Using Heterogeneous Annotation and Visual Information for the Benchmarking of Image Retrieval Systems

Henning Müller[a], Paul Clough[b], William Hersh[c], Thomas Deselaers[d], Thomas M. Lehmann[d], Bruno Janvier[e], Antoine Geissbuhler[a]

[a]University and Hospitals of Geneva, Medical Informatics, Geneva, Switzerland
[b]Sheffield University, UK
[c]Oregon Health and Science University, Portland, OR, USA
[d]Aachen University of Technology (RWTH), Germany
[e]Computer Vision and Multimedia Lab, University of Geneva, Switzerland

ABSTRACT

Many image retrieval systems, and the evaluation methodologies of these systems, make use of either visual or textual information only. Only few combine textual and visual features for retrieval and evaluation. If text is used, it is often relies upon having a standardised and complete annotation schema for the entire collection. This, in combination with high–level semantic queries, makes visual/textual combinations almost useless as the information need can often be solved using just textual features. In reality, many collections do have some form of annotation but this is often heterogeneous and incomplete. Web–based image repositories such as FlickR even allow collective, as well as multilingual annotation of multimedia objects.

This article describes an image retrieval evaluation campaign called ImageCLEF. Unlike previous evaluations, we offer a range of realistic tasks and image collections in which combining text and visual features is likely to obtain the best results. In particular, we offer a medical retrieval task which models exactly the situation of heterogenous annotation by combining four collections with annotations of varying quality, structure, extent and language. Two collections have an annotation per case and not per image, which is normal in the medical domain, making it difficult to relate parts of the accompanying text to corresponding images. This is also typical of image retrieval from the web in which adjacent text does not always describe an image. The ImageCLEF benchmark shows the need for realistic and standardised datasets, search tasks and ground truths for visual information retrieval evaluation.

1. INTRODUCTION

Content–Based Image Retrieval (CBIR) and Visual Information Retrieval (VIR) have been an extremely active area of research over the last 20 years.[1–3] This is mainly due to the need for tools to manage and access the rising amount of digital multimedia data produced, for example, by consumers with cheap digital cameras. Approaches for information retrieval range from using purely visual features[4] to purely textual approaches based on associated text annotations,[5] with combined methods somewhere in between. Many restricted domains have been identified where specialised retrieval can have high impact, such as trademark retrieval[6] and the retrieval of medical images.[7,8] On the other hand, more general image repositories such as the web, should also be considered. Systems such the FlickR* photo management and sharing tool generate specific types of annotation including comments whereby anyone can add text in any language. This results in annotations that are often short and fairly emotional, which can be a problem for automatic retrieval algorithms.[9] As purely visual retrieval itself has not created entirely satisfying results, the trend has gone towards semi–automatic annotation and the linking of visual features with textual keywords.[10]

Corel[11] and the images provided by the University of Washington are commonly used databases for visual information retrieval. Although images in these datasets are accompanied by textual annotations, the annotations

Further author information: (Correspondence to Henning Müller) henning.mueller@sim.hcuge.ch, tel. ++41 22 372 61 75, fax ++41 22 372 8680

*http://www.flickr.com/

Internet Imaging VII, edited by Simone Santini, Raimondo Schettini, Theo Gevers,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6061, 606105, © 2006 SPIE-IS&T · 0277-786X/06/$15

are limited to simple keywords which are often not taken into account for retrieval. Datasets like these can be used to test visual–only approaches in which users are assumed to perform query–by–example (QBE) searches. However, this is not typical of many search tasks in which users formulate their needs using descriptive text, e.g. journalists searching stock photographs or users searching the web. This leads to the evaluation of purely visual information retrieval with unrealistic search topics being used. Major search engines on the web such as Google[†] and Yahoo![‡] serve millions of search requests each day, offering large–scale multimedia search on hyperlinked networks to a global audience. For example, at the time of writing Yahoo! searches over 1.6 billion images. Web image search is based on various types of associated text including: text adjacent to an image, the image caption, textual information extracted from associated HTML (e.g. URLs, ALT tags, anchor text), and information extracted from the link structure of the web. For example, Harmandas et al.[12] showed that associated text from art gallery web sites was well-suited for image retrieval over a range of query types. The majority of web image search is text–based and the success of such approaches often depends on reliably identifying relevant text associated with a particular image. FlickR[§], on the other hand, is a large–scale online photo management tool containing over five million freely accessible images. These are annotated by their authors with freely chosen keywords in a naturally multilingual manner: most authors use keywords in their native language; some combine more than one language. In addition, photos have titles, descriptions, collaborative annotations, and natural language comments. For retrieval, FlickR presents a number of challenges including: different types of associated text (e.g. keywords, titles, comments and description fields), collective classification and annotation using freely selected keywords (known as folksonomies) resulting in non–uniform and subjective categorization of images and annotations in multiple languages.

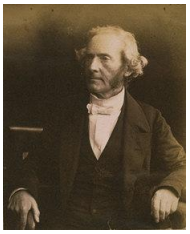
Whereas the text–based IR community started to evaluate systems using standardised methodologies as far back as the 1960s,[13] CBIR was often criticised for its lack of evaluation standards and comparability of systems.[11] For CBIR evaluation, standardised models such as those used in TREC[14,15] – the Text REtrieval Conference – were often proposed as role models. However, most initiatives such as the Benchathlon[¶] and the IAPR benchmarks[‖] never managed to compare actual systems, although many important aspects of benchmarks were discussed. Since 2003, ImageCLEF[**] has created a platform for benchmarking image retrieval applications in the context of CLEF[††] (Cross Language Evaluation Forum) in a style similar to TREC. Databases are distributed to participants annually, followed by search queries (topics) that are based (where possible) on realistic examples gathered using, for example, surveys of real users. After the submission of results by participating groups, the ground truths are generated enabling the calculation and comparison of effectiveness for submitted systems. A post–evaluation workshop is organised where participants can present their results in oral or poster form and compare their techniques with those used by other participants. Many have commented to us that they appreciate access to large datasets including ground truths, as well as the possibility to judge the performance of their system compared to others.

This article describes the databases used and topics generated for ImageCLEF 2005,[16] with experiences from 2004,[17] focusing on heterogeneity of annotation. We focus mainly on the ImageCLEF medical retrieval task because this offers the most heterogeneous and challenging dataset with respect to annotations. Section 4 will explain the ways that participants combined visual and textual cues as well as how they used the extremely heterogeneous annotation. We also present the non–medical ad–hoc retrieval and automatic annotation tasks which are also offered to participants in ImageCLEF and present further types of annotations and tasks.

2. DATABASES AND ANNOTATIONS

ImageCLEF 2005 offered three main search tasks: non–medical ad–hoc retrieval, medical ad–hoc retrieval, and automatic annotation of medical images (we grouped all medical retrieval under the title ImageCLEFmed).

[†] http://images.google.com
[‡] http://search.yahoo.com/search/images
[§] http://www.flickr.com/
[¶] http://www.benchathlon.net
[‖] http://www.cs.cityu.edu.hk/~leung/TC-12/benchmark.htm
[**] http://ir.shef.ac.uk/imageclef/
[††] http://www.clef-campaign.org



Short title: Rev William Swan.
Long title: Rev William Swan.
Location: Fife, Scotland
Description: Seated, 3/4 face studio portrait of a man.
Date: ca.1850
Photographer: Thomas Rodger
Categories: `[ ministers ][ identified male ][ dress -- clerical ]`
Notes: ALB6-85-2 jf/ pcBIOG: Rev William Swan ( ) ADD: Former owners of album:
A Govan then J J? Lowson. Individuals and other subjects indicative of St Andrews provenance.
By T. R. as identified by Karen A. Johnstone " Thomas Rodger 1832-1883. A biography
and catalogue of selected works".

Figure 1. An example image and caption from the St. Andrews collection.

These are predominantly system–centered evaluation tasks where, for example, algorithms are compared in a standardised manner. A further user–centered task[18] was also offered to participants where alternative aspects of the retrieval system are evaluated, e.g. how well users are able to formulate queries or judge the relevance of documents. For the medical ad–hoc retrieval task, the image collection consists of four completely separate databases mainly chosen due to their availability to us in a research context and because all are real–world collections. Although several medical datasets are available on the Internet via MIRC[‡‡] (Medical Image Resource Center), we often did not obtain the right to distribute the images for an evaluation campaign. Finally, we were able to distribute a total of more than 50,000 medical images making it an extremely valuable resource.

2.1. The non–medical ad–hoc task

The St Andrews collection has been used for the past three years at ImageCLEF[17] for an ad–hoc retrieval task: a system is expected to match a user's one–time query against a more or less static collection (i.e. the set of documents to be searched is known prior to retrieval, but the search requests are not). The task is bilingual where queries typical to this kind of historic collection have been generated in English and translated into several languages. The St Andrews collection is typical of many photographic collections found in the cultural heritage domain whereby specialists (e.g. historians or librarians) annotate images with specific attributes such as the name of the photographer, a date and description of the image for archival purposes. All captions in the St Andrews collection follow the same pre–defined semi–structured format. This contrasts with less structured collections such as the web, shared photographic collections such as FlickR and personal photographs.

Search requests have been based on factors such as log file analysis, communication with curators of the St Andrews collection, previous research on retrieval from photographic collections, identification of query dimensions to test the capabilities of a CLIR and CBIR system. Typically, queries based on abstract semantic concepts rather than visual features are a predominant, e.g. "humpback bridge on a country road", "cathedrals in St Andrews" and "children playing by the sea". This limits the effectiveness of using only visual retrieval methods, as either these concepts cannot be extracted using visual features and require extra semantic knowledge (e.g. name of the photographer), or images with different visual properties may be relevant to a search (e.g. "different views of Rome"). Queries in 2005 were aimed to reflect more visual topics, e.g. "woman wearing a white dress".[19]

As a retrieval task, cross–language image retrieval encompasses two main research areas: (1) image retrieval, and (2) cross–language information retrieval (CLIR). The St Andrews collection is particularly challenging for visual–based image retrieval techniques because of the variety in composition and predominately non–colour appearance (see Figure 1). For CLIR, challenges include: captions which are short in length increasing the likelihood of vocabulary mismatch, captions with text not directly associated with the visual content of an image (e.g. expressing something in the background), and the use of colloquial and domain–specific language in the caption (i.e. British English). Participants have used a variety of retrieval techniques based on visual and textual features, but text–based methods have continued to dominant this task. In particular because of the semi–structured format of the captions and the existence of named entities in the collection and queries.[20]

[‡‡] `http://mirc.rsna.org/`

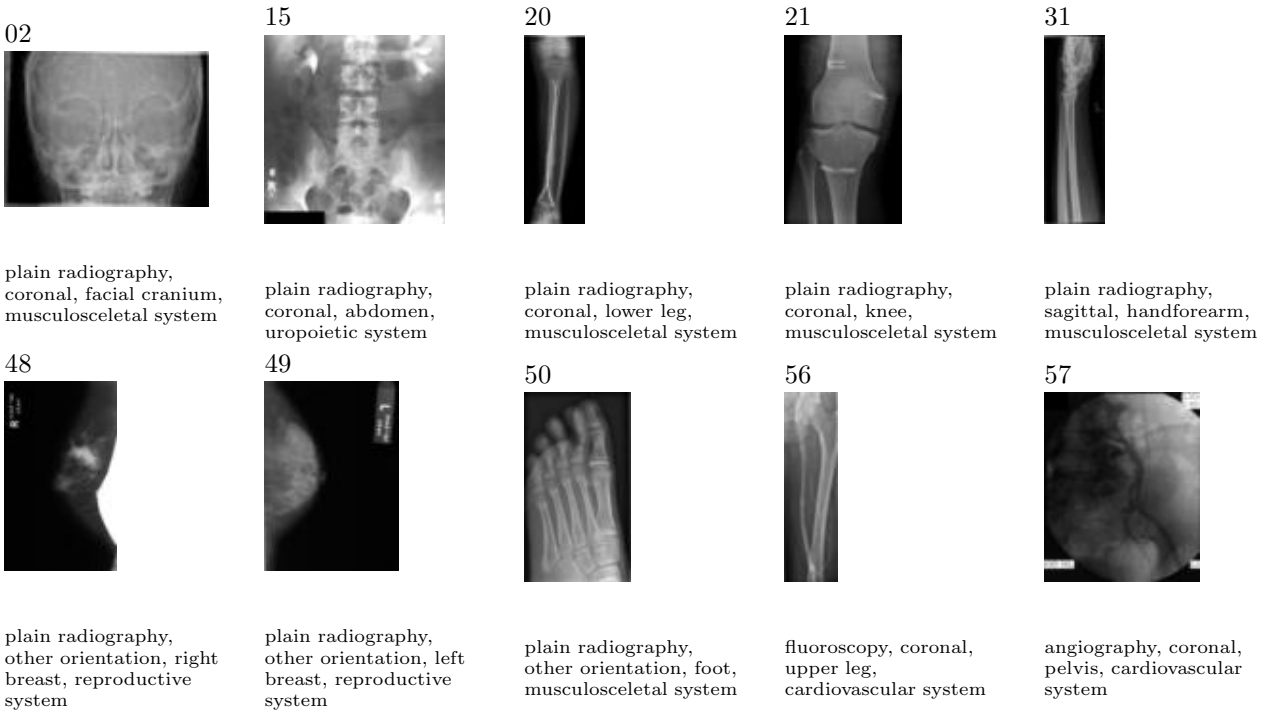


Figure 2. Example images of the IRMA database with classes and textual annotation

2.2. The automatic annotation task

The aim of the automatic annotation task was to compare state–of–the–art methods for automatic annotation, that is classification, of mainly medical radiographs. Automatic annotation and classification can be used in a variety of applications:

- for automatic parameterisation of image analysis and segmentation procedures that are frequently applied to medical image data;
- for consistency checking of DICOM headers, as a portion of DICOM headers contain errors;
- for the generation of query texts for image retrieval queries, i.e. given an image, a textual query is generated to search an annotated database based on visual features and text.

In the 2005 CLEF/ImageCLEF evaluation, the IRMA database* was used for the automatic annotation task. This database consists of 9,000 training images and 1,000 test images. Although only 57 simple class numbers were provided for ImageCLEFmed 2005. The images are annotated with complete IRMA code, a multi-axial code for image annotation. The code is currently available in English and German. It is planned to use the results of such automatic image annotation tasks for further, textual image retrieval tasks in the future. Some example images together with the class numbers and textual annotation are given in Figure 2.

In total 26 groups registered for participation in the automatic annotation task. All groups downloaded the data but only 12 groups submitted runs. Each group had at least two different submissions. In total, 41 runs were submitted. Several of the groups used common CBIR techniques to deduce the classes from the images, but also some object recognition and classification methods from other fields of computer vision were successfully applied. Most of the groups could clearly outperform the common baseline result of a nearest neighbour classifier using Euclidean distance. This baseline classifier achieves an error rate of 36.8%, the best error rates achieved are between 10 and 15% and quite a few results are in the area of 20%.

*`http://www.irma-project.org`

Description: A large hypoechoic mass is seen in the spleen. CDFI reveals it to be hypovascular and distorts the intrasplenic blood vessels. This lesion is consistent with a metastatic lesion. Urinary obstruction is present on the right with pelvo-caliceal and uretreal dilatation secondary to a soft tissue lesion at the junction of the ureter and baldder. This is another secondary lesion of the malignant melanoma. Surprisingly, these lesions are not hypervascular on doppler nor on CT. Metastasis are also visible in the liver.
Diagnosis: Metastasis of spleen and ureter, malignant melanoma
Clinical presentation: Workup in a patient with malignant melanoma. Intravenous pyelography showed no excretion of contrast on the right.
Comment: Splenic metastasis occur most commonly in malignant melanoma, lymphoma, and leukemia but can also occur in carcinoma of the ovary, breast, lung, and stomach. Metastasis are usually hypoechoic and can be hyper or more commonly hypovascular. Liver metastasis are frequent in malignant melanoma. These lesions are often hypervascular and can be extremely hemorrhagic.

Figure 3. An example case from the casimage collection.

It becomes clear that automatic annotation of images can achieve a quality that is sufficient to be used for the creation of textual information that can further–on be used to achieve a superior retrieval performance than using the visual information alone.

2.3. Datasets for the medical retrieval task

2.3.1. Casimage

The casimage[†] dataset contains almost 9,000 images of 2,000 cases and was already used in 2004.[21,22] Images present in the data set include mostly radiographs, but also present are photographs, powerpoint slides and illustrations. Cases are mainly in French, with around 20% being in English.

Figure 3 shows a case with only two images and a limited annotation. The full annotation can contain more fields, some of them administrative in nature and are not shown here (e.g. the physician's name and the date of inclusion). Often, references to articles are copied to the comments section. In the diagnosis section, ACR (America College of Radiology) codes are often used. The shown case is in pure English but many cases are mixed containing English references and French comments or even comments in several languages. Some cases are completely empty and contain only automatically added fields such as dates. In 2004, several groups tried language detection on the database but had an error rate that was only very little above random. Many fields contain spelling errors and unusual abbreviations.

2.3.2. Mallinkrodt Institute of Radiology Nuclear Medicine teaching file

The nuclear medicine database of MIR (Mallinkrodt Institute of Radiology[‡]),[23] was made available to us for ImageCLEF 2005. This dataset contains over 2,000 images of 400 cases mainly from nuclear medicine with annotations per case and in English.

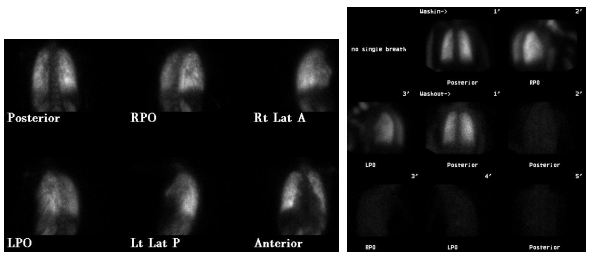
Figure 4 shows an example case from the MIR collection with two images. The database mainly contains nuclear medicine images but also a few other imaging modalities. Annotation is almost unstructured and the XML of ImageCLEF kept it all in one large block as unstructured text, although the cases do have an internal structure. Text quality is high and language is always English.

2.3.3. PathoPic

The PathoPic[§] collection (Pathology images[24]) was used for the first time in 2005. It contains 9,000 images, almost all from pathology, with an extensive annotation per image in German. Only part of the German annotation exists in English.

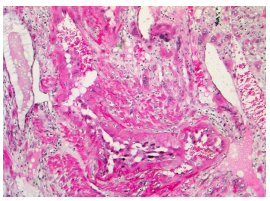
Figure 5 shows an example image from Pathopic with the German and shorter English annotation. Some images contain a much longer annotation but this example shows well the difference between German and English

[†] http://www.casimage.com/
[‡] http://gamma.wustl.edu/home.html
[§] http://alf3.urz.unibas.ch/pathopic/intro.htm



Text: Diagnosis: Patent Ductus Arteriosus with Secondary Eisenmenger's Syndrome
Full history: The patient has hypoxemia and dyspnea on exertion. He has long standing pulmonary hypertension secondary to a large patent ductus arteriosus. Because of his hypoxemia, the patient's hematocrit is increased. He is treated with periodic phlebotomy.
Radiopharmaceutical: 13.1 mCi Xe-133 gas by inhalation and 4.2 mCi Tc-99m MAA i.v.
Findings: The Xe-133 ventilation exam is normal. The perfusion images show no pulmonary defects. However, there is abnormal activity in the kidneys, spleen and bone marrow. These findings are consistent with a anatomic right to left shunt. The amount of activity in the bone marrow is unusual but is presumably related to a hyperactive bone marrow due to the patient's periodic phlebotomies.
Discussion: The increased red cell turn-over due to the patient's periodic phlebotomies has caused bone marrow hyperplasia. The increased blood flow to the hyperplastic bone marrow is demonstrated by the increased MAA activity.
ACR Codes and Keywords:
General ACR code: 51

Figure 4. An example case from the Mallinkrodt (MIR) collection.



Diagnose: Gravidität — (5206)
Synonyme:intrauterine schwangerschaft
Beschreibung: Zytotrophoblastzellen des extravillösen Zytotrophoblasten mit grossen hyperchromatischen Kernen invadieren nicht nur das Myometrium, sondern auch die Spiralarterien der Dezidua. Fetale Zellen sind im Lumen der mütterlichen Spiralarterie nachweisbar.
Klinik:11. Schwangerschaftswoche. Normale Schwangerschaft

Diagnosis: pregnancy — (5206)
Description: Trophoblast cells invading the myometrium and spiral arteries.

Figure 5. An example image with annotations from the Pathopic collection.

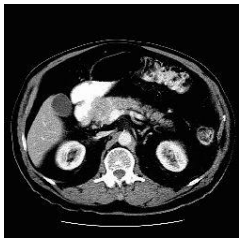
annotation with respect to completeness. It becomes clear that for successful retrieval from this collection at least a partly use of the German annotation seems necessary.

2.3.4. PEIR and HEAL

We also used the PEIR[¶] (Pathology Education Instructional Resource[25]) database using annotation from the HEAL[‖] project (Health Education Assets Library, mainly pathology images[26]). This dataset contains over 33,000 images with English annotation in XML per image. This is the largest subset that was used within ImageCLEF and also the most varied with respect to images included. Although it is called a Pathologic teaching resource, many images concern modalities other than microscopic images or photographs.

Figure 6 shows an example image with part of the annotation. The full HEAL annotation schema is much larger, and includes many more fields than shown here. Most cases in our dataset have only a very small number of fields completed: mainly the title and description fields. The example shown actually has a large amount of text compared to most other annotations.

[¶] http://peir.path.uab.edu/
[‖] http://www.healcentral.com/



Title: PANCREAS.
Source collection: PEIR - University of Alabama at Birmingham Department of Radiology
Description: PANCREATIC NEUROENDOCRINE TUMOR.
73 year old man with a pancreatic mass by outside CT scan. There is a heterogenously enhancing mass within the head of the pancreas measuring approximately 5 cm. The mass abuts the SMV and portal vein but does not appear to encase them. Another heterogeneous mass measuring 4 x 6 cm is seen in the suprapancreatic/gastrohepatic ligament region compatible with nodal metastasis.
The neuroendocrine tumors of the pancreas are derived from the islet cells.
The functioning islet cell tumors include: insulinoma, gastrinoma, glucagonoma, VIPoma, and somatostatinoma. Some islet cell tumors are non-functioning.

Figure 6. An example image from the PEIR collection with the HEAL annotation.

database	size	unit	language	problems
Casimage	8725	case	French, English	spelling errors, language mix
MIR	1177	case	English	very unstructured
PathoPic	7805	image	German, English	English very short
PEIR	32319	image	English	often extremely short text

Table 1. Comparison of the databases of the medical tasks.

2.3.5. Heterogeneity of the annotation

The heterogeneity of the database concerns many more details than can be shown in four short examples (see also Table 1). Still, these examples give a hint of the variability within the dataset. The datasets used are all from the real world and although it is in the restricted domain of medical image retrieval, many of the problems of heterogeneous annotation also occur on the web.

The dataset is *multilingual* with two databases being in English, one mostly in French with a little English, and one in German with a less extensive annotation in English. The *unity of annotation* is twice the case (several images) and twice the image itself. The *structuredness* of annotation is very different from one database to another. Whereas the MIR data set only has one large free text field, the other databases have many more fields with often little content. The *size* of the annotation is also very variable from few keywords to long sentences and descriptions of the situation. Most of the time, the image *content* itself is not described but rather the *context* in which the image was taken.

Other problems are mainly specific to the medical domain, but again much is also similar in other domains. Many *abbreviations* are used in the datasets, often in a non–standard way: particular use to a specific community of users. Several annotations contain a large amount of *spelling errors*. Then of course, much of the *vocabulary* is specific to the medical domain and unlikely to be found within most general–purpose dictionaries or work with usual stemmers. On the other hand, many of the terms are almost language–independent.

3. TOPIC CREATION FOR RETRIEVAL SYSTEM EVALUATION

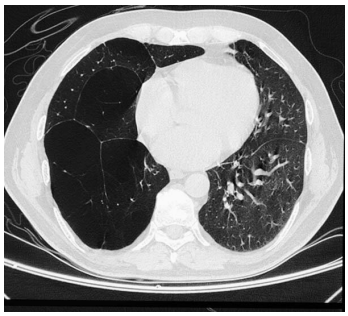
In both the non–medical and medical tasks, the goal has been to create realistic search tasks based on corresponding real–world information needs. This aims to narrow the gap between how a system performs in the evaluation versus how it might perform in practice. In this section we describe the creation of search topics for the medical ad–hoc task. These were based on a survey conducted among medical professionals asking them about their image use and image search behaviour.[27] Based on the responses we developed several concepts and axes for topics.

3.1. Axes of chosen topics

Most information needs according to the survey are along four specific axes:

- modality of the image (x–ray, CT, MRI, gross pathology, photo, ...);
- anatomic region (knee, hand, lung, ...);
- pathology (emphysema, leukemia, ...);
- visual observation or abnormality (enlarged heart, visible vessels in the liver, ...).

Among the pathologies that we choose for the topics, we preferred those in a list of the most common diseases. For the other axes we based our choice on their frequency in clinical practice and the amount of time that particular images were chosen in the survey. Much is also based on a constraint to have as many dimensions of the axes covered as possible.



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

Figure 7. An example of a visual query. The use of the annotation can augment retrieval quality.

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

Figure 8. A query that requires more than visual retrieval but visual features can deliver hints to good results.

3.2. Different groups of topics for evaluation

A clear goal of ImageCLEF is to encourage the use of multimodal retrieval based on combined visual and textual features. Based on this goal we formulated three groups of topics to promote mixed techniques for retrieval:

- *visual topics*: where purely visual algorithms are likely successful;
- *mixed topics*: where visual and semantic information is likely required to solve the information need;
- *semantic topics*: where visual algorithms alone are expected to be unsuccessful.

Despite this categorisation of Still, even the visual topics were harder than in ImageCLEF 2004. It was shown by the submissions that even for these queries textual retrieval can be as good as visual retrieval. For the semantic queries on the other hand the visual retrieval results were not performing well at all. We had 11 visual, 11 mixed and 3 semantic topics in 2005 as the semantic topics were meant to be a test for systems. In 2006, we plan to have an equal number of topics for each of the three classes.

Figure 7 shows an example of a visual query. A query topic that will require more than purely visual features can be seen in Figure 8. We observe that local spots are important in these images, whereas most of the common visual retrieval engines use mainly global features.

4. RESULTS OF THE PARTICIPANTS AND TECHNIQUES USED

In the medical ad–hoc task, participants used an extremely wide variety of techniques for visual and textual retrieval. We provided groups with the possibility of using visual retrieval results (ranked lists) made available by us using the GIFT[**] system and this was used by a number of groups enabling them to concentrate mainly on text-based IR. Combining text and visual features obtained the overall best performance.

Overall, purely visual retrieval had mixed results with the best systems achieving a mean average precision (MAP) of 0.15, but only after manual training using a pre–ordering of the images according to the query topics. The best fully automatic visual system was GIFT with a MAP of 0.09. Main differences between visual retrieval submissions were the type of visual features which ranged from simple smaller representations of the images (also quad trees), to wavelet filters responses and also Gabor filter responses. Other features included color histograms, Tamura texture measures, and features based on co–occurrence matrices. A further main characteristic between visual systems was in the distance measure/weighting function/classification algorithm used. GIFT uses a weighting function based on text IR methods involving the frequencies of features in the collection and the query. Other groups used simple distance measures such as Euclidean distance within a vector space model. As no training data was available for the database, the algorithms for classification and machine learning could not demonstrate their strengths completely. Only very few visual systems used manual relevance feedback.

As this article mainly concentrates on the heterogenous annotation, this part of the participants' submission is extremely important for us. Groups used a variety of methods to pre–process the collection text and query itself to cope with the heterogeneity of the annotation. Several word stemmers for the different languages were used, as well as language–specific stopword lists. Some groups created a single index to include text from all languages and this did not seem to alter results significantly in 2004[22] and 2005.[28] Jensen et al.[28] were the only group to manually optimise the query itself by reformulating the query text, adding keywords and mixing keywords from all three languages. This achieved overall good results (MAP=0.2116). An even better technique seemed to be the use of ontologies adapted for the specific domain of medical IR, such as MeSH (Medical Subject Headings) or UMLS (Unified Medical Language System). The extraction of MeSH terms from text in various languages also partly resolves the multilingual problems that can be encountered as the ontologies exist in several languages. The extraction of MeSH terms was already used in 2004[22] with very good results. Ruiz et al.[29] extracted UMLS concepts (that actually include MeSH) from queries, then expanded these queries into several languages with good results (MAP=0.1746). Best overall results were obtained by.[30] This group created a limited ontology for the task based on several axes of MeSH. Then, the query itself was analysed and parts of the query mapped to the axes of the ontology (modality, body region, ...). The resulting subqueries were executed and negative query expansion was added. A question for a certain modality does, for example, exclude other modalities and similar assumptions are true for the other axes. The best textual–only run had a MAP of 0.2139.

There are several areas of improvement one could envisage in approaches used for retrieval across the heterogeneous datasets of ImageCLEF. Firstly, an improvement in retrieval is likely to result from normalisation with respect to the size of the annotation as this varies widely between datasets. Secondly, the combination of visual and textual features are mainly linear. No approaches perform an analysis of the query itself to select relevant feature weights. Jensen et al.[28] used a simple linear combination of visual and textual results with a slightly reduced MAP as a result. Ruiz et al.[29] used the visual results we supplied for query expansion, then used a linear combination of visual and textual results resulting in an improvement of MAP from 0.1746 to 0.2358. Chevallet et al.[30] also improved results when combining both textual and visual features. The best result, again, is a linear combination of visual and textual values obtained improving the MAP score from 0.2139 to 0.2821 and underlining the complementary nature of the two retrieval techniques. Other combination techniques use either text or images as a main list and then use the other ranked results list to re–rank the first list. Results of this in 2004 were good but in 2005 were not among the best submissions.

In general, the heterogeneity of the annotation and the mix of visual information with textual information for the query tasks does not cause problems for participating groups. Many different techniques and approaches have been used so far in ImageCLEF. In particular, the combination of visual and textual results still offers huge

[**] texttthttp://www.gnu.org/software/gift/

potential. Already the best submitted runs are combining the two, but currently most combinations are of a simple linear nature. Having more complex methods that can judge the influence of visual and textual parts of a query based on the query images and text promise even better results.

5. FUTURE WORK

The clear goal of ImageCLEF is to promote the use of mixed visual/textual retrieval mainly in a multilingual or language–independent context. Making available visual retrieval results of the GIFT system helped many groups without visual retrieval experience or systems. For 2006 it is planned to make also textual retrieval results available using the Lucene retrieval system to give this same possibility of combinations to groups that want to specialise in visual retrieval only.

A major change in 2006 will be the inclusion of an interactive image retrieval task using a collection from FlickR. Until now, the interactive task has been separated from ImageCLEF (despite using the St Andrews collection). However, the use of images from FlickR will allow us to create an extremely interesting interactive task based on truly heterogenous annotations (that will in turn hopefully attract more participants). Using images from within a web environment is definitely realistic and allows many important research questions to be addressed from a quickly developing field. User–centered studies are required within image retrieval, but are often neglected as they require more effort and time from participating groups than a system–centered comparison that can often be run without human intervention. Still, user–centered evaluation cannot be replaced and the influence of the user interaction on the results is in general stronger than the influence of the system itself.

The St Andrews collection will be replaced for the non–medical ad–hoc task in 2006 by the IAPR collection,[31] a set of personal photographs containing currently over 30,000 images annotated in English, German, and Spanish.[32] Many of the pictures are taken from popular holiday resorts and model often typical retrieval from personal collections. Tasks are planned for multilingual retrieval with more semantic–based queries, in addition to a pilot experiment using images for query–by–example searches to motivate participation from the visual retrieval community.

The visual annotation task is planned to increase in complexity in 2006 by augmenting the number of classes. The goal is to allow annotation on the level of the fine–grained IRMA (Image Retrieval in Medical Applications) code. The future goal is to have a two step query process: in the first step a certain number of images must be annotated, and in the second step the annotation can be used as the query to a different (but related) collection.

A pilot study is planned for 2006 on a non–medical automatic annotation task using a database made available by LookThatUp[††]. The training dataset currently contains over 70,000 images of over 300 object classes. As a new task, images must be classified based on whether they contain such objects or not.

6. CONCLUSIONS

The workshops of ImageCLEF 2003–2005 have shown the need for publicly–accessible datasets, topics, and ground truths to compare techniques in the field of visual information retrieval. Participation in ImageCLEF is rising strongly and comments show that realistic tasks are wanted in future evaluations. To really have added value, topics must be more than a means of verifying an existing algorithm. Added value of such a benchmarking event can be showing the advances of systems over the years: a comparison that would not be possible without these datasets and evaluation campaigns.

Many real–world multimedia collections exist (particularly on the web), but also in all kinds of digital libraries. Collections that grow over time usually have extremely varying annotations and in which people add their own text in very different and subjective ways. This heterogeneity needs to be taken into account when creating datasets and tasks for comparing systems to prevent the creation of unrealistic "laboratory" conditions for retrieval evaluation. Mixing several databases with heterogeneous annotation partly fulfils this goal and allows an evaluation that is more realistic to tasks based on web environments and image search. The heterogeneous annotations in ImageCLEF 2005 were not found to cause any major problems to participants, especially in the medical ad–hoc retrieval task where the challenge was the greatest.

[††] `http://www.ltutech.com/`

For 2006, several new tasks and pilot experiments are planned to take into account the comments of existing ImageCLEF participants. The goal is to create real–world tasks on realistic datasets that can be distributed to participants without copyright restrictions. The main potential for improvement has been identified as techniques that combine visual and textual features for image retrieval and a challenge for the research community is to generate suitable benchmarks which promote this.

7. 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 grants 632-066041 and 205321-109304/1. 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.

REFERENCES

1. 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**, pp. 1349–1380, 2000.
2. A. del Bimbo, *Visual Information Retrieval*, Academic Press, 1999.
3. 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**, pp. 644–655, September 1998. (Special Issue on Segmentation, Description, and Retrieval of Video Content).
4. C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region–based image indexing and retrieval," in *Third International Conference on Visual Information Systems (VISUAL'99)*, D. P. Huijsmans and A. W. M. Smeulders, eds., *Lecture Notes in Computer Science*, pp. 509–516, Springer, (Amsterdam, The Netherlands), June 2–4 1999.
5. P. G. B. Enser, "Pictorial information retrieval," *Journal of Documentation* **51**(2), pp. 126–170, 1995.
6. M. E. Graham and J. P. Eakins, "Artisan: A prototype retrieval system for trademark images," *Vine* **107**, pp. 73–80, 1998.
7. 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**, pp. 1–23, 2004.
8. T. M. Lehmann, M. O. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. B. Wein, "Content–based image retrieval in medical applications," *Methods of Information in Medicine* **43**, pp. 354–361, 2004.
9. C. Jörgensen, "Retrieving the unretrievable in electronic imaging systems: emotions, themes and stories," in *Human Vision and Electronic Imaging IV*, B. Rogowitz and T. N. Pappas, eds., *SPIE Proceedings* **3644**, (San Jose, California, USA), January 23–29 1999. (SPIE Photonics West Conference).
10. C.-F. Tsai, K. McGarry, and J. Tait, "Qualitative evaluation of automatic assignment of keywords to images," *Information Processing and Management* , 2005 – to appear.
11. H. Müller, S. Marchand-Maillet, and T. Pun, "The truth about corel – evaluation in image retrieval," in *Proceedings of the International Conference on the Challenge of Image and Video Retrieval (CIVR 2002)*, (London, England), July 2002.
12. V. Harmandas, M. Sanderson, and M. D. Dunlop, "Image retrieval by hypertext links," in *Proceedings of the 20th International Conference on Research and Development in Information Retrieval*, pp. 296–303, September 1997.
13. C. W. Cleverdon, "Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems," tech. rep., Aslib Cranfield Research Project, Cranfield, USA, September 1962.
14. H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content–based image retrieval: Overview and proposals," *Pattern Recognition Letters* **22**, pp. 593–601, April 2001.
15. J. R. Smith, "Image retrieval evaluation," in *IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'98)*, pp. 112–113, (Santa Barbara, CA, USA), June 21 1998.

16. P. Clough, H. Müller, T. Deselaers, M. Grubinger, T. Lehmann, J. Jenser, and W. Hersh, "The CLEF 2005 cross–language image retrieval track," in *Working Notes of the 2005 CLEF Workshop*, (Vienna, Austria), September 2005.
17. P. Clough, H. Müller, and M. Sanderson, "Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004," in *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, C. Peters, P. D. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, eds., *Lecture Notes in Computer Science*, Springer–Verlag, (Bath, England), 2005.
18. J. Gonzalo, P. Clough, and A. Vallin, "Overview of the clef 2005 interactive track," in *Working Notes of the 2005 CLEF Workshop*, (Vienna, Austria), September 2005.
19. M. Grubinger, C. Leung, and P. Clough, "Towards a topic complexity measure for cross language image retrieval," in *Working Notes of the 2005 CLEF Workshop*, (Vienna, Austria), September 2005.
20. V. Painado, López-Ostenero, J. Gonzalo, and F. Verdejo, "Uned at imageclef 2004: Detecting named entities and noun phrases for autoamtic query expansion and structuring," in *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, C. Peters, P. D. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, eds., *Lecture Notes in Computer Science*, pp. 643–652, Springer–Verlag, (Bath, England), 2005.
21. A. Rosset, H. Müller, M. Martins, N. Dfouni, J.-P. Vallée, and O. Ratib, "Casimage project – a digital teaching files authoring environment," *Journal of Thoracic Imaging* **19**(2), pp. 1–6, 2004.
22. H. Müller, P. Ruch, and A. 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.
23. 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), pp. 1520–1527, 1995.
24. 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**, pp. 394–399, 2003.
25. K. N. Jones, R. Kreisle, R. Geiss, J. Holliman, P. Lill, and P. G. Anderson, "Group for research in pathology education online ressources to facilitate pathology instruction," *Archives of Pathology and Laboratory Medicine* **126**, pp. 346–350, 2002.
26. C. S. Candler, S. H. Uijtdehaage, and S. E. Dennis, "Introducing HEAL: The health education assets library," *Academic Medicine* **78**(3), pp. 249–253, 2003.
27. W. Hersh, H. Müller, P. Gorman, and J. 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.
28. J. R. Jensen and W. R. Hersh, "Manual query modification and automatic translation to improve cross–laguage medical image retrieval," in *Working Notes of the 2005 CLEF Workshop*, (Vienna, Austria), September 2005.
29. M. E. Ruiz and S. B. Southwick, "UB at CLEF 2005: Medical image retrieval task," in *Working Notes of the 2005 CLEF Workshop*, (Vienna, Austria), September 2005.
30. J.-P. Chevallet, J.-H. Lim, and S. Radhouani, "Using ontology dimentsions and negative expansion to solve precise queries in clef medical task," in *Working Notes of the 2005 CLEF Workshop*, (Vienna, Austria), September 2005.
31. M. Grubinger and C. Leung, "Incremental benchmark development and administration," in *Proceedings of the Conference on Visual Information Systems (VISUAL 2004)*, (San Francisco, CA, USA), 2004.
32. M. Grubinger, C. Leung, and P. Clough, "The IAPR benchmark for assessing image retrieval performance in cross–language evaluation tasks," in *Proceedings MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, pp. 33–50, (Vienna, Austria), September 2005.