Overview of the VISCERAL Retrieval Benchmark 2015

Oscar Alfonso Jiménez–del–Toro[1], Allan Hanbury[2], Georg Langs[3], Antonio Foncubierta–Rodríguez[4], Henning Müller[1]

[1]University of Applied Sciences Western Switzerland (HES–SO), Switzerland
[2]Vienna University of Technology (TUW), Austria
[3]Medical University of Vienna (MUW), Austria
[4]Swiss Federal Institute of Technology (ETH) Zurich, Switzerland

Abstract. The results of the VISCERAL 3D case retrieval benchmark were presented during the Multimodal Retrieval in the Medical Domain (MRMD) 2015 workshop in Vienna, Austria on March 29, 2015. The main task for the participanta was to find and rank similar medical cases from a large multimodal (semantic RadLex terms extracted from text and visual 3D data) data set using a query case as input. The approaches that integrated information from both the RadLex terms and the 3D volumes provided in the benchmark obtained the best results based on 5 standard evaluation metrics. The benchmark set up, data set description and result analysis from the benchmark are presented for all the submitted methods.

Keywords: Medical Content–Based Retrieval, Multimodal Retrieval, Information Retrieval Infrastructures

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 [7]. Going through additional resources such as medical image repositories and inter–patient radiologic reports for medical case–based retrieval is currently inefficient and is not 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) Retrieval benchmark[1] aims 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 the following paper the 2015 Retrieval benchmark data set, evaluated task and results from the submitted approaches are presented.

[1] http://www.visceral.eu/benchmarks/retrieval-benchmark/, as of 1st may 2015.

1.1 Data Set

The VISCERAL Retrieval data set includes 2311 patient volumes obtained from computed tomography (CT) scans and T1– or T2–weighted magnetic resonance imaging (MRI). For a subset of these volumes (1813), a list of anatomy–pathology RadLex terms (APterms), in German, is also provided. RadLex is a unified language of radiology terms that can be used for standardized indexing and retrieval of radiology information resources [6]. 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, an example list is shown to illustrate the naming convention and the file content specifications. For each row of anatomy

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. Sample Anatomy/Pathology RadLex term list from the 2015 VISCERAL Retrieval data set. The lists are organized by columns and rows and each term is separated by a comma. From left to right in each row the following elements: anatomical structure radlex term (AnatRID), name of the structure in German (Anatomy), corresponding pathological radlex term (PathoRID), pathology name and negation (Neg). The pathological term is negated when the negation element is 1.

terms found in the report, the corresponding pathology is stated and marked if it was positive(0) or negative(1). For example, a positive report of liver cirrhosis will appear as: RID58,Leber,RID3822,Zirrhose,0. Table 1 shows an overview of the number of volumes per modality in the data set, as well as the number of APterms lists.

1.2 Content–Based Medical Image Retrieval

The general benchmark task was to evaluate the retrieval ranking of relevant medical cases from the data set taking 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 in this 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:

Table 1. Retrieval data set.

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**

1. a patient's 3D volume (CT, MRI);
2. a binary 3D bounding box of the region of interest containing the radiological signs of the pathology;
3. a manually annotated 3D binary mask of the main organ affected;
4. the APterms list extracted from the radiologic report.

The participants then had to develop an algorithm that finds clinically relevant (related) cases given a query case (imaging and text data), but without having access to the final diagnosis of the case.

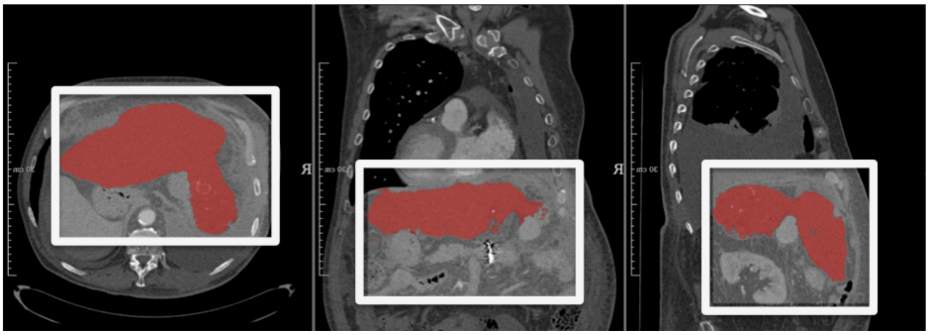

Fig. 2. Sample visual data provided per query case. The white block in the image represents the region of interest for the given case. The manually annotated organ with the main diagnosis is shown in red in the image

1.3 Evaluation

Relevance Judgements Evaluation of the submitted results by the participants was made with an interface using the Crowdflower platform[2]. This choice

[2] `http://www.crowdflower.com/`, as of 1st May 2015

was made following the suggestions of [4, 2] and as the interfce can both be used internally without payment and using the crowd workers. The evaluation task was divided into two parts: a task based on RadLex terms before the submissions and task based on pooling after the submissions.

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 after 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 of 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 judgements was that a case had to be relevant for differential diagnosis for the query case.

Metrics The trec_eval [3] tool 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 and has been 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).

1.4 Participants

Four research groups submitted results for the benchmark the benchmark after thirteen groups initially registered for the task:

[3] http://trec.nist.gov/trec_eval/, as of 1st May 2015

Spanier et al. [8] 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 an automatic 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 including text, visual and mixed retrieval, differentiated by the type of imaging used in the database cases, pathologic findings, region of interest or all these features together.

Zhang et al. [9] 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 only text approaches. The terms were weighted computing the frequencyinverse document frequency(TF–IDF) or with probabilistic latent semantic analysis(pLSA) to generate a probability distribution of the terms. For the only visual approach, the scale invariant feature transform (SIFT) was used to generate content descriptors for a bag–of–visual–words and was refined with a 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.

Choi [1] submitted runs for text, visual and mixed 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 imaging 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 weighted Borda–fuse method.

Jiménez del Toro et al. [5] 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 texture features with covariance descriptors, the local visual image similarity is added to the text information as an additional weight.

The information that the participants provided about their techniques is summarized in Table 2:

1.5 Results and Discussions

The following results of the Retrieval benchmark 2015 were presented at the *Multimodal Retrieval in the Medical Domain (MRMD) 2015* workshop, as part of the 37th European Conference on Information Retrieval (ECIR) 2015. The visualization of the results is structured as follows: For each run, the selected

Table 2. VISCERAL Retrieval algorithms in the submitted runs.

RunID	Group	Type	External training	Input	Language	Topics
BxcvfH_1	HebrewUniv	Mixed	No	Automatic	Ger/Eng	03–10
BxcvfH_2	HebrewUniv	Mixed	No	Automatic	Ger/Eng	03–10
BxcvfH_3	HebrewUniv	Mixed	No	Automatic	Ger/Eng	03–10
BxcvfH_4	HebrewUniv	Mixed	No	Automatic	Ger/Eng	03–10
BxcvfH_5	HebrewUniv	Mixed	No	Automatic	Ger/Eng	03–10
s5l55Q_01	MedGIFT	Mixed	No	Semi–auto	German	01–10
SNUMedinfo_01_SURF	SNUMedinfo	Image	No	Automatic	N/P	01–10
SNUMedinfo_02_SURF	SNUMedinfo	Image	No	Automatic	N/P	01–10
SNUMedinfo_03_SURF	SNUMedinfo	Image	No	Automatic	N/P	01–10
SNUMedinfo_04_Heur	SNUMedinfo	Text	No	Automatic	N/P	01–10
SNUMedinfo_05_HeSU	SNUMedinfo	Mixed	No	Automatic	N/P	01–10
SNUMedinfo_06_HeSU	SNUMedinfo	Mixed	No	Automatic	N/P	01–10
SNUMedinfo_07_HeSU	SNUMedinfo	Mixed	No	Automatic	N/P	01–10
SNUMedinfo_08_HeSU	SNUMedinfo	Mixed	No	Automatic	N/P	01–10
SNUMedinfo_09_HeSU	SNUMedinfo	Mixed	No	Automatic	N/P	01–10
SNUMedinfo_10_HeSU	SNUMedinfo	Mixed	No	Automatic	N/P	01–10
hNcmJn_BoVW	USYD	Image	No	Automatic	English	01–10
hNcmJn_fusion	USYD	Mixed	No	Automatic	English	01–10
hNcmJn_iter	USYD	Image	No	Automatic	English	01–10
hNcmJn_plsa	USYD	Text	No	Automatic	English	01–10
hNcmJn_tfidf	USYD	Text	No	Automatic	English	01–10

five evaluation metrics of trec_eval are provided as averages for all the topics contained within each run (num_q : number of queries, 10 total). Participants

Table 3. Scores from participant's runs using only textual information.

RunID	Type	MAP	GM–MAP	bpref	P10	P30
SNUMedinfo_04_Heur	Text	0.1942	0.1806	0.3221	0.5700	0.4967
hNcmJn_plsa	Text	0.0944	0.0697	0.1830	0.4100	0.3800
hNcmJn_tfidf	Text	0.0810	0.0582	0.1623	0.3700	0.2767

could submit a maximum of 10 runs and up to 300 ranked cases from the full data set per query topic. The runs are divided according to the techniques used for the query (textual, visual and mixed). The four teams submitted a total of 21 runs, with results for all the ten query topics, except for the approach of Spanier et al. which submitted results for 8 out of the 10 query topics. There were two

Table 4. Scores from participant's runs using only visual information.

RunID	Type	MAP	GM–MAP	bpref	P10	P30
hNcmJn_iter	Image	0.0828	0.0541	0.1881	0.3300	0.3300
hNcmJn_BoVW	Image	0.0783	0.0572	0.1900	0.0000	0.0333
SNUMedinfo_03_SURF	Image	0.0672	0.0474	0.1647	0.2700	0.3267
SNUMedinfo_02_SURF	Image	0.0661	0.0485	0.1671	0.2200	0.2633
SNUMedinfo_01_SURF	Image	0.0462	0.0188	0.1430	0.1400	0.1867

groups (Spanier et al. and Jiménez del Toro et al.) who submitted only mixed runs, using text and visual information. It is not straightforward to compare the influence of the visual or textual features based only on this results to the participants (Choi and Zhang et al.) who did submit results using only textual features or only visual features. However, these last two groups obtained higher scores using only textual features than their mixed runs. 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. On the other hand, 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. Overall, the best scores from the benchmark were obtained with mixed technique runs from Spanier et al. Both the best text only runs and best visual only runs were obtained by Choi. The text only runs by this participant had better scores than their mixed approach.

Table 5. Scores from participant's runs using a mixed (text and visual) technique.

RunID	Type	MAP	GM–MAP	bpref	P10	P30
BxcvfH_5	Mixed	0.2831	0.2308	0.3897	0.6875	0.6375
BxcvfH_2	Mixed	0.2625	0.2205	0.3720	0.6375	0.6208
BxcvfH_1	Mixed	0.2610	0.2183	0.3690	0.6875	0.6292
s5l55Q_01	Mixed	0.2367	0.2016	0.3664	0.5700	0.5533
SNUMedinfo_05_HeSU	Mixed	0.1875	0.1722	0.3082	0.5400	0.4600
SNUMedinfo_08_HeSU	Mixed	0.1867	0.1721	0.3099	0.5300	0.4533
SNUMedinfo_09_HeSU	Mixed	0.1861	0.1700	0.3143	0.4300	0.4700
SNUMedinfo_06_HeSU	Mixed	0.1858	0.1697	0.3102	0.4500	0.4633
SNUMedinfo_07_HeSU	Mixed	0.1857	0.1688	0.3097	0.3900	0.4567
SNUMedinfo_10_HeSU	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	0.0584	0.0024	0.0755	0.3625	0.3250
BxcvfH_4	Mixed	0.0282	0.0013	0.0731	0.0000	0.0208

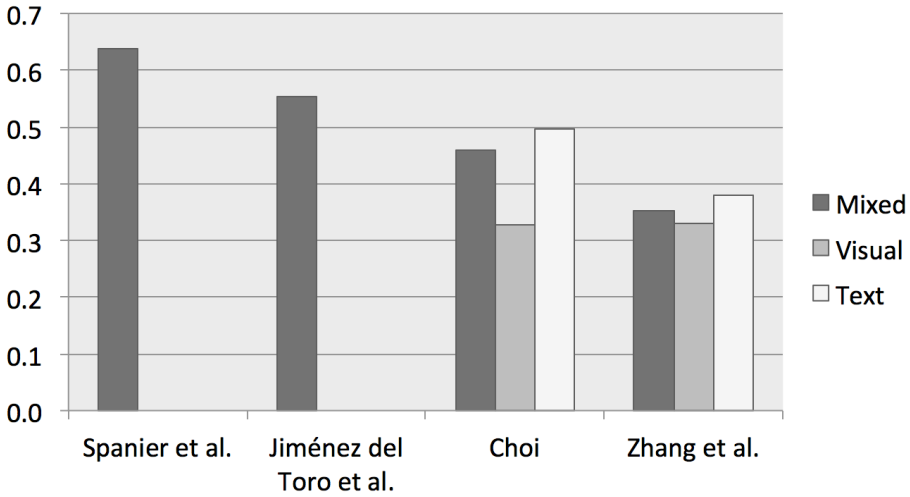

Fig. 3. P30 score obtained by the best run from each participant in the different techniques: text, visual and mixed.

2 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 an innovative 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 ofr the tasks. The results were compared using standard retrieval evaluation metrics. Multimodal approaches (using text+visual information) obtained the best results when compared to the gold standard relevance judgments performed by clinical experts. The discussion of the results and analysis during the MRMD2015 workshop with the attending groups helped to address the current challenges of medical information retrieval. This feedback could in turn, target the development of future benchmarks with common goals from the research community in this field.

3 Acknowledgments

This research was funded by the EU 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. García Seco de Herrera, A.: Use Case Oriented Medical Visual Information Retrieval & System Evaluation. Ph.D. thesis, University of Geneva (2015)
4. 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)
5. 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)
6. Langlotz, C.P.: Radlex: A new method for indexing online educational materials. Radiographics 26(6), 1595–1597 (2006)
7. 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)
8. 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)
9. 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)