Variation of Relevance Assessments for Medical Image Retrieval

Henning Müller[1], Paul Clough[2], Bill Hersh[3], and Antoine Geissbühler[1]

[1] University and Hospitals of Geneva, Medical Informatics, Geneva, Switzerland
henning.mueller@sim.hcuge.ch
[2] Department of Information Studies, Sheffield University, Sheffield, UK
p.d.clough@sheffield.ac.uk
[3] Biomedical Informatics, Oregon Health and Science University, Portland, OR, USA
hersh@ohsu.edu

Abstract. Evaluation is crucial for the success of most research domains, and image retrieval is no exception to this. Recently, several benchmarks have been developed for visual information retrieval such as TRECVID, ImageCLEF, and ImagEval to create frameworks for evaluating image retrieval research. An important part of evaluation is the creation of a ground truth or gold standard to evaluate systems against. Much experience has been gained on creating ground truths for textual information retrieval, but for image retrieval these issues require further research. This article will present the process of generating relevance judgements for the medical image retrieval task of ImageCLEF. Many of the problems encountered can be generalised to other image retrieval tasks as well, so the outcome is not limited to the medical domain. Part of the images analysed for relevance were judged by two assessors, and these are analysed with respect to their consistency and potential problems. Our goal is to obtain more information on the ambiguity of the topics developed and generally to keep the variation amongst relevance assessors low. This might partially reduce the subjectivity of system-oriented evaluation, although the evaluation shows that the differences in relevance judgements only have a limited influence on comparative system ranking. A number of outcomes are presented with a goal in mind to create less ambiguous topics for future evaluation campaigns.

1 Introduction

Visual information retrieval has been an extremely active research domain for more than 20 years [1]. It includes several diverse research areas such as information retrieval, computer vision, image analysis, and pattern recognition. Despite the enormous research effort spent on analysing and retrieving images, still many questions remain and visual retrieval has still not become part of consumer or industrial applications in the same way that text retrieval has. Of all similar research domains, text retrieval is probably the one with the most realistic benchmarks and evaluation scenarios. Since the 1960s, standardised testing and comparisons between research systems and methods has been common [2],

S. Marchand-Maillet et al. (Eds.): AMR 2006, LNCS 4398, pp. 233–247, 2007.

and TREC[1] (TExt Retrieval Conference) has become the standard 'model' for large–scale evaluation of different aspects of information access [3]. Besides running several benchmarks in an annual cycle of data release, topic release, submissions, ground truthing, evaluation and workshop, TREC has also managed to analyse many of TREC submissions from participating systems. In addition, analysis of the relevance judgements (or ground truths) have been undertaken by researchers to obtain a better idea of the statistical properties required to accurately and reliably compare systems [4]. Subjectivity in judgements was show to exist but also to have only a very limited influence on comparative system rankings.

In image retrieval evaluation was neglected for a long time, although a few proposals and initiatives did exist [5,6,7], such as the Benchathlon[2]. Over the past few years, several visual information retrieval benchmarks have shown that a strong need exists to evaluate visual information retrieval in a standardised manner. TRECVID, for example started as a task in TREC but has since become an independent workshop on the evaluation of video retrieval systems [8]. The strong participation has also made this benchmark important for image retrieval where evaluation can be performed on extracted video key frames. Another initiative is ImagEval[3], financed by the French research foundation and with participants mainly from the French research community. INEX[4] (INiative for the Evaluation of XML retrieval) has also started a multimedia retrieval task in 2006. A fourth benchmarking event is ImageCLEF [9,10]. This event is part of the Cross–Language Evaluation Forum (CLEF) campaign to evaluate and compare multilingual information retrieval systems [11]. ImageCLEF concentrates on the retrieval of images from multilingual repositories and combining both visual and textual features for multimodal retrieval. A strong participation in ImageCLEF over the past two years has shown the need for standardised system comparison and the importance of creating an infrastructure to support the comparisons in this way. This can dramatically reduce the effort required by researchers to compare their approaches: able to concentrate on developing novel methods rather than issues associated with evaluation.

This article will first present an overview of ImageCLEF, its collections, topics, participants, and results. Following this, a closer look at the relevance judgements is undertaken, and in particular at the judgements for the topics assessed by two judges. The conclusions summarise our findings and provide ideas for future development of information needs (or topics).

2 ImageCLEFmed 2005

This section describes the main components of the medical ImageCLEF benchmark in 2005: ImageCLEFmed.

[1] http://trec.nist.gov/
[2] http://www.benchathlon.net/
[3] http://www.imageval.org/
[4] http://inex.is.informatik.uni-duisburg.de/2006/

2.1 Collections Used

A total of four collections were used for ImageCLEFmed 2005, all with separate annotations in a wide variety of XML formats containing a large variety of images. The Casimage[5] dataset [12] contains almost 9'000 images (all modalities, photographs, illustrations, etc.) of 2'000 cases with annotations mainly in French, but also in part in English. Each case can contain one to several different images of the same patient (or condition). The PEIR[6] (Pathology Education Instructional Resource) database uses annotations based on the HEAL[7] project (Health Education Assets Library, mainly Pathology images [13]). This dataset contains over 33'000 images (extremely varied but a majority of pathology images) with English annotations. Each image has an associated annotation rather than per case as in the Casimage collection. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology[8] [14], was also made available to us. This dataset contains over 2'000 images mainly from nuclear medicine with annotations in English per case. Finally, the PathoPic[9] collection (Pathology microscopic images [15]) was part of our benchmark's dataset. It contains 9'000 images, each with extensive annotations in German (and parts translated into English).

This provided a heterogeneous database of more than 50'000 images in total, with annotations in three different languages (although the majority in English). Through an agreement with the copyright holders, we were able to distribute these images to participating research groups of ImageCLEF free of charge. Challenges of the data with respect to text include: different structures and formats, incomplete or partial annotations with a large number of empty cases, domain-specific (i.e. medical) vocabulary and images, unusual abbreviations and spelling errors. Even with a consistent XML structure, not all fields were filled in correctly with many of the fields containing free–text. Visual challenges include the large variety of data sources and sorts of images used and a considerable variation of images of the same modality or anatomic region as the images were taken and processed by a large number of different programs and machines. Image size and quality vary also strongly. Another challenge is of course the combination of visual and textual data as input for a query.

2.2 Topics

The image topics were based on a small survey administered to clinicians, researchers, educators, students, and librarians at Oregon Health & Science University (OHSU)[16]. Based on this survey, topics for ImageCLEFmed were developed along one or more of the following axes:

- Anatomic region shown in the image;
- Image modality (x–ray, CT, MRI, gross pathology, ...);

[5] http://www.casimage.com/
[6] http://peir.path.uab.edu/
[7] http://www.healcentral.com/
[8] http://gamma.wustl.edu/home.html
[9] http://alf3.urz.unibas.ch/pathopic/intro.htm

– Pathology or disease shown in the image;
– abnormal visual observation (eg. enlarged heart).

The goal of topic development was also to create a mix of topics to test different aspects of visual and textual retrieval. To this end, three topics groups were created: visual topics, mixed topics and purely semantic topics. The grouping of topics into these categories was performed manually based upon the assumption that visual topics would perform well with visual–only retrieval, mixed topics would require semantic text analysis together with visual information, and the semantic topics were expected not to profit at all from visual analysis of the images. The topics were generated by the ImageCLEF organisers and not by the relevance judges. A total of 25 topics (11 visual, 11 mixed and 3 semantic) were distributed to the participants. All topics were in three languages: English, French, German. Each topic was accompanied by one to three example images of the concept and one topic also contained a negative example image. In this context topics means a specific information need of a possible user that is described by multimodal means. It was verified through tests with a visual and a textual retrieval system that all topics had at least three relevant images.

2.3 Participants Submissions

In 2004 the medical retrieval task was entirely visual and 12 participating groups submitted results. In 2005, as a mixture of visual and non-visual retrieval, 13 groups submitted results. This was far less than the number of registered participants (28). We send a mail to all registered groups that did not submit results to ask for their reasons. There non–submission was partly due to the short time span between delivery of the images and the deadline for submitting results. Another reason was that several groups registered very late, as they did not have information about ImageCLEF beforehand. They were mainly interested in the datasets and future participation in ImageCLEF. All groups that did not submit results said that the datasets and topics were every valuable resource for their research. In total, 134 ranked lists from different systems (runs) were submitted from the twelve research groups, among them 128 automatic runs that had no manual adaptation or feedback and only very few (6) manual runs that could include relevance feedback, query reformulation, or manual optimisations of feature weights based on the collection.

2.4 Pooling and Constraints for the Judgement Process

Relevance assessments were performed by graduate students who were also physicians in the OHSU biomedical informatics program. A simple interface was used from previous ImageCLEF relevance assessments. Nine judges, eight medical doctors and one image processing specialist with medical knowledge, performed the relevance judgements. Half of the images for most topics were judged in duplicate to enable the analysis of assessor–subjectivity in the judgement process.

In large collections, it is impossible to judge all documents to establish their relevance to an information need or search topic. Therefore a method called pooling where assessors judge "pools" of documents rather than all documents in a collection [17]. In our case the unity for judgement was the image and not the case, also to make the task harder for pure text retrieval. To obtain these pools the first 40 images from the top of each submitted run were collected and used to create pools resulting in an average pool size of 892 images. The largest pool size was 1.167 and the smallest 470. We aimed to have less than 1'000 images to judge per topic to reduce effort. Even so, it was estimated to take on average three hours to judge all images in a pool for a single topic. Compared to the purely visual topics from 2004 (around one hour of judgement per topic with each pool containing an average of 950 images), the judgement process was found to take almost three times as long. This is likely due to the use of "semantic" topics requiring the judges to view the associated annotations to verify relevance, and/or the judges needing to view an enlarged version of the image. The longer assessment time may have also been due to the fact that in 2004 all images were pre–marked as irrelevant, and only relevant images required a change. In 2005, we did not have images pre–marked. Still, this process was generally faster than the time required to judge documents in previous text retrieval [18], and irrelevant images could be established very quickly. In text retrieval, however, checking documents for irrelevance takes longer and requires more cognitive effort.

2.5 Outcome of the Evaluation

The results of the benchmark showed a few clear trends. Very few groups submitted runs involving manual relevance feedback, most likely due to the requirement of more resource to do this. Still, relevance feedback has shown to be extremely useful in many retrieval tasks and its evaluation is extremely important. The ImageCLEF interactive retrieval task suffered from similar problems with a small number of participants. Surprisingly, in the submitted runs relevance feedback did not appear to offer much improvement compared to the automatic runs. In the 2004 tasks, runs with relevance feedback were often significantly better than without feedback.

The results also showed that purely textual systems (best run: Mean Average Precision (MAP)=0.2084) had better overall performance than purely visual systems (best run: MAP=0.1455). For the visual topics, the visual and textual or mixed systems gave comparable performance. By far the best results were obtained when combining visual and textual features (MAP=0.2821) [19]. The best system actually separated the topics into their main axes (anatomy, modality, pathology) and performed a query along these axes with the supplied negative feedback concepts (if an MRI is searched for, all other modalities can be fed back negatively).

3 Analysis of the Relevance Judgements and Their Variations

This section analyses our relevance judgement process of 2005 with the goal to find clues for reducing the subjectivity among relevance judges in future tasks.

3.1 The Relevance Judgement Process

In 2005 we used the same relevance judgement tool as in 2004. We used a ternary judgement scheme that allows assessors to mark images as relevant, partially relevant and non–relevant. The judges received a detailed explanation on the judgement process including the fact that partially relevant was only to be used if it cannot be outruled that the image might correspond to the concept. If only a part of the concept was fulfilled (i.e. an x–ray with emphysema when the search was for a CT with emphysema) the image had to be regarded as non–relevant. Judges had the possibility to read the text that came with the images and they also had the possibility to enlarge the images on screen to see more detail. This relevance definition is somewhat different from the relevance definition used in TREC, where a document is regarded as relevant even if only a small part of it is relevant. Much more on relevance can be found in [20,21]. The judges were given a description of relevance but no explicit description with respect to where the limits of relevance were. They could ask questions when they were unsure, which happened several times.

As the judgement tool (see Figure 1) was web–based, the judges were able to perform relevance judgements at will. In total, three weeks were foreseen for the judgement process and topics were distributed among the 8 judges, with each person responsible for three topics (and one person doing four). The image processing judge did a single topic, only. No time constraint was given on judging topics or that they had to finish judgements for one topic in one go. This was to allow for breaks in between finishing topics. Participating judges told us that a judgement took an average of three hours, but no further details were asked about the process. This is slightly more than in 2004, where visual topics took an average of one hour per topic with a slightly larger number of images per topic. After the single judgements were finished we asked judges to judge the first half the images of three more topics. Some judges did not have the time for the double judgements and so only part of the topics are double–judged. Only the first topic was entirely judged by two judges. For the other topics the first half of the images was double–judged to have a maximum of relevant images double–judged. Indeed, as the images to be judged were ordered by the numbers of runs that they were included in, the first half contains many more relevant images than the second have resulting in most relevant images being judged twice in this process.

The images were shown on screen starting with those images that most runs had in their 1'000 submitted results. The goal of this was to have a concentration of relevant documents at the beginning when the judge is (hopefully) more

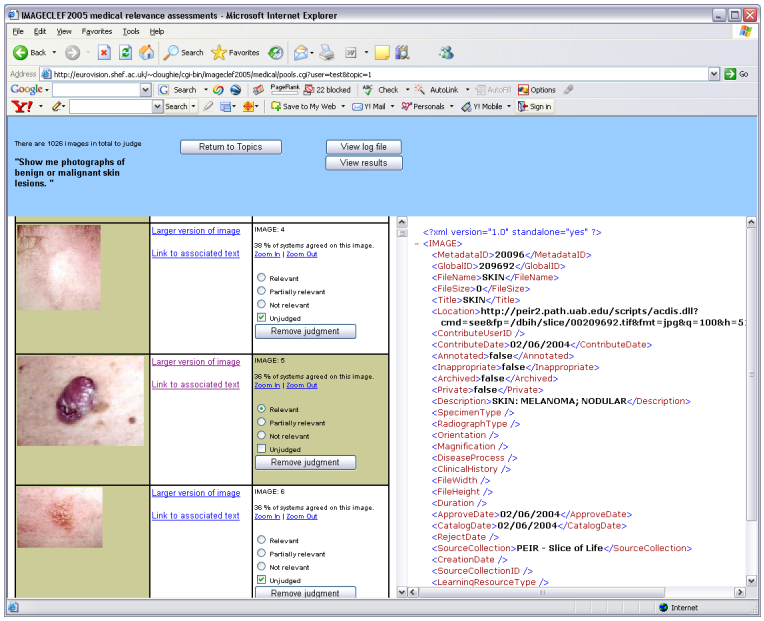


Fig. 1. A screen shot of the tool for acquiring the relevance judgements

attentive and less likely to be suffering from fatigue. However, this could lead to later images being judged less carefully as there are less relevant items.

3.2 Selection of Judges and Problems Encountered

One important point of a domain-specific benchmark is that the judges need to have a sufficient knowledge of the domain to judge topics correctly. On the other hand, this also limits the depth of the topics that can be constructed even if the judges are knowledgeable. We choose students from the OHSU graduate student program in medical informatics. All of the eight chosen students are also physicians and can thus be regarded as domain experts for the medical topics constructed in a rather general medical context. No knowledge on specific diseases was necessary as the text of the images was regarded as sufficient.

Several problems were encountered in the process. One of the problems was with respect to the relevance judgement tool itself. As it showed all images on a single screen it took fairly long to build the page in the browser (containing around 1'000 images). Another problem was that the tool required to specifically modify the settings of the browser to enable JavaScript and disable all caching so the changes were stored directly in the database. As many different browsers under Linux, Mac OS X and Windows were used, some problems with browsers occurred that lead to a loss of some judgements that afterwards had to be repeated. Unfortunately, browser-based environments still seem to suffer from differences from one environment to another.

Table 1. Differences encountered in the topics judged twice

Topic	#	same	different	+/+	0/0	-/-	+/0	-/0	+/-
1	1018	916	102 (10.02%)	193	3	720	19	50	33
2	440	372	68 (15.45%)	49	8	315	30	23	15
3	441	361	80 (18.14%)	75	1	285	8	41	31
4	383	356	27 (7.05%)	59	8	289	9	16	2
8	491	471	20 (4.07%)	14	1	456	14	5	1
9	550	517	33 (6.00%)	79	33	405	23	10	0
10	235	226	9 (3.83%)	6	0	220	1	0	8
11	492	487	5 (1.02%)	23	0	464	1	2	2
12	326	281	45 (13.80%)	10	2	269	5	22	18
13	484	338	146 (30.17%)	214	7	117	49	34	63
14	567	529	38 (6.70%)	51	0	478	22	1	15
15	445	438	7 (1.57%)	29	0	409	3	0	4
16	467	460	7 (1.50%)	1	0	459	0	1	6
17	298	224	74 (24.83%)	15	2	207	11	27	36
18	403	394	9 (2.23%)	1	0	393	0	7	2
19	441	439	2 (0.45%)	11	0	428	0	1	1
20	608	314	294 (48.35%)	1	11	392	236	26	22
21	401	276	125 (31.17%)	131	4	141	30	48	47
22	448	395	53 (11.83%)	36	3	356	11	24	18
23	472	454	18 (3.81%)	24	0	430	1	3	14
total	9'410	8'238	1'072 (11.39%)	1'212 (12.87%)	83 (0.88%)	7'233 (76.87%)	473 (5.03)	341 (3.62%)	338 (3.60%)

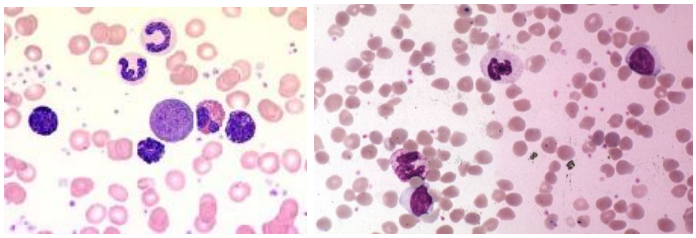
Sometimes, the text available with images made it hard to judge semantic topics that required assessors to also read the annotation text. For these topics, where the user was not sure about the results and could not choose from the image itself, we recommended selecting a partially relevant judgement.

Most of the comments and questions received from judges during the assessment process were with respect to the partially relevant relevance level. Generally, relevance and non-relevance could be determined fairly quickly, whereas they contacted us when not sure about the outcome.

3.3 Differences Per Topic

In Table 1 we can see for each topic how many double judgements were available, how many times the judges agreed and disagreed and then, how many times what kind of difference between the judges occurred. The three different section in the table are for visual topics, mixed topics and semantic topics. As notation we have *+* for a relevant judgement, *0* for a partially relevant judgement and *–* for a non–relevant judgement. Combinations such as *–/+* mean that one judge judged the image relevant and another one non–relevant.

It can be seen that, fortunately, the agreement between the judges is fairly high. In our case the judges agree in 88.61% of their judgements. A more common measure for inter–judge agreement is the Kappa score. In our case the Kappa



Show me microscopic pathologies of cases with chronic myelogenous leukemia.
Zeige mir mikroskopische Pathologiebilder von chronischer Leukämie (Chronic myelogenous leukemia, CML).
Montre-moi des images de la leucémie chronique myélogène.

Fig. 2. Topic 20, where the judges disagreed the most strongly

score using three categories is 0.679, which indicates a good agreement and is for example much higher than in the similar Genomics TREC [18] where it is usually around 0.5.

It becomes clear that there is a difference with respect to which categories were judged incorrectly, when limiting ourself to only the images and topics judged twice. From 15145 negative judgements, only 4.48% are in agreement. From the 3235 positive judgements, already 25.07% are in disagreement and the worst are the partially relevant judgements, where 814 of 980 (83.06%) are not in agreement.

When looking at topic groups (visual, mixed, semantic) it is clearly visible that we cannot judge the semantic topics as only a single topic was judged twice, which is statistically insufficient. The mixed topics on the other hand have a much higher average disagreement than the visual topics. The four topics with the highest disagreement among judges are from this category although a few mixed topics with high agreement do exist. For topic 20, the disagreement among relevant items is actually next to 0%, meaning that these topics will need to be avoided in the future or additional instructions for the judges are required.

The various forms of disagreement (relevant/non–relevant, partially/relevant, partially/non–relevant) occur in similar quantities, and underline the fact that determining irrelevance is easy, relevance is harder, and with the partially relevant items much disagreement exists.

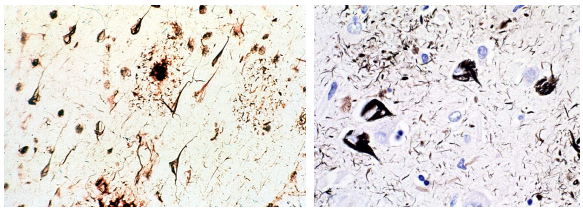
Another tendency that can be seen is that most topics with a very high disagreement have a large number of relevant items. Topics with a very small number of relevant items seem clearer defined and have less ambiguity.

3.4 Ambiguous and Non–ambiguous Topics

After having looked at the table it becomes clear that a per topic analysis needs to be done as differences are large. Here, the two most agreed upon and the two least agreed upon topics are discussed.

Show me all x–ray images showing fractures.
Zeige mir Röntgenbilder mit Brüchen.
Montres–moi des radiographies avec des fractures.

Fig. 3. Topic 21, where the judges disagreed the second most strongly

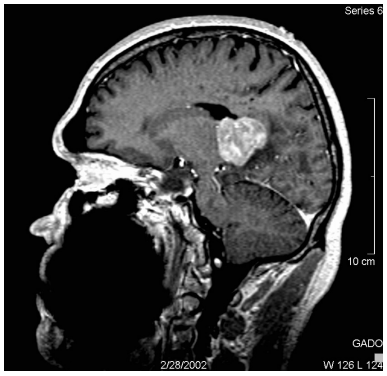


Show me microscopic pathologies of cases with Alzheimers disease.
Zeige mir mikroskopische Pathologiebilder von Fällen der Alzheimer Krankheit.
Montre-moi des images microscopiques de cas avec Alzheimer.

Fig. 4. Topic 19, where the judges agreed the most strongly

Figure 2 shows the topic with the strongest disagreement among judges. It becomes apparent that two of the experts must have interpreted this description in different ways. It is possible that one of the judges marked any case with leukemia whereas another judge marked the same sort of images with no further specification as chronic and myelogenous in the text as partially relevant. These sort of topics can profit from describing not only what is relevant but also clearly what can not be regarded as relevant.

In Figure 3 the topics the second most often disagreed upon is shown. This topic actually seems very surprising as it seems extremely well defined with very clear example images. It is only imaginable that one person actually searched the images for micro fractures or searched the text for the word fracture as well whereas the second judge only took into account very clearly visible fractures. For example, an image can show a healed fracture, when fracture appears in the text but is not anymore visible in the image.



Show me sagittal views of head MRI images.
Zeige mir sagittale Ansichten von MRs des Kopfes.
Montre–moi des vues sagittales dIRMs de la tête.

Fig. 5. Topic 11, where the judges agreed the second most strongly

Figure 4 shows the least ambiguous topic. It is very clear that for this topic it was necessary to read the text and find the word Alzheimer, so no purely visual identification of relevance was possible. This finally lead to a very homogeneous judgement. The number of relevant items is also very small and thus well defined. Looking for such a simple keywords seems well–defined, and excluding non pathology images should also be quick simply by visual identification.

Figure 5 is finally the second least ambiguous topic. Again, it is very well defined as such views (sagittal) only occur on MRI and mixing up CT and MRI seems impossible in this case. The view also leads to a small number of finally relevant images.

Unfortunately, it is not easy to find a few determining factors to identify ambiguous or non-ambiguous topics. Topic creation needs to include several people to review topics and the descriptions to the judges also need to be defined extremely well to limit subjectivity in the judgement process.

3.5 Influence of Varying Judgements on the Results

When looking at the agreement table it is clear that topics with an extreme disagreement exist and we have to inspect this closer to find out whether this agreement can influence the final results. Still, for the official evaluation, only the primary judge was taken into account and all partially relevant were also regarded as relevant. We finally generated several sets of relevance judgements based on all judgements and including the double judgements. For images with a single judgement, only the primary judge was taken into account.

- strict – when the primary judge judges images as relevant, only, the final results is relevant;
- Lenient – when the primary judge says relevant or partially relevant it is relevant (default for system ranking);

- AND strict – when both judges say relevant;
- AND lenient – if both judges say relevant or partially relevant;
- OR strict – if any one judge says relevant;
- OR lenient – if any one judge says relevant or partially relevant;

The evaluations of all runs were repeated and the systems re–ranked. The absolute number of relevant items changes strongly according to this rule. It becomes very quickly clear that the absolute differences in performance occur but that the ranking of systems changes basically not at all. A few systems are ranked several positions lower but only very few systems gain more than two ranks and if they do so, the absolute differences are very small. A per topics analysis on the influence of judgements on performance is currently in preparation and would be too much for this paper.

4 Discussion and Conclusions

It becomes clear very quickly that the relevance judgement process for visual information retrieval evaluation is extremely important. Although many classification or computer vision tasks try to simulate users and automatically create judgements [22], in our opinion such a process needs to include real users. Only for very specific tasks can automatic judgements be generated, e.g. completely classified collections [23].

A few important guidelines need to be taken into account when creating new topics that are to be judged:

- a relevance judgement tool has to be easy to use, based on simple web technologies to work in every browser;
- the judgement tool could include the possibility to query visually or by text to examine also images not covered by the pools;
- the description of topics for judges needs to be as detailed as possible to accurately define the topic; it needs to also include negative examples and a description of what is regarded as partially relevant;
- trying to target a limited number of relevant images for the topics as a large number increases both the subjectivity and also increases the risk that the pool is lacking some relevant images;
- work on realistic topics as judges can more easily relate to these topics and imagine the result that they would expect;
- limit the judgement process to a certain maximum time in a row, describe how pauses should be done to have more stable and reproducible conditions for the judgement process;

Our first step to improve the judgement process is the judgement tool. The goal is to have a tool that only shows a limited number of images on screen and is thus faster to use. Access to an enlarged image and the full text of the images needs to be quick. The possibility to search for visually similar images or to search the database by keywords needs to be possible. This can improve the relevance sets by adding images that have not been in the judgement pools.

A simple change to ease evaluation after the campaign is to have the same number of topics in the three categories visual, mixed and semantic. Our goal for 2006 is to have ten topics of each category to get more of an idea about how this influences the judgement process.

When creating these new topics we have now a larger basis for creating realistic scenarios. Besides two user survey among medical professionals, the log files of the health on the net[10] HONmedia search engine were developed to create realistic topics. This should make it easier for the judges to have an idea about the desired outcome. At the same time a clearer definition of *relevance* in out context is needed as this has been less studied for images. Along with this, a clearer topic definition for the judges is needed that does not only describe when an image must be judged as relevant, but also gives examples of non–relevant and partially relevant images. Particularly important is the partially relevant level because judges were less sure about this level which has led to lower agreement. This could be improved by a more formal definition of partially relevant. It still seems important for us to have a category for partially relevant as this can help us to identify problematic areas for a particular topic. It is important to verify afterwards that the final system ranking is not significantly influenced by the diversity of the relevance judgements. Several judgement sets for more strict or rather lenient judgements will be created for this. We still have to decide whether we really want to have stronger constraint for the judges such as a limit of one hour for judging to avoid fatigue or even choose the place for the judgements in a lab. This might improve the results but it also bears a risk to limit the motivation of the judges by giving them too many constraints.

Another very simple thing to employ is the reduction of the number of relevant items. We simply need to perform test queries ahead of topic release to make sure that the number of relevant items stays limited. A rough number of a maximum of 100 relevant items seems reasonable. Although this cannot be solved exhaustively ahead of time some simple constraint can improve the judgement process.

It is becoming clear that evaluation of visual information retrieval system is starting to grow. Standardised evaluation and use of standard datasets is becoming increasingly common and at the main multimedia conferences systems become comparable through these standard datasets such as TRECVID. Still, to better create topics and adapt the entire evaluation process to the needs of visual data, much work is needed. Whereas text retrieval has 30 years of experience, for visual retrieval much work is still needed to better define the concepts of relevance and particularly real application scenarios than can make the techniques usable for real users.

Acknowledgements

Part of this research was supported by the Swiss National Science Foundation with grant 205321-109304/1. We also acknowledge the support of the EU FP6 project SemanticMining (IST NoE 507505).

[10] `http://www.hon.ch/`

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content–based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22 No 12** (2000) 1349–1380
2. Cleverdon, C.W.: Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, Aslib Cranfield Research Project, Cranfield, USA (1962)
3. Voorhees, E.M., Harmann, D.: Overview of the seventh Text REtrieval Conference (TREC–7). In: The Seventh Text Retrieval Conference, Gaithersburg, MD, USA (1998) 1–23
4. Zobel, J.: How reliable are the results of large–scale information retrieval experiments? In Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J., eds.: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, ACM Press, New York (1998) 307–314
5. Smith, J.R.: Image retrieval evaluation. In: IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'98), Santa Barbara, CA, USA (1998) 112–113
6. Leung, C., Ip, H.: Benchmarking for content–based visual information search. In Laurini, R., ed.: Fourth International Conference on Visual Information Systems (VISUAL'2000). Number 1929 in Lecture Notes in Computer Science, Lyon, France, Springer–Verlag (2000) 442–456
7. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Performance evaluation in content–based image retrieval: Overview and proposals. Pattern Recognition Letters **22** (2001) 593–601
8. Smeaton, A.F., Over, P., Kraaij, W.: TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In: Proceedings of the international ACM conference on Multimedia 2004 (ACM MM 2004), New York City, NY, USA (2004) 652–655
9. Clough, P., Müller, H., Sanderson, M.: Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004. In Peters, C., Clough, P.D., Jones, G.J.F., Gonzalo, J., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. LNCS 3491, Bath, England, Springer–Verlag (2005) 597–613
10. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross–language image retrieval track. In: Springer Lecture Notes in Computer Science (LNCS), Vienna, Austria (2006 – to appear)
11. Savoy, J.: Report on CLEF–2001 experiments. In: Report on the CLEF Conference 2001 (Cross Language Evaluation Forum), Darmstadt, Germany, Springer LNCS 2406 (2002) 27–43
12. Müller, H., Rosset, A., Vallée, J.P., Terrier, F., Geissbuhler, A.: A reference data set for the evaluation of medical image retrieval systems. Computerized Medical Imaging and Graphics **28** (2004) 295–305
13. Candler, C.S., Uijtdehaage, S.H., Dennis, S.E.: Introducing HEAL: The health education assets library. Academic Medicine **78** (2003) 249–253
14. Wallis, J.W., Miller, M.M., Miller, T.R., Vreeland, T.H.: An internet–based nuclear medicine teaching file. Journal of Nuclear Medicine **36** (1995) 1520–1527
15. Glatz-Krieger, K., Glatz, D., Gysel, M., Dittler, M., Mihatsch, M.J.: Web-basierte Lernwerkzeuge für die Pathologie – web–based learning tools for pathology. Pathologe **24** (2003) 394–399

16. Hersh, W., Müller, H., Gorman, P., Jensen, J.: Task analysis for evaluating image retrieval systems in the ImageCLEF biomedical image retrieval task. In: Slice of Life conference on Multimedia in Medical Education (SOL 2005), Portland, OR, USA (2005)
17. Sparck Jones, K., van Rijsbergen, C.: Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge (1975)
18. Hersh, W., Bhupatiraju, R.T.: Trec genomics track overview. In: Proceedings of the 2003 Text REtrieval Conference (TREC), Gaithersburg, MD, USA (2004)
19. Chevallet, J.P., Lim, J.H., Radhouani, S.: Using ontology dimentsions and negative expansion to solve precise queries in clef medical task. In: Working Notes of the 2005 CLEF Workshop, Vienna, Austria (2005)
20. Saracevic, T.: Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science **November/December** (1975) 321–343
21. Schamber, L., Eisenberg, M.B., Nilan, M.S.: A re–examination of relevance: Toward a dynamic, situational definition. Information Processing and Management **26 No 6** (1990) 755–775
22. Vendrig, J., Worring, M., Smeulders, A.W.M.: Filter image browsing: Exploiting interaction in image retrieval. In Huijsmans, D.P., Smeulders, A.W.M., eds.: Third International Conference on Visual Information Systems (VISUAL'99). Number 1614 in Lecture Notes in Computer Science, Amsterdam, The Netherlands, Springer–Verlag (1999) 147–154
23. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The IRMA code for unique classification of medical images. In: Medical Imaging. Volume 5033 of SPIE Proceedings., San Diego, California, USA (2003)