A reference data set for the evaluation of medical image retrieval systems

Henning Müller, Antoine Rosset, Jean–Paul Vallée, François Terrier,
Antoine Geissbuhler
University Hospitals of Geneva, Service for Medical Informatics
Rue Micheli-du-Crest 24, 1211 Geneva 14, Switzerland
tel ++41 22 372 6175, fax ++41 22 372 8680
henning.mueller@sim.hcuge.ch

April 6, 2004

Abstract

Content–based image retrieval is starting to become an increasingly important factor in medical imaging research and image management systems. Several retrieval systems and methodologies exist and are used in a large variety of applications from automatic labeling of images to diagnostic aid and image classification. Still, it is very hard to compare the performance of these systems as the used databases often contain copyrighted or private images and are thus not interchangeable between research groups, also for patient privacy. Most of the currently used databases for evaluating systems are also fairly small which is partly due to the high cost in obtaining a gold standard or ground truth that is necessary for evaluation. Several large image databases, though without a gold standard, start to be available publicly, for example by the NIH (National Institutes for Health).

This article describes the creation of a large medical image database that is used in a teaching file database containing more than 8700 varied medical images. The images are anonymized and can be exchanged free of charge and copyright. Ground truth (a gold standard) has been obtained for a set of 26 images being selected as query topics for content–based query by image example. To reduce the time for the generation of ground truth, pooling methods well known from the text or information retrieval field have been used. Such a database is a good starting point for comparing the current image retrieval systems and to measure the retrieval quality, especially within the context of teaching files, image case databases and the support of teaching. For a comparison of retrieval systems for diagnostic aid, specialized image databases, including the diagnosis and a case description will need to be made available, as well, including gold standards for a proper system evaluation.

A first evaluation event for image retrieval is foreseen at the 2004 CLEF conference (Cross Language Evaluation Forum) to compare text– and content–based access mechanism to images.

keywords: medical image retrieval, content–based search, evaluation, reference database, gold standard, ground truth

1 Introduction

Content–based access to multimedia data is a very active research area and a number of specialized applications have been developed whether for audio data [1], videos [2] or for images [3–5]. The need for such new access methods is generated by the enormous amount of multimedia data that is being created and often made available on the Internet. Digital cameras at consumer prizes also resulted in a large amount of visual data being produced that is unavailable for current search tools other than by searching for a filename. Various research prototypes have been presented such as MITs Photobook [6] or Berkeley's Blobworld [7] and also several commercial solutions are available such as IBMs QBIC or Virage [8, 9] but there has not yet been a general technological breakthrough. One of the reasons for this stagnation is the unavailability of objective measures for a comparison of systems and techniques to objectively identify the best–performing technologies. A few databases for system comparison have recently been made available such as the database of the University of Washington[1] and methods for the evaluation of retrieval systems are described in a large number of publications [10–14] and can be discussed at the Benchathlon[2] conference.

In the medical domain, the amount of audio–visual data produced and stored in digital form is also increasing quickly. The knowledge contained in these data is currently barely used and the access to the PACS (Picture Archiving and Communication System) archives is mainly done by patient identification. The need for alternative access methods to medical image data such as content–based retrieval has also been described in a large number of scientific articles [15–19]. Many research projects exist and the number of publications in the field has been rising strongly in recent years. Unfortunately, only few projects are properly evaluated like the *ASSERT* (Automatic Search and Selection Engine with Retrieval Tools) [20–22] and *IRMA* (Image Retrieval in Medical Applications) [23, 24] projects. For the ASSERT project, even a real user test on diagnostic quality in radiology has been performed which shows a significant improvement in diagnostic quality when using their system as decision support [22]. Unfortunately, the datasets for evaluation of these systems are not publicly available and thus no comparisons between several systems can be performed and it is impossible to prove that another technique is inferior or superior to an existing one.

Retrieval systems exist for a variety of applications from access to large varied data collections such as they exist in PACS systems [24–27] to very specialized collections, mainly for diagnostic aid [28–31] in special domains such as pathology, dermatology and mammography. Both are equally important for a proper management of medical image data. This article describes the generation

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

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

of an image dataset for the evaluation of systems that perform visual similarity retrieval on varied, PACS–like databases containing a very varied set of images. Goal is as well, to develop specialized databases for diagnostic aid and made them available free of charge including ground truth for a system evaluation.

A comprehensive review of content–based medical information retrieval and example systems can be found in [32].

2 Medical image databases available

Most of the databases currently used for image retrieval are small and copyrighted so they cannot be taken for retrieval system evaluation if reliable measures are wanted. It is clear that medical images are very sensible and databases can only be made publicly available when they are anonymized and when it is impossible to find out the identity of the patient with the accompanying information. Some of these rules are described in the RSNA (Radiological Society of North America) MIRC (Medical Imaging Resource Center[3]) that creates a standard for radiological teaching databases that can be made available as one large source and can be searched in parallel. Their resources currently include seven data sets that can be searched by text and/or patient attributes such as sex, year of birth etc. when available.

Some image databases currently being used for content–based image retrieval even contain simulated images which is good for showing the efficiency (speed) of an algorithm but is not possible to be used for an evaluation of retrieval quality [33, 34]. Some examples for rather small databases used for retrieval are [35], using 15 PET studies and [36] using 41 biopsies slides. Such small databases are not able to show any statistically significant results for retrieval system evaluation or comparison. A different approach was taken by [37] where a collection was gathered by accessing various medical sources on the web and thus freely accessible images. This allows a larger dataset and allows an exchange of the data set with other research groups which is important for system comparison.

The National Institutes of Health (NIH) in the United States is making several of its resources available such as an image database on medical history[4], a database on Spine X–Rays [38] as well

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

[4] http://wwwihm.nlm.nih.gov/

as the well–known visible human project.

A number of web pages also offer access to varied medical image resources on the Internet such as at Creighton University[5]. The Karolinska institute maintains a very extensive listing of medical image resources[6] which contains references to, for example, the Bristol biomedical image database that is searchable by the image content, references to anatomic atlases and links to image sources ordered by anatomic regions. Several initiatives aim to generate medical reference image databases for evaluation on a European level [39] and at the NIH[7].

Unfortunately, these databases can not be used straight away for the evaluation of image retrieval systems as neither query topics nor a gold standard exist. As the generation of ground truth is without question the hardest (=most expensive, most time–consuming) part of system evaluation, mainly small, local databases are being used and the large image repositories available are being discarded.

3 A database for the evaluation of medical image retrieval systems

The image database described in this paper can be used on the web in the *CasImage* [8] system. The database contains 8751 images from various fields of medical imaging with a strong focus on the radiology department. A total of around 2000 cases is in the database with several selected images per case. Textual descriptions of most of the cases exist in varying quality with some cases being in French and others in English. Figure 1 shows how varied the image database is, containing images that are mainly from the radiology department but also a few color images such as pathologic cuts or microscopic images and also dental pictures. The displayed images are the chosen query topics for evaluation.

Much of the procedure of choosing query topics and generating ground truth for the evaluation of retrieval systems is taken from the experience of the text retrieval community. TREC (Text REtrieval Conference[9]) is the standard for benchmarks in the field and published several articles

[5] `http://www.hsl.creighton.edu/hsl/Guides/Gd-Images.html`

[6] `http://www.mic.ki.se/Medimages.html`

[7] `http://www3.cancer.gov/bip/lidc_comm.htm`

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

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

about their methodology of evaluation [40, 41]. Many of the techniques for evaluation even range further back to the SMART text retrieval system in the 1960s [42]. Other important evaluation iniatives in the textual domain such as NTCIR (National institute of informatics Test Collection for Information Retrieval) and also for cross language image retrieval [43] within CLEF[10] (Cross Language Evaluation Forum) use the same techniques. In 2003, CLEF started an image retrieval track called imageCLEF[11] to evaluate image retrieval methods but no group actually used the visual data and all retrieved the images by text only. 2004 will see the first medical image retrieval subtask of imageCLEF using the database described in this paper. Query topics will be images only, without text. Still, query expansion and relevance feedback can use the multilingual textual data.

3.1 Query topics

Among the 8751 images in the databases, a set of 26 images was chosen by a radiologist, an expert, as query tasks or topics. These images are meant to well represent the image database, and the contained images cover a wide spectrum of possible query tasks that are interesting for data management within a PACS system. This includes a variation in modalities, anatomic regions and radiologic protocols used. Some initially chosen query topics were removed from the final set of topics as no visually similar images existed in the database. Figure 1 shows the images finally chosen as queries.

(Figure 1)

3.2 A gold standard

One of the most costly and also most important tasks for evaluation is the definition of a gold standard or ground truth as it is called in information retrieval. It needs to be defined what a perfect system performance would be like. Then, the responses of the retrieval system can be compared with the gold standard. Within medical systems, this is only a small part of the evaluation, though. After the validation of the algorithms, it still needs to be thought about the inclusion of human factors and finally the evaluation of clinical effects [44, 45]. Medical image

[10] `http://www.clef-campaign.org/`

[11] `http://ir.shef.ac.uk/imageclef2004/index.html`

retrieval is still in its infancy and is a research domain with clinical applications being sparse as of yet. For this reason it is very important to at least have data for the validation of the algorithms which is a first step and needs to be followed by clinical evaluations.

3.2.1 Using a pooling method for ground truthing

In domains such as information retrieval, systematic evaluation of retrieval quality has been done since more than 40 years [46]. As the collections need to correspond to real world databases, TREC, for example has millions of documents in the database that need to be indexed and searched compared with often only a few hundred images or less in databases for medical image search. This necessitates techniques to obtain ground truth or a gold standard other than judging *all* documents in the collection. Such methods have been adopted by most evaluation iniatives such as TREC, CLEF and NTCIR.

In [47, 48], a method was first proposed for not having to judge all documents but to still have a fair comparison of the systems under test. All systems have to send in the highest–scoring N (*i.e.* = 200) documents that they regard as the most relevant results of a query. The first N documents of all participating systems are thus the pool that needs to be judged by a domain expert. This method is likely to miss a few relevant documents and the recall results and precision/recall (PR) graphs of systems can be altered slightly by the not completely correct gold standard. Test on the number of missed documents have been performed in TREC. In [41], it is mentioned that this can change the end of a precision/recall graph. The results were altered starting from around 40% recall. Changes cite were in the order of 10% difference between the pooled evaluation and full evaluation. Still, this is very similar for all participating systems and thus a fair way to obtain relevance judgments for comparing systems. No systematic bias towards systems was recognized in several other studies, too. An even more profound analysis of this pooling and the effects on the ground truth, incomplete judgment and the evaluation results can be found in [49].

In the tests to generate the database described in this article, it was not possible to query various substantially different retrieval systems but several color quantizations and several texture descriptors were used to obtain a total of 3 result sets that are included into the pooling. The different color/grey level features and the texture descriptor correspond well to the characteristics of many current image retrieval systems. Whether Gabor filter responses or features based on cooccurence matrices are used should not extremely change the retrieval results as the same information is modeled in principal. The first $N = 500$ images of a query for each image and each system are included into the pool. Three different quantizations of the color space HSV (Hue, Saturation, Value) were used as follows (HSV/gray=18,3,3/4;9,2,2/64;9,2,2/32), and two different set ups of the Gabor filters using three scales and four to six directions. As color space, the HSV space was chosen as it is proven to correspond better to human perception than other common spaces such as RGB (Red, Green, Blue) [50]. As most of the images in the database are grey scale anyways, this should not matter extremely much.

The results show that some very similar images are on the top of every result set but that the result sets of the three systems compared entirely are indeed fairly different. The pooling set as a result of the first 500 retrieved images of each system contains an average of 846 images with 549 being the smallest number of relevant images and 1011 being the highest number, meaning the query where the three systems delivered the most differing results. This shows that the result sets of the system are varied and can be taken to represent three different systems in our tests. With the imageCLEF task on the same database we hope to get results from more systems and consequently a better ground truth than with simply feature space variations of one single system. The use of this pooling technique allows us to reduce the time for generating the relevance judgments to less than 10% of the time that it would have been if all images were controlled for relevance. As the images are ordered based on their visual similarity with respect to three different retrieval techniques, the risk to fail in finding all relevant images is slightly limited. It does not influence a comparison of these techniques as has been shown in [41, 49].

It is clear that incomplete ground truthing is a disadvantage of the used pooling technique. On the other hand, large databases for evaluation can be made available relatively quickly with ground truth. This avoids to use the same database all the time and having systems extremely optimized for one single database.

3.2.2 Performing the ground truthing

The images in the pooling sets were simply displayed on a web page that was accessible for the radiologist who could judge thumbnails of the images quickly and who could have access to the full size images and a case description on a mouse click if further inspection was necessary. For most of the images, visual inspection is quick and no further verification is necessary after inspecting the thumbnail. The internal URLs of the images were used as identifications for generating the gold standard and thus the outcome were text files containing all images that were regarded as a relevant result for a given query image.

Further, manual checks had the goal to find missing images or simple errors when selecting the images. All result sets were controlled once again manually to not include irrelevant images. Then, a query was executed with all the images from a result set. The first 100 of the returned images not in the relevance set were then again inspected manually to avoid missing any relevant images. Further small changes in the relevance sets might still be possible when finding missing images.

Relevance itself is an often discussed topic in information retrieval evaluation [51, 52]. Whereas TREC and CLEF regard documents as relevant as soon as a little part of it is regarded relevant, our definition for images being relevant to a query image is slightly stricter. Our definition of relevance is that the retrieved image is in modality, radiologic protocol used and anatomic region shown the same as the query image. It still has to be found out how much room there is for subjectiveness of users as it does exist for textual information retrieval [53]. This process is thus not on the level of including medical diagnosis as a relevance factor which is desirable as a tool for diagnostic aid but it can still show how well techniques can retrieve visually similar images based on modality, anatomic region and protocol used. For the evaluation of tools for diagnostic aid it is important to have specific databases for a certain purpose and specific ground truth in the form of diagnoses.

The person obtaining the ground truth was further asked to note the following for each query image under test:

- The time it took to judge all the images for one query;
- the position of the last relevant image to estimate the chance of not finding all the relevant images.

3.2.3 Resulting gold standard

As a result of the ground truthing, 26 relevance sets were obtained, one for each query topic. These relevance sets contain an average of 85 relevant images with the maximum being at 383 and the minimum at 2 relevant images. The expert evaluators took an average of 17 minutes to perform the judgment of each of the 26 topics, resulting in a total of 2361 relevant images chosen from the 22,000 images shown for judgment. This comes down to more than 7 hours of relevance judgments compared with 80 hours it would have taken to judge all the images, not included the reduced speed after having done part of the judgments due to fatigue and lack of concentration. The longest time to judge one topic was 30 minutes and the shortest time 1 minute, which was a data set containing only two relevant images. Most data sets took around 15 minutes to be judged. The average position of the last relevant image found was at 639 with a maximum and minimum at 2 and 956. This shows that the large majority of relevant images will have been found but it also means that often, images were found rather late. As the images are ordered by visual similarity averaged over three systems, only a very small fraction of images is likely to be missed in the process.

All the relevance sets, a list of the query images and the images themself can be ordered by email from the principal author. When using the image set for publications it should be cited in a proper way so it is clearly visible which image data set was used. Other than that there are no restrictions with respect to the use of the images for evaluation purposes.

4 An example evaluation

This chapter shows a short example evaluation for one retrieval system called *medGIFT* [12] [54] in three different configurations using various grey level quantizations and texture descriptors as has been used for the pooling. The performance measures used for evaluation are those used in benchmarks for information retrieval such as TREC and also for content–based image retrieval such as the Benchathlon. The measures are easy to interpret and the variety of measures assures the evaluation of various system parameters and aspects.

4.1 Methodology

The methodology for evaluation follows that described in [12] for the evaluation of image retrieval systems. Queries are performed in a completely automated way for all the query topics that were

[12] http://www.dim.hcuge.ch/medgift/

chosen in the form of example images. In our case, there are thus 26 single–image–queries. Such Query by Example (QBE) is the most commonly used query paradigm for content–based image retrieval. The retrieval results of the three system configurations are then compared with the gold standard. Various measures can be calculated to see how close the system response was compared to a domain expert and how quickly the system was responding to allow interactive working and querying.

The following measures are used for system evaluation:

- $Rank_1$, $\overline{Rank}$ and $\widetilde{Rank}$: rank at which first relevant image is retrieved, average rank and normalized average rank of relevant images (see below and Equation 1).
- $P(20)$, $P(50)$ and $P(N_R)$: *precision* after 20, 50 and N_R images are retrieved, where N_R is the number of relevant images. This corresponds in general well to the numbers of images that users look at on screen.
- $R_P(.5)$ and $R(100)$: *recall* at *precision* .5 and after 100 images are retrieved.
- *Precision* versus *recall* graph.
- Time it takes to execute the query and retrieve the results.

A simple average rank is difficult to interpret, since it depends on both the collection size N and the number of relevant images N_R for a given query. Consequently, it was normalized by these numbers to propose the *normalized average rank*, $\widetilde{Rank}$:

$$\widetilde{Rank} = \frac{1}{NN_R}\left(\sum_{i=1}^{N_R} R_i - \frac{N_R(N_R+1)}{2}\right) \tag{1}$$

where R_i is the rank at which the ith relevant image is retrieved. This measure is 0 for perfect performance, and approaches 1 as performance worsens. Random retrieval will result in 0.5.

As relevance feedback is a very important factor and the performance with feedback is often seen as even more important than the results for single–image queries [55], the query performance with relevance feedback was also evaluated. To generate relevance feedback, the first 50 query results were taken and all relevant images were fed back as positive feedback and all non–relevant images as negative feedback in the same way as we would expect a real user to mark the images. The same performance measures are used for the evaluation of the performance with relevance feedback as without.

To have a better idea of the overall performance of a certain system configuration, all performance measures are averaged over all 26 queries.

For the execution of the evaluation, a benchmark based on Perl scripts was used that sends a query to the server, receives the results and then, calculates the performances measures. Here, the advantage is to have *MRML* (Multimedia Retrieval Markup Language[13]), an access method to the query engine that allows such a fully automatic evaluation.

4.2 *medGIFT*

medGIFT is a medical adaptation of the *GIFT* [14] (GNU Image Finding Tool) which is an open source image retrieval framework and thus available free of charge. Currently, main changes are with respect to the color space used where the number of gray levels is augmented to 32/64 and the number of texture measures is also enlarged. This is achieved through simply adding six instead of four directions to the Gabor filters and also with tests for several scales. *medGIFT* also offers a new user interface that shows the diagnosis of a case in the interface with a thumbnail image and features a link to the full–size image and the textual case description in the *CasImage* system. It is planned to implement a larger set of visual image descriptors and evaluate them on various databases to obtain their performance differences for certain query tasks. This was originally the main reason for creating this database because subjective impressions of the retrieval quality do often not correspond to results of objective evaluation.

4.3 Evaluation results

These evaluation results show indeed a slightly unexpected behavior. The basic gift system without adaptations in the feature space has a surprisingly good behavior in the first query step. Table 1 shows that in the first query step, the system has a slightly better performance in early precision and is even much better with respect to average rank measures. This can be linked to the fact that a few of the queries do use color characteristics and that the system performs better on these queries. Another explication could be that a larger number of grey levels becomes too specific for image retrieval and misses out on a few groups of images that only have a fixed grey level offset.

[13] `http://www.mrml.net/`

[14] `http://www.gnu.org/software/gift/`

This could be circumvented with a grey level normalization before the feature extraction.

After feedback, the performance difference between the three systems is fairly small as can be seen in the various performance measures. With respect to the position of the first relevant image the two systems with larger numbers of grey levels perform significantly better than the system with only four grey levels. This means that there is at least one query where the system with the smaller number of grey levels does not retrieve any relevant images within the first 50 results that are used for feedback.

Evaluations are performed on an old Pentium III computer with 500 MHz 384 MB RAM and a hard disk with 5400 RPM. This partly explains the slow response times shown in Table 1. For the evaluation, very large response lists containing all 8751 image and thumbnail URLs are transmitted each time. For a real user who usually displays between 20 and 50 images on screen response times are much faster because only 20–50 URLs need to be transmitted.

Table 1 also shows that the systems with a larger number of grey levels performs significantly better for the average position of the first relevant image retrieved. This means that the system with only four grey levels has at least one really poorly performing query in the first step. The query time with a larger number of grey levels is slightly faster. All the precision and recall measures are the best for the system with only four grey levels but the differences are not very strong. With relevance feedback, the precision and recall results are basically the same which means that a larger number of grey levels leads to a stronger gain in performance than a small number. The average rank measure in contrast are significantly better for a small number of grey levels, even after relevance feedback.

(Table 1)

The precision vs. recall graphs in Figure 2 underline the observations of the other performance measures. The system with the smallest number of grey levels leads to very good results in the first query step. Then, with the use of relevance feedback, the system performance of the retrieval system with a larger number of grey levels get better. This can lead to a retrieval strategy where all one–image–queries are performed with a smaller number of grey levels and once more information is available in the form of feedback images, a system with a larger number of grey levels can be used for perfect performance.

(Figure 2)

Although subjective observations suggested that the retrieval using a larger number of grey levels leads to better image retrieval result, only a proper evaluation can show which features are really performing well and which ones are less performant. Only a proper comparison based on the same data allows to objectively judge whether a technique is performing better than another one or not. Standardized performance evaluation has already shown surprising results several times such as the Cranfield II studies [56] that showed that automatic full text retrieval performs as well as manually attached keywords from experts. These findings had a strong influence on the further development of text retrieval systems. Standardized evaluation can do the same thing for content–based image retrieval.

5 Conclusions

This article gives an introduction into the field of content–based medical image retrieval and the evaluation of image retrieval systems for the medical domain, including data sets and evaluation methods. Such proper evaluations are necessary to compare systems and to make the technology a success and gain acceptance for its use. For evaluation, standard databases are needed including ground truth data that can be used to evaluate a system without having to gather large amounts of data and create new relevance judgments every time. An initiative to create reference databases for the evaluation of medical image processing algorithms has also been started by the European Federation for Medical Informatics (EFMI) [57] and we are working closely with this working group.

The entire process of the creation of an evaluation collection is taken to build a medical image database for the benchmarking of medical content–based image retrieval systems. The database is available free of charge and free of copyright. It can be ordered from the principle author including the query topics and relevance judgments which will ease significantly the evaluation and comparison of image retrieval projects for the use in teaching files and PACS systems. The methodology to choose query topics and obtain the relevance judgments with respect to these topics is taken for the text retrieval domain that has more than forty years of experience in evaluating retrieval system performance. The methodology is therefore proven and experienced and the theoretical foundations are sound. Basic assumptions such as the need to have a large image collection and a large number of query topics to average the system performance for statistically significant evaluation are taken into account. The number of query topics can of course still be enlarged for further evaluations.

Of course, not all retrieval systems can be evaluated with such a database. Many retrieval systems are destined to be used as a very specific diagnostic aid and thus are working on very specialized databases in very small domains. Such systems can use different features and specialized distance metrics, others than those to manage images in a PACS system or a radiology teaching file. For the evaluation of these systems, other, specialized databases need to be developed and ground truth in the form of proven diagnoses can be taken. Still, the management of varied medical image databases is important and the use of the technology can well be learned in the environment of a teaching file where visually similar cases using the same modality, radiologic protocol and anatomic region are destined to be retrieved.

Goal of this article is also to foster the comparison of various techniques and feature sets for image retrieval. It is important to find out which visual features perform well for what kind of image retrieval tasks. System comparisons and benchmarks are an important factor in this. The performance of text retrieval systems improved dramatically in the first few years that TREC took place as systems were offered an evaluation basis and data including relevance judgments to compare their retrieval algorithms on. Several key findings in text retrieval are due to benchmarks.

A benchmarking event for the medical domain where system performance can be compared and technologies and visual features can be discussed would be even better, just in the spirit of the "friendly" TREC meeting where discussion of the used techniques is regarded more important than the pure performance of the systems. A first event into this direction can be the imageCLEF track that uses a database with medical images. A comparison of visual retrieval techniques and textual multi–lingual retrieval is one goal of this event. A strong participation at such a benchmarking event can really help to advance the field of medical image retrieval and compare a large number of competing techniques. Visual information retrieval will become an important factor in the medical field and standardized evaluation plus standard databases can strengthen this development and identify promising techniques.

References

[1] 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.

[2] D. Schonfeld and D. Lelescu. VORTEX: Video retrieval and tracking from compressed multimedia databases – Template matching from MPEG2 video compression standard. In C.-C. J. Kuo, S.-F. Chang, and S. Panchanathan, editors, *Multimedia Storage and Archiving Systems III (VV02)*, volume 3527 of *SPIE Proceedings*, pages 233–244, Boston, Massachusetts, USA, November 1998. (SPIE Symposium on Voice, Video and Data Communications).

[3] 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.

[4] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Past, present and future. In M. Liao, editor, *Proceedings of the International Symposium on Multimedia Information Processing*, Taipei, Taiwan, December 1997.

[5] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, May 1997.

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

[7] 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–Verlag.

[8] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu. The Virage image search engine: An open framework for image management. In I. K. Sethi and R. C. Jain, editors, *Storage & Retrieval for Image and Video Databases IV*, volume 2670 of *IS&T/SPIE Proceedings*, pages 76–87, San Jose, CA, USA, March 1996.

[9] 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.

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

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

[13] 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.

[14] 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.

[15] H. J. Lowe, I. Antipov, W. Hersh, and C. Arnott Smith. Towards knowledge–based retrieval of medical images. the role of semantic indexing, image content representation and knowledge–based retrieval. In *Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA)*, pages 882–886, Nashville, TN, USA, October 1998.

[16] C. Traina Jr, A. J. M. Traina, R. R. dos Santos, and E Y. Senzako. A support system for content–based medical image retrieval in object oriented databases. *Journal of Medical Systems*, 21(6):339–352, 1997.

[17] A. Rosset, O. Ratib, A. Geissbuhler, and J.-P. Vallée. Integration of a multimedia teaching and reference database in a PACS environment. *RadioGraphics*, 22(6):1567–1577, 2002.

[18] E. G. M. Petrakis. Content–based retrieval of medical images. *International Journal of Computer Research*, 11(2):171–182, 2002.

[19] H. D. Tagare, C. Jaffe, and J. Duncan. Medical image databases: A content–based retrieval approach. *Journal of the American Medical Informatics Association*, 4(3):184–198, 1997.

[20] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick. ASSERT: A physician–in–the–loop content–based retrieval system for HRCT image databases. *Computer Vision and Image Understanding (special issue on content–based access for image and video libraries)*, 75(1/2):111–132, July/August 1999.

[21] J. G. Dy, C. E. Brodley, A. Kak, C.-R. Shyu, and L. S. Broderick. The customized–queries approach to CBIR using using EM. In *Proceedings of the 1999 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'99)*, pages 400–406, Fort Collins, Colorado, USA, June 23–25 1999. IEEE Computer Society.

[22] A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, and A. Marchiori. Automated storage and retrieval of thin–section CT images to assist diagnosis: System description and preliminary assessment. *Radiology*, 228:265–270, 2003.

[23] D. Keysers, J. Dahmen, H. Ney, B. B. Wein, and T. M. Lehmann. A statistical framework for model–based image retrieval in medical applications. *Journal of Electronic Imaging*, 12(1):59–68, 2003.

[24] T. M. Lehmann, M. O. Güld, C. Thies, B. Fischer, M. Keysers, D. Kohnen, H. Schubert, and B. B. Wein. Content–based image retrieval in medical applications for picture archiving and communication systems. In *Medical Imaging*, volume 5033 of *SPIE Proceedings*, San Diego, California, USA, February 2003.

[25] H. Qi and W. E. Snyder. Content–based image retrieval in PACS. *Journal of Digital Imaging*, 12(2):81–83, 1999.

[26] C. Le Bozec, E. Zapletal, M.-C. Jaulent, D. Heudes, and P. Degoulet. Towards content–based image retrieval in HIS–integrated PACS. In *Proceedings of the Annual Symposium of the*

American Society for Medical Informatics (AMIA), pages 477–481, Los Angeles, CA, USA, November 2000.

[27] E. El-Kwae, H. Xu, and M. R. Kabuka. Content–based retrieval in picture archiving and communication systems. *Journal of Digital Imaging*, 13(2):70–81, 2000.

[28] M. E. Mattie, L. Staib, E. Stratmann, H. D. Tagare, J. Duncan, and P. L. Miller. PathMaster: Content–based cell image retrieval using automated feature extraction. *Journal of the American Medical Informatics Association*, 7:404–415, 2000.

[29] P. Schmidt-Saugeon, J. Guillod, and J.-P. Thiran. Towards a computer–aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics*, 27:65–78, 2003.

[30] S. Baeg and N. Kehtarnavaz. Classification of breast mass abnormalities using denseness and architectural distorsion. *Electronic Letters on Computer Vision and Image Analysis*, 1(1):1–20, 2002.

[31] L. H. Y. Tang, R. Hanka, and H. H. S. Ip. A review of intelligent content–based indexing and browsing of medical images. *Health Informatics Journal*, 5:40–49, 1999.

[32] H. Müller, N. Michoux, D. Bandon, and A. 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.

[33] E. G. M. Patrakis and C. Faloutsos. Similarity searching in medical image databases. *IEEE Transactions on Knowledge and Data Engineering*, 9(3):435–447, 1997.

[34] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas. Fast and effective retrieval of medical tumor shapes. *IEEE Transactions on Knowledge and Data Engineering*, 10(6):889–904, 1998.

[35] W. Cai, D. D. Feng, and R. Fulton. Content–based retrieval of dynamic PET functional images. *IEEE Transactions on Information Technology in Biomedicine*, 4(2):152–158, 2000.

[36] F. Schnorrenberg, C. S. Pattichis, C. N. Schizas, and K. Kyriacou. Content–based retrieval of breast cancer biopsy slides. *Technology and Health Care*, 8:291–297, 2000.

[37] A. Mojsilovis and J. Gomes. Semantic based image categorization, browsing and retrieval in medical image databases. In *IEEE International Conference on Image Processing (ICIP' 2000)*, Rochester, NY, USA, September 2000.

[38] S. Antani, L. R. Long, and G. R Thoma. A biomedical information system for combined content–based retrieval of spine x–ray images and associated text information. In *Proceedings of the 3rd Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2002)*, Ahamdabad, India, December 2002.

[39] A. Horsch, M. Prinz, S. Schneider, O. Sipilä, K. Spinnler, J.-P. Vallée, I. Verdonck-de Leeuw, R. Vogl, T. Wittenberg, and G. Zahlmann. Establishing an international reference image database for research and development in medical image processing. *Methods of Information in Medicine*, 2004 - to appear.

[40] E. M. Voorhees and D. Harmann. Overview of the seventh Text REtrieval Conference (TREC–7). In *The Seventh Text Retrieval Conference*, pages 1–23, Gaithersburg, MD, USA, November 1998.

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

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

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

[44] T. Bürkle, E. Ammenwerth, H.-U. Prokosch, and J. Dudeck. Evaluation of clinical information systems. what can be evaluated and what cannot. *Journal of Evaluation in Clinical Practice*, 7(4):373–385, 2001.

[45] A. W. Kushniruk and V. L. Patel. Cognitive and usability engineering methods for the evaluation of clinical information systems. *International Journal of Biomedical Informatics*, 37(1):56–76, 2004.

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

[47] 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.

[48] K. Sparck Jones and C. J. Van Rijsbergen. Progress in documentation. *Journal of Documentation*, 32:59–75, 1976.

[49] J. 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.

[50] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[51] T. Saracevis. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, November/December:321–343, 1975.

[52] L. Schamber, M. B. Eisenberg, and M. S. Nilan. A re–examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, 26 No 6:755–775, 1990.

[53] E. M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[54] H. Müller, A. Rosset, J.-P. Vallée, and A. Geissbuhler. Integrating content–based visual access methods into a medical case database. In *Proceedings of the Medical Informatics Europe Conference (MIE 2003)*, St. Malo, France, May 2003.

[55] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content–based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, September 1998. (Special Issue on Segmentation, Description, and Retrieval of Video Content).

[56] C. W. Cleverdon, L. Mills, and M. Keen. Factors determining the performance of indexing systems. Technical report, ASLIB Cranfield Research Project, Cranfield, 1966.

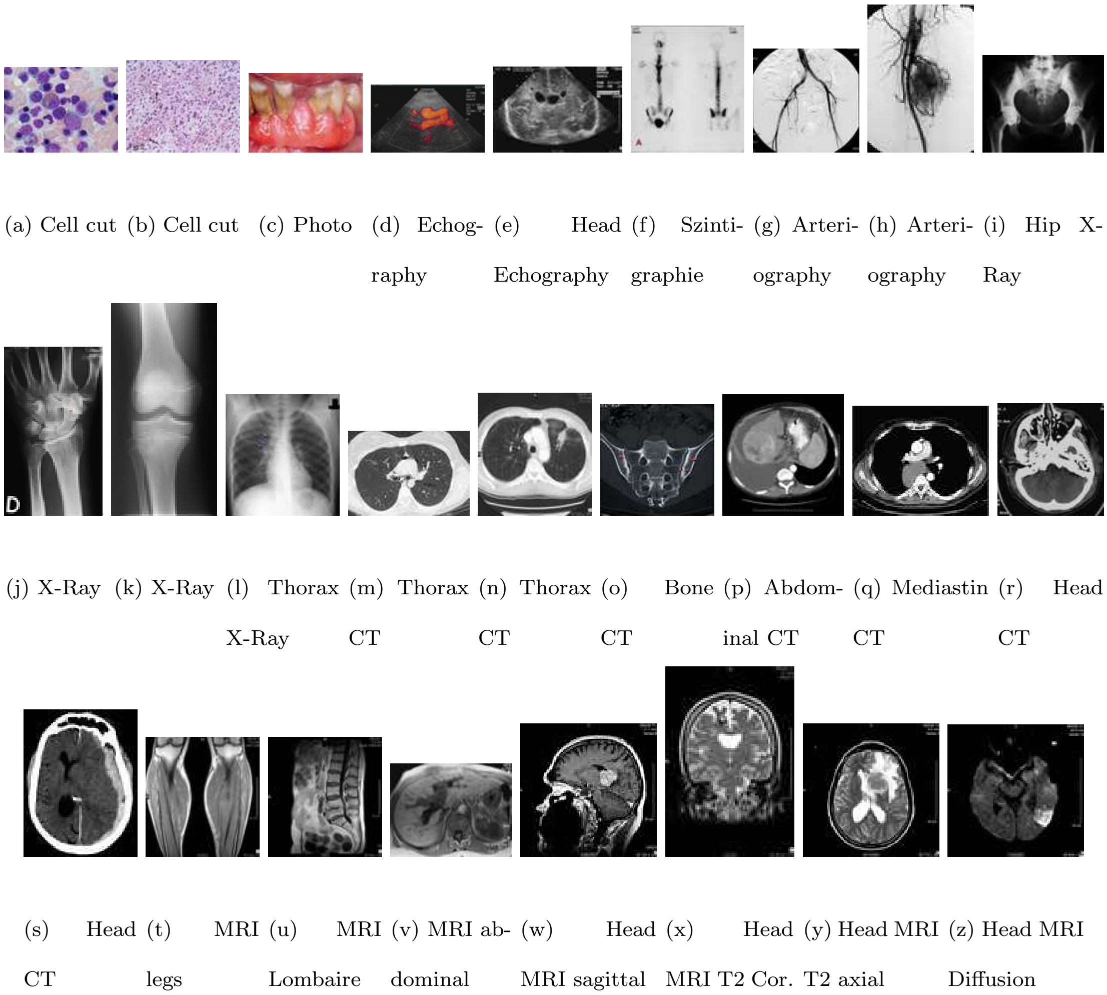
[57] A. Horsch and R. Thurmayr. How to identify and assess tasks and challenges of medical image processing. In *Proceedings of the Medical Informatics Europe Conference (MIE 2003)*, St. Malo, France, May 2003.

Summary

This article describes a methodology for the evaluation of medical image retrieval systems and provides a database that is available free of charge from the authors for evaluation purposes. The databases includes 8751 images of around 2000 cases. For the evaluation, 26 query topics were chosen and ground truth (a gold standard) was generated by domain experts. The methodology for choosing the query topics and for generating the ground truth is from the closely related discipline of text or information retrieval where evaluation and benchmarking events have largely contributed to a strong performance gain. The domain also has more than 40 years of experience in generating document collections and performing standardized performance comparisons.

Content–based image retrieval, not only in the medical domain, currently has the problem that no standard measures and databases are available to compare systems on the same grounds. Thus it is impossible to compare any two techniques and to identify the best–performing ones. Such databases need to be generated and objective performance measures need to be created and accepted to prove a systems performance. Only with a proof of performance will systems get more accepted as a clinical tool and an aid for research and teaching.

The generated database only covers part of the image retrieval applications in that it uses a database with a large variety of medical images. For these systems that use PACS–like databases such a freely available data set with query topics and a gold standard is a valuable tool for system comparison and evaluation. It eases significantly the barrier to evaluate a system properly as there is no costly generation of an image database and the even more costly part of generating a gold standard with the help of an expert.

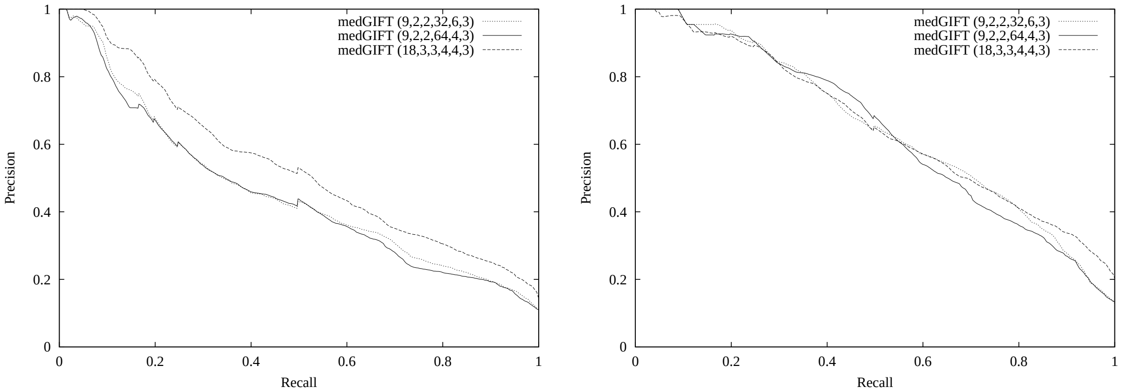


(a) Cell cut (b) Cell cut (c) Photo (d) Echography (e) Head Echography (f) Szintigraphie (g) Arteriography (h) Arteriography (i) Hip X-Ray

(j) X-Ray (k) X-Ray (l) Thorax X-Ray (m) Thorax CT (n) Thorax CT (o) Bone CT (p) Abdominal CT (q) Mediastin CT (r) Head CT

(s) Head CT (t) MRI legs (u) MRI Lombaire (v) MRI abdominal (w) Head MRI sagittal (x) Head MRI T2 Cor. (y) Head MRI T2 axial (z) Head MRI Diffusion

Figure 1: Some example images of the evaluation database.



(a) first query step

(b) with relevance feedback

Figure 2: PR graphs without and with feedback using various color quantizations.

	without feedback			with feedback		
Measure	(9,2,2,32,6,3)	(9,2,2,64,4,3)	(18,3,3,4,4,3)	(9,2,2,32,6,3)	(9,2,2,64,4,3)	(18,3,3,4,4,3)
N_r	85.15	85.15	85.15	85.15	85.15	85.15
t	14.4	14.8	15.3	29.1	25.6	24.6
$Rank_1$	4.42	3.19	8.58	1.03	1	2.42
R(P(0.5))	0.36	0.37	0.45	0.58	0.53	0.50
$\overline{Rank}$	590	820	306	540	683	287
$\widetilde{Rank}$	0.062	0.089	0.030	0.056	0.073	0.029
P(20)	0.55	0.54	0.60	0.69	0.69	0.68
P(50)	0.42	0.43	0.45	0.52	0.52	0.53
P(N_r)	0.42	0.43	0.48	0.58	0.58	0.59
R(100)	0.48	0.47	0.51	0.58	0.60	0.60

Table 1: Performance measures for various color quantizations with and without relevance feedback.

The numbers in parentheses correspond to (hues, saturations, values, gray levels, directions, scales).