Evaluation Axes for Medical Image Retrieval Systems — The ImageCLEF Experience

Henning Müller[1], Paul Clough[2], William Hersh[3], Thomas Deselaers[4] ,
Thomas Lehmann[4], Antoine Geissbuhler[1]
[1]Medical Informatics, University and Hospitals of Geneva, Switzerland
[2]Information Studies, Sheffield University, United Kingdom
[3]Biomedical Informatics, OHSU, Portland Oregon, USA
[4]Computer Science and Medical Informatics, RWTH Aachen, Germany
henning.mueller@sim.hcuge.ch

ABSTRACT

Content–based image retrieval in the medical domain is an extremely hot topic in medical imaging as it promises to help better managing the large amount of medical images being produced. Applications are mainly expected in the field of medical teaching files and for research projects, where performance issues and speed are less critical than in the field of diagnostic aid. Final goal with most impact will be the use as a diagnostic aid in a real–world clinical setting. Other applications of image retrieval and image classification can be the automatic annotation of images with basic concepts or the control of DICOM header information.

ImageCLEF is part of the Cross Language Evaluation Forum (CLEF). Since 2004, a medical image retrieval task has been added. Goal is to create databases of a realistic and useful size and also query topics that are based on real–world needs in the medical domain but still correspond to the limited capabilities of purely visual retrieval at the moment. Goal is to direct the research onto real applications and towards real clinical problems to give researchers who are not directly linked to medical facilities a possibility to work on the interesting problem of medical image retrieval based on real data sets and problems. The missing link between computer science research departments and clinical routine is one of the biggest problems that becomes evident when reading much of the current literature on medical image retrieval. Most databases are extremely small, the treated problems often far from clinical reality, and there is no integration of the prototypes into a hospital infrastructure. Only few retrieval articles specifically mention problems related to the DICOM format (Digital Imaging and Communications in Medicine) and the sheer amount of data that needs to be treated in an image archive (> 30.000 images per day in the Geneva radiology).

This article develops the various axes that can be taken into account for medical image retrieval system evaluation. First, the axes are developed based on current challenges and experiences from ImageCLEF. Then, the resources developed for ImageCLEF are listed and finally, the application of the axes is explained to show the bases of the ImageCLEFmed evaluation campaign. This article will only concentrate on the medical retrieval tasks, the non-medical tasks will only shortly be mentioned.

MM'05, November 6–11, 2005, Singapore.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Performance, Human Factors, Algorithms

Keywords

Benchmarking, evaluation, image retrieval, medical image retrieval, content–based image retrieval

1. INTRODUCTION

Content–based visual information retrieval (CBVIR) or content–based image retrieval (CBIR) is an extremely active domain in the multimedia and computer vision fields [1, 2, 3, 4]. An ever–increasing amount of multimedia data (images, video, music, ...) is produced and made available in digital form. Almost every modern computer user has most of its hard disk filled with multimedia data (images, video clips, mp3 music, ...) but tools to manage these data well are scarce. Most web pages become increasingly mixed–media documents integrating images, animations, texts, etc. The medical field is no exception to this trend. There is an increasing amount and variety of visual data being produced for the diagnostic process and the role of images in the diagnostic process is increasing. Currently, these visual or multimedia data are mainly used for the treatment of a single patient, only. Much of the diagnostic process of medical doctors (MDs) is based on comparing a current case with experience from past cases. To support the memory concerning images, many medical doctors store interesting or typical cases with a textual description and the images on

their hard disk or in a teaching file such as *myPACS*[1] or casimage [2] [5]. Having a larger source of images and descriptions available for all medical doctors can make this stored information and experience available to a larger audience, but the rising number of images requires good tools to not only store the data. Quick search and retrieval tools are needed for these growing databases to find relevant information quickly. Then of course, tools are necessary to anonymise the images as the use of images out of the pure diagnostic or treatment planning process is often not permitted, even within a single institution.

The potential and need for medical image retrieval has been defined fairly early [6, 7, 8, 9, 10]. Still, only very few real applications derived from these first ideas, and most applications remained pure research prototypes that were evaluated on extremely small datasets or even only ideas defining a need. One of the few projects evaluated in a clinical setting is the ASSERT project of Purdue University [11], that shows an improvement in diagnostic quality when using their tool to diagnose lung CTs, especially among less experienced radiologists. Another active project is IRMA[3] [12], where an image retrieval framework was created. Overviews on medical image retrieval, systems and techniques can be found in [13, 14] but new projets develop rapidly. It is also hard to judge the real quality of systems as there are no comparisons of visual retrieval systems based on the same tasks and databases. Most databases are not available and cannot be exchanged between institutions for privacy reasons.

Systematic evaluation of information retrieval systems is a very strong point of the text retrieval domain, where first evaluation databases and performance measures where studied systematically already in the 1960s [15, 16]. With TREC[4], an important evaluation event was created in 1992 that stimulated research groups and managed to show a strong improvement in retrieval quality of extremely large text databases [17]. TREC is really THE standard evaluation event in information retrieval with a yearly circle of resource generation (data sets), topic generation, system results submission, evaluation and a final workshop to discuss results. Many subtasks that started in TREC afterwards created their on evaluation workshops such as CLEF[5] (Cross Language Evaluation Forum) and TRECVID [18] on video retrieval, which both have had a strong success. ImageCLEF started in 2003 as part of CLEF with an image retrieval task based on multilingual text, where the query was in a different language than the image collection, which exists only in English [19]. In 2004, a medical retrieval task was started [20]. This task also required for the first time to use visual features for querying as the query itself was an image only, whereas the collection contained French and English annotation that could be used for query expansion and relevance feedback. Another current image retrieval initiative is ImagEval[6]. Many publications on the evaluation of image retrieval systems exist as well [21, 22, 23, 24].

This article will present the motivation for and the reasoning behind the ImageCLEF 2005 query tasks and the evaluation resources created for the evaluation as well as the axes for evaluation covered in the task and those not covered. These axes are described in more detail to further explain our reasoning behind the task and topic creation for ImageCLEF.

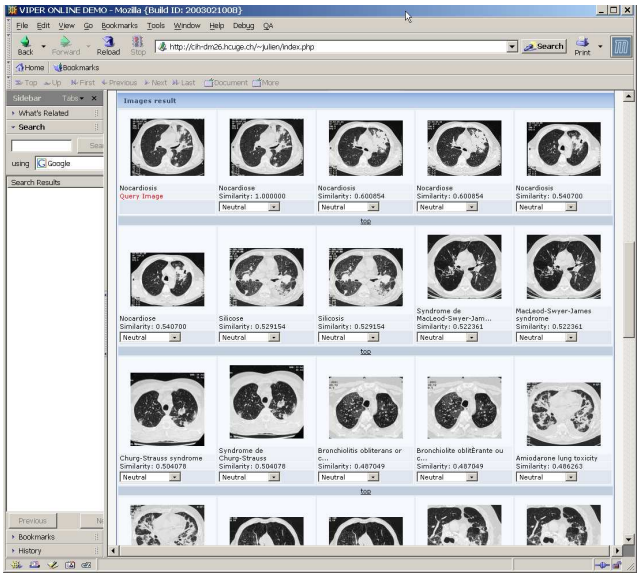


Figure 1: A screen shot of a typical web interface of a medical image retrieval system that allows query by example(s).

2. AXES OF RETRIEVAL EVALUATION

This section explains several of the axes that we regard as important for creating the tasks for ImageCLEF to satisfy various research directions but also to stick to our goal by creating a research environment to prepare medical image retrieval for the use in a real–world setting. Much of the outline and form of the ImageCLEF evaluation is based on the experiences of the TREC workshops and will not be detailed in this article.

2.1 User– vs. system–centered evaluation

User–centered evaluation is evaluating how a user judges the results of an information retrieval system. This includes more than only technical aspects as the user judges what he receives as a result interactively, and a large number of factors together influence the user's judgement on the entire retrieval system. Query speed and ease of use and layout of the interface are extremely important (an example interface for visual queries can be seen in Figure 1). On the other hand, the evaluation can be subjective as several users might judge the same result in a different way. Even the same user might judge the same result differently at different times [25]. User–centered evaluation is also relatively "expensive" as it does include the time of real system users and cannot be automated. Each new setting of parameters requires a new interaction circle with the users.

System–centered evaluation is less costly as it can be automated and does not necessarily require user interaction. Normally, query topics are formulated in advance, and then system developers can tune their system and submit re-

[1] http://www.mypacs.net/
[2] http://www.casimage.com/
[3] http://www.irma-project.org/
[4] http://trec.nist.gov/
[5] http://www.clef-campaign.org/
[6] http://www.imageval.org/

sults that are subsequently evaluated against a ground truth, which is usually created after submission. This means that a large number of system variations can be evaluated with low cost but on the other hand only a part of the system parameters is taken into account, the technical parameters, and important parts such as query speed and the user interface are not analysed at all. Both TREC and CLEF run mainly system–centered tasks.

2.2 Visual vs. textual vs. mixed retrieval

One of the first questions regarding image retrieval is to choose whether a purely textual image retrieval based on available meta data [26] is planned or whether visual data is to be used for the retrieval [1]. Based on the chosen application scenario, only one or the other is really possible. If only very limited meta data is available for retrieval and if many images do not contain any annotation, a keyword search will not be successful but a search with an image example can allow navigation in the database. If good meta data is available text allows to search for semantics and concepts which is usually what a user is looking for. Purely visual retrieval is currently limited to extremely simple concepts and a fairly limited number of concepts as well. On the other hand, visual content and textual context of the images are most often very complementary [27]. Even if the query is only in one media, the other media can be used in a combined visual/textual approach to improve the final results [28].

2.3 Multilingual vs. monolingual retrieval

Most experience in information retrieval is definitely available on monolingual and mostly on English retrieval. Still, in fields such as web search a large number of users exist who might want to use a query language other than English but still retrieve English documents. Most image collections are actually understandable without the text, so searching in a multilingual collection for images is also possible, even if the language can not be understood. In multi–lingual environment such as the European Union or Switzerland, multi–lingual information retrieval is indispensable.

2.4 Classification vs. Information retrieval

An often discussed topic is whether information retrieval is basically the same thing as classification or not. Often, we can see an information retrieval problem as a two–class problem with the class of relevant and the class of non–relevant items maybe with a third class of partially relevant items, and without having any learning data. Still, in most cases, when we think about information retrieval, we have very large collections in mind on which we do not have have much information concerning the content, groups of images or documents, etc. Then, we would like to satisfy the information need of a user and find documents that (s)he is interested in for a particular search. Through the use of frequency–based feature weights some information on the distributions of words or features within the database are extracted in an automated fashion. Judgement of the entire collection for relevance is often impossible due to the large size, so incomplete relevance sets are often based on pooling methods [16].

Most often for classification, information on class membership of the entire collection is known and well defined, which allows the use of machine learning techniques and system optimisations. An example for images belonging to the same classes is Figure 2 taken from ImageCLEF 2005. To evaluate algorithms there are several methods that are commonly used based on the available training data. Leaving–one–out means that algorithm training is done on all images but the image under test, making available a maximum of test data. The process is repeated such that all images serve once as tests, and the mean error rate over all experiments can be determined. Classification error rate can be used as performance measure for these completely annotated databases [29].

Figure 2: Images representing one of the smallest classes in the IRMA task of ImageCLEF 2005.

2.5 Object recognition vs. visual appearance

These two fields are both very active in the domain of computer vision for a variety of application, and both can be very beneficial for image retrieval. Whereas object recognition tries to identify a generally limited number of concepts or objects in an image and label them by techniques such as template matching.

Similarity search by visual appearance in contrast to this takes into account either global features representing the entire image or features representing the layout of an image such as a smaller representation of the image itself. Segmentation can also give access to visual appearance search based on regions [30, 31].

2.6 Real world applications vs. controlled lab conditions

Of course, it is often desirable to have very controlled conditions for testing an algorithm and for the creation of large datasets with little cost. For testing the invariance of visual features for example, artifically created test sets can help well to test these properties [32, 33]. Thus, most systems that create invariances test their algorithms on artificially created test sets. This is justified where the invariance is really needed for the final retrieval system. This can be the case for trademark retrieval as well as for many industrial quality control algorithms. Testing on artificial collections has the disadvantage that not much can be said on the quality of retrieval when applied in a real setting. Only rarely, a user will query with a rotated version of an image in the database. In the medical domain, images are often taken under comparable conditions, so invariances are not always useful because orientation can play an important role. Almost every invariance also results in an information loss.

In the domain of textual information retrieval the system evaluation is rather defined by real world conditions, and often tests such as surveys with real users are performed before tasks are defined on how to evaluate algorithms. This is also the case in the medical domain [34]. The Genomics TREC[7] for example creates realistic queries every year for participants as do most other tracks within TREC.

2.7 Other factors for retrieval evaluation

Of course an evaluation event such as ImageCLEF has to correspond to the needs of the participants, as only the participants can make the event useful. Several propositions for evaluation initiatives did not lead to any system comparison such as the Benchathlon[8] and other initiatives [21, 24, 35]. For ImageCLEF it is thus extremely important to answer concrete needs of the participants and obtain resources to create real–world tasks and evaluations based on these needs. The evolution of ImageCLEF from 4 registrants in 2003, to 18 in 2004 and 36 registrations in 2005 shows that there is an important need in image retrieval evaluation and that the resources made available are appreciated.

The yearly circle of data distribution, task creation, submissions of participants, evaluation and a workshop with discussions follows the successful TREC schedule. This also helps participants to reserve time for participation in a yearly repeating schedule. Pure evaluation as proposed as a service without workshop for discussion by the Council on Library and Information resources (CLIR)[9] has in our opinion only a limited impact as the comparison of results and discussion with participants can lead to much better future outcomes. Evaluation results need to be discussed in a broader forum and also the planning for the following evaluation campaign to react to the participant's comments and needs. Evaluation is not static but a continuously moving process. We are also trying to ease participation by groups that work only on visual or only on textual retrieval by supplying baseline results sets of open source retrieval systems for visual retrieval since 2004 (GNU Image Finding Tool, GIFT[10]) and textual retrieval (Lucene[11]) for 2006.

Of course, there are a large number of other aspects that cannot be discussed in such a short paper, such as the *relevance* model used for the generation of ground truth [36, 37]. TREC uses mainly a relevant vs. non–relevant model. ImageCLEF uses a tertiary scheme with relevant, partially relevant and non–relevant for the ground truthing. For the evaluation, we then use a relevant/non–relevant scheme to ease the calculation of measures such as recall and precision. Still, often several relevance sets are made available and systems can be compared on these varying relevance sets.

Then, there is a question of the performance measures to use. Of course, precision and recall have been criticised frequently, and there is reason for this [38]. Still, they are easy to calculate and easily understandable, so we still stick with them for the time being. We use the mean average precision (MAP) as a lead measure which can also be discussed. For most users it is not relevant whether an image was retrieved at position 500 or 900, so mainly the precision after 20–50 images can be regarded as important for a user, unless high recall is required as in trademark retrieval. In general it is important to calculate a mix of measures and take a closer look at systems based on all the measures. One lead measure is still needed for a final ranking but this ranking should not be taken as too important.

3. RESOURCES MADE AVAILABLE

One of the biggest problems when working on medical image analysis is the access to data. As all images are patient data, we need to be careful with them to respect their privacy and everything used for research needs to be anonymised carefully. The advent of the digital radiology and cheap storage capacities have made the exchange and sharing of images much easier than in the film–based days. Teaching files are created in many medical institutions and quite a few of these are made available publicly. One of the larger initiatives to publish images on the Internet is the MIRC[12] (Medical Image Resource Center) project. In this project, a common access method to teaching files is created based on the XML standard. Software for clients and servers is made available free of charge and cross–platform in the form of a Java program. Currently, more than 15 databases are accessible in this format to be searched by keywords via the MIRC web page. Still, often images are only stored on local hard disks and much knowledge could be extracted from these images if they were available.

One of the databases that is accessible via MIRC is the casimage dataset that contains almost 9.000 images of 2.000 cases and that was used in the ImageCLEFmed 2004 competition [5, 39]. It is also part of the 2005 collection. Images present in the data set include mostly the radiology department, but also photographs, powerpoint slides and illustrations. Cases are mainly in French, with around 20% being in English. We were also allowed to use the PEIR[13] (Pathology Education Instructional Resource) database using annotation from the HEAL[14] project (Health Education Assets Library, mainly pathology images [40]). This dataset contains over 33.000 images with English annotation, with the annotation being in XML per image and not per case as casimage. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology[15] [41], was as well made available to us for ImageCLEF. This dataset contains over 2.000 images mainly from nuclear medicine with annotations per case and in English. Finally, the PathoPic[16] collection (Pathology images [42]) was included into our dataset. It contains 9.000 images with an extensive annotation per image in German. Part of the German annotation is translated into English, but it is still incomplete. This means, that a total of more than 50.000 images was made available with annotations in three different languages. Two collections have case–based annotations whereas two collections have image image–based annotations. Only through the access to the data by the copyright holders, we were able to distribute these images to the participating research groups.

The automatic annotation task was organised by the IRMA group and based on their dataset [43]. This database is an-

[7] http://ir.ohsu.edu/genomics/
[8] http://www.benchathlon.net/
[9] http://www.clir.org/pubs/reports/trant04.html
[10] http://www.gnu.org/software/gift/
[11] http://lucene.apache.org/
[12] http://mirc.rsna.org/
[13] http://peir.path.uab.edu/
[14] http://www.healcentral.com/
[15] http://gamma.wustl.edu/home.html
[16] http://alf3.urz.unibas.ch/pathopic/intro.htm

Show me all x–ray images showing fractures.
Zeige mir Röntgenbilder mit Brüchen.
Montres–moi des radiographies avec des fractures.

Figure 3: A query that requires more than visual retrieval but visual features can deliver some hints to good results as well.

notated according to the four–axes IRMA code. To simplify the task in the first year of existence, a subset of 57 classes was chosen that all have at least 5 images in the class. The database contains a total of 10.000 images. 9.000 images representing the 57 classes were given out with class labels as training data. The remaining 1.000 images were given to participants without a class label for classification. The IRMA code in English and German was also made available to the participants.

4. TASKS FOR IMAGECLEFMED 2005

This article will focus on the two medical tasks but there is also a strong evolution concerning the other tasks of ImageCLEF, the ad–hoc and interactive tasks. More on these can be found in [44].

4.1 Medical retrieval task

The medical retrieval tasks evolved this year from a task with a visual start using an image only, to a multilingual retrieval task making all topics available in English, French and German as can be seen in Figure 3. The scenario is a medical doctor searching information in a collection of teaching files to illustrate a course. The information need is described in text and with one to three images. One query also contains a negative feedback image for visual retrieval to test out the use of negative feedback and whether the use makes sense for more tasks in 2006. Each research group could chose which information made available to take into account, including images and one or several annotations in the different languages.

The queries were created according to several axes planned to be evaluated. First goal was to evaluate visual as well as mixed and rather semantic text–based queries. These three sorts of queries were marked in the topic description containing 11 visual queries (example see Figure 6), 11 mixed queries (see Figure 3) and 3 purely semantic queries. Goal is to evaluate these query types separately to see whether different strategies lead to success for these groups of queries. Other axes taken into account were partly derived from the result of a survey on information needs performed:

Figure 4: Example for the most frequent class of chest x–rays.

- search for imaging modality (CT, MIR, x–ray, gross pathology, micro pathology, ...);
- search for anatomic region (lung, liver, heart, ...);
- search for pathology (emphysema, chronic myelogenous leukemia, ...);
- search for visual observations and findings (such as large blood vessels in the liver, enlarged heart, ...).

A total of 28 groups registered for this task and finally 13 groups submitted results. Many research groups did finally not submit results, but some of them said that they had a lack of resources and most said that the resources made available were still very useful to test their system and that they will participate in 2006. Several groups also stated that the tasks were extremely hard and that the training data from the 2004 task was very different and in consequence not extremely useful. This should be different in 2006 when a similar database and similar query tasks are planned.

4.2 The IRMA task

A completely new task is the automatic image annotation task (IRMA task). A dataset with 9.000 images containing class labels of 57 classes was given out to the participants (see Figure 4 for an example). A new, unlabelled dataset with 1.000 images had to be annotated/classified with the correct labels learned from the training dataset based on visual means, only. 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.

The class labels actually correspond to a simplification of the full 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 for example 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 [45].

A total of 22 groups inscribed for the IRMA task and 12 finally submitted results. Best classification results had an error rate of 12.6% for the classification of 1000 images into 57 classes, which is a very good result.

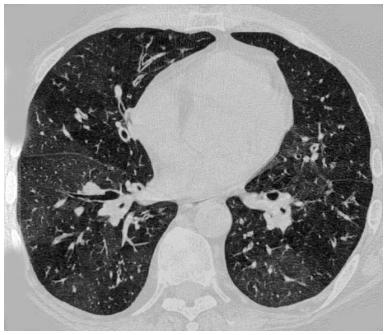
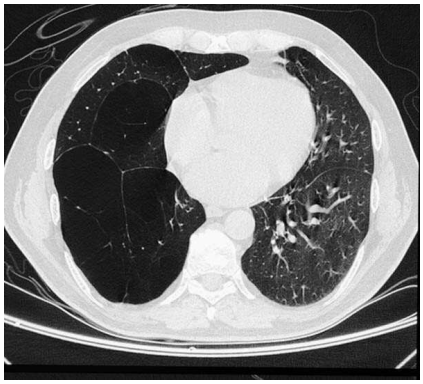


Figure 5: An example query from the 2004 medical task, with the goal to retrieve all images of the same anatomic region, viewing angle and modality. Here, all lung CTs independent of the pathology are expected as result.



Show me chest CT images with emphysema.
Zeige mir Lungen CTs mit einem Emphysem.
Montre–moi des CTs pulmonaires avec un emphysème.

Figure 6: An example of a query that is solvable visually, using image and text as query. Still, the use of the annotation can augment the retrieval quality. The query text is presented in three languages.

4.3 Application of the axes

4.3.1 User vs. system–centered

ImageCLEF has an interactive (user–centered, non–medical) task since 2004, but participation is still fairly low containing 2–5 submissions, mostly due to the high cost of user involvement and the lack of experience in this domain. The task measures how many steps a user needs to find several images by keyword search and using relevance feedback. Still, most of the tasks are clearly system-centered, and all the medical tasks currently are.

4.3.2 Textual vs. visual vs. mixed

ImageCLEF covers all three fields but has a main focus on mixed retrieval as this is a field where still a lot of research is needed and much less experience is currently available. To ease such a combination, visual retrieval results were made available and in the next year it is planned to make also textual retrieval results available to all topics for participants mainly working in one of the two fields. In 2004, the medical task had an image as query, only, as shown in Figure 5, whereas the ad hoc query task was a text accompanied by a single image. In 2005, a purely visual medical image annotation task was added (IRMA task). On the other hand, the medical retrieval task contains one or several images plus text in three languages (English, French, German) and has thus a small visual component. Several topics are expected to be solvable with a visual system such as as the example in Figure 6, whereas other topics are more semantic and text processing appears to be necessary. This focus towards more semantic queries was based on critics in 2004 with the goal to have more realistic topics that are useful in a clinical setting. The 2005 topics are based on a real user survey among medical professionals [46].

4.3.3 Multilingual vs. monolingual

The medical task in 2005 models the scenario of a collection in several languages, currently English, French and German. This is also a fairly common and realistic scenario as medical doctors often annotate their cases in their mother tongue, whereas they might understand enough in another language as well to use the images of a case. Thus, for the medical retrieval task 2005, query topics were made available in the same three languages as the collection, and queries also contain one or several query images (Figure 6).

Techniques for multilingual retrieval include the translation of the queries to a unique language, translations of the documents or the extraction of concepts in multilingual ontologies such as MeSH (Medical Subject Headings) [28].

4.3.4 Classification vs. information retrieval

In the context of ImageCLEF, the classification task is actually called automatic annotation task, which is a very similar problem because the classes actually correspond to a text that can be added to the image collection. The IRMA code [45] to which the classes correspond actually exists in several languages, so such a classification and annotation can further–on be used for multilingual retrieval as well. We distribute a learning set of images and then an evaluation set that the evaluation is performed on, so participants have no idea about class memberships of the images to be categorised but can use the entire training data for system optimisation.

The main retrieval task is a typical information retrieval task with 25 query topics that correspond to an information need of a user from a very large data set. The relevance judgements are done on the first $N = 40$ images of all system submissions so results stay comparable even if relevance is not judged on the entire dataset. As training data, only the topics from 2004 were made available that were not really corresponding to the 2005 topics and underline the character of an information retrieval task.

4.3.5 Object recognition vs. visual appearance

In ImageCLEF 2005, both of these techniques have very useful applications and can well improve retrieval quality. A typical example for an object recognition topic can be seen in Figure 7, where all images showing faces are wanted as a response. For several other queries, object recognition can be useful through very specific detectors but in general the variability of medical images in our database and the variability of query topics is extremely large and constructing one detector per topic is tedious. Thus, for most of the topics, query by visual appearance can deliver overall

Show me photographs of a face.
Zeige mir Fotos eines Gesichtes.
Montre-moi des photos d'un visage

Figure 7: An example for a topic where object recognition would work well and in a limited way visual appearance.

acceptable results in addition and as complement to the textual queries, although query by visual appearance is much less specific. Many of the queries are very hard for object recognition as well as for search by visual appearance, which makes the use of text important to complement the two.

Whereas object recognition can be important if almost no annotation is available to extract semantics, the visual appearance is important where textual information is available. This can for example be used to rank images within a group of semantically related images, such as ranking all images with a text containing the word emphysema based on the similarity with a lung CT.

4.3.6 Lab conditions vs. real–world

We started in late 2004 to survey medical doctors [46] and first results of this survey influenced the way that the tasks were formulated for 2005. This responds also to the main critics of the 2004 task, that some participants regarded as a rather academic problem of limited practical interest. If image retrieval algorithms are foreseen to be applied in real settings, it is extremely important to create resources including databases and topics that are realistic, although this will make the optimisation of algorithms and the testing of particular system parameters harder as a larger number of parameters influences final results. It is important to direct image retrieval research towards such realistic tasks now.

The goal of a realistically–sized database has definitely been reached with the combination of four databases to make available more than 50.000 images to participants. Even if this does not correspond to the size of a PACS (Picture Archival and Communication system) it is bigger than the teaching files of most institutions and it can not easily be overseen, so "cheating" of research groups can be limited.

4.4 Future ideas for ImageCLEF

There are many ideas for possible future tasks. One is the evaluation of an interactive medical task in addition to the other interactive image retrieval tasks. To do so, more groups need to be attracted to such interactive performance evaluations, which currently seems hard to do. The IRMA task can also be made harder by supplying the entire hierarchy of the IRMA code. Then, systems can run the classification up to the level in the code where they feel confident. This would definitely make the task significantly harder and allow image classification algorithms to test the calculation of a confidence for the classification they perform.

Another goal is to add new databases to the campaign creating an even larger and even more realistic approach. With the availability of training data and some experience with the data set, it should be possible to significantly improve retrieval results. Still, the database enlargement is not a priority as the current collection can still give us challenging tasks for at least another year or two.

In the longer run and based on availability, we could also imagine a task that models the scenario of image retrieval as a diagnostic aid in a limited domain such as lung CT retrieval or melanoma classification. This depends particularly on the availability of good datasets and a ground truthing of the datasets.

Most other new ideas for ImageCLEF concern the non–medical tasks, where a new dataset is planned to be used containing holiday pictures of a picture agency in several languages. This models the realistic scenario of multilingual holiday picture retrieval, which will be on the rise with prizes of digital cameras still falling

The realization of much of this also depends on the funding situation of ImageCLEF. Evaluation and the creation of datasets is expensive as is the organisation of a benchmarking event. Without sufficient funding progress seems limited, but it is easier to obtain funds for new research projects than for an evaluation campaign, so the community is asked to take part in the evaluation campaign, make available datasets, realistic topic descriptions, and also help with the ground truthing.

5. CONCLUSIONS

Medical image retrieval has the potential to become a very important factor in clinical medical data management. Still, much research is necessary before these applications can reach a sufficient performance with respect to speed and quality for being accepted in the clinical domain, where time is precious and every decision can have drastic consequences. To advance this challenging field, we need to foster the evaluation of techniques to identify promising approaches and show advances in system performance to convince users. Only standardised evaluation can bring a prove of performance and confidence into such systems. For the creation of tasks and query topics, we need to take into account the information needs of real users and to do so, we need to continue surveys among clinicians to identify important information needs and translate these needs into topics for the tasks. The most expensive part is currently the relevance judgement process and resources need to be found in the research community to support this process and share the charges. The possibility to distribute the datasets generated by medical professionals to the participating research groups is also extremely important. This makes medical datasets available to non–medical research groups and the effort in creating the databases is not only limited to a small number of people of one single research group.

Most comments with respect to ImageCLEF have been extremely positive, especially by the participating research groups. It is important to keep such a workshop event, where algorithms can be compared based on the same data and where data sets can be made accessible to a large number of participants. To prepare multimedia tools for the real–world use, they need to be tested and evaluated based on real–world tasks, and this is the main goal of ImageCLEF.

6. ACKNOWLEDGEMENTS

This work has been funded by the EU FP6 within the Bricks project (IST 507457), the SemanticMining project (IST NoE 507505), and the Swiss National Science Foundation with grant 632-066041. We also acknowledge the support of National Science Foundation (NSF) grant ITR-0325160. The establishment of the IRMA database was funded by the German DFG with grant Le 1108/4.

7. REFERENCES

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

[2] Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad 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).

[3] Serge Belongie, Chad Carson, Hayit Greenspan, and Jitendra Malik. Color– and texture–based image segmentation using EM and its application to content–based image retrieval. In *Proceedings of the International Conference on Computer Vision (ICCV'98)*, pages 675–682, Bombay, India, 1998.

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

[5] Antoine Rosset, Henning Müller, Martina Martins, Natalia Dfouni, Jean-Paul Vallée, and Osman Ratib. Casimage project – a digital teaching files authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.

[6] Henry J. Lowe, Ilya Antipov, William Hersh, and Catherine 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.

[7] Stelios C. Orphanoudakis, Catherine E. Chronaki, and Despina Vamvaka. *I²Cnet*: Content–based similarity search in geographically distributed repositories of medical images. *Computerized Medical Imaging and Graphics*, 20(4):193–207, 1996.

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

[9] Wesley W. Chu, Alfonso F. Cárdenas, and Ricky K. Taira. KMED: A knowledge–based multimedia distributed database system. *Information Systems*, 19(4):33–54, 1994.

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

[11] Alex M. Aisen, Lynn S. Broderick, Helen Winer-Muram, Carla E. Brodley, Avinash C. Kak, Christina Pavlopoulou, Jennifer Dy, Chi-Ren Shyu, and Alan Marchiori. Automated storage and retrieval of thin–section CT images to assist diagnosis: System description and preliminary assessment. *Radiology*, 228:265–270, 2003.

[12] Thomas M. Lehmann, Marc Oliver Güld, Christian Thies, Benedikt Fischer, Klaus Spitzer, Daniel Keysers, Hermann Ney, Michael Kohnen, Henning Schubert, and Berthold B. Wein. Content–based image retrieval in medical applications. *Methods of Information in Medicine*, 43:354–361, 2004.

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

[14] Lilian 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.

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

[16] 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.

[17] Donna 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.

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

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

[20] Paul Clough, Mark Sanderson, and Henning Müller. A proposal for the CLEF cross language image retrieval track (ImageCLEF) 2004. In *The Challenge of Image and Video Retrieval (CIVR 2004)*, Dublin, Ireland, July 2004. Springer LNCS 3115.

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

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

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

[24] Eero Sormunen, Marjo Markkula, and Kalervo 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.

[25] Farzin Mokhtarian, Sadegh Abbasi, and Josef Kittler. Efficient and robust retrieval by shape content through curvature scale space. In A. W. M. Smeulders and R. Jain, editors, *Image Databases and Multi–Media Search*, pages 35–42, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands, August 1996. Amsterdam University Press.

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

[27] Thijs Westerveld. Image retrieval: Content versus context. In *Recherche d'Informations Assistée par Ordinateur (RIAO'2000) Computer–Assisted Information Retrieval*, volume 1, pages 276–284, Paris, France, April12–14 2000.

[28] Henning Müller, Antoine Geissbuhler, and Patrick Ruch. ImageCLEF 2004: Combining image and multi–lingual search for medical image retrieval. In *Cross Language Evaluation Forum (CLEF 2004)*, Springer Lecture Notes in Computer Science (LNCS), Bath, England, 2005.

[29] Thomas Deselaers, Daniel Keysers, and H. Ney. Classification error rate for quantitative evaluation of content–based image retrieval systems. In *Internation Conference on Pattern Recognition (ICPR)*, volume II, pages 505–508, Cambridge, United Kingdom, August 2004.

[30] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. Blobworld: A system for region–based image indexing and retrieval. In Dionysius P. Huijsmans and Arnold 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.

[31] Alexandre Winter and Chahab Nastar. Differential feature distribution maps for image segmentation and region queries in image databases. In *IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'99)*, pages 9–17, Fort Collins, Colorado, USA, June 22 1999.

[32] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, May 1997.

[33] Andrew Zisserman, David A. Forsyth, Joseph L. Mundy, Charlie Rothwell, Jane Liu, and Nic Pillow. 3D object recognition using invariance. *Artificial Intelligence*, 78(1–2):239–288, 1995.

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

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

[36] Tefko Saracevic. 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.

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

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

[39] Henning Müller, Patrick Ruch, and Antoine Geissbuhler. Enriching content–based medical iamge retrieval with automatically extracted mesh terms. In *Jahrestagung der deutschen Gesellschaft für medizinische Informatik (GMDS 2004)*, Innsbruck, Austria, sep 2004.

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

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

[42] K Glatz-Krieger, D. Glatz, M. Gysel, M. Dittler, and M. J. Mihatsch. Webbasierte Lernwerkzeuge für die Pathologie – web–based learning tools for pathology. *Pathologe*, 24:394–399, 2003.

[43] Thomas M. Lehmann, Mark O. Güld, Thomas Deselaers, Henning Schubert, Klaus Spitzer, Hermann Ney, and Berthold B. Wein. Automatic categorization of medical images for content–based retrieval and data mining. *Computerized Medical Imaging and Graphics*, 29:143–155, 2005.

[44] Paul Clough, Henning Müller, and Mark Sanderson. Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul D. Clough, Gareth J. F. Jones, Julio 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.

[45] Thomas M. Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B. Wein. The IRMA code for unique classification of medical images. In *Medical Imaging*, volume 5033 of *SPIE Proceedings*, San Diego, California, USA, February 2003.

[46] William Hersh, Henning Müller, Paul Gorman, and Jeffery Jensen. Task analysis for evaluating image retrieval systems in the ImageCLEF biomedical image retrieval task. In *Slice of Life conference on Multimedia in Medical Education (SOL 2005)*, Portland, OR, USA, June 2005.