www.visceral.eu

Result analysis for Competition 2

Deliverable number	*D4.4*
Dissemination level	*Public*
Delivery date	*18 May 2015*
Status	*Final*
Author(s)	*Antonio Foncubierta Rodríguez, Orcun Göksel, Allan Hanbury, Bjoern Menze, Henning Müller, Abdel Aziz Taha, Oscar Alfonso Jiménez del Toro*

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

Abstract

This document describes the three Benchmarks that were organised in the final year of the VISCERAL project and presents the results of these Benchmarks. The Benchmarks are:

- **Anatomy3:** Segmentation of organs in MRI and CT volumes (Section 2)
- **Detection:** Detection of lesions in MRI and CT volumes (Section 3)
- **Retrieval:** Retrieval of relevant cases given a query, using both text and visual information (Section 4)

For each Benchmark, a short description is provided, followed by a summary of the results of the submitted algorithms.

The Benchmarks have continued to be organised using the VISCERAL Cloud Evaluation Infrastructure. The Anatomy3 Benchmark is a continuation of the Anatomy1 and Anatomy2 Benchmarks, with the main distinguishing characteristics being the use of a larger training set and an online leaderboard. The Detection and Retrieval Benchmarks were run for the first time. These benchmarks have resulted in large amounts of annotated medical imaging data, which can continue to be used for further Benchmarks beyond the VISCERAL project.

Table of Contents

1 Introduction

This document describes the three Benchmarks that were organised in the final year of the VISCERAL project and presents the results of these Benchmarks. The Benchmarks are:

- **Anatomy3:** Segmentation of organs in MRI and CT volumes (Section 2)
- **Detection:** Detection of lesions in MRI and CT volumes (Section 3)
- **Retrieval:** Retrieval of relevant cases given a query, using both text and visual information (Section 4)

A workshop for the Anatomy3 Benchmark and Detection Benchmark was held at the ISBI 2015 conference.[1] The proceedings, including papers from participants describing their approaches in more detail, will appear in the CEUR proceedings series.

A workshop for the Detection Benchmark was held at the ECIR 2015 conference.[2] The proceedings of this workshop will appear in Springer LNCS.

2 Anatomy3 Benchmark

The Anatomy3 Benchmark continued the Anatomy1 and Anatomy2 Benchmarks. The main changes were the availability of more annotated data as well as an online leaderboard.

2.1 Description of the Benchmark

In this challenge, a set of annotated medical imaging data was provided to the participants, along with a powerful complimentary cloud-computing instance (8-core CPU with 16GB RAM) where participant algorithms can be developed and evaluated. The available data contains segmentation of several different anatomical structures in different image modalities, e.g. CT and MRI. Annotated structures in the training and testing data corpus included the segmentations of left/right kidney, spleen, liver, left/right lung, urinary bladder, rectus abdominis muscle, 1st lumbar vertebra, pancreas, left/right psoas major muscle, gallbladder, sternum, aorta, trachea, left/right adrenal gland.

As training, 20 volumes each were provided for 4 different image modalities and field-of-views, with and without contrast enhancement, which add up to 80 volumes in total. In each volume, up to 20 structures were segmented. The missing annotations are due to poor visibility of the structures in certain image modalities or due to such structures being outside the field-of-view. Accordingly, in all 80 volumes, a total of 1295 structures are segmented. A breakdown of annotations per anatomy can be seen in Figure 1.

[1] http://www.visceral.eu/workshops/anatomy-grand-challenge-workshop/

[2] http://www.visceral.eu/workshops/mrmd-2015/

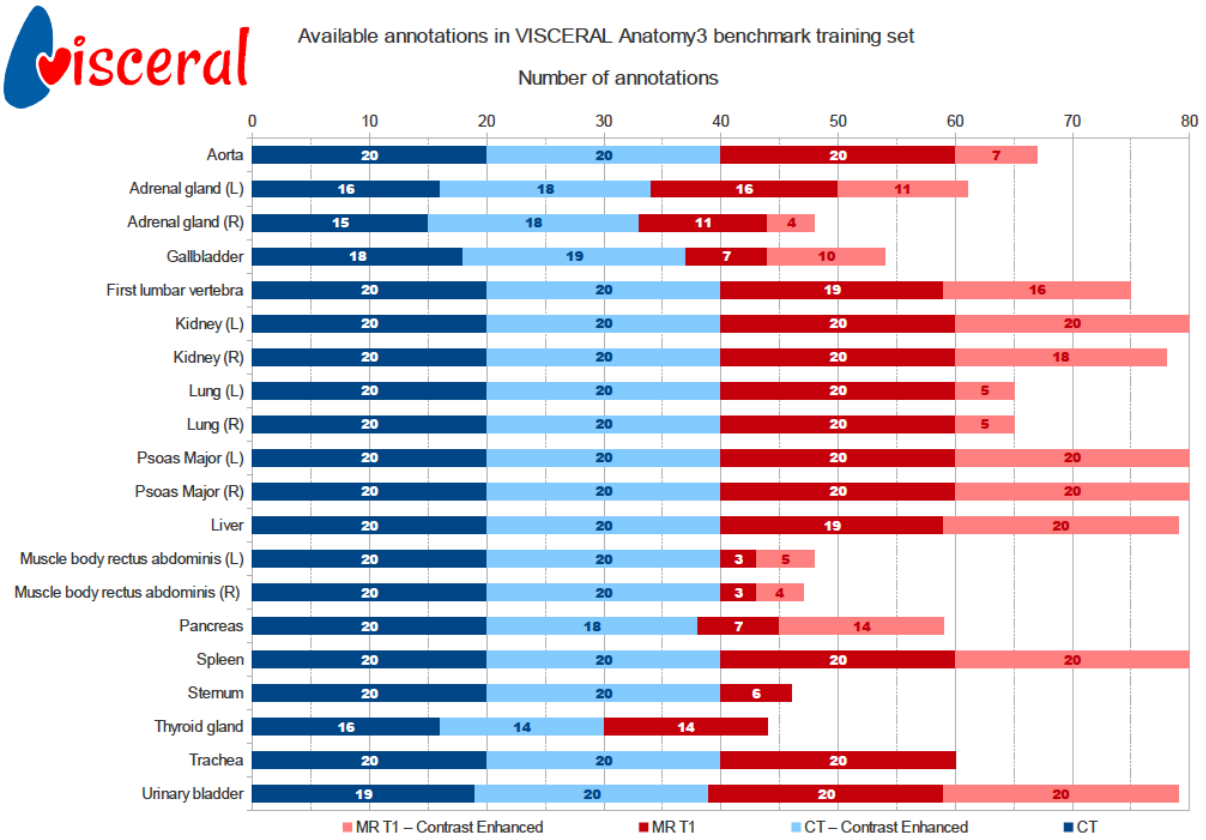

Figure 1 : Breakdown of annotations per anatomy

The participants need not address and segment all the structures involved in such data, but rather they could attempt any sub-problem thereof. For instance, an algorithm that could segment all organs in all the modalities was evaluated in those given categories for which it output any results. Accordingly, our results were presented at per-anatomy, per-modality evaluation result depending on the nature of participating algorithms and the attempted image analysis tasks. This is, indeed, inline with the VISCERAL vision of creating a single, large, and multi-purpose medical image dataset, on which different research groups can test their specific applications and solutions.

Participants for registered for a benchmark account at the VISCERAL registration website. For the options during registration, they had to pick "Anatomy 3 Benchmark" and their choice of operating system (Linux, Windows, etc.) for the virtual machine (VM), in order to get access to the VM and the data. Having signed the data usage agreement and uploaded it to the participant dashboard, they could then access the VM for development and the training data therein. They could additionally access the training dataset via FTP and download it for offline training.

They developed and installed their algorithms in the VM, while adapting and testing them on the training data, using the guidance of the Anatomy3 Guidelines for Participation that was published by us. They then prepared their executable on the VM according to the announced input/output specifications, and submitted their VMs (through "Submit VM" button in the dashboard) for the evaluation on the test data. We then ran their VM on the test data, and computed the relevant metrics.

This evaluation process could be performed several times during the training phase, nevertheless, we limited submissions to 1 per week, in order to prevent the participants “training on the test data”. The participants received feedback from their evaluations in a private leaderboard and had the option to make their results publicly available on the online leaderboard, if they wished.

2.2 Benchmark Results

Detailed results can be seen in the online leaderboard, accessible at:

http://visceral.eu:8080/register/Leaderboard.xhtml

Results are summarized in Table 1 and Table 2, for DICE coefficient and the mean surface distance. These are the most commonly used segmentation evaluation metrics. The former is an overlap metric, describing how well an algorithm estimates the target anatomical region. The latter is a surface distance metric, summarizing the overall surface estimation errors by a given algorithm.

The participant row refers to the citation for the publication contribution to Anatomy3 proceedings. In the Dice results that are categorized by image modality, the highest ranking methods are marked in bold. Any other method within 0.01 (1%) Dice of this are also considered a winner (or a tie) due to the insignificance of the difference. Dice values below a threshold are considered potentially unsuccessful results, even though depending on application they can still be useful. This cutoff is selected as a Dice value of 0.6, based on the gap in the reported participant results.

The results corresponding to the same bold values are also marked in the mean surface distance table, in order to facilitate comparison of the surface results for the best methods in terms of Dice metric. For successfully segmented organs (defined by the empirical 0.6 cutoff), both metrics agree on the results for all structures and modalities – except for the first lumbar vertebra in CT. The reader should note that the mean surface distances are presented in voxels, therefore the values between modalities (e.g. MRce and CT) are not directly comparable in the latter table.

Table 1: DICE Coefficient results for Anatomy3

DICE coefficient

Participant	Heinrich et	Jimenez et	He et al.	Cid et al.	Kahl et al.	Jimenez et	He et al.	Cid et al.
Modality	MRce	CTce	CTce	CTce	CT	CT	CT	CT
L Kidney	**0.862**	**0.91**	**0.91**		**0.934**	0.784		
R Kidney	**0.855**	0.889	**0.922**		**0.915**	0.79		
Spleen	**0.724**	0.73	**0.896**		**0.87**	0.703	**0.874**	
Liver	**0.837**	0.887	**0.933**		**0.921**	0.866	**0.923**	
L Lung		0.959	**0.966**	**0.974**	**0.972**	**0.972**	0.952	**0.972**
R Lung		**0.963**	**0.966**	**0.973**	**0.975**	**0.975**	0.957	**0.974**
Bladder	0.494	**0.679**			**0.763**	0.698		
Pancreas		0.423			0.383	0.408		
Gallbladder		0.484			0.19	0.276		
Thyroid		0.41			0.424	0.549		
Aorta		**0.721**			**0.847**	0.761		
Trachea		**0.855**			**0.931**	0.92		
Sternum		**0.762**			**0.83**	0.753		
Lumbar1		0.523			**0.775**	0.718		
L Adrenal G		0.331			0.282	0.373		
R Adrenal G		0.342			0.22	0.355		
L Psoas Maj	**0.801**	**0.794**			**0.861**	0.806		
R Psoas Maj	**0.772**	**0.799**			**0.847**	0.787		
L Rectus Abd		0.474			**0.746**	0.551		
R Rectus Abd		0.453			**0.679**	0.519		

Table 2: Mean Surface Distance results for Anatomy3

Mean Surface Distance [voxels]

Participant	Heinrich et	Jimenez et	He et al.	Cid et al.	Kahl et al.	Jimenez et	He et al.	Cid et al.
Modality	MRce	CTce	CTce	CTce	CT	CT	CT	CT
L Kidney	**0.251**	**0.172**	**0.171**		**0.147**	1.209		
R Kidney	**0.3**	0.243	**0.131**		**0.229**	1.307		
Spleen	**1.138**	2.005	**0.385**		**0.534**	1.974	**0.36**	
Liver	**0.935**	0.514	**0.203**		**0.299**	0.78	**0.239**	
L Lung		0.071	**0.069**	**0.05**	**0.045**	**0.043**	0.101	**0.05**
R Lung		**0.065**	**0.078**	**0.052**	**0.043**	**0.038**	0.094	**0.046**
Bladder	2.632	**1.879**			**1.057**	1.457		
Pancreas		3.804			4.478	5.521		
Gallbladder		3.603			9.617	5.938		
Thyroid		3.337			2.163	1.466		
Aorta		**0.899**			**0.542**	0.938		
Trachea		**0.223**			**0.083**	0.103		
Sternum		**1.094**			**0.798**	1.193		
Lumbar1		4.504			**2.424**	1.953		
L Adrenal G		3.115			3.298	2.672		
R Adrenal G		2.66			7.046	3.445		
L Psoas Maj	**0.493**	**0.742**			**0.443**	0.595		
R Psoas Maj	**0.569**	**0.757**			**0.55**	0.775		
L Rectus Abd		6.068			**1.614**	3.55		
R Rectus Abd		6.6			**1.922**	4.032		

According to these tables, there are different algorithms suitable and performing well for different anatomies, as one would anticipate. In contrast-enhanced MR modality, we had only a single participant, Heinrich et al., due to the difficulty in segmentation from this modality. In CTce, He et al. performed the best for the six structures they participated in, with some ties with Jimenez et al. The latter group segmented all the given structures in CTce, some of them with satisfactory accuracy, while for the others with potentially unusable results.

We had the most participants for the CT modality, in which the lungs – an almost solved segmentation problem – were segmented well by most participants; most likely at the accuracy of inter-subject annotations. For most other structures for which successful segmentations were achieved, Kahl et al. achieved the best results. Nevertheless, for structures where lower fidelity segmentations (defined by 0.6 cutoff) were attained, Jimenez et al. are seen to provide better structure estimations, likely due to the atlas-based approach they used. It is also observed that, despite the relatively good contrast of CT, several structures (prominently the pancreas, gallbladder, thyroid, and adrenal glands) are still quite challenging to segment from CT – potentially due to the lower sensitivity of CT to those structures also complicated by the difficult-to-generalize shapes of these anatomies.

3 Detection Benchmark

The Detection Benchmark was the first VISCERAL benchmark to consider pathology instead of anatomy. The goal of the benchmark is to automatically detect lesions in images acquired in clinical routine.

3.1 Description of the Benchmark

We distributed training data with a large number of expert annotated lesions for training of detection algorithms. This medical imaging data (CT, MRI) contains various lesions in anatomical regions such as the bones, liver, brain, lung, or lymph nodes. Some examples are shown in Figure 1. There are about 300 annotated lesions in the dataset.

The data set comprised different types of lesions that were visible in either the CT or the MRT T2 images: For multiple myeloma, a blood cancer that leads to lesions in the bone marrow, a total of 911 focal bone lesions showing signs of osteolysis had been annotated in CT. A set of 540 had also indicated bone marrow affection in MRI. The involvement of lymph nodes is always a sign of metastastatic tumor growth. In addition, lymph nodes may show signs of lymphoma, a primary cancer of the lymph nodes. Here, about 50 cases had been annotated in MR and CT. For other tumors, such as lung, liver, or brain, both primary tumors and metastasis had been annotated.

The image volumes used for the Detection Benchmark are the same as those used in the anatomy benchmark and, as a consequence, anatomical annotations are available for them as well. Overall there are about 1600 annotated lesions from 100 patients in the data-set, shown in Table 3.

During the training phase, participants were given imaging data and annotations in the form of lesion center position, and for large lesions, annotations that indicate the radius. During the benchmark phase, the algorithms had to return locations and type of the lesions (see Section 3.2).

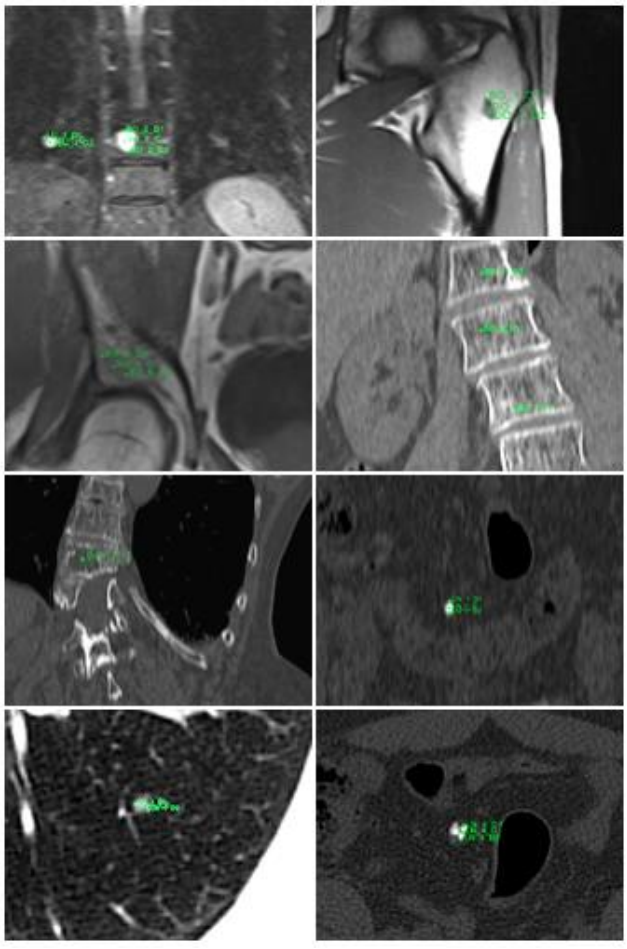

Figure 2: Annotations of different lesions. Shown are bone lesion from multiple myeloma patients, lymph nodes that show metastatic involvement, lung lesion

Table 3: Overview of the full data set comprising both training and testing data

# annotated volumes	modality	bone lesion	lung lesions	liver lesions	lymph node lesion	brain lesion	TOTAL
51	CT_wb	911	24	27	48	2	**1012**
50	MRT2_wb	541	5	44	1	6	**597**

3.2 Detection Metrics

3.2.1 Evaluation

In the detection task, an annotated lesion, L, is represented by three points, namely the center of the lesion, C_i, and two other points, $D1_i$ and $D2_i$, indicating the diameter of the lesion. Participating algorithms are expected to provide per lesion exactly one point, P_i, as near as possible to the center of the lesion, C_i.

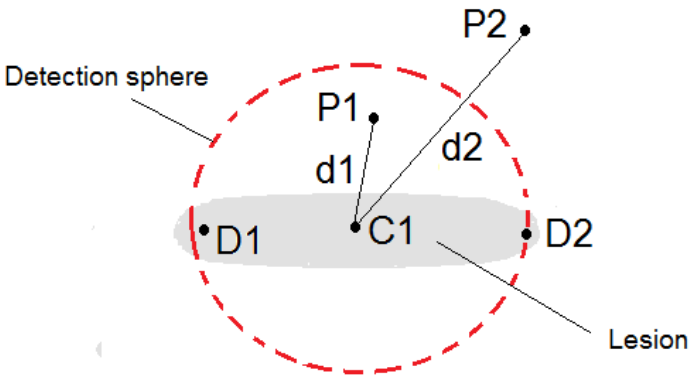

Figure 3: Schematic representation of a lesion annotated by the center C1 and the two diameter points D1 and D2. The points P1 and P2 are retrieved by an algorithm. P1 lies within the detection sphere and is thus considered as detected in contrast of the point P2.

The evaluation of the detection task takes place at three different levels:

i. **Lesion level**: for each annotated lesion, two values are measured, namely

- Minimum Euclidean distance, min(d_i): for each annotated lesion, the distance to the nearest point retrieved by the participating algorithm is measured as shown in Figure 3. This distance is provided for each annotated lesion, regardless of whether the lesion is considered as detected or not.
- Detection: a lesion is considered as detected if the point P_i, provided by the algorithm, is within the imaginary sphere centered on C_i and has the diameter given by the points D1 and D2. In particular, a radius of the sphere, r, is considered, which is equal to the distance between the center C_i and the farthest of the points D1 and D2. That is, a lesion is detected iff min(d) < r. In Figure 3, the point P1 is detected and P2 is not detected.

ii. **Volume level**: The confusion matrix (true positives, false positives, true negatives and false negatives) is calculated per volume, based on the detection values calculated in (i). From this confusion matrix, the precision (Percentage of correctly detected lesions), and the recall

(Percentage of total lesions detected) are calculated for each volume and participating algorithm. As it is expected that algorithms provide exactly one point per lesion, all other points that may be retrieved are considered as false positives.

iii. **Structure average level**: To test whether the scores of lesion detection are generally dependant on the structure, we calculate score averages (the Euclidean distance) for each structure over all volumes/participants.

3.2.2 Regions of Interest (Masks)

As mentioned above, it is expected that exactly one point per lesion is retrieved by each participating algorithm. To penalize algorithms that may try to improve evaluation results by providing many points, all other points retrieved are considered as false positives.

However, annotators have looked at specific regions of the volume, which means that one cannot be sure that other regions are free of lesions. In other words, participating algorithms could detect lesions that have not been annotated. To avoid penalizing such lesions, binary masks are used for each volume, which masks only those regions that have been annotated. Retrieved points that lie outside the mask are not considered in the confusion matrix.

3.3 Difficulty of the Benchmark

A key objective of the Lesion Detection Benchmark was to complement the anatomical information that we generated in the Silver and Gold Corpus by additional information about the diseases that were underlying the imaging data used in the annotation and retrieval benchmark. By this, we wanted to enhance the "application dimension" of the algorithms developed and tested during the annotation and retrieval benchmark. Vice versa, we saw a potential to have data sets for a number of 2-3 diseases with high radiological impact, that do not only comprise annotations of the lesions, but also a broad annotation of other anatomical structures, a unique feature for lesion detection data sets. As the localization of the disease organ (liver, bone, lung) is always the first step in such organs, this information is of high relevance in wide field of view image data sets.

As such, we saw the highest potential for participation through those groups that had previously annotated the anatomical structures in the Anatomy 1-3 challenges. While we initially received positive feedback by several of these groups, none of them participated at the end.

Somewhat disappointingly, we did not have participants from other groups for this particular challenge either. Originating from the Anatomy and Retrieval data, we obtained a data set that was rather diverse in terms of the diseases and lesions visible from them. This may have led to difficulties in communicating key objectives of the challenge to other groups that would have potentially have had the algorithms for addressing some of the detection subtasks. Moreover, while detecting abnormalities is possible when a good anatomical reference is given (e.g., obtained through algorithms from Anatomy1-3), it is significantly harder to detect them without this context, and related pattern recognition and machine learning algorithms need significantly larger data bases. Only bone and liver lesion detection tasks would have had significant numbers for such a "lesion detection only" algorithm here.

We see that new detection benchmarks we will have to a) be much more targeted in terms of the disease that is addressed, preferably only addressing a single disease or diagnostic task; b) we should establish data sets that are homogenous with respect to the image modalities used and if there are two or more, then they should define separate subtasks; and c), we will have to compile a ***significantly*** larger data set with hundreds to thousands of lesions that capture the full diversity of lesions and that would allow them to be detected without previous anatomical annotation.

4 Retrieval Benchmark

The Retrieval Benchmark considered a different scenario in medical imaging, namely the retrieval of images relevant to a query based on both visual and textual information.

4.1 Description of the Benchmark Evaluation

The retrieval of relevant medical cases based on a query case is evaluated. It serves the following scenario: a user is assessing a query case in a clinical setting, e.g. a CT volume with a dubious diagnosis, and is searching for cases that are relevant to this assessment. The participant's algorithm has to find cases that are relevant in a large data base of cases. Each topic (query case) is composed of:

- The patient 3D imaging data (CT, MRI)
- 3D bounding box region of interest containing the main radiological signs of the pathology
- Binary mask of the main organ affected
- Radiologic report extracted anatomy-pathology terms in form of csv files.

Volumes are in NIFTI file format.

The submitted approaches must find clinically relevant (related) cases given a query case (imaging and text data) without information on the final diagnosis. For each topic, the algorithms should generate a ranked list of search results out of the VISCERAL Retrieval dataset (containing imaging data and text data).

A set of ten test query cases (topics) were used to evaluate the result rankings of the algorithms. The database contains both imaging data and corresponding text data. There are two query scenarios:

- Image data and ROI considered for the query
- Image data, ROI and text data (anatomy-pathology RadLeX terms) considered for the query.

While the first case is of immediate clinical relevance, we expect also the second case to be valuable in evaluating specific retrieval algorithms. Therefore the use of text information is optional during retrieval evaluation. The algorithms were evaluated in three groups corresponding to the information used for the retrieval [image+ROI], [image+ROI+text] or [mixed].

All submissions are summarised in Table 4. The information that the participants provided about their techniques is below (linked to the result lists in the Appendix by the abbreviations in square brackets). Proceedings of the workshop containing papers describing the methods in more detail will be published by Springer.

[SNUMedinfo]

Sungbin Choi

Seoul National University

Multimodal medical case-based retrieval on the radiology image and report: SNUMedinfo at VISCERAL Retrieval Benchmark

We extracted low-level visual feature (SURF) from image and trained query-specific SVM classifier for imaging retrieval. For textual retrieval, we estimated relevance with anatomy-pathology paired RadLexID similarity function. In mixed retrieval, we combined them using weighted Borda-fuse method.

[s5l55Q]

Oscar Alfonso Jiménez del Toro, Pol Cirujeda, Yashin Dicente Cid, Adrien Depeursinge and Henning Müller

University of Applied Sciences Western Switzerland (HES-SO)

Case-based medical image retrieval: Clinical and image texture similarities

A retrieval method for medical cases that uses both textual and visual features was used. It defines a weighting scheme that combines the RadLex terms anatomical and clinical correlations with the information from local texture features obtained from the region of interest in the query cases. The method implementation uses an innovative 3D Riesz wavelet texture analysis and an approach to generate a common spatial domain to compare medical images. The proposed method obtained overall competitive results in the VISCERAL Retrieval benchmark and could be seen as a tool to perform medical case based retrieval in large clinical data sets.

[BxcvfH]

Assaf B. Spanier and Leo Joskowicz

School of Computer Science and Engineering, the Hebrew University of Jerusalem

Medical case-based retrieval of patient records using the RadLex hierarchical lexicon

We use a new method for the retrieval radiological cases from a database of clinical cases described by terms from the RadLex lexicon. The input is a database of cases and a query consisting of the patient volumetric scan, a user defined region of interest in it, and a list of RadLex terms from the radiological report. The output is list of the most relevant cases from the database in decreasing order. Our method uses the RadLex terms and their hierarchical representation to define a similarity metric between terms based on their relative location in the hierarchy. For this purpose, we develop the Augmented RadLex Graph, a data structure that augments the RadLex hierarchy with links derived from the terms in the case reports, and a search algorithm that ranks case similarity based on the link distance between the terms in the graph. Our method was evaluated in the VISCERAL Retrieval Benchmark Challenge on 8 queries and a database of 1,813 cases. It ranked first in 6 out of the 8 cases tested.

[hNcmJn]

Fan Zhang, Yang Song, Weidong Cai, Adrien Depeursinge and Henning Müller

School of Information Technologies, University of Sydney

USYD/HES-SO in the VISCERAL Retrieval Benchmark

Given a query case, the cases with highest similarities in the database were retrieved. Five runs were submitted for the ten queries provided in the task, of which two were based on the anatomy-pathology terms, two were based on the visual image content, and the last one was based on the fusion of the aforementioned four runs.

Table 4: VISCERAL Retrieval benchmark query set up from algorithms

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
s5155Q_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

4.2 Retrieval Metrics

The trec_eval[1] tool was used for the evaluation of participants' submissions in the Retrieval benchmark. This program uses the standard NIST (US National Institute of Standards and Technology) evaluation procedures and has been previously extensively used for the Text Retrieval Conference (TREC). It's currently a commonly used program for comparing different information retrieval techniques, particularly for text documents, but that can be applied also to images and cases.

Medical experts performed relevance assessment for the top 300 ranked cases by each approach, to judge the quality of the retrieval systems. The main evaluation measures considered for the evaluation were the precision of the top ranked cases. The precision for top ranked 10 and 30 cases (P@10, P@30), mean uninterpolated average precision (MAP), the bpref measure, and the Rprecision were included in the evaluation.

4.3 Results

The average results for each of the participant runs is presented in the Appendix. The analysis of the medical case-based retrieval benchmark is structured as follows: For each run, all the evaluation metrics from trec_eval are provided as averages for all the topics addressed in each run (num_q : number of queries, 10 total). Participants could submit a maximum of 10 runs and up to 300 ranked cases from the full dataset per query topic. The results for the text-only, visual-only and mixed

[1] http://trec.nist.gov/trec_eval/, as of 29 April 2015

submissions are shown in Table 5, Table 6, and Table 7 respectively. Figure 4 summarises the best P@30 results for each participant.

The four teams submitted a total of 12 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 groups (Spanier et al. and Jimenez 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 these 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, Jimenez 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 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

Table 6: 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

Table 7: 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

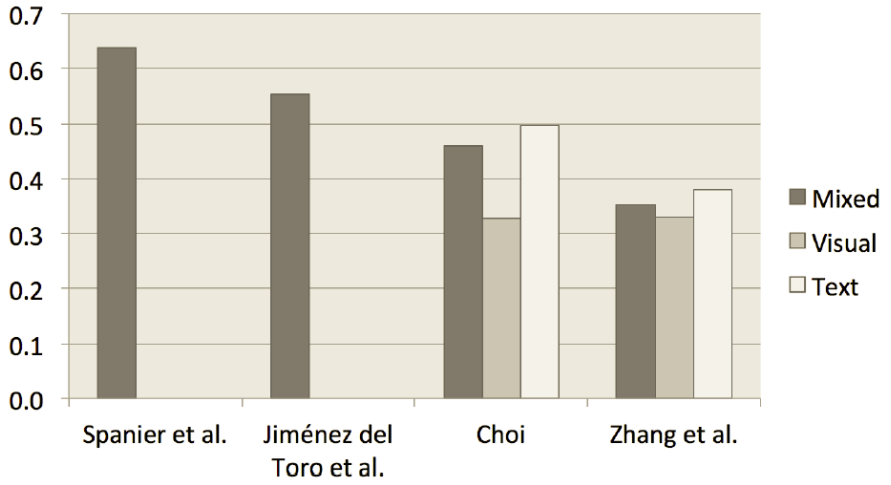

Figure 4: P@30 score obtained by the best run from each participant in the different techniques: text, visual and mixed.

5 Conclusion

During the last year of the VISCERAL project, three Benchmarks were organised in parallel: Anatomy3, Detection and Retrieval. The organisation of these Benchmarks led to the creation of large amounts of annotated medical imaging data, which will continue to be available beyond the end of the VISCERAL project. The Detection and Retrieval benchmarks represented new types of Benchmarks in the medical imaging domain, and were therefore potentially seen as unusual and rather difficult. Based on experience gained in organising these Benchmarks, we plan to improve them and continue running variants on them in the future, based on annotated imaging data already created.

6 Appendix: Detailed Results from the Retrieval Benchmark

[SNUMedinfo]

runid	all	SNUMedinfo_01
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	495
map	all	0.0462
gm_map	all	0.0188
Rprec	all	0.1552
bpref	all	0.1430
recip_rank	all	0.1439
iprec_at_recall_0.00	all	0.2778
iprec_at_recall_0.10	all	0.2405
iprec_at_recall_0.20	all	0.1204
iprec_at_recall_0.30	all	0.0000
iprec_at_recall_0.40	all	0.0000
iprec_at_recall_0.50	all	0.0000
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.1000
P_10	all	0.1400
P_15	all	0.1600
P_20	all	0.1600
P_30	all	0.1867
P_100	all	0.2380
P_200	all	0.1870
P_500	all	0.0990
P_1000	all	0.0495

---

runid	all	SNUMedinfo_02
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	581
map	all	0.0661
gm_map	all	0.0485
Rprec	all	0.1851
bpref	all	0.1671
recip_rank	all	0.3528
iprec_at_recall_0.00	all	0.4672
iprec_at_recall_0.10	all	0.3018
iprec_at_recall_0.20	all	0.2081
iprec_at_recall_0.30	all	0.0033

iprec_at_recall_0.40	all	0.0000
iprec_at_recall_0.50	all	0.0000
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.2000
P_10	all	0.2200
P_15	all	0.2600
P_20	all	0.2750
P_30	all	0.2633
P_100	all	0.2840
P_200	all	0.2280
P_500	all	0.1162
P_1000	all	0.0581

---

runid	all	SNUMedinfo_03
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	575
map	all	0.0672
gm_map	all	0.0474
Rprec	all	0.1839
bpref	all	0.1647
recip_rank	all	0.4479
iprec_at_recall_0.00	all	0.5355
iprec_at_recall_0.10	all	0.3154
iprec_at_recall_0.20	all	0.2009
iprec_at_recall_0.30	all	0.0027
iprec_at_recall_0.40	all	0.0000
iprec_at_recall_0.50	all	0.0000
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.2400
P_10	all	0.2700
P_15	all	0.2533
P_20	all	0.3050
P_30	all	0.3267
P_100	all	0.2820
P_200	all	0.2255
P_500	all	0.1150
P_1000	all	0.0575

---

runid	all	SNUMedinfo_04
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	943

map	all	0.1942
gm_map	all	0.1806
Rprec	all	0.3355
bpref	all	0.3221
recip_rank	all	0.7778
iprec_at_recall_0.00	all	0.8384
iprec_at_recall_0.10	all	0.5165
iprec_at_recall_0.20	all	0.4764
iprec_at_recall_0.30	all	0.3870
iprec_at_recall_0.40	all	0.2473
iprec_at_recall_0.50	all	0.0242
iprec_at_recall_0.60	all	0.0164
iprec_at_recall_0.70	all	0.0126
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.6200
P_10	all	0.5700
P_15	all	0.5600
P_20	all	0.5300
P_30	all	0.4967
P_100	all	0.4350
P_200	all	0.3675
P_500	all	0.1886
P_1000	all	0.0943

---

runid	all	SNUMedinfo_05
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	953
map	all	0.1875
gm_map	all	0.1722
Rprec	all	0.3098
bpref	all	0.3082
recip_rank	all	0.7250
iprec_at_recall_0.00	all	0.8100
iprec_at_recall_0.10	all	0.5040
iprec_at_recall_0.20	all	0.4491
iprec_at_recall_0.30	all	0.3679
iprec_at_recall_0.40	all	0.1968
iprec_at_recall_0.50	all	0.0891
iprec_at_recall_0.60	all	0.0228
iprec_at_recall_0.70	all	0.0067
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.5800
P_10	all	0.5400
P_15	all	0.4733
P_20	all	0.4500
P_30	all	0.4600
P_100	all	0.4380

P_200	all	0.3750
P_500	all	0.1906
P_1000	all	0.0953

---

runid	all	SNUMedinfo_06
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	955
map	all	0.1858
gm_map	all	0.1697
Rprec	all	0.3105
bpref	all	0.3102
recip_rank	all	0.6833
iprec_at_recall_0.00	all	0.7850
iprec_at_recall_0.10	all	0.4993
iprec_at_recall_0.20	all	0.4659
iprec_at_recall_0.30	all	0.3621
iprec_at_recall_0.40	all	0.1928
iprec_at_recall_0.50	all	0.0293
iprec_at_recall_0.60	all	0.0249
iprec_at_recall_0.70	all	0.0079
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.4600
P_10	all	0.4500
P_15	all	0.4600
P_20	all	0.4700
P_30	all	0.4633
P_100	all	0.4460
P_200	all	0.3680
P_500	all	0.1910
P_1000	all	0.0955

---

runid	all	SNUMedinfo_07
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	957
map	all	0.1857
gm_map	all	0.1688
Rprec	all	0.3092
bpref	all	0.3097
recip_rank	all	0.6583
iprec_at_recall_0.00	all	0.7745
iprec_at_recall_0.10	all	0.4888
iprec_at_recall_0.20	all	0.4562
iprec_at_recall_0.30	all	0.3619
iprec_at_recall_0.40	all	0.1926
iprec_at_recall_0.50	all	0.0846
iprec_at_recall_0.60	all	0.0248
iprec_at_recall_0.70	all	0.0081

iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.3800
P_10	all	0.3900
P_15	all	0.4200
P_20	all	0.4450
P_30	all	0.4567
P_100	all	0.4470
P_200	all	0.3675
P_500	all	0.1914
P_1000	all	0.0957

---

runid	all	SNUMedinfo_08
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	967
map	all	0.1867
gm_map	all	0.1721
Rprec	all	0.3092
bpref	all	0.3099
recip_rank	all	0.7833
iprec_at_recall_0.00	all	0.8417
iprec_at_recall_0.10	all	0.5108
iprec_at_recall_0.20	all	0.4439
iprec_at_recall_0.30	all	0.3565
iprec_at_recall_0.40	all	0.2364
iprec_at_recall_0.50	all	0.0819
iprec_at_recall_0.60	all	0.0195
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.6000
P_10	all	0.5300
P_15	all	0.4667
P_20	all	0.4900
P_30	all	0.4533
P_100	all	0.4320
P_200	all	0.3760
P_500	all	0.1934
P_1000	all	0.0967

---

runid	all	SNUMedinfo_09
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	971
map	all	0.1861
gm_map	all	0.1700
Rprec	all	0.3172
bpref	all	0.3143

recip_rank	all	0.7583
iprec_at_recall_0.00	all	0.8139
iprec_at_recall_0.10	all	0.4935
iprec_at_recall_0.20	all	0.4650
iprec_at_recall_0.30	all	0.3535
iprec_at_recall_0.40	all	0.2402
iprec_at_recall_0.50	all	0.0831
iprec_at_recall_0.60	all	0.0195
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.4800
P_10	all	0.4300
P_15	all	0.4467
P_20	all	0.4600
P_30	all	0.4700
P_100	all	0.4360
P_200	all	0.3715
P_500	all	0.1942
P_1000	all	0.0971

---

runid	all	SNUMedinfo_10
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	974
map	all	0.1845
gm_map	all	0.1681
Rprec	all	0.3122
bpref	all	0.3110
recip_rank	all	0.6833
iprec_at_recall_0.00	all	0.7819
iprec_at_recall_0.10	all	0.4815
iprec_at_recall_0.20	all	0.4515
iprec_at_recall_0.30	all	0.3535
iprec_at_recall_0.40	all	0.2412
iprec_at_recall_0.50	all	0.0829
iprec_at_recall_0.60	all	0.0197
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.3600
P_10	all	0.3900
P_15	all	0.3933
P_20	all	0.4250
P_30	all	0.4500
P_100	all	0.4370
P_200	all	0.3705
P_500	all	0.1948
P_1000	all	0.0974

[s5l55Q]

runid	all	s5l55Q
num_q	all	10
num_ret	all	3000
num_rel	all	2462
num_rel_ret	all	1077
map	all	0.2367
gm_map	all	0.2016
Rprec	all	0.3572
bpref	all	0.3664
recip_rank	all	0.5421
iprec_at_recall_0.00	all	0.7839
iprec_at_recall_0.10	all	0.6437
iprec_at_recall_0.20	all	0.5537
iprec_at_recall_0.30	all	0.4455
iprec_at_recall_0.40	all	0.2501
iprec_at_recall_0.50	all	0.1345
iprec_at_recall_0.60	all	0.0800
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.5200
P_10	all	0.5700
P_15	all	0.5333
P_20	all	0.5400
P_30	all	0.5533
P_100	all	0.5110
P_200	all	0.4140
P_500	all	0.2154
P_1000	all	0.1077

[BxcvfH]

runid	all	BxcvfH_01
num_q	all	8
num_ret	all	2400
num_rel	all	1916
num_rel_ret	all	722
map	all	0.2610
gm_map	all	0.2183
Rprec	all	0.3619
bpref	all	0.3690
recip_rank	all	0.9375
iprec_at_recall_0.00	all	0.9688
iprec_at_recall_0.10	all	0.6875
iprec_at_recall_0.20	all	0.5153
iprec_at_recall_0.30	all	0.4076

iprec_at_recall_0.40	all	0.2072
iprec_at_recall_0.50	all	0.1837
iprec_at_recall_0.60	all	0.1062
iprec_at_recall_0.70	all	0.1062
iprec_at_recall_0.80	all	0.0792
iprec_at_recall_0.90	all	0.0179
iprec_at_recall_1.00	all	0.0000
P_5	all	0.7750
P_10	all	0.6875
P_15	all	0.7000
P_20	all	0.6625
P_30	all	0.6292
P_100	all	0.4512
P_200	all	0.3631
P_500	all	0.1805
P_1000	all	0.0903

---

runid	all	BxcvfH_02
num_q	all	8
num_ret	all	2400
num_rel	all	1916
num_rel_ret	all	736
map	all	0.2625
gm_map	all	0.2205
Rprec	all	0.3647
bpref	all	0.3720
recip_rank	all	0.9375
iprec_at_recall_0.00	all	0.9688
iprec_at_recall_0.10	all	0.6912
iprec_at_recall_0.20	all	0.5253
iprec_at_recall_0.30	all	0.4233
iprec_at_recall_0.40	all	0.2131
iprec_at_recall_0.50	all	0.1896
iprec_at_recall_0.60	all	0.1062
iprec_at_recall_0.70	all	0.1062
iprec_at_recall_0.80	all	0.0792
iprec_at_recall_0.90	all	0.0179
iprec_at_recall_1.00	all	0.0000
P_5	all	0.7250
P_10	all	0.6375
P_15	all	0.6583
P_20	all	0.6125
P_30	all	0.6208
P_100	all	0.4525
P_200	all	0.3688
P_500	all	0.1840
P_1000	all	0.0920

---

runid	all	BxcvfH_03
num_q	all	8
num_ret	all	1568
num_rel	all	1916
num_rel_ret	all	221

map	all	0.0584
gm_map	all	0.0024
Rprec	all	0.0787
bpref	all	0.0755
recip_rank	all	0.4429
iprec_at_recall_0.00	all	0.5005
iprec_at_recall_0.10	all	0.3124
iprec_at_recall_0.20	all	0.1677
iprec_at_recall_0.30	all	0.0000
iprec_at_recall_0.40	all	0.0000
iprec_at_recall_0.50	all	0.0000
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.4250
P_10	all	0.3625
P_15	all	0.3250
P_20	all	0.3313
P_30	all	0.3250
P_100	all	0.2475
P_200	all	0.1381
P_500	all	0.0553
P_1000	all	0.0276

---

runid	all	BxcvfH_04
num_q	all	8
num_ret	all	1568
num_rel	all	1916
num_rel_ret	all	221
map	all	0.0282
gm_map	all	0.0013
Rprec	all	0.0787
bpref	all	0.0731
recip_rank	all	0.0227
iprec_at_recall_0.00	all	0.2197
iprec_at_recall_0.10	all	0.1960
iprec_at_recall_0.20	all	0.1258
iprec_at_recall_0.30	all	0.0000
iprec_at_recall_0.40	all	0.0000
iprec_at_recall_0.50	all	0.0000
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.0000
P_10	all	0.0000
P_15	all	0.0000
P_20	all	0.0000
P_30	all	0.0208
P_100	all	0.1900

P_200	all	0.1381
P_500	all	0.0553
P_1000	all	0.0276

---

runid	all	BxcvfH_05
num_q	all	8
num_ret	all	2400
num_rel	all	1916
num_rel_ret	all	788
map	all	0.2831
gm_map	all	0.2308
Rprec	all	0.3869
bpref	all	0.3897
recip_rank	all	0.9375
iprec_at_recall_0.00	all	0.9688
iprec_at_recall_0.10	all	0.6958
iprec_at_recall_0.20	all	0.5199
iprec_at_recall_0.30	all	0.4561
iprec_at_recall_0.40	all	0.2789
iprec_at_recall_0.50	all	0.2085
iprec_at_recall_0.60	all	0.2075
iprec_at_recall_0.70	all	0.1062
iprec_at_recall_0.80	all	0.0792
iprec_at_recall_0.90	all	0.0179
iprec_at_recall_1.00	all	0.0000
P_5	all	0.7750
P_10	all	0.6875
P_15	all	0.6833
P_20	all	0.6750
P_30	all	0.6375
P_100	all	0.4600
P_200	all	0.3775
P_500	all	0.1970
P_1000	all	0.0985

[hNcmJn]

runid	all	hNcmJn_BoVW
num_q	all	10
num_ret	all	2940
num_rel	all	2462
num_rel_ret	all	603
map	all	0.0783
gm_map	all	0.0572
Rprec	all	0.2061
bpref	all	0.1900
recip_rank	all	0.6260
iprec_at_recall_0.00	all	0.6660
iprec_at_recall_0.10	all	0.3170
iprec_at_recall_0.20	all	0.2251
iprec_at_recall_0.30	all	0.0573
iprec_at_recall_0.40	all	0.0397

iprec_at_recall_0.50	all	0.0246
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.3000
P_10	all	0.2500
P_15	all	0.2667
P_20	all	0.2950
P_30	all	0.2833
P_100	all	0.2930
P_200	all	0.2370
P_500	all	0.1206

---

runid	all	hNcmJn_fusion
num_q	all	10
num_ret	all	2864
num_rel	all	2462
num_rel_ret	all	688
map	all	0.1101
gm_map	all	0.0766
Rprec	all	0.2343
bpref	all	0.2070
recip_rank	all	0.4685
iprec_at_recall_0.00	all	0.6179
iprec_at_recall_0.10	all	0.3990
iprec_at_recall_0.20	all	0.3266
iprec_at_recall_0.30	all	0.0774
iprec_at_recall_0.40	all	0.0455
iprec_at_recall_0.50	all	0.0366
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.4400
P_10	all	0.4200
P_15	all	0.4000
P_20	all	0.3700
P_30	all	0.3533
P_100	all	0.3540
P_200	all	0.2860
P_500	all	0.1376
P_1000	all	0.0688

---

runid	all	hNcmJn_iter
num_q	all	10
num_ret	all	2940
num_rel	all	2462
num_rel_ret	all	595
map	all	0.0828
gm_map	all	0.0541

Rprec	all	0.1938
bpref	all	0.1881
recip_rank	all	0.5389
iprec_at_recall_0.00	all	0.6103
iprec_at_recall_0.10	all	0.3478
iprec_at_recall_0.20	all	0.2345
iprec_at_recall_0.30	all	0.0333
iprec_at_recall_0.40	all	0.0293
iprec_at_recall_0.50	all	0.0253
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.3600
P_10	all	0.3300
P_15	all	0.3067
P_20	all	0.3400
P_30	all	0.3300
P_100	all	0.3170
P_200	all	0.2380
P_500	all	0.1190
P_1000	all	0.0595

---

runid	all	hNcmJn_plsa
num_q	all	10
num_ret	all	2896
num_rel	all	2462
num_rel_ret	all	614
map	all	0.0944
gm_map	all	0.0697
Rprec	all	0.1999
bpref	all	0.1830
recip_rank	all	0.4812
iprec_at_recall_0.00	all	0.6113
iprec_at_recall_0.10	all	0.3879
iprec_at_recall_0.20	all	0.1852
iprec_at_recall_0.30	all	0.1036
iprec_at_recall_0.40	all	0.0316
iprec_at_recall_0.50	all	0.0316
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.4200
P_10	all	0.4100
P_15	all	0.3733
P_20	all	0.3600
P_30	all	0.3800
P_100	all	0.3360
P_200	all	0.2440
P_500	all	0.1228

P_1000	all	0.0614

---

runid	all	hNcmJn_tfidf
num_q	all	10
num_ret	all	2896
num_rel	all	2462
num_rel_ret	all	528
map	all	0.0810
gm_map	all	0.0582
Rprec	all	0.1808
bpref	all	0.1623
recip_rank	all	0.5193
iprec_at_recall_0.00	all	0.6396
iprec_at_recall_0.10	all	0.3189
iprec_at_recall_0.20	all	0.1611
iprec_at_recall_0.30	all	0.0714
iprec_at_recall_0.40	all	0.0323
iprec_at_recall_0.50	all	0.0323
iprec_at_recall_0.60	all	0.0000
iprec_at_recall_0.70	all	0.0000
iprec_at_recall_0.80	all	0.0000
iprec_at_recall_0.90	all	0.0000
iprec_at_recall_1.00	all	0.0000
P_5	all	0.3200
P_10	all	0.3700
P_15	all	0.3400
P_20	all	0.3150
P_30	all	0.2767
P_100	all	0.2160
P_200	all	0.2025
P_500	all	0.1056
P_1000	all	0.0528