Benchmarking image retrieval applications

H. Müller, A. Geissbuhler
University Hospitals of Geneva
Rue Micheli-du-Crest 24,
CH-1211 Geneva 14

S. Marchand–Maillet
University of Geneva,
Rue du Général Dufour 24,
CH-1211 Geneva 4

P. Clough*
University of Sheffield
211 Portobello Street
Sheffield, S1 4DP UK

Abstract

Since the early 1990s, content–based visual information retrieval has been an important research topic in computer vision. A large number of systems have been developed as research prototypes, as well as commercial and open source systems. However, still no general breakthrough in performance has emerged and important real–world applications stay rare. The 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 cameras and the increasing digitisation in fields such as medicine will still increase this need.

One of the reasons for the impossibility to show an increase in performance is the fact that there is no standard for evaluating the performance of content-based retrieval systems. In the last years a rising number of proposals have been made on how to evaluate or not 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 describes work carried out by the University of Geneva on benchmarking visual information retrieval systems. A special emphasis will be on the Benchathlon and ImageCLEF evaluation events, their methodologies and outcomes.

1 Introduction

Ideas for content–based retrieval (CBR) in image or multimedia databases (DB) date back to the the early 1980s. Serious applications started in the early 1990s and the most well–known systems are maybe IBM's QBIC [5] and MIT's Photobook [20]. Content–based image retrieval (CBIR) became an extremely active research area with hundreds of systems and several hundred publications. A good overview article is [23]. Although active in research, only very little effort was put into comparing and evaluating the performance of systems. Small, copyrighted DBs were used that made comparisons betwen systems almost impossible and the shown graphs and measures problematic. The related field of text retrieval (TR) already did systematic evaluation and creation of datasets since the early 1960s with the Cranfield studies [1] and SMART [22].

The MIRA (Evaluation frameworks for interactive and multimedia information retrieval (IR) applications) project first focused on visual IR evaluation starting from 1996 [25]. A first article on benchmarking CBIR algorithms was published in 1997 [19]. New measures for evaluation were created but no example evaluation nor a DB was shown. In [24], the TR community and the TREC conference were first mentioned as a role model for visual retrieval evaluation. Leung and Ip [13] mention some minimum requirements with respect to the number of images and methodology used, but still no common DB or ground truth was used. In [12], 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 [9] 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 very good, but the comparison of retrieval results across DBs is still problematic. The Benchathlon network for evaluation is described in [7]. This includes concrete measures of performance effectiveness, justification for them and a literature review. Müller et al. [16] describes a more general framework for evaluation and includes a literature review as well as an example evaluation with an openly accessible DB. A more recent review is [10].

Many researchers have been critical of current benchmarking initiatives [6]. Part of the criticism is that current retrieval systems do not perform well enough to realistically benchmark them and that they are too separate from real user needs for results to be meaningful to end-user applications. This is not without reason. The current low level features correspond only sometimes to concepts that users are looking for. It is important, therefore, to evaluate systems based on real user needs, i.e. 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 CBIR are available but now is the time to find out which technology works for what kind of queries.

*Part of this work was carried out within the Eurovision project at Sheffield University funded by the EPSRC (Eurovision: GR/R56778/01).

2 Benchmarking components

An complete benchmark will include several components. The most important of these is the creation and availability of standard or common DBs, of typical search tasks, and ground truths for these tasks against which to compare and evaluate new systems. Following this, one can discuss and compare system performance at an organised evaluation campaign.

2.1. Data sets

Currently, the de–facto standard for image retrieval are still the Corel Photo CDs. However, there are problems with these including: they are fairly expensive, copyrighted and not available as a public resource, and they are now unavailable on the market-place. A request from our University to Corel for using lower–resolution images for benchmarking was not answered. A DB that is available free of charge and copyright and is used for evaluation is that of the Uni. of Washington. It contains around 1000 images that are clustered by regions. Other DBs are available for computer vision research but only rarely for image retrieval. The Benchathlon also created a test DB, but currently without search tasks and ground truths. In specialised domains such as medical imaging, there are DBs available. The National Institute for Health (NIH) publishes free of charge all the DBs gathered. A medical DB used for retrieval is that of casimage[1] [18]. In TR, the need for DBs was, again, identified very early on and test sets have been for years at the very core of evaluation [26]. For images, there is an effort to create annotated DBs [11] 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 DBs [14, 4] 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 [27]. 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 the major evaluation campaigns and much literature is available on statistical significance testing and problems when using pooling schemes to reduce the number of documents that the relevance assessors will have to watch [28].

2.4. Evaluation measures

A good review of performance measures used for image retrieval can be found in [17]. 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 DB and ground truth available which is the current problem. Simple measures based on precision and recall, and especially precision vs. recall graphs seem to be the accepted standard for CBIR.

2.5. Benchmarking events

TR used to have several standard DBs that were used for evaluation since the 1960s [1]. Still, the single big event that showed a significant increase in performance was TREC[2] (Text REtrieval Conference) starting from 1992 [8]. TREC is a "friendly" benchmarking event for which large data sets and sets of seach tasks are generated, and systems compared based on this new data each year. Several subtasks have become independent conferences in the meantime as they grew bigger and more important (e.g. CLEF, TRECVID). Unfortunately a request to include CBIR into TREC was denied with the explication that there were no DBs available that could be distributed and were judged large enough.

Image retrieval does need a benchmarking event such as TREC to meet and discuss technologies based on a variety of DBs and specialised tasks (medical image retrieval, trademark retrieval, consumer pictures, ...)! This will allow having standard datasets, 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. Events for visual information retrieval

3.1. TRECVID

TRECVID 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.

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

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

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 created a meeting point where technologies and their influences on retrieval can be discussed and compared based on the same datasets. Test collections have been created and can be used to optimise the system performance for future tasks.

3.2. The Benchathlon

The Benchathlon[3] was created in the context of the SPIE Photonics West conference, one of the important conferences for CBIR. 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 [7] stating performance measures and their justification. An interactive evaluation methodology based on the Multimedia Retrieval Markup Language (MRML[4]) was presented [15] to allow interactive evaluation of systems. This was supposed to take into account the importance of relevance feedback (RF) 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 [7] and discussions among participants. In 2002 a first workshop with five presentations was held and this number raised to 8 in 2003. Unfortunately, the goal to really compare the systems' performance was not reached. Efforts included the generation of a DB containing a few thousand private pictures and a partly annotation of these [21]. 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[5]) started as a subtask of TREC to allow IR over languages where for example the queries are in a different language than the documents. CLEF began in 2000 taking two days and listing over 25 papers in the proceedings. In 2003, one of the subtasks included was ImageCLEF[6], for the evaluation of cross language image retrieval systems [2]. ImageCLEF started with 4 participants using a DB of approximately 30,000 historic photographs from St. Andrews University Images have English annotations and typical search requests were created in a variety of languages. The queries include one query image plus a textual description of the query.

Figure 1. Examples St. Andrews collection.

Figure 1 shows images of the DB. 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 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 and an interactive task has also been added [3] to include some more user–centred evaluation. Figure 2 shows some example images from this DB that contains a total of almost 9000 medical images [18] of a medical teaching file including annotations in French and English.

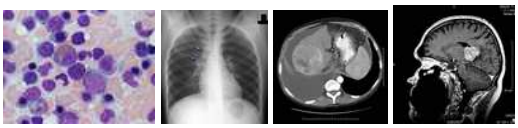

Figure 2. Examples medical collection.

Query topics (26 in total) were chosen by a radiologist to represent the entire DB. Ground truthing is performed by radiologists. The search task is expressed as an image only, but within the DB, images are accompanied by texts describing medical conditions in French or English. This makes the task cross-language, but also gives particular potential to visual IR. Automatically extracted visual information is inherently insensitive to language and can thus be an important aid to cross–language IR. On the other hand, the combination of textual and visual cues can also deliver important results for the visual IR community as it adds semantics which are not easily derived from the image itself. With this, both the cross language and image retrieval communities can profit from the other to improve system performance for certain search tasks and obtain new insight into this particular type of IR. The 2004 competition has 10 participants for a set of search tasks based on the St. Andrews data, and 10 for tasks based on the medical data. This improvement from 4 in 2003 to 20 in 2004 shows the perceived importance of image retrieval within the context of cross–language IR. Entries vary widely from those using purely textual methods to those using purely visual ones. A large number of entries have also experimented with combining text and visual methods to increase performance. Further techniques such as automatic

[3] http://www.benchathlon.net/
[4] http://www.mrml.net
[5] http://www.clef-campaign.org/
[6] http://ir.shef.ac.uk/ImageCLEF2004/

query expansion and manual RF have been submitted by participants, as well as the use of various translation resources.

4. Conclusions

The CBIR 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 DBs and of cross–language IR 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 visual IR.

References

[1] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative effi ciency of indexing systems. Technical report, Aslib Cranfi eld Research Project, Cranfi eld, USA, 1962.

[2] P. Clough and M. Sanderson. The CLEF 2003 cross language image retrieval task. In *Proceedings of the Cross Language Evaluation Forum*, 2004.

[3] P. Clough, M. Sanderson, and H. Müller. A proposal for the CLEF cross language image retrieval track 2004. *The Challenge of Image and Video Retrieval*, Dublin, Ireland, 2004.

[4] P. G. B. Enser. Pictorial information retrieval. *Journal of Documentation*, 51(2):126–170, 1995.

[5] M. Flickner, et al. Query by Image and Video Content: The QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[6] D. A. Forsyth. Benchmarks for storage and retrieval in multimedia databases. In *Storage and Retrieval for Media Databases*, vol 4676, pp 240–247, San Jose, USA, 2002.

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

[8] D. Harman. Overview of the fi rst Text REtrieval Conference. In *Proceedings of the fi rst Text REtrieval Conf.*, pp 1–20, Washington DC, USA, 1992.

[9] D. P. Huijsmans and N. Sebe. Extended performance graphs for cluster retrieval. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition*, pp 26–31, Kauai, Hawaii, USA, 2001.

[10] I. Jermyn, C. Shaffrey, and N. Kingsbury. The methodology and practice of the evaluation of image retrieval systems and segmentation methods. CNRS Rapport de recherche ISRN I3S/RR-2003-05-FR, Sophia antipolis, 2003.

[11] C. Jörgensen. Towards an image testbed for benchmarking image indexing and retrieval systems. In *Proceedings of the Int. Workshop on Multimedia Content– Based Indexing and Retrieval*, Rocquencourt, France, 2001.

[12] M. Koskela, J. Laaksonen, S. Laakso, and E. Oja. Evaluating the performance of content–based image retrieval systems. In *Int. Conf. On Visual Information Systems*, LNCS 1929, Lyon, France, 2000.

[13] C. Leung and H. Ip. Benchmarking for content–based visual information search. In *Int. Conf. On Visual Information Systems*, LNCS 1929, pp 442–456, Lyon, France, 2000.

[14] 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*, Newcastle upon Tyne, 1998.

[15] 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 Int. Conf. on Multimedia*, pp 50–54, Ottawa, Canada, October 2001.

[16] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun. A framework for benchmarking in visual information retrieval. *Int. Journal on Multimedia Tools and Applications*, 21:55–73, 2003.

[17] 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*, 22(5):593–601, 2001.

[18] 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.

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

[20] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content–based manipulation of image databases. *Int. Journal of Computer Vision*, 18(3):233–254, 1996.

[21] T. Pfund and S. Marchand-Maillet. Dynamic multimedia annotation tool. In *Internet Imaging III*, vol 4672, pp 216–224, San Jose, USA, 2002.

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

[23] 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*, 22 No 12:1349–1380, 2000.

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

[25] 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*, Glasgow, 14–16 April 1999.

[26] K. Sparck Jones and C. 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, Uni. of Cambridge, 1975.

[27] J. Vendrig, M. Worring, and A. W. M. Smeulders. Filter image browsing: Exploiting interaction in image retrieval. In *Third Int. Conf. On Visual Information Systems*, LNCS 1614, pp 147–154, Amsterdam, The Netherlands, 1999.

[28] J. Zobel. How reliable are the results of large–scale information retrieval experiments? In *Proceedings of the 21st Annual Int. Conf. on Research and Development in Information Retrieval*, pp 307–314, Melbourne, Australia, August 1998.