Chapter 1
Retrieval of Medical Cases for Diagnostic Decisions: VISCERAL Retrieval Benchmark

Oscar Alfonso Jiménez–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 the large amount of medical records generated daily in hospitals is only very rarely done, limiting the re–usability 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 data set (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 5 standard evaluation metrics. The benchmark set up, data set description and result analysis are presented.

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

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

Antonio Foncubierta-Rodríguez
Swiss Federal Institue of Technology (ETH), Zürich, Switzerland, e-mail: antonio.foncubierta@vision.ee.ethz.ch

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

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

1.1 Introduction

The majority of diagnostic and treatment decisions taken by clinicians in their daily routine are based on acquired textbook knowledge and their experience [11]. Going through additional resources such as medical image repositories and inter–patient 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 [4].

The VISual Concept Extraction challenge in RAdioLogy (VISCERAL) project was oriented towards improving medical image analysis tools through the evaluation on big data sets [10], and by running benchmarks in the cloud it aims to bring the algorithms and computation to the data [3]. 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 data set of multimodal clinical data (text and images) for the evaluation of medical retrieval and analysis approaches. In this chapter, the 2015 Retrieval Benchmark data set, evaluated task and results from the submitted approaches are presented.

1.2 Data Set

The VISCERAL Retrieval data set includes 2311 patient volumes obtained from computed tomography (CT) scans and T1– or T2–weighted magnetic resonance (MR) imaging (see Table 1.1). 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 data set 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 [**?**] for automatic extraction of terms characterizing pathological findings and their anatomy in radiology reports. RadLex is a unified language of radiology terms that can be used for standardized indexing and retrieval of radiology information resources [9]. These terms were extracted automatically from the German radiology reports and were marked in the list as negated if they were explicitly negated in the 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. In Figure 1.1 an example list is shown to illustrate the naming convention and the file content specifications.

For each row of anatomy terms found in the report, the corresponding pathology is stated and marked if it was positive (0) or negative (1). For example, a positive

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

Table 1.1 Statistics from the VISCERAL Retrieval benchmark data set: patient volumes and Radlex term lists.

Body region	Modality	Volumes	RadLex APterms lists
Abdomen	CT	336	213
	MR T1	167	114
	MR T2	68	18
Thorax + Abdomen	CT	86	86
Thorax	CT	971	699
Whole body	CT	410	410
Unknown	MR T1	24	24
	MR T2	38	38
TOTAL		**2311**	**1813**

report of liver cirrhosis will appear as: RID58,Leber,RID3822,Zirrhose,0. Around 1600 different pathology terms are included in the data set.

1.3 Medical Case–based Retrieval

The general benchmark task was to evaluate the retrieval ranking of relevant medical cases from the data set 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 were judged by medical experts to generate the gold standard against which the algorithms were evaluated. Each topic (query case) included the following (Fig. 1.2):

- 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

AnatRID	Anatomy	PathoRID	Pathology	Neg
RID480	Aorta	RID5227	Sklerose	0
RID58	Leber	RID3822	Zirrhose	0
RID1384	Mediastinum	RID3798	Lymphadenopathie	1
RID1327	Oberlappen der linken Lunge	RID3953	Granulom	0
RID1362	Pleura	RID4872	Erguss	0
RID1315	Unterlappen der rechten Lunge	RID28493	Atelektase	0

Fig. 1.1 Sample anatomy-pathology RadLex term list from the VISCERALRetrieval data set. Each row includes the following elements: anatomical structure RadLex term (AnatRID), name of the structure in German (Anatomy), corresponding pathologic RadLex term (PathoRID), pathology name and if the pathologic term is negated (Neg). The pathologic term is negated when the negation element is 1. The number of rows varies according to the radiology report from the medical cases.

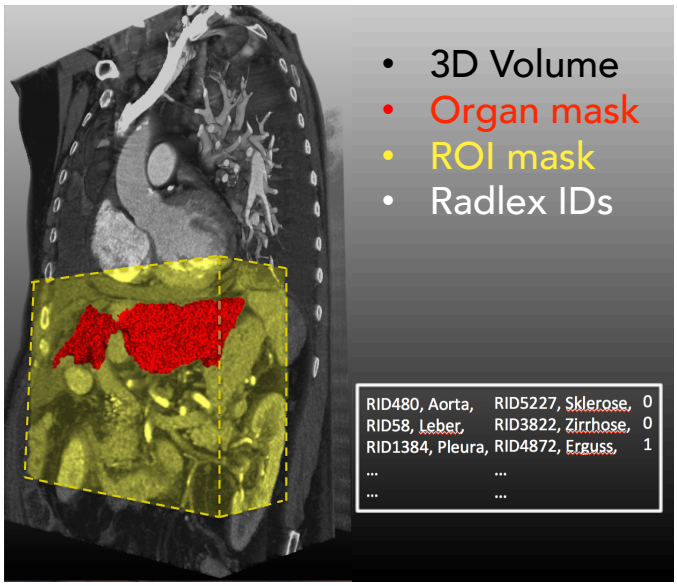


Fig. 1.2 Graphic representation of the provided data per query case. Each query topic included text information as a list of RadLex anatomy–pathology terms and a 3D volume of the patient. The manually annotated organ mask with the target diagnosis was a binary mask volume (red). The yellow block represents 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.

- 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 to the final diagnosis of the case.

1.4 Evaluation

1.4.1 Relevance Judgements

Evaluation of the submitted results by the participants was performed with an interface using the Crowdflower platform[2]. This choice was made following the suggestions of [2, 5] as the interface can both be used internally without payment or with paid crowd workers. The evaluation task was divided into two parts: a task based on RadLex terms before the submissions and a task based on pooling after the submissions.

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

Relevance judgments 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 judgments 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, there is no need to individually evaluate all the retrieved cases that contain this term. This can produce a reduction of 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 analyzing the number of judgments received during the preliminary task, the average decision time for each pair diagnosis–RadLex terms is 5 seconds.

The second task consisted in judging the relevance of the cases retrieved by the participants. A pool with the top 100 retrieved cases by all submitted runs is built and the already judged cases based on the preliminary task are removed from the pool. In this case, each individual judgment required an average of 11 to 29 seconds depending on the topic.

The relevance criterion for the relevance judgments was that a case had to be useful for differential diagnosis of a given query case.

Table 1.2 Query topics of the VISCERAL Retrieval benchmark. For each topic the following features are shown: 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 judgments.

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

1.4.2 Metrics

The trec_eval tool [3] was used to compute several evaluation metrics from the participants' results. This program uses the standard NIST (US National Institute of Standards and Technology) evaluation procedures is used for the Text Retrieval Conference (TREC). Although multiple evaluation metrics were computed with trec_eval, the five main evaluation metrics considered for the Retrieval Benchmark were:

- mean average precision (MAP);
- geometric mean average precision (GM-MAP);
- binary preference (bpref);
- precision after 10 cases retrieved (P10);
- precision after 30 cases retrieved (P30).

1.5 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 data set 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 like 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 data set. 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. [12] 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

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

Table 1.3 Submitted runs of the VISCERAL Retrieval benchmark.

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

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. [13] 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 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.

The information that the participants provided about their techniques is summarized in Table 1.3.

1.6 Results

The results of the retrieval benchmark were presented at the *Multimodal Retrieval in the Medical Domain (MRMD) 2015* workshop, as part of the 37th European Confer-

Table 1.4 Scores of 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

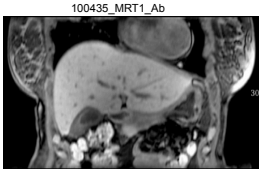
ence on Information Retrieval (ECIR) 2015. Participants could submit a maximum of 10 runs and a ranked list of up to 300 cases per query topic. The 300 case threshold was defined based on experience from previous ImageCLEF benchmarks [5] where no more than 200 results were selected as relevant in the relevance judgments. In this VISCERAL benchmark a few runs did have more relevant results. However, as all the participant algorithms shared this submission 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.

The runs are divided into three subtasks according to the techniques used for the query: text, visual and mixed. 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 score of all the topics within each run (num_q : number of queries, 10 total). 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]. The complete list of results is shown in Tables 1.4, 1.5 and 1.6.

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 MAP 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 from the 10 query topics. 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 than techniques using only visual features.

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 1.5). Topic 01–Gallbladder 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

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
...	...	...	...	...

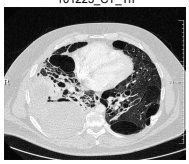
Top match
101159_MRT1_Ab

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
...	...	...	...	...

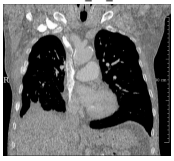
Topic_09
102423_CT_Th

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. 1.3 Four sample query topics (left column) and their corresponding top match (right column) from the algorithm with the best MAP in the 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.

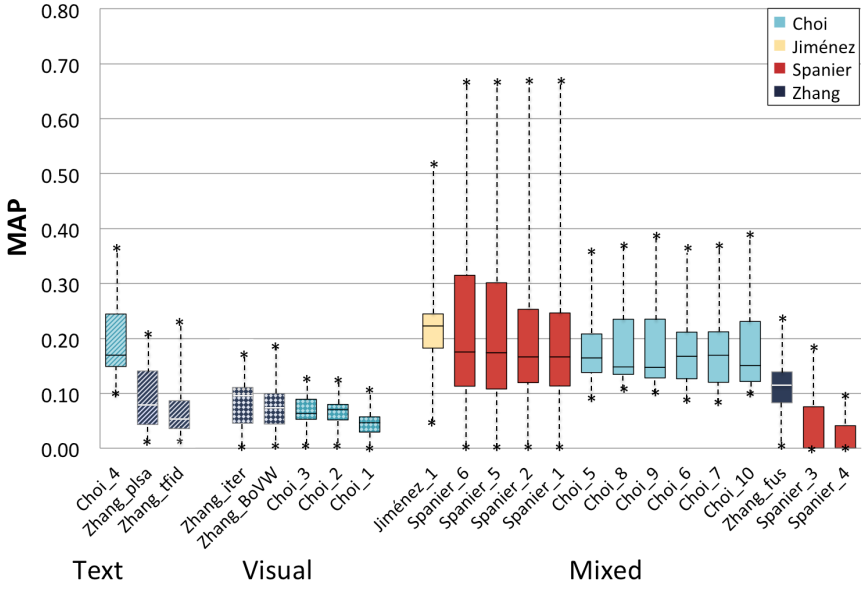


Fig. 1.4 Mean average precision (MAP) of the 22 runs in the Retrieval benchmark. Each run is represented by a box that is extended from the first to the third quartile of the query topic scores. The median score is shown as an horizontal line inside the box. The minimum and maximum scores obtained per run are shown as asterisks below and above their corresponding boxes. The runs are ordered per technique (only text, only visual and mixed) and per descending score order. The color of the boxes is defined by the submitting group as shown in the upper right legend. The color is stripped in only text runs, only visual runs are checkered and mixed runs are in solid color.

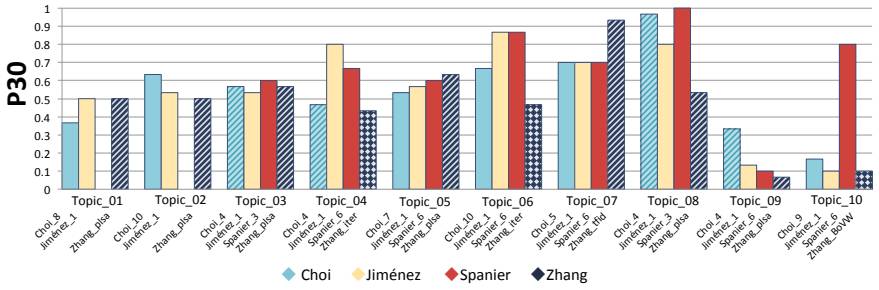


Fig. 1.5 P30 score obtained by the best run of each group, including text, visual and mixed, in the various query topics. The color from text only runs is stripped, visual only runs are checkered and mixed runs are in solid color bars. The name of the selected runs is shown below the corresponding bar.

Table 1.5 Scores from 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 1.6 Scores from 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

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 [8].

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 the mixed technique of Jiménez–del–Toro et al. This method also obtained the best MAP score in 6 out of the 10 query topics. showed promising results for most of the query topics. However, the runs from Spanier et al., specially those using both imaging modalities and all the pathological findings in the RadLex term lists (i.e. Spanier_6), showed promising results for most 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 all the benchmark (0.6758) and a P30 of 0.8. Jiménez del Toro et al. included the visual information in a late fusion with the tex-

tual 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.

1.7 Conclusions

The Retrieval benchmark was the first medical case–based retrieval benchmark using a large data set of 3D volumes and anatomy–pathology RadLex term lists. The data set 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 judgments performed by clinical experts. The organization and result analysis from the benchmark helps to address the current challenges in medical information retrieval and target the development of future benchmarks with common goals in this field.

1.8 Acknowledgments

This research was funded by the European Commission via the FP7 VISCERAL project (318068).

References

1. Choi, S.: Multimodal Medical Case–Based Retrieval on the Image and Report: SNUMedinfo at VISCERAL Benchmark. In: Multimodal Retrieval in the Medical Domain. Lecture Notes in Computer Science, vol. 9059. Springer (2015)
2. Foncubierta-Rodríguez, A., Müller, H.: Ground Truth Generation in Medical Imaging: A Crowdsourcing Based Iterative Approach. In: Workshop on Crowdsourcing for Multimedia, ACM Multimedia (oct 2012)
3. Hanbury, A., Müller, H., Langs, G., Weber, M.A., Menze, B.H., Fernandez, T.S.: Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis. In: CLEF conference. Springer Lecture Notes in Computer Science (2012)
4. García Seco de Herrera, A.: Use Case Oriented Medical Visual Information Retrieval & System Evaluation. Ph.D. thesis, University of Geneva (2015)
5. García Seco de Herrera, A., Foncubierta-Rodríguez, A., Markonis, D., Schaer, R., Müller, H.: Crowdsourcing for Medical Image Classification. In: Annual Congress SGMI 2014 (2014)
6. Jiménez-del-Toro, O.A., Cirujeda, P., Dicente Cid, Y., Müller, H.: RadLex Terms and Local Texture Features for Multimodal Medical Case Retrival. In: Multimodal Retrieval in the Medical Domain. Lecture Notes in Computer Science, vol. 9059. Springer (2015)
7. Jiménez-del-Toro, O.A., Hanbury, A., Langs, G., Foncubierta-Rodríguez, A., Müller, H.: Overview of the VISCERAL retrieval benchmark 2015. In: Multimodal Retrieval in the Medical Domain (MRMD) 2015. Lecture Notes in Computer Science, vol. 9059. Springer (2015)

8. Kurtz, C., Beaulieu, C.F., Napel, S., Rubin, D.L.: A hierarchical knowledge–based approach for retrieving similar medical images described with semantic annotations. Journal of biomedical informatics (2014)
9. Langlotz, C.P.: Radlex: A new method for indexing online educational materials. Radiographics 26(6), 1595–1597 (2006)
10. Langs, G., Müller, H., Menze, B.H., Hanbury, A.: Visceral: Towards large data in medical imaging — challenges and directions. Lecture Notes in Computer Science 7723, 92–98 (2013)
11. Quellec, G., Lamard, M., Bekri, L., Cazuguel, G., Roux, C., Cochener, B.: Medical case retrieval from a committee of decision trees. IEEE Transactions on Information Technology in Biomedicine 14(5), 1227–1235 (2010)
12. Spanier, A.B., Joskowicz, L.: Medical Case–Based Retrieval of Patient Records Using the Radlex Hierarchical Lexicon. In: Multimodal Retrieval in the Medical Domain. Lecture Notes in Computer Science, vol. 9059. Springer (2015)
13. Zhang, F., Song, Y., Cai, W., Depeursinge, A., Müller, H.: USYD/HES–SO in the VISCERAL Retrieval Benchmark. In: Multimodal Retrieval in the Medical Domain. Lecture Notes in Computer Science, vol. 9059. Springer (2015)