BENCHMARKING IMAGE RETRIEVAL APPLICATIONS

Henning Müller, Antoine Geissbuhler

University Hospitals of Geneva
Service of Medical Informatics
Rue Micheli-du-Crest 24
1211 Geneva 14
henning.mueller@sim.hcuge.ch

Stéphane Marchand–Maillet, Paul Clough[*]

University of Geneva
Computer Vision and Multimedia Lab
Rue du Général Dufour 24
1211 Geneva 4
marchand@cui.unige.ch

ABSTRACT

Content–based visual information retrieval is an important research topic in the computer vision field sind the early 1990s. A large number of systems have been developed as research prototypes as well as commercial and open source systems. Still, there has not been a general breakthrough in performance yet and important real–world application stay fairly rare. The very large amount of available multimedia information creates a need to develop new tools to explore and retrieve within mixed media databases. The replacement of analog films by digital consumer cameras and the increasing digitisation in several fields such as medicine will still increase this need.

One of the reasons for the impossibility to show an increase in performance is the simple fact that there is no standard for evaluating the performance of systems. In the last years a rising number of proposals have been made on how to evaluate or not to evaluate the performance of visual information retrieval systems which underlines the importance of the issue. Several benchmarking events such as the Benchathlon, TRECVID and imageCLEF have been started, with varying success. This article described mainly the work of the University of Geneva on benchmarking of visual information retrieval systems. A special emphasis will be on the Benchathlon and imageCLEF evaluation events and their methodology and outcome.

1. INTRODUCTION

Ideas for content–based retrieval in image or multimedia databases dates back to the the early 1980s [1]. Serious applications started in the early 1990s and the most well–known systems are maybe IBM's QBIC [2] and MIT's Photobook [3]. Content–based image retrieval became an extremely active research area with most likely hundreds of systems and several hundred publications. A good overview article can be found with [4].

Although active in research, only very little effort was put into comparing and evaluating the system performance. Small, copyrighted databases were used that made any comparison impossible and the shown graphs and measures basically useless. The closely related field of text retrieval already did systematic evaluation and creation of databases since the early 1960s with the Cranfield studies [5] and the SMART system [6].

The MIRA (Evaluation frameworks for interactive and multimedia information retrieval applications) project first focused on visual information retrieval evaluation starting from 1996 [7]. A first article on benchmarking content–based image retrieval algorithms was published in 1997 [8]. New measures for evaluation were created but no example evaluation nor a database was shown or made accessible. In [9], the text retrieval community and the TREC conference were first mentioned as a role model for visual retrieval evaluation. [10] mentions some minimum requirements with respect to the number of images and methodology used. Still, no database or ground truth was used to underline the evaluation ideas. In [11], the evaluation was reduced to one single performance measure which might be convenient for comparisons but will not be a good indicator to compare systems based on various aspects. Huijsmans [12] describes very interesting graphs that include measures such as the collection size and size of the ground truth into precision vs. recall graphs to eliminate the retrieval of relevant documents simply by chance. This is definitely good but the comparison of retrieval results with differing databases has many more problems. Currently, only results obtained for the same database can really be compared.

The Benchathlon network for retrieval system evaluation was described in [13]. This includes concrete measure and a justification for them as well as a literature review. [14] describes a more general framework for evaluation and includes a literature review as well as an example evaluation with an openly accessible database. A more recent review can be found in [15].

There is also a lot of critics with respect to current benchmarking initiatives [16]. Part of the critics is that current retrieval systems do not perform well enough to really benchmark them and that they are too far away from real user needs to be evaluated. This is not without reason. The current low level features correspond only sometimes to the concepts that users are looking for (and this might actually be by chance). Still, it is important to evaluate the system based on real user needs, on what a real user is looking for. Only systematic evaluation can show system improvements. Not evaluating at all does not advance any system. The basic technologies for content–based image retrieval are available but now is the time to find out which technology works well for what kind of images and queries.

2. PARTS OF A BENCHMARK

An image retrieval benchmark will need to include a variety of components. The most important one is currently the creation of standard databases including search task, query topics and ground truth for these topics. Afterwards, the evaluation measures can be discussed to actually perform system comparisons. Important for

[*] Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP UK

the multimedia retrieval community is definitely an event where systems can be compared and experiences can be exchanged.

2.1. Data sets

Currently, the de–facto standard for image retrieval are still the Corel Photo CDs despite their many problems. They are fairly expensive, copyrighted and are not available on the market anymore. A request from our University to Corel for using lower–resolution images as basis for a benchmarking event was not answered.

A database that is available free of charge and copyright and is used for evaluation is that of the University of Washington [1]. It contains a few 1000 images that are clustered by regions. Other databases are available for computer vision research but only rarely for image retrieval[2]. The Benchathlon also created a test database but currently without query tasks and ground truth. In specialised domains such as medical imaging, there are several databases available free of charge. The National Institute for Health (NIH), for example publishes free of charge all the databases gathered. A medical database used for retrieval is that of casimage[3] [17].

In text retrieval, the need for databases was, again, identified very early and test sets have been created for years [18]. For images, there is an effort to create annotated databases [19] that can further on be used for system evaluation.

2.2. Query tasks and topics

The first question when evaluating a system should actually be *"What do we want to evaluate?"*. The goal for evaluation should be based on real user needs and not a computer vision expert's interest. Some studies have been performed on how real users query image databases [20, 21] but too few and they are currently all based on users searching with text.

Normally, there should be a selection of query tasks based on real–world user queries and then, images or textual formulations should be taken to select evaluation topics that can be used to compare systems. This will deliver results that correspond to what a user would expect from a system, and systems can consequently be optimised for these goals.

2.3. Ground truth

Of course, users can for simplicity be simulated to asses the system performance [22]. Like this, the system developer can define noise levels and as a consequence the system performance. Real ground truth or a gold standard will need to include real users that assess the system performance for each query task and topic. This is expensive and involves much work. It has successfully been done in TREC and much literature is available on statistical significance and problems when using pooling schemes to reduce the number of documents that the relevance assessors will have to watch [23].

2.4. Evaluation measures

A good review of performance measures used for image retrieval can be found in [24]. Although good descriptors that are easy to interpret are important for retrieval system evaluation, this is not the main problem at the moment. The measures can only be as good as the database and ground truth available which is definitely the current problem. Simple measure based on precision and recall, and especially precision vs. recall graphs seem to be the accepted standard for content–based image retrieval at the moment.

2.5. Benchmarking events

Text retrieval used to have several standard databases that were used for evaluation since the 1960s [5]. Still, the single big event that showed a significant increase in performance was TREC[4] (Text REtrieval Conference) starting from 1992 [25]. TREC is a "friendly" benchmarking event for which large data sets are generated, and systems are compared based on these new data sets every year. Several subtasks have become independent conferences in the meantime as they grew bigger and more important (CLEF, TRECVID). Unfortunately a request to include content–based image retrieval into TREC was denied with the explication that there were no databases available that could be distributed and were judged large enough for the task.

Image retrieval does need a benchmarking event such as TREC to meet and discuss technologies based on a variety of databases and specialised tasks (medical image retrieval, trademark retrieval, consumer pictures, ...)! This will allow the community to have standard datasets and to identify good and less good techniques as well as performant interaction schemes. System improvements can be shown over time with such an event.

3. BENCHMARKING INITIATIVES FOR VISUAL INFORMATION RETRIEVAL

Currently, there are few real benchmarking events for visual information retrieval. TRECVID is an exception as it deals with video data whereas the Benchathlon and imageCLEF deal with images.

3.1. TRECVID

TRECVID[5] was introduced as a TREC task in 2001 with subtasks in shot–boundary detection and search tasks, mainly based on a textual description. Data sets in 2003 contain more than 130 hours of video in total. Video is different from images in that the speech and captions can be translated into text and thus, more that low–level visual descriptors can be used for semantic queries. The number of participants for TRECVID has grown steadily from 12 in 2001 to 24 in 2003. The number of subtasks has also grown and includes now story segmentation and classification as well as higher level feature extraction. This can be the recognition of a group of people etc. TRECVID is a success and has given to the community a meeting point where technologies and their influences on retrieval can be discussed and compared based on the same datasets. Test collection have been created and these can be used to learn and optimise the system performance for future tasks.

3.2. The Benchathlon

The Benchathlon[6] was created in the context of the Internet Imaging session at the SPIE Photonics West conference, one of the im-

[1] `http://www.cs.washington.edu/research/imagedatabase/groundtruth/`

[2] `http://homepages.inf.ed.ac.uk/cgi/rbf/CVONLINE/entries.pl?TAG363`

[3] `http://www.casimage.com/`

[4] `http://trec.nist.gov/`

[5] `http://www.itl.nist.gov/iaui/894.02/projects/trecvid/`

[6] `http://www.benchathlon.net/`

portant conferences for content–based image retrieval. The goal was to create a workshop where benchmarking and evaluation could be discussed among researchers and industry and where a benchmarking event for image retrieval was to be started. An evaluation methodology was developed [13] stating performances measures and there justifications. An interactive evaluation methodology based on the Multimedia Retrieval Markup Language (MRML[7]) was presented [26] to allow interactive evaluation of systems. This was supposed to take into account the importance of relevance feedback for the evaluation of image retrieval systems. Based on real user ground truth, the behaviour on marking positive/negative feedback can be automised and used for evaluations.

2001 saw the first Benchathlon with basically a presentation of the outline document [13] and discussions among participants. In 2002 a first workshop with five presentations was held and this number raised to 8 presentations in 2003. Unfortunately, the goal to really compare the systems' performance was not reached. Efforts included the generation of a databases containing a few thousand private pictures and a partly annotation of these [27]. Ground truth has not yet been generated for query topics to evaluate system performance. The proposed architecture for automatic evaluation was not accepted by many research groups either, although efforts were taken write tools for participants and help them to install an MRML–based system access.

3.3. imageCLEF

The Cross Language Evaluation Forum (CLEF[8]) started as a subtask of TREC to allow information retrieval over languages where for example the queries are in a different language than the documents. 2000 saw the first independant CLEF conference taking two days and listing over 25 papers in the proceedings. One of the subtasks that was developed within CLEF is imageCLEF[9], on the evaluation of cross language image retrieval [28]. ImageCLEF started in 2003 with 6 participants using a database of 26.000 images from St. Andrews University with English annotation and queries in a variety of languages. The queries include one query image plus a textual description of the query.

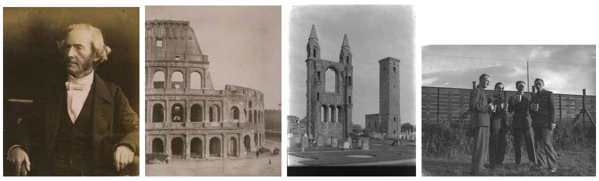

Figure 1: Some example images of the St. Andrews database.

Figure 1 shows some images of the database. The fact that most images are in grey or brown scales also explains why, in 2003, there was no use of visual retrieval algorithms in the competition. The kind of query topics are also very hard to answer visually as they are not based on the visual content but the semantics of the image. For this reason, in 2004, a more visual retrieval task will be added to imageCLEF in the domain of medical images. Figure 2 shows some example images from this database that contains a total of almost 9000 medical images [17] of a medical teaching file including annotations in French and English.

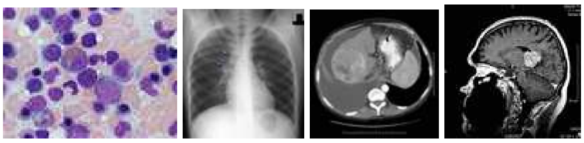

Figure 2: Some example images of the medical database.

Query topics (26 in total) were chosen by a radiologist to well represent the spectrum of the entire database and not based on on real user queries. Ground truthing will also be done by radiologists. The query is an image only but for the database the images and the accopagnying textual information in French or English are available. This stresses the character of a cross–language retrieval task but it also gives a particular potential to visual information retrieval. Automatically extracted visual information is inherently insensitive to language and can thus be an important aid to cross–language information retrieval. On the other hand, the combination of textual and visual cues can also deliver important results for the visual information retrieval community as it adds the longtime–missed semantics. Like this both communities can profit from the other to improve the performance and get new insights into information retrieval. The 2004 competition counts 10 participants for the St. Andrews and 10 for the medical task. This improvement from 6 in 2003 to 20 in 2004 shows the important of image retrieval also in the context of cross–language information retrieval. A large variety from purely textual, to mixed visual/textual and purely textual retrieval have been used. Techniques such as automatic query expansion and manual relevance feedback have also been submitted by several participants.

4. CONCLUSIONS

The content–based visual information retrieval community needs a common effort to create and make available datasets/query topics and ground truth to be able to compare the performance of various techniques. A benchmarking event is needed more than ever to give a discussion forum for researchers to compare techniques and identify promising approaches. Especially the use of multi–modal databases and of cross–language information retrieval on the evaluation of image retrieval algorithms is important as many real–world collections such as the Internet have exactly these characteristics. Strong participation in events such as TRECVID and imageCLEF shows that there is a need to share data and results to advance the science of visual information retrieval.

5. REFERENCES

[1] N.-S. Chang and K.-S. Fu, "Query–by–pictorial–example," *IEEE Transactions on Software Engineering*, vol. SE 6 No 6, pp. 519–524, 1980.

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC system," *IEEE Computer*, vol. 28, no. 9, pp. 23–32, September 1995.

[3] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Tools for content–based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, June 1996.

[7] http://www.mrml.net

[8] http://www.clef-campaign.org/

[9] http://ir.shef.ac.uk/imageclef2004/

[4] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content–based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22 No 12, pp. 1349–1380, 2000.

[5] C. W. Cleverdon, "Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems," Aslib Cranfield Research Project, Cranfield, USA, Tech. Rep., September 1962.

[6] G. Salton, *The SMART Retrieval System, Experiments in Automatic Document Processing*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1971.

[7] E. Sormunen, M. Markkula, and K. Järvelin, "The perceived similarity of photos – seeking a solid basis for the evaluation of content–based retrieval algorithms," in *Final MIRA Conference*, ser. Electronic Workshops in Computing. Glasgow: The British Computer Society, 14–16 April 1999.

[8] A. D. Narasimhalu, M. S. Kankanhalli, and J. Wu, "Benchmarking multimedia databases," *Multimedia Tools and Applications*, vol. 4, pp. 333–356, 1997.

[9] J. R. Smith, "Image retrieval evaluation," in *IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'98)*, Santa Barbara, CA, USA, June 21 1998, pp. 112–113.

[10] C. Leung and H. Ip, "Benchmarking for content–based visual information search," in *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, ser. Lecture Notes in Computer Science, R. Laurini, Ed., no. 1929. Lyon, France: Springer–Verlag, November 2000, pp. 442–456.

[11] M. Koskela, J. Laaksonen, S. Laakso, and E. Oja, "Evaluating the performance of content–based image retrieval systems," in *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, ser. Lecture Notes in Computer Science, R. Laurini, Ed., no. 1929. Lyon, France: Springer–Verlag, November 2–4 2000, pp. 430–441.

[12] D. P. Huijsmans and N. Sebe, "Extended performance graphs for cluster retrieval," in *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2001)*. Kauai, Hawaii, USA: IEEE Computer Society, December 9–14 2001, pp. 26–31.

[13] N. J. Gunther and G. Beretta, "A benchmark for image retrieval using distributed systems over the internet: BIRDS–I," HP Labs, Palo Alto, Technical Report HPL–2000–162, San Jose, Tech. Rep., 2001.

[14] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun, "A framework for benchmarking in visual information retrieval," *International Journal on Multimedia Tools and Applications*, vol. 21, pp. 55–73, 2003, (Special Issue on Multimedia Information Retrieval).

[15] I. Jermyn, C. Shaffrey, and N. Kingsbury, "The methodology and practice of the evaluation of image retrieval systems and segmentation methods," Laboratoire Informatique, signaux et sytmes de sophia antipolis, CNRS Rapport de recherche ISRN I3S/RR-2003-05-FR, 2003.

[16] D. A. Forsyth, "Benchmarks for storage and retrieval in multimedia databases," in *Storage and Retrieval for Media Databases*, ser. SPIE Proceedings, vol. 4676, San Jose, California, USA, January 21–22 2002, pp. 240–247, (SPIE Photonics West Conference).

[17] H. Müller, A. Rosset, A. Geissbuhler, and F. Terrier, "A reference data set for the evaluation of medical image retrieval systems," *Computerized Medical Imaging and Graphics*, 2004 (to appear).

[18] K. Sparck Jones and C. van Rijsbergen, "Report on the need for and provision of an ideal information retrieval test collection," Computer Laboratory, University of Cambridge, British Library Research and Development Report 5266, 1975.

[19] C. Jörgensen, "Towards an image testbed for benchmarking image indexing and retrieval systems," in *Proceedings of the International Workshop on Multimedia Content–Based Indexing and Retrieval*, Rocquencourt, France, September 2001.

[20] M. Markkula and E. Sormunen, "Searching for photos – journalists' practices in pictorial IR," in *The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval*, ser. Electronic Workshops in Computing, J. P. Eakins, D. J. Harper, and J. Jose, Eds. Newcastle upon Tyne: The British Computer Society, 5–6 February 1998.

[21] P. G. B. Enser, "Pictorial information retrieval," *Journal of Documentation*, vol. 51, no. 2, pp. 126–170, 1995.

[22] J. Vendrig, M. Worring, and A. W. M. Smeulders, "Filter image browsing: Exploiting interaction in image retrieval," in *Third International Conference On Visual Information Systems (VISUAL'99)*, ser. Lecture Notes in Computer Science, D. P. Huijsmans and A. W. M. Smeulders, Eds., no. 1614. Amsterdam, The Netherlands: Springer–Verlag, June 2–4 1999, pp. 147–154.

[23] J. Zobel, "How reliable are the results of large–scale information retrieval experiments?" in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds. Melbourne, Australia: ACM Press, New York, August 1998, pp. 307–314.

[24] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content–based image retrieval: Overview and proposals," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 593–601, April 2001.

[25] D. Harman, "Overview of the first Text REtrieval Conference (TREC–1)," in *Proceedings of the first Text REtrieval Conference (TREC–1)*, Washington DC, USA, 1992, pp. 1–20.

[26] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun, "A web–based evaluation system for content–based image retrieval," in *Proceedings of the 9th ACM International Conference on Multimedia (ACM MM 2001)*. Ottawa, Canada: The Association for Computing Machinery, October 2001, pp. 50–54.

[27] T. Pfund and S. Marchand-Maillet, "Dynamic multimedia annotation tool," in *Internet Imaging III*, ser. SPIE Proceedings, G. Beretta and R. Schettini, Eds., vol. 4672, San Jose, California, USA, January 21–22 2002, pp. 216–224, (SPIE Photonics West Conference).

[28] P. Clough, M. Sanderson, and H. Müller, "A proposal for the clef cross language image retrieval track (imageclef) 2004," in *The Challenge of Image and Video Retrieval (CIVR 2004)*. Dublin, Ireland: Springer LNCS, July 2004.