.4500	0.4633
Choi_7	Mixed	0.1857	0.1688	0.3097	0.3900	0.4567
Choi_10	Mixed	0.1845	0.1681	0.3110	0.3900	0.4500
hNcmJn_fusion	Mixed	0.1101	0.0766	0.2070	0.4200	0.3533
BxcvfH_3	Mixed	0.0467	0.0444	0.0604	0.2900	0.2600
BxcvfH_4	Mixed	0.0225	0.0220	0.0584	0.0000	0.0167

Table 8.6 Top scores obtained by participant runs per topic for four evaluation metrics: MAP, bpref, P_10 and P_30. When more than 1 run obtained the highest score in the Retrieval Benchmark, all the runs with the same score are shown

	MAP			bpref			P_10			P_30		
1	Zhang_fus	Mixed	0.239	Jiménez_1	Mixed	0.504	Zhang_fus Zhang_tfid	Mixed Text	0.600	Jiménez_1 Zhang_plsa	Mixed Text	0.500
2	Jiménez_1	Mixed	0.223	Jiménez_1	Mixed	0.347	Jiménez_1	Mixed	0.600	Choi_10_He	Mixed	0.633
3	Jiménez_1	Mixed	0.223	Jiménez_1	Mixed	0.347	Choi_4_He	Text	0.900	Spanier_3	Miixed	0.600
4	Jiménez_1	Mixed	0.250	Jiménez_1	Mixed	0.405	Spanier_5 Spanier_2 Spanier_1	Mixed	1.000	Jiménez_1	Mixed	0.800
5	Jiménez_1	Mixed	0.195	Jiménez_1	Mixed	0.354	Spanier_5 Spanier_2 Spanier_1	Mixed	0.800	Zhang_plsa	Text	0.633
6	Jiménez_1	Mixed	0.388	Jiménez_1	Mixed	0.491	Jiménez_1	Mixed	0.900	Spanier_5 Spanier_2 Spanier_1 Jiménez_1	Mixed	0.867
7	Choi_5_He	Mixed	0.306	Choi_5_He	Mixed	0.465	Choi_5_He Zhang_tfid	Mixed Text	1.000	Zhang_tfid	Text	0.933
8	Jiménez_1	Mixed	0.513	Jiménez_1	Mixed	0.631	Spanier_3 Choi_4_He	Mixed Text	1.000	Spanier_3	Mixed	1.000
9	Choi_4_He	Text	0.220	Choi_4_He	Text	0.325	Choi_8_He	Mixed	0.400	Choi_4_He	Text	0.333
10	Spanier_1	Mixed	0.676	Spanier_1	Mixed	0.825	Spanier_5 Spanier_2 Spanier_1	Mixed	0.700	Spanier_5 Spanier_2 Spanier_1	Mixed	0.800
All	Jiménez_1	Mixed	0.237	Jiménez_1	Mixed	0.3664	Jiménez_1	Mixed	0.570	Spanier_5	Mixed	0.638

the mixed technique of Jiménez–del–Toro et al. This method also obtained the best AP score in 6 out of the 10 query topics. However, the runs from Spanier et al., especially those using both imaging modalities and all the pathological findings in the RadLex term lists (i.e. `Spanier_6`), obtained high scores for the majority of the query topics. This was best exemplified in Topic 10–Rib fracture, where the algorithms by Spanier et al. obtained the highest MAP scores from the whole benchmark (0.6758) and a P30 of 0.8. Jiménez del Toro et al. included the visual information in a late fusion with the textual features as an additional weighting in the final ranking score. On the other hand, Spanier et al. included the visual information early in their method for the selection of the main RadLex terms in the lists from the query cases.

8.7 Conclusion

The Retrieval Benchmark was the first medical case-based retrieval benchmark using a large dataset of 3D volumes and anatomy-pathology RadLex term lists. The dataset was hosted in a cloud infrastructure with the objective to provide access to a large number of medical cases to the participants. Four research groups submitted a variety of techniques (22 in total) for the tasks. The results were compared using standard retrieval evaluation metrics. Multimodal techniques (mixed) obtained the best results

when compared to the gold standard relevance judgements performed by clinical experts. The organization and result analysis from the benchmark helps address the current challenges in medical information retrieval and target the development of future benchmarks with common goals in this field.

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. Choi S (2015) Multimodal medical case-based retrieval on the radiology image and report: SNUMedinfo at VISCERAL retrieval benchmark. In: Müller H, Jimenez del Toro OA, Hanbury A, Langs G, Foncubierta Rodríguez A (eds) Multimodal retrieval in the medical domain. LNCS, vol 9059. Springer, Cham, pp 124–128. doi:10.1007/978-3-319-24471-6_11
2. Foncubierta-Rodríguez A, Müller H (2012) Ground truth generation in medical imaging: a crowdsourcing based iterative approach. In: Workshop on crowdsourcing for multimedia. ACM Multimedia, New York, pp 9–14
3. García Seco de Herrera A (2015) Use case oriented medical visual information retrieval & system evaluation. Ph.D. thesis, University of Geneva
4. García Seco de Herrera A, Foncubierta-Rodríguez A, Markonis D, Schaer R, Müller, H (2014) Crowdsourcing for medical image classification. In: Annual congress SGMI 2014
5. Hofmanninger J, Krenn M, Holzer M, Schlegl T, Prosch H, Langs G (2016) Unsupervised identification of clinically relevant clusters in routine imaging data. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W (eds) MICCAI 2016. LNCS, vol 9900. Springer, Cham, pp 192–200. doi:10.1007/978-3-319-46720-7_23
6. Jiménez-del-Toro OA, Cirujeda P, Cid YD, Müller H (2015) RadLex terms and local texture features for multimodal medical case retrieval. In: Müller H, Jimenez del Toro OA, Hanbury A, Langs G, Foncubierta Rodríguez A (eds) Multimodal retrieval in the medical domain. LNCS, vol 9059. Springer, Cham, pp 144–152. doi:10.1007/978-3-319-24471-6_14
7. Jiménez-del-Toro OA, Hanbury A, Langs G, Foncubierta-Rodríguez A, Müller H (2015) Overview of the VISCERAL Retrieval Benchmark 2015. In: Müller H, Jimenez del Toro OA, Hanbury A, Langs G, Foncubierta Rodríguez A (eds) Multimodal retrieval in the medical domain. LNCS, vol 9059. Springer, Cham, pp 115–123. doi:10.1007/978-3-319-24471-6_10
8. Jimenez-del-Toro O, Müller H, Krenn M, Gruenberg K, Taha AA, Winterstein M, Eggel I, Foncubierta-Rodríguez A, Goksel O, Jakab A, Kontokotsios G, Langs G, Menze B, Salas Fernandez T, Schaer R, Walleyo A, Weber MA, Dicente Cid Y, 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
9. Kurtz C, Beaulieu CF, Napel S, Rubin DL (2014) A hierarchical knowledge-based approach for retrieving similar medical images described with semantic annotations. J Biomed Inform 49:227–244
10. Langlotz CP (2006) Radlex: a new method for indexing online educational materials. Radiographics 26(6):1595–1597
11. Langs G, Hanbury A, Menze B, Müller H (2013) VISCERAL: towards large data in medical imaging — challenges and directions. In: Greenspan H, Müller H, Syeda-Mahmood T (eds) MCBR-CDS 2012. LNCS, vol 7723. Springer, Heidelberg, pp 92–98. doi:10.1007/978-3-642-36678-9_9

12. Peters C, Braschler M, Clough P (2012) Multilingual information retrieval: from research to practice. Springer, New York, pp 129–169
13. Quellec G, Lamard M, Bekri L, Cazuguel G, Roux C, Cochener B (2010) Medical case retrieval from a committee of decision trees. IEEE Trans Inform Technol Biomed 14(5):1227–1235
14. Spanier AB, Joskowicz L (2015) Medical case-based retrieval of patient records using the RadLex hierarchical lexicon. In: Müller H, Jimenez del Toro OA, Hanbury A, Langs G, Foncubierta Rodríguez A (eds) Multimodal retrieval in the medical domain. LNCS, vol 9059. Springer, Cham, pp 129–138. doi:10.1007/978-3-319-24471-6_12
15. Voorhees EM, Ellis A (eds) (2015) In: Proceedings of the twenty-fourth text REtrieval conference, TREC, Gaithersburg, Maryland, USA, 17–20 Nov 2015, vol Special Publication 500–319. National Institute of Standards and Technology (NIST)
16. Zhang F, Song Y, Cai W, Depeursinge A, Müller H (2015) USYD/HES-SO in the VISCERAL retrieval benchmark. In: Müller H, Jimenez del Toro OA, Hanbury A, Langs G, Foncubierta Rodríguez A (eds) Multimodal retrieval in the medical domain. LNCS, vol 9059. Springer, Cham, pp 139–143. doi:10.1007/978-3-319-24471-6_13