IMAGECLEF 2004–2005: RESULTS, EXPERIENCES AND NEW IDEAS FOR IMAGE RETRIEVAL EVALUATION

Henning Müller[a], *Paul Clough*[b]

[a]Medical Informatics, University of Geneva
24, Rue Micheli-du-Crest,
1211 Geneva 14, Switzerland
henning.mueller@sim.hcuge.ch

William Hersh[c], *Antoine Geissbuhler*[a]

[b]Information Studies, University of Sheffield,
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP, UK
p.d.clough@sheffield.ac.uk

ABSTRACT

The *ImageCLEF* image retrieval benchmark was established in 2003 as part of the *CLEF* (Cross Language Evaluation Forum) to evaluate the retrieval of images from multilingual document collections or retrieval where a query is formulated in a language different from the language of the collection. In 2004, a visual retrieval task was added from the medical domain (using a mixed French/English annotation collection) because the use of visual information has one big advantage: it is inherently language independent.

This article describes the achievements of *ImageCLEF* 2004 by describing its tasks, goals, and the submissions received. The key findings will be explained, which will lead to further ideas to improve retrieval system performance. These will also lead to new ideas for *ImageCLEF* 2005 that will also be described in this paper, together with the evaluation goals and challenges for *ImageCLEF* 2005. Systematic performance evaluation is extremely important to show the progress of research in a domain, and there is an important lack of standardised evaluation in image retrieval. *ImageCLEF* is trying to fill this gap by supplying document collections, image retrieval tasks and topics based on user needs and ground truth for evaluating systems. This creates resources that can subsequently be used to advance research in image retrieval. *ImageCLEF* also provides a forum in which mixed–media information retrieval researchers can exchange ideas and technical details through an annual workshop to stimulate discussions and further research.

Goal of this article is to motivate research groups from the multimedia retrieval area to participate at *ImageCLEF* and help propose interesting tasks for further evaluation campaigns and shape the future of *ImageCLEF*. Only with a wide participation of the research communities can such a workshop remain interesting and fruitful for research. New databases and realistic topics based on current user requirements are needed regularly to keep the tasks challenging and advance the research domain of image retrieval to a level where it becomes important in real application domains.

This work has been funded in part by the EU Sixth Framework Programme (FP6) within the Bricks project (IST contract number 507457) as well as the SemanticMining project (IST NoE 507505). We also acknowledge the generous support of National Science Foundation (NSF) grant ITR-0325160.

[c]Oregon Health and Science University (OHSU), Department of Medical Informatics and Clinical Epidemiology, Portland, OR, USA

1. INTRODUCTION

Content–based image retrieval (CBIR) or visual information retrieval (VIR) has been on of the most active research domains in the fields of computer vision and image processing for the past 20 years [3, 28]. Hundreds of systems have been developed as well as commercial programs [7] and as research prototypes [2], and even open source systems are available [32] (GIFT[1]). However, progress has somewhat stalled and no general breakthrough has been achieved as of yet. Problems such as the semantic gap (distance between low level feature representation of images and high level "semantic" search tasks) still remain. In part, these limited improvements are due to the lack of a common benchmark with which to evaluate and compare systems. At computer vision conferences, it is clear that a large number of systems presented evaluate their techniques on different datasets, which makes all comparison impossible. Only the Corel dataset is used by several systems for evaluation [21]. However, it is used inconsistently by researchers (different subsets, etc.) making the comparison of retrieval systems very difficult if not impossible, and the dataset is not available anymore.

The closely related domain of text retrieval has successfully been performing systematic evaluation since the early 1960s [4], even though computing resources were limited. Significant advances in retrieval quality have been shown since then, especially TREC (Text REtrieval Conference[2]) [12], which has furthered research enormously and made

[1] http://www.gnu.org/software/gift/
[2] http://trec.nist.gov/

available large–scale document collections and realistic query tasks. Since 1992, new collections and tasks have been created annually in a cycle that includes a phase of *inscription* for participants, *data distribution*, delivery of *query tasks* on the data, *ground truthing*, *evaluation* and a *workshop* in which participants can present their retrieval methods and discuss results.

Similar ideas have been proposed for image retrieval from early on [29] following the TREC methodology. Other early publications on the evaluation of retrieval systems include [17, 22, 23, 30] where a variety of methodologies and needs for image retrieval evaluation were defined, including several performance measures. A benchmarking event was even started at the SPIE Photonics West conference called the Benchathlon[3] [11]. However, the event never went beyond discussions among the participants and no systematic comparison between systems was started. Evaluation remained a topic that was often discussed and in general people agreed on its importance for image retrieval [10, 15, 20], although negative voices on evaluation persist [8] stating that current systems are not good enough to be evaluated. This also lead to two evaluation initiatives that are currently in the starting phase called ImageEval[4] in France and one in the US for the Council on Library and Information resources (CLIR)[5]. Whereas ImageEval rather comes from an algorithmic view point using altered images to retrieve the originals and other simple tasks, the CLIR initiative rather tries to solve typical information retrieval problems with respect to images that occur in libraries. Another important and accepted evaluation campaign on moving images is TRECVID[6] [26, 27]. The main concentration is on videos, but it also includes the retrieval of key frames and most techniques used are basically the same as for image retrieval. Several of the tasks for visual analysis are on a simple semantic level, now, and the strong participation at TRECVID shows its importance for the field.

Multilingual information retrieval, which is the main focus of *CLEF*, is a domain well–rooted in text retrieval. Systematic evaluation for Cross–Language Information Retrieval (CLIR) began at TREC in 1997. In 2000, however, *CLEF* was established as an independent entity [25] and has had a growing number of participants every year since. Images are still not completely integrated into other fields of textual information retrieval, although the number of images produced and made available in digital form is rapidly increasing through the low cost of digital cameras and the ease with which images can be published on the Internet. *ImageCLEF* 2003 was the answer to this rising demand [6]. The realistic task (similar to the web) was modelled where an annotated collection of images exists in English but users want to query this collection in languages different from English. Tasks were based on real user queries performed at the St. Andrews library.

[3] http://www.benchathlon.net/
[4] http://www.imageval.org/
[5] http://www.clir.org/pubs/reports/trant04.html
[6] http://www-nlpir.nist.gov/projects/trecvid/

2. IMAGECLEF 2004

In 2003, there was already an image retrieval track with several query languages, but none of the participants actually used visual features for the retrieval of images (even though it was envisaged that combined with text retrieval this would achieve highest retrieval performance). The goal for 2004 was to strengthen the visual aspect of the task and motivate research groups from visual information retrieval to participate. More information about retrieval methods used by participants and results of *ImageCLEF* 2004 can be found in [5].

2.1. Tasks and datasets

Participants of *ImageCLEF* 2004 could register for one or all of the following three tasks:

- an ad–hoc multilingual task based on an image collection with English captions and queries in a variety of languages (no visual retrieval was required and queries included a short text in a language different than English, plus an example relevant image);
- a medical retrieval retrieval task based on an image collection with mixed language French/English texts and queries consisting of images only (visual query processing was necessary as the query was an image only, but visual results from an open source retrieval system were made accessible);
- a user–centered retrieval task based on the same collection as the ad–hoc task (no visual retrieval necessary).

The goal was clearly to motivate groups to use visual analysis to enrich queries and improve retrieval performance. To ease the use of visual features for research groups (particularly from the text retrieval community), an open source visual retrieval system called GIFT was made available to participants via a Web interface. Results lists from this system for all query images were also made available, to start query processing and use text for the results construction through automatic query expansion, for example in a second step.

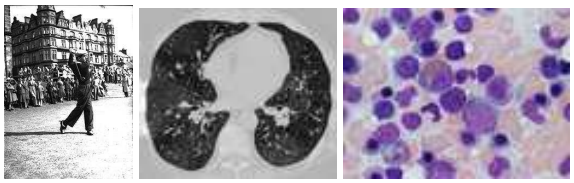
Two databases were used: the St. Andrews collection containing 28,133 historical photographs from St. Andrews library [7] and the casimage[8] medical collection containing

[7] http://www-library.st-andrews.ac.uk/
[8] http://www.casimage.com/

Title: Old Tom Morris, golfer, St Andrews.
Short title: Old Tom Morris, golfer.
Location: Fife, Scotland
Description: Portrait of bearded elderly man in tweed jacket, waistcoat with watch chain and flat cap, hand in pockets; painted backdrop.
Date: ca.1900
Photographer: John Fairweather
Categories: `[golf -- general]`, `[identified male]`, `[St. Andrews, Portraits]`,`[Collection -- G M Cowie]`
Notes: GMC–F202 pc/BIOG: Tom Morris (1821–1908) Golf ball and clubmaker before turning professional, later Custodian of the Links, St Andrews; golfer and four times winner of the Open Championship; father of Young Tom Morris (1851–1875).DETAIL: Studio portrait.

Tab. 1. Example caption from the St. Andrews collection.

8,725 medical images of different kinds including various modalities and anatomic regions. Images derive from 2,078 medical cases meaning that several images usually have the same textual annotation. Figure 1 shows one example from the St. Andrews collection (left) and two from casimage.



(a) (b) (c)

Fig. 1. Example images from the ImageCLEF collections.

Table 1 shows one example caption of the St. Andrews collection, which has all captions in English. The casimage collection has around 70% of the captions in French, 20% in English and almost 10% of the images have empty captions, only. A few cases have mixed English/French annotations

The query topics for the St. Andrews collection (25 in total) were based on real user queries from the owners of the collection at St. Andrews University Library and contained a small descriptive text that was translated into several languages plus one image (e.g. "golfers swinging their clubs"). The query topics for the medical task (26 in total) where single images only. Results needed to contain images of the same modality (CT, MRI, x–ray, etc.), same anatomic region, same view direction and sometimes the same radiologic protocol (or grey level setting). Representative query images were chosen by a radiologists familiar with the dataset to well represent the database. No training data was available for these tasks.

The goal of the interactive task was to find a given image in the database with as few iterations as possible. Experiments had to be performed by the participating groups themselves who sent in their results and strategies used.

2.2. Ground truths

A crucial part of the evaluation is having a reliable gold standard or set of ground truths. Judging the relevance of every image in a large collection for each query is practically infeasible due to limited resources and time. Therefore, we used a method called *document pooling*, which is described in [31] and used for most large–scale evaluation experiments such as TREC and CLEF. We took the first N retrieved images of every run submitted to the competition ($N = 50$ for the ad–hoc and $N = 60$ for the medical task). By computing the union of submitted results, we formed a document pool for every query topic that could then be judged for relevance. A total of three assessors were asked to judge the relevance of all images in the document pool. The judges rated the relevance of each image using a ternary scheme: relevant, partially relevant and non–relevant (although in practice only relevant and non-relevant images were indicated). Differences between the three judges were surprisingly high and we also had a large number of images judged as partially relevant.

Given multiple assessments per query, there are several ways to generate ground truths. For final evaluation and system ranking we used a set of images which were judged as relevant or partially relevant by the topic creator and at least one other assessor (called *pisec-total*). All ground truths were provided to participants for their own evaluation. We used trec–eval to evaluate all runs of the participants and the Mean Average Precision (MAP) was used as the lead measure for comparing and ranking systems (although a variety of measures were calculated and distributed, e.g. recall and precision at several points).

2.3. Participants

In total, 17 research groups participated in the three tasks (all groups that inscribed initially, really participated). Two groups submitted runs to the interactive task, 12 to the ad–hoc task and 11 to the medical image retrieval task. A total of 50 runs were submitted for the medical task and 190 for the ad–hoc task. Table 2 gives a further overview of the participating groups and the runs submitted.

2.4. Results

Results of the submissions were extremely varied, as well as techniques used for both tasks. The ad–hoc task was evidently more suited for textual information retrieval; whereas the medical tasks yielded very good results for visual–only retrieval. On the other hand, the medical retrieval task was not possible without a first visual step. Textual inclusion could on the other hand improve the results.

For the adhoc retrieval task the best monolingual (English) runs obtained a MAP of 0.58. Best results for other

Group	Country	Medical	Ad Hoc	Int.
National Taiwan University	Taiwan		5	
I–Shou University	Taiwan	3	4	X
University of Sheffi eld	UK		5	
Dublin City University	Ireland		79	
Imperial College	UK	1		
University of Montreal	Canada		11	
Oregon Health and Sci. U.	USA	1		
State University of New York	USA	3		
Michigan State University	USA		4	X
University of Alicante	Spain		27	
Daedalus	Spain	4	40	
UNED	Spain		5	
University of Geneva	Switzerland	14	2	
Medical Informatics, Aachen	Germany	2		
Computer Science, Aachen	Germany	8	4	
University of Tilburg	Netherlands	1		
CWI	Netherlands	4		
CEA	France	2	4	
Total		43	190	2

Tab. 2. Participating Groups in ImageCLEF.

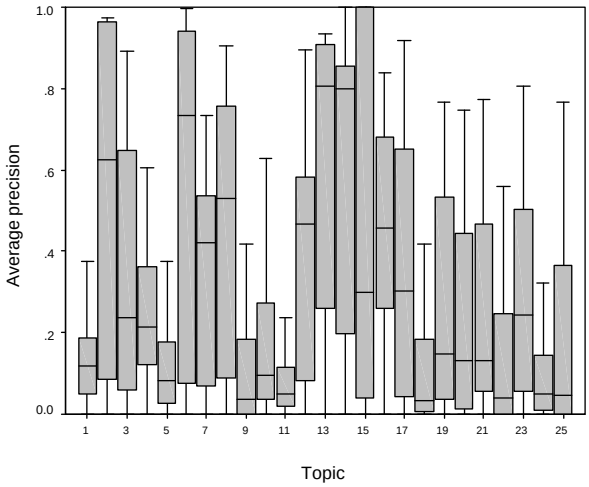


Fig. 2. Quality of responses per query topic for the adhoc task, showing the distribution of average precision.

languages were German, Spanish and French with MAP scores around 0.5. The worst results in the test were Japanese and Finnish with MAP=0.2 and purely visual retrieval with MAP=0.1.

In the medical tasks, the best automatic visual system had a MAP of 0.386, whereas the best system using textual and visual information was only slightly better at 0.39. With relevance feedback the results get better, reaching 0.43 for visual only retrieval and 0.476 for visual and textual retrieval combined. The combination of visual and textual features can actually be extremely important, especially with feedback. This shows when comparing several systems that used the same visual retrieval engine (GIFT) but varying text retrieval techniques with a strong variety in results.

Interestingly, the distribution of retrieval quality varies enormously among query topics. Figure 2 shows the distribution of retrieval quality defi ned by the MAP score (average precision) for the ad–hoc task. It shows that certain topics (topic 11) are clearly harder than others (topic 13), but for several topics the span between the best and the worst system is enormous, with a larger distance than for the medical task.

For the medical task the span of the retrieval quality is less strong as shown in Figure 3. Topics such as topic 7 can clearly be defi ned as an easy task and topic 11 clearly as an extremely hard task. Then, there are several topics with a large span in retrieval quality. The smaller variety can be explained with, in general, more similar visual retrieval techniques used. The variety of the database is also also smaller containing images from a limited number of modalities and anatomic regions. An interesting comparison is shown in [14] that compares various visual features and their performance for retrieval. Although Gabor fi lters seem to deliver best results as a single feature, only a combination of various features delivers really satisfying results as several features have some extremely good and some extremely bad queries.

Some of the main fi ndings of the ImageCLEF 2004 benchmark include:

- the use of visual features can improve the results even for rather "textual" or semantic tasks if used with care;
- visual queries can also be improved through careful use of textual information through automatic query expansion;
- manual relevance feedback lead to strongly improved results;
- there is no single visual feature set that leads to good results for all topics, rather a good combination is important for an overall good retrieval result;
- as a single feature set, Gabor fi lter responses seem to cover general visual characteristics best [14].

2.5. Comments from participants

The workshop went very well and stimulated much discussion among participants (although a certain distance between the cross–language retrieval research and image retrieval groups was visible). Through many small sessions and social events, discussions could be stimulated and it is important to move these two very specialised fi elds closer together as each can profi t from the knowledge of the other fi eld. Results show that visual and textual retrieval are very complementary.

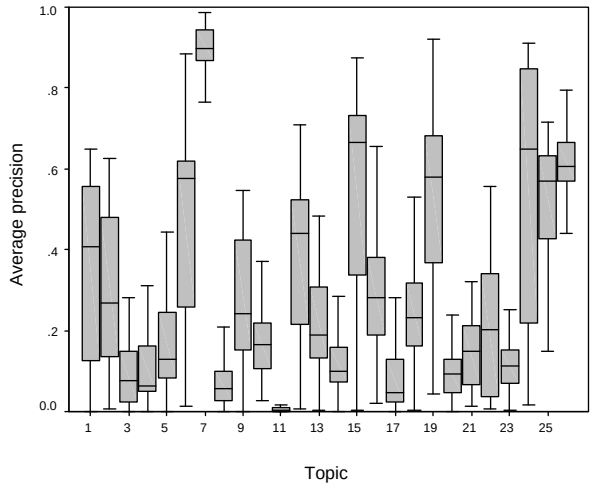


Fig. 3. Quality of responses per query topics for the medical task, showing the distribution of average precision.

Most comments were very positive about the opportunity to test retrieval systems on a common testbed within a comparative evaluation. However, a lack of training data was one of the main criticisms as many systems made use of classification strategies that require the use of training data, and so it was hard to optimise systems. Several groups were also not sure about what a good system performance would be like for the medical task as they did not have any medical specialists in their group. Medical specialists on the other hand criticised the topics as being not close enough to medical reality for a real system use. Another proposition for 2005 was to have a shorter time span between the time of topic release and the moment of results submissions. The reason for this proposition was the possibility of optimising systems too much by hand tuning.

3. IMAGECLEF 2005

ImageCLEF 2005[9] includes several of the comments of the participants into the organisation. The relevance sets of 2004 are made available as training data, so groups can use machine learning techniques to optimise their systems, although some of the tasks have changed significantly. A new automatic image annotation task has been added that provides 9,000 classified images as training data and requires participants to classify 1,000 new images into one of the classes that correspond to an IRMA code [16]. This task particularly stresses the importance of training data for the visual classification of images.

[9] http://ir.shef.ac.uk/imageclef2005/

3.1. Datasets

The datasets in 2005 have evolved from 2004, only the St. Andrews collection remains the same for one more year for the ad–hoc and the interactive tasks.

The casimage collection [24] is enlarged by three other medical teaching files: the PEIR[10] (Pathology Education Instructional Resource) database using annotation from the HEAL[11] collection (Health Education Assets Library, mainly pathology images [1]), the nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology[12] [33], and the PathoPic[13] collection (Pathology images [9]). This means that over 50,000 images are used instead of the 8,725 images previously used in 2004 for the medical task. This also changes the content of the dataset significantly. In 2004, only images of full medical cases were used and annotation was based on a case (not on a single image) and of very varying quality. In 2005 the annotations and content of the three datasets vary significantly, especially the HEAL collection, which contains more illustrations than typical medical images used in clinical routine. This is particularly good for the evaluation of image retrieval in teaching practice. The collection becomes also more multilingual as a much larger number of English annotations exist now as well as almost 10.000 images that are annotated in German and over 5,000 images annotated in French.

A second new database is the IRMA[14] collection. This collection contains a total of 10,000 images annotated with the IRMA code. This is a four–axes code, with axes for modality, body part, viewing direction and biological system examined. The IRMA code currently exists in English and German. A typical IRMA code is in the following form: **TTTT-DDD-AAA-BBB**, where T, D, A and B mean respectively technical, anatomical and biological axis. The code **1123-211-520-3a0** corresponds to "x–ray, projection radiography, analog, high energy – sagittal, left lateral decubitus,inspiration – chest, lung – respiratory system, lung". A complete description of the IRMA code and several examples can be found in [16]. A subset of the data (9,000 images) will be given to the participants with the complete IRMA code. The remaining 1,000 images will then need to be classified with the correct code. This year, the classification will not include the entire IRMA code as several categories of the code are visually almost indistinguishable. The task is limited to 57 classes and these class labels will be distributed to participants. The IRMA database contains mainly radiographs.

[10] http://peir.path.uab.edu/
[11] http://www.healcentral.com/
[12] http://gamma.wustl.edu/home.html
[13] http://alf3.urz.unibas.ch/pathopic/intro.htm
[14] http://www.irma-project.org/

3.2. Tasks

The ad–hoc task will change slightly in 2005 to better accommodate visual retrieval systems. More topics will be included that can lead to good results with visual retrieval. Experiments are also planned with respect to supplying more than one image as example with the textual query as well as purely visual topics "show me images similar to this one". To avoid subjective relevance judgements, the query topics are described in a very clear fashion also stating the limits of relevance for a particular query.

The interactive retrieval tasks will be promoted more to attract a larger number of participants. User–centered evaluation of retrieval system performance is extremely important, especially for the search of images. It is planned to give more room to this tasks as well at the workshop to promote the evaluation of interactive systems and stress the importance of query interfaces.

The medical retrieval task, on the other hand, will contain textual as well as image queries, in contrast to the purely visual task of the IRMA task. The collection will still be multilingual, but with a larger part being in English this year. The query texts have been translated into three languages (French, German, and English). The tasks for 2005 are based on responses from a survey that was carried out at Oregon Health & Science University (OHSU) in Portland, Oregon. Researchers, clinicians, and educators from OHSU were asked about their information needs regarding images, the tools that they currently use, and tools they would like to have available to them in the future. The tasks were based on the four axes of possible queries, including pathology, anatomic region and modality or imaging technique used. This is also a response to generate topics that will correspond better to real world information needs of clinicians [13].

A completely new task is the automatic image annotation task (IRMA task). A dataset that contains class labels will be given out to the participants. Each label represents an exact annotation of the fours axes. A new, unlabelled dataset will then need to be labelled with the correct labels learned from the training dataset. This is a fairly realistic task that can be used to obtain knowledge about collections that have not been annotated at all. An application can be the automatic correction of errors in DICOM headers by scanning images before being stored in a medical picture archive.

3.3. Further ideas and expectations

The amount of images and videos among data currently produced is growing steadily. The Internet has made available an extremely large variety of data, and currently only textual access is possible to most of these data. This does not only include images but also videos and sound, where a large research community exists [18, 19].

We are currently thinking about realistic tasks for multilingual image retrieval in various domains. One collection is the Web itself or part of it. In 2005, CLEF will contain a Web track and a natural extension for *ImageCLEF* would be to address multilingual image retrieval from the Web. With increased usage of digital cameras by domestic users and popularity in sharing pictures with a global audience, another realistic domain is that of multilingual access to personal image collections, e.g. holiday pictures. We have been provided with access to such a database, which already includes multilingual captions in English, French, German and Spanish. An automatic annotation task for medical images will begin in 2005 and one can imagine such an annotation task for non–medical images. A major advantage of annotation vs. low–level image features is that the extracted concepts can be easily translated into a variety of languages.

In 2005, there will also be a pre–CLEF workshop on the evaluation of visual retrieval systems in a more general fashion. This one–day–workshop will be in cooperation with the ImageEval evaluation effort and the MUSCLE[15] Network of Excellence (Multimedia Understanding through Semantics, Computation and Learning). The goal is to coordinate evaluation efforts in the domain and work on realistic tasks for evaluation as well as on the sharing of visual resources for evaluation purposes to advance the field of content–based image retrieval.

Through surveys among image users and comments from participants, we are hoping to obtain further ideas and suggestions to enable us to create a useful evaluation environment for image retrieval. Realistically–sized test collections are another important point to consider. Whereas medical teaching files are relatively simple to obtain, it is much harder to obtain permission to make freely available large image databases. We are currently in contact with image agencies on the Web to convince them of the importance of such a benchmark not only for the research community, but also for themselves. The use of smaller–sized images (thumbnails) or of image URLs instead of a complete database in an archive may be a solution to solve some of the copyright constraints, although it is important to keep the collections as evaluation resource accessible in the future.

4. CONCLUSIONS AND DISCUSSION

ImageCLEF is establishing itself as an important benchmark for both textual and visual methods of image retrieval after just two years of existence. Evaluation is extremely important for research and the benefits of such a benchmarking event cannot be underestimated. In the past, many researchers have talked about large–scale image retrieval evaluation similar to those already well–established in the

[15] http://www.muscle-noe.org/

text retrieval domain, such as TREC. The *ImageCLEF* evaluation campaign begins to address this and provides resources, which can be shared and used by the image retrieval community as a whole. Image retrieval is starting to become a commercially interesting domain as well, so a proof of performance can be an important factor for or against certain systems. However, *ImageCLEF* will remain an informal benchmarking event where the focus is not just about comparing system performance, but rather on providing resources for subsequent evaluation and a framework in which researchers can exchange their ideas. Commercial benefi ts of *ImageCLEF* are also foreseen in the future where companies using images can profi t from the outcome of the evaluation campaign to build systems that can combine the techniques shown to perform best. The success of *ImageCLEF* is dependent on the contributions from participants as only their involvement makes comparative evaluation a success.

5. REFERENCES

[1] C. S. Candler, S. H. Uijtdehaage, and S. E. Dennis. Introducing HEAL: The health education assets library. *Academic Medicine*, 78(3):249–253, 2003.

[2] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region–based image indexing and retrieval. In D. P. Huijsmans and A. W. M. Smeulders, editors, *Third International Conference on Visual Information Systems (VISUAL'99)*, number 1614 in Lecture Notes in Computer Science, pages 509–516, Amsterdam, The Netherlands, June 2–4 1999. Springer.

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

[4] 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, September 1962.

[5] P. Clough, H. Müller, and M. Sanderson. Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004. In C. Peters, P. D. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, Lecture Notes in Computer Science, Bath, England, 2005. Springer–Verlag.

[6] P. Clough and M. Sanderson. The CLEF 2003 cross language image retrieval task. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2003)*, 2004 – to appear.

[7] 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*, 28(9):23–32, September 1995.

[8] D. A. Forsyth. Benchmarks for storage and retrieval in multimedia databases. In *Storage and Retrieval for Media Databases*, volume 4676 of *SPIE Proceedings*, pages 240–247, San Jose, California, USA, January 21–22 2002. (SPIE Photonics West Conference).

[9] K. Glatz-Krieger, D. Glatz, M. Gysel, M. Dittler, and M. J. Mihatsch. Webbasierte lernwerkzeuge fr die pathologie – web–based learning tools for pathology. *Pathologe*, 24:394–399, 2003.

[10] M. Grubinger and C. Leung. Incremental benchmark development and administration. In *Proceedings of the Conference on Visual Information Systems (VISUAL 2004)*, San Francisco, CA, USA, 2004.

[11] 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.

[12] D. Harman. Overview of the fi rst Text REtrieval Conference (TREC–1). In *Proceedings of the first Text REtrieval Conference (TREC–1)*, pages 1–20, Washington DC, USA, 1992.

[13] W. R. Hersh and D. H. Hickam. How well do physicians use electronic information retrieval systems? *Journal of the American Medical Association*, 280(15):1347–1352, 1998.

[14] P. Howarth, A. Yavlinsky, D. Heesch, and S. Rüger. Visual features for content–based medical image retrieval. In C. Peters, P. D. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, Lecture Notes in Computer Science, Bath, England, 2005. Springer–Verlag.

[15] 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, Laboratoire Informatique, signaux et sytèmes de Sophia Antipolis, 2003.

[16] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. B. Wein. The IRMA code for unique classifi - cation of medical images. In *Medical Imaging*, volume 5033 of *SPIE Proceedings*, San Diego, California, USA, February 2003.

[17] C. Leung and H. Ip. Benchmarking for content–based visual information search. In R. 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.

[18] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proceedings of the second International Conference on Multimedia and Exposition (ICME'2001)*, pages 952–955, Tokyo, Japan, August 2001. IEEE Computer Society, IEEE Computer Society.

[19] S. Marchand-Maillet. Content–based video retrieval: An overview. Technical Report 00.06, CUI - University of Geneva, Geneva, Switzerland, 2000.

[20] H. Müller, A. Geissbuhler, S. Marchand-Maillet, and P. Clough. Benchmarking image retrieval applications. In *Proceedings of the Conference on Visual Information Systems (VISUAL 2004)*, San Francisco, CA, USA, 2005.

[21] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel – evaluation in image retrieval. In *Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002)*, London, England, July 2002.

[22] 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, April 2001.

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

[24] A. Rosset, H. Müller, M. Martins, N. Dfouni, J.-P. Vallée, and O. Ratib. Casimage project – a digital teaching fi les authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.

[25] J. 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.

[26] A. F. Smeaton, W. Kraaij, and P. Over. Trecvid 2003: An overview. In *Proceedings of the TRECVID 2003 conference*, December 2003.

[27] A. F. Smeaton, P. 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.

[28] 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.

[29] J. 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.

[30] 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*, Electronic Workshops in Computing, Glasgow, 14–16 April 1999. The British Computer Society.

[31] 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, University of Cambridge, 1975.

[32] D. M. Squire, W. Müller, H. Müller, and T. Pun. Content–based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)*, 21(13-14):1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.

[33] J. W. Wallis, M. M. Miller, T. R. Miller, and T. H. Vreeland. An internet–based nuclear medicine teaching fi le. *Journal of Nuclear Medicine*, 36(8):1520–1527, 1995.