Problems with Running a Successful Multimedia Retrieval Benchmark [*]

Henning Müller[1], Thomas Deselaers[2], Michael Grubinger[3],
Paul Clough[4], Allan Hanbury[5], William Hersh[6]

[1]Medical Informatics, University and Hospitals of Geneva

[2]Computer Science, RWTH Aachen, Aachen Germany

[3]Victoria University, Melbourne, Australia

[4]Sheffield University, Sheffield, England

[5]Vienna University of Technology, Austria

[6]OHSU, Portland, Oregon, USA

e-mail: henning.mueller@sim.hcuge.ch

Abstract

Content-based image retrieval (CBIR) and multimedia retrieval are at the point where they are ready to leave the pure research status and become integrated into commercial prototypes and products. This requires techniques not only to be interesting as theoretical approaches but also to be comparable with respect to performance obtained. Similar to the text retrieval domain many years ago, several evaluation events or benchmarks have started these past years to compare multimedia retrieval techniques with a varying focus. TRECVID focuses on video, INEX Multimedia on structured data, ImageEval on visual retrieval and classification, and ImageCLEF on multimodal and multilingual data access. Running such a benchmark poses many problems and difficulties. This article summarises some of the major problems encountered in the organisation of ImageCLEF from 2003 to 2007 and tries to find solutions for at least part of the problems.

1 Introduction

Visual information is ubiquitous and the amount produced with cheap digital cameras is rising strongly. To better manage this information content-based images retrieval has been proposed proposed for general image retrieval [16, 6] as well as in specialised domains [11]. Many techniques have been developed for image retrieval but one of the problems is that most approaches

[*]This work was partially supported by the Swiss National Science Foundation (FNS) with grant 205321–109 and the American National Science Foundation (NSF) with grant ITR–0325160.

are very difficult to compare to each other as varying databases, performance measures, and methodologies are used [12].

In recent years several multimedia retrieval benchmarks with a varying focus have been created and run. Many ideas on benchmarking multimedia were already presented early [17, 8, 12] but the Benchathlon[1] was the first event generating a wider discussion in the community. TRECVID[2], the first real benchmark, started as a task in TREC but has since become an independent workshop on the evaluation of video retrieval systems [15]. 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 and mainly on visual retrieval of images and image classification. INEX[4] (INitiative for the Evaluation of XML retrieval) has also started a multimedia retrieval task in 2006. A fourth benchmarking event is ImageCLEF[5] [3, 2]. This event is part of the Cross-Language Evaluation Forum (CLEF) campaign to evaluate and compare multilingual information retrieval systems [14]. 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 three years has shown the need for standardised system comparison and the importance of creating an infrastructure to support the comparisons in this way. The connection of multimedia benchmarks with events such as TREC, CLEF, and INEX seems necessary to obtain a critical mass and limit the administration overhead concerning the management of document collections, copyright issues, organising a workshop, reserving rooms, registering participants, etc.

Such multimedia retrieval benchmarks can dramatically reduce the effort required by researchers to compare their approaches and start working on optimising the technical part of their work. They are able to concentrate on developing novel methods rather than issues associated with pure evaluation such as defining a methodology and obtaining a database.

This article describes experiences from running the ImageCLEF campaign over five years and tries to propose ways to limit these problems to a minimum.

2 Problems

This section summarises the largest problems encountered in five years of organising the ImageCLEF image retrieval benchmark [4, 5, 2, 10, 1].

2.1 Funding

To organise a benchmark with a yearly rhythm of:

[1] http://www.benchathlon.net/
[2] http://www-nlpir.nist.gov/projects/t01v/
[3] http://www.imageval.org/
[4] http://inex.is.informatik.uni-duisburg.de/2006/
[5] http://ir.shef.ac.uk/imageclef/

- participant registration,
- data release,
- results submission,
- topic definition,
- ground truthing,
- system evaluation,
- and workshop organisation.

takes much time of the organisers. Justifying the time spent on this benchmark is often hard for researchers as evaluation is often seen as part of technical work and not as research by itself. This means that these coordination efforts are sometimes tolerated but often not encouraged and much of the time spent is rather on the "free" time of researchers. Advantages for the organising researchers are the possibility to have an increased visibility in the community, a possibility to influence important research directions, and to publish on the obtained results. Real impact by an evaluation event can particularly be obtained through heavily funded initiatives such as TREC and TRECVID (both funded by NIST). Such funding guarantees a professional organisation, a sufficient marketing and a good evaluation of obtained data and results to optimise impact and to advance the research field. CLEF started as part of TREC and since its independence in 2000 has received minor funding, mainly in an indirect way through research projects funded by the European Union. This has allowed the running of the benchmark for several years but still required much personal work by the organisers and has the risk of creating a non-sustainable structure.

Lack of funding can result in the following problems:

- smaller participation of researchers in the benchmark due to a lack of marketing and confidence in the organisers;
- problems to get access to important data collections as high quality data can be expensive;
- problems with the ground truthing as specialists to judge the documents are usually expensive, and someone performing the ground truthing needs an incentive; bad data quality limits the validity of results;
- lack of analysing the results data obtained and thus a lack of creating new knowledge from available data;
- shortcomings in the organisation of the benchmark can frustrate participants and limit future participation;
- problems in creating a sustainable structure for benchmarking for long term support.

The only way to get out of this is to motivate funding agencies to finance research on evaluation in the same way as technical research. Benefits of benchmarking need to be taken into account as benchmarking is an infrastructure activity for research and can be a multiplier of research results in several domains.

2.2 Getting Access to Proper Data Sets

Image data in good quality is usually expensive and web sites selling images have sprung up over the last years (Corbis[6], Getty[7]). It is correct that image sharing sites such as FlickR[8] have proliferated as well and give access to many images with less restrictive licences. Still, many of these sites contain images that are not put there by the original copyright owners and using these for a benchmarking test collection can cause major problems. In the scientific domain it becomes clear that evaluations on small datasets are not an option and that it becomes important to share images among research groups to limit efforts [19, 13]. Some funding agencies such as the National Institutes of Health (NIH) in the US even require funded research to make their datasets available and similar initiatives exist in other domains.

Still, currently most visual datasets are copyrighted and their use for an evaluation campaign is often difficult. For ImageCLEF we have so far mainly taken image collections from institutions (medical teaching files), from libraries, personal photographic collections, and image collections available on the Internet through MIRC[9] (Medical Imaging Resource Centre). An access to a wider collection of images could strongly influence what exactly can be evaluated in benchmarks.

2.3 Advertise the Benchmark and Motivate Participants

One of the hardest problems in benchmarking is advertising such an event and motivating partners to participate as most researchers are always busy and do not like to read advertisement mails. In the multimedia domain this is particularly true as several research domains overlap in this field and it is hard to address all fields at the same time. Information retrieval, computer vision, machine learning, databases, information systems, digital libraries, e-learning and image processing have all their own methodologies for evaluation and their own conferences to present research results. All of them could address content-based image retrieval.

The main way to motivate research groups is through the reputation of researchers organising a benchmark and through personal contacts by inviting certain groups of researchers and then by mouth to mouth propaganda. Only reputation can create trust that the benchmark will be objective and unbiased. Through contacts between researchers the group of participants becomes larger and through presentations of the benchmark at conferences an even larger group can be informed and motivated to participate.

[6] http://www.corbis.com/

[7] http://www.gettyimages.com/

[8] http://www.flickr.com/

[9] http://mirc.rsna.org/

Once many groups register for benchmarks it still remains hard to motivate these groups to submit results and have their results compared with the other submissions. Without submissions the influence of these benchmarks will be very limited — the percentage of registrants submitting results in ImageCLEF is still unfortunately less than 50%. Incentives are the possibility to present results to a larger audience at the workshop and to publish on the data in proceedings of good quality (Springer Lecture Notes in Computer Science for CLEF). Still, several groups prefer using the data and task to see how their techniques work and then publish at other occasions. Organising a benchmark together with conferences in the field (such as ECDL for CLEF) can help to soften the problem of limited travel funds and time of participants. Only a high participation leads to important discussions among participants.

Another big problem is the fear of researchers to obtain poor results for their research and thus get problems with potential funding agencies that want to fund only the best technologies. Thus, benchmarking results cannot be taken out of a context and it does not have to be taken as a pure competition. Established techniques might obtain better short-term results but have less potential than some new approaches. This needs to be highlighted to participants to reduce the fear. Selection of oral presentation is at ImageCLEF not based on performance but rather based on interest and novelty of the technique.

2.4 Partners from Professional Companies

Partners from companies are important for multimedia retrieval benchmarks in two ways. Help with the organisation can professionalise research through indirect funding and publicity of companies, a field where they have more experience than most researchers do. They can also focus research towards real problems and realistic user models with realistic datasets through connections with their product development. The advantages can be on both sides: the companies get access to the newest technology and get ideas on how well these techniques work, while researchers get access to realistic tasks and maybe even commercial contacts to fund future projects.

Another problem are participants from companies at the events and a comparison of their techniques with those of other participants. Several companies cannot publish details on their algorithms as the algorithms are sometimes patented or should at least assure the advantage over competitors. As a consequence sometimes companies participate but give no details on the techniques they use, but instead broad descriptions (this is being practised by TRECVID). Some companies, mainly startups are even afraid that bad results would bring their products into discredit (or even reduce venture capital) and thus they would like to be able to remove their results from the final comparison if they turn out to be poor. Such an approach is currently being tested by ImageCLEF with the goal to improve the framework for commercial participation.

2.5 Realistic Tasks and User Models

The definition of tasks and topics depends mainly on the databases available and a very clear user model needs to be defined before tasks can be developed. This can help to tackle real problems and requirements but poorly defined topics can also limit the results of tasks completely.

Typically, realistic tasks can be gained from expert knowledge [5], from log files of system use [4, 1] or through interviews with experts [7].

When observing these information sources, one of the problems with multimedia retrieval is that most information needs are not formulated visually as only few running systems are currently employed. To develop visual tasks from text can be a hard task and requires time for selection. Another problem is the need to have an idea whether and how many relevant images for a certain topic exist in the database. Very broad topics can lead to an extremely large number of relevant items with the risk to miss some of them in the pooling process. If no relevant images exist, the task should also be omitted.

With a document collection and a source to define tasks, the user model can relatively easily be derived.

2.6 Ground Truthing or Gold Standards

Ground truthing for image retrieval evaluation is an expensive task and this is thus linked to the funding problems of many benchmarks. High quality annotations for many specialised tasks can only be performed by domain specialists, whereas some tasks such as image search tasks on personal collections can also be performed by the organisers themselves. INEX even lets participants judge documents of the pools, which limits the effort but creates a slight risk of a bias towards images one is sure the own system would find. To control the quality of relevance judgements several people can be asked to judge the same topics and then a kappa score on agreement between them can be calculated. Variations of judgements have been reported in several domains but they do not in general influence the evaluation results strongly [9, 20].

When using expert judges it is extremely important to define the topics well as human interpretation of seemingly clear information needs can vary strongly [9]! Variation among judges can be reduced through supplying a narrative with the topic explaining in more detail what is regarded as relevant and particularly what is not regarded as relevant. A description of non-relevance has to be highlighted to obtain high-quality results. In ImageCLEF a ternary judgement scheme is used: relevant, partially relevant and non-relevant. Despite the fact that we explain to judges to use partially relevant only when it is impossible to determine relevance, a significant proportion of judgements is in this category.

Another judgement problem concerns multilingual collections such as ImageCLEFmed. If judges are primarily familiar with one language and the judgement process requires to read the text, then a bias towards the native language can appear. A translation of the main terms or a mapping of multilingual text onto an ontology can help to limit the problem. It will only rarely be possible to have judges that are familiar with all languages.

In general, no complete judgement of a test collection is possible and thus a pooling technique has to be applied to judge the most important parts of the dataset based on the results submissions to not bias evaluation towards any system [20, 18]. There is a compromise to be made with respect to how many images to judge. The more images are judged the more time and money it takes and the better the results obtained can be, although it can also increase the fatigue of the judges.

2.7 Organisational Issues

Many benchmarks have a fairly similar model of organisation and a yearly cycle of events. To automate at least part of the process from registration, to document delivery, query submission relevance judgements and evaluation every benchmark seems to develop its own methodology depending on the domain. Within information retrieval TREC has helped massively to standardise at least part of the evaluation. Packages such as trec_eval[10] to evaluate runs based on a particular format of the participants' runs and the relevance judgements have helped to use the same measures and avoid calculation errors that can appear when developing software from scratch.

Even after several years of organisation of a benchmarking event, there are still many small errors happening in ImageCLEF. Among them are those in this short list of some of the main problems:

- errors or inconsistencies in the distributed data collections as no exhaustive tests were performed beforehand, and participants usually discover them at some point;
- incorrect submissions from participants that need to be corrected for correct evaluation, although formats were described and examples made available;
- incompletely or incorrect description of the techniques used for certain runs;
- incomplete descriptions for relevance judges due to time limitations resulting in lower quality judgements;
- delays due to other tasks of the organising researchers;
- problems with software for results submission or relevance judgements resulting in a loss of time for participants or judges.

2.8 Proving Advances and Benefits of Benchmarks

One of the most important parts in "selling" the utility of benchmarks is to show the improvement that they have brought to the domain. Again, this can be linked with funding and impact. When manpower is available it is much easier to prove the utility than when the manpower is lacking to analyse outcomes of benchmarks over time. TREC has shown that through detailed analysis of the results many important points can be shown such as the lack of a bias when using pooling techniques [20] or the fact that changing relevance judges generally does not change the ranking of performing systems significantly; measures such as B-Pref results also from TREC research. Benchmarks with less funding have a harder time doing these in depth analyses and will only be able to achieve minor impacts.

An easy way to prove performance is to measure the use of the created collections, topics and relevance judgements. Unfortunately, authors often reuse the resources for other publications but do not inform the organisers on this although it is requested from participants. Research on the web can bring up some of the publications but will always be incomplete. Through

[10] http://trec.nist.gov/trec_eval/

the number of reuses, the saved time of researchers can be estimated. Still, the most important part is the comparability of approaches and this is difficult to be measured: the comparability of techniques and focusing of researchers on promising techniques avoiding typical mistakes of the past.

Another way to show how a research field is improving is to run older techniques on new data and show where they are with respect to current techniques or to run newer techniques on older data. In ImageCLEF this shows well that the performance of participating systems has significantly improved over time.

3 Conclusions

Benchmarks in the multimedia field have enormously advanced the techniques developed in research labs. Re-creation of small datasets and the impossibility of comparing approaches have been reduced, and at several conferences approaches can now be compared on the same datasets making it possible to have a clear idea of advantages and disadvantages of various approaches. Instead of spending much time and money on the creation of datasets, research groups can start with standard datasets and participate in evaluation campaigns.

One of the main criticisms of benchmarks is a sort of standardisation of research and the tendency to reuse well-performing techniques and make minor modifications instead of developing completely new techniques that might have more potential for the future (“Do benchmarks kill innovation?”). Some of these criticisms are true and thus benchmarks cannot be used for completely new research domains but rather in domains where an established set of techniques has already been developed and that is at the point to be ready for a use in real prototype systems. Another point to soften this criticism is to attempt a quickly changing set of benchmarks to avoid running the same sort of tests every year. Similar to TREC where many tracks run between 2 and 5 years it is important to have changes in the types of tasks. Another important part is to include recommendations and new people from the community in organising new tasks to avoid the impression of an elitist organisation and to adapt running benchmarks towards real and up-to-date tasks of the users, which is the research community.

References

[1] Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In *CLEF 2006 Proceedings*, Springer Lecture Notes in Computer Science, pages 579–594, 2007. 2, 6

[2] Paul Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeffery Jensen, and William Hersh. The CLEF 2005 cross–language image retrieval track. In *Cross Language Evaluation Forum (CLEF 2005)*, Springer Lecture Notes in Computer Science, pages 535–557, September 2006. 2

[3] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul Clough, Julio Gonzalo, Michael Jones, Gareth J. F.and Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 597–613, Bath, UK, 2005. Springer. 2

[4] Paul Clough and Mark Sanderson. The CLEF 2003 cross language image retrieval task. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2003)*, 2004. 2, 6

[5] Paul Clough, Mark Sanderson, and Henning Müller. The clef cross language image retrieval track (ImageCLEF) 2004. In *The Challenge of Image and Video Retrieval (CIVR 2004)*, Springer Lecture Notes in Computer Science, pages 243–251, July 2004. 2, 6

[6] A. del Bimbo. *Visual Information Retrieval*. Academic Press, 1999. 1

[7] William Hersh, Jeffery Jensen, Henning Müller, Paul Gorman, and Patrick Ruch. A qualitative task analysis for developing an image retrieval test collection. In *ImageCLEF/MUSCLE workshop on image retrieval evaluation*, pages 11–16, Vienna, Austria, 2005. 6

[8] Clement Leung and Horace Ip. Benchmarking for content–based visual information search. In Robert Laurini, editor, *Fourth International Conference on Visual Information Systems (VISUAL'2000)*, number 1929 in Lecture Notes in Computer Science, pages 442–456, Lyon, France, November 2000. Springer–Verlag. 2

[9] Henning Müller, Paul Clough, William Hersh, and Antoine Geissbuhler. Variations of relevance assessments for medical image retrieval. In *Adaptive Multimedia Retrieval (AMR)*, volume 4398 of *Springer Lecture Notes in Computer Science (LNCS)*, pages 233–247, 2007. 6

[10] Henning Müller, Thomas Deselaers, Thomas Lehmann, Paul Clough, Eugene Kim, and William Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In *CLEF 2006 Proceedings*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, 2007 – to appear. Springer. 2

[11] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content–based image retrieval systems in medicine – clinical benefits and future directions. *International Journal of Medical Informatics*, 73:1–23, 2004. 1

[12] Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet, and Thierry Pun. Performance evaluation in content–based image retrieval: Overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, April 2001. Special Issue on Image and Video Indexing. 2

[13] Guiseppe Sasso, Hugo Raul Marsiglia, Francesca Pigatto, Antonio Basilicata, Mario Gargiulo, Andrea Francesco Abate, Michele Nappi, Jenny Pulley, and Francesco Silvano Sasso. A visual query–by–example image database from chest CT images: Potential role as a decision and educational support tool for radiologists. *Journal of Digital Imaging*, 18(1):78–84, March 2005. 4

[14] Jacques Savoy. Report on CLEF–2001 experiments. In *Report on the CLEF Conference 2001 (Cross Language Evaluation Forum)*, pages 27–43, Darmstadt, Germany, 2002. Springer LNCS 2406. 2

[15] Alan F. Smeaton, Paul Over, and W. Kraaij. TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In *Proceedings of the international ACM conference on Multimedia 2004 (ACM MM 2004)*, pages 652–655, New York City, NY, USA, October 2004. 2

[16] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Armarnath Gupta, and Ramesh Jain. Content–based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 No 12:1349–1380, 2000. 1

[17] John R. Smith. Image retrieval evaluation. In *IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'98)*, pages 112–113, Santa Barbara, CA, USA, June 21 1998. 2

[18] K. Sparck Jones and C.J. van Rijsbergen. 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. 6

[19] Michael W. Vannier and Ronald M. Summers. Sharing images. *Radiology*, 228:23–25, 2003. 4

[20] Justin Zobel. How reliable are the results of large–scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York. 6, 7