Advancing Biomedical Image Retrieval: Development and Analysis of a Test Collection

William Hersh, MD[1], Henning Müller, PhD[2], Jeffery Jensen, BS[1], Jianji Yang, MS[1], Paul Gorman, MD[1], Patrick Ruch, PhD[2]
[1]Oregon Health & Science University, Portland, OR, USA
[2]University & Hospitals of Geneva, Geneva, Switzerland

Abstract

Objective: Develop and analyze results from an image retrieval test collection.
Methods: After participating research groups obtained and assessed results from their systems in the image retrieval task of Cross-Language Evaluation Forum, we assessed the results for common themes and trends. In addition to overall performance, results were analyzed on the basis of topic categories (those most amenable to visual, textual, or mixed approaches) and run categories (those employing queries entered by automated or manual means as well as those using visual, textual, or mixed indexing and retrieval methods). We also assessed results on the different topics and compared the impact of duplicate relevance judgments.
Results: A total of 13 research groups participated. Analysis was limited to the best run submitted by each group in each run category. The best results were obtained by systems that combined visual and textual methods. There was substantial variation in performance across topics. Systems employing textual methods were more resilient to visually oriented topics than those using visual methods were to textually oriented topics. The primary performance measure of mean average precision (MAP) was not necessarily associated with other measures, including those possibly more pertinent to real users, such as precision at 10 or 30 images.
Conclusions: We developed a test collection amenable to assessing visual and textual methods for image retrieval. Future work must focus on how varying topic and run types affect retrieval performance. Users studies also are necessary to determine the best measures for evaluating the efficacy of image retrieval systems.

Introduction

Image retrieval is a poor stepchild to other forms of information retrieval (IR). Whereas a broad spectrum of Internet users, from lay persons to biomedical professionals, perform text searching routinely {Rice, 2005 #3815}, fewer (though a growing number) search for images on a regular basis. Image retrieval systems generally take two approaches to indexing and retrieval of data. One is to perform indexing and retrieval of the textual annotations associated with images {Rui, 1999 #3016}. A number of commercial systems employ this approach, such as Google Images (images.google.com) and Flickr (www.flickr.com). A second approach, called visual or content-based, is to employ image processing techniques to features in the images, such as color, texture, shape, and segmentation {Müller, 2004 #2610}.

Each approach to indexing and retrieval of images has its limitations. Little research has assessed the optimal approaches or limitations to text-based indexing of images. Greenes has noted one problem particular to biomedicine, which is the "findings-diagnosis continuum" that leads images to be described differently based on the amount of diagnostic inference the interpreter of the images is applying {Greenes, 1992 #457}. One effort to improve the discipline of image indexing has been the Health Education Assets Library (HEAL) project, which aims to standardize the metadata associated with all medical digital objects, but its adoption remains modest at this time {Candler, 2003 #1971}.

Visual indexing and retrieval also have their limitations. In a recent review article of content-based image retrieval applied in biomedicine, Müller et al. noted that image processing algorithms to automatically identify the conceptual content of images have not been able to achieve the performance of IR and extraction systems applied to text {Müller, 2004 #2610}. Visual image indexing systems have only been able to discern primitive elements of images, such as color (intensity and sets of color), texture (coarseness, contrast, directionality, linelikeness, regularity, and roughness), shape (types present), and segmentation (ability to recognize boundaries).

Another problem plaguing all image retrieval research has been the lack of robust test collections and realistic query tasks that allow comparison of system performance {Horsch, 2004 #3022; Müller, 2004 #2610}. A few initiatives exist for certain types of visual information retrieval (e.g., TRECVID for retrieval of video news broadcasts {Smeaton, 2005 #3581}), but none focus on the biomedical domain.

The lack of useful test collections is one of the motivations for the ImageCLEF initiative, which aims to build test collections for image retrieval research. ImageCLEF has a lineage from several of the "challenge evaluations" that have been developed over the years to assess performance of IR systems. The foci within these initiatives is usually driven by the interests of the participating research groups. ImageCLEF arose from the Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org), a challenge evaluation for IR from diverse languages {Braschler, 2004 #2997}, when a group of researchers developed an interest in evaluating retrieval of images annotated in a variety of different languages. Some participants in ImageCLEF expressed an interest in retrieval of biomedical images, which led to the image retrieval task described in this paper. CLEF itself is an outgrowth of the Text Retrieval Conference (TREC, trec.nist.gov), the original forum for evaluation of text retrieval systems. TREC and CLEF, along with their outgrowths, operate on an annual cycle of test collection development and distribution, followed by a conference where results are presented and analyzed.

The goals of TREC and CLEF are to build realistic test collections that simulate real-world retrieval tasks and enable researchers to assess and compare system performance {Sparck-Jones, 1995 #1930}. The goal of test collection construction is to assemble a large collection of *content* (documents, images, etc.) that resemble collections used in the real world. Builders of test collections also seek a sample of realistic *tasks* to serve as *topics* that can be submitted to systems as *queries* to retrieve content. The final component of test collections is *relevance judgments* that determine which content is relevant to each topic. A major challenge for test collections is to develop a set of realistic topics that can be judged for relevance to the retrieved items. Such benchmarks are needed by any researcher or developer in order to evaluate the effectiveness of new tools.

Test collections usually measure how well systems or algorithms retrieve relevant items. The most commonly used evaluation measures are recall and precision. *Recall* is the proportion of relevant documents retrieved from the database whereas *precision* is the proportion of relevant documents retrieved in the search. Often there is a desire to combine recall and precision into a single aggregate measure. Although many approaches have been used for aggregate measures, the most frequently used one in TREC and CLEF has been the mean average precision (MAP) {Buckley, 2005 #3538}. In this measure, which can only be used with ranked output from a search engine, precision is calculated at every point at which a relevant document is obtained. The average precision for a topic is then calculated by averaging the precision at each of these points. MAP is then calculated by taking the mean of the average precision values across all topics in the run. MAP has been found to be a stable measure for combining recall and precision, but suffers from its value arising from being a statistical aggregation and having no real-world meaning {Buckley, 2000 #1020}.

Test collections have been used extensively to evaluate IR systems in biomedicine. A number of test collections have been developed for document retrieval in the clinical domain {Hersh, 1994 #391;Hersh, 2001 #1054}. More recently, focus has shifted to the biomedical research domain in the TREC Genomics Track {Hersh, 2006 #3257}. Test collections are also being used increasingly for image retrieval outside of medicine {Clough, 2005 #3527}. This paper provides an extended analysis of the results reported in the ImageCLEF 2005 overview paper {Clough, 2005 #3527}.

Methods

As noted above, test collections consist of three components: content items that actual users are interested in retrieving, topics that represent examples of their real information needs, and relevance judgments that denote which content is relevant (i.e., should be retrieved) to which topic. For the content of our collection, we set out to develop one of realistic size and scope. We aimed to use collections that already existed and did not intend to modify them (e.g., improve them with better metadata) other than organizing them into a common structure for the experiments. We obtained four collections of images that varied in both subject matter and existing annotation. Consistent with the nature of CLEF, they were annotated in different languages.

Tables 1 and 2 describe the collections used in the 2005 task. The Casimage collection consists of clinical case descriptions with multiple association images of a variety of types, including radiographs, gross images, and microscopic images {Rosset, 2004 #2570}. While most of the case descriptions are in French, some are in English. None are in more than one language. The Mallinckrodt Institute of Radiology (MIR) collection consists of nuclear medicine images, annotated around cases in English {Wallis, 1995 #2998}. The Pathology Education Instructional Resource (PEIR) is a large collection of pathology images (gross and microscopic) that are tagged using the HEAL format in English {Jones, 2002 #3520}. PathoPIC is another pathology collection that has all images annotated in longer German and shorter English versions {Glatz-Krieger, 2003 #2999}.

*** Table 1

*** Table 2

Images and annotations were organized into a single library, which was structured as shown in Figure 1. The entire library consists of multiple *collections*. Each collection is organized into *cases* that represent one or more related *images* and *annotations*. Each case consists of a group of images and an optional annotation. Each image is part of a case and has optional associated annotations, which consist of metadata and/or a textual annotation.

*** Figure 1

We developed 25 topics for the test collection consisting of a textual information needs statement and an index image. The topics were classified based on *topic categories* reflecting whether they were more amenable to retrieval by visual, textual, or mixed algorithms. Eleven topics were visually oriented (1-11), 11 topics were mixed (12-22), and three topics were semantically oriented (23-25). Because the images were variously annotated in English, German, or French, the topics were translated into all three languages. (See Figure 2 for an example of one topic and the Appendix for all the topics.)

*** Figure 2

The experimental process was conducted by providing each group with the collection and topics. They then carried out *runs*, consisting of the same retrieval approach applied to all 25 topics. Groups were allowed to submit as many runs as possible, but were required to classify them based on whether the run used manual modification of topics (automatic vs. manual) and whether the system used visual retrieval, text retrieval, or both (visual vs. textual vs. mixed). The two categories of topic modification and three categories of retrieval system type led to six possible *run categories* to which a run could belong (automatic-visual, automatic-textual, automated-mixed, manual-visual, manual-textual, and manual-mixed).

For systems using textual techniques, runs were designated as using manual modification if the topics were processed in any way by humans before being entered as queries into systems. Otherwise the processing of topics was deemed to be automatic, and could consist of such techniques, for example, as (automatically) mapping text into controlled terminologies, expanding words with synonyms, or translating words into different languages. Systems could use either the translations provided in the topic statements or translate across languages using their own approaches. Any manual translation of topics would require the run to be categorized as manual.

The final component of the test collection was the relevance judgments. As with most challenge evaluations, the collection was too large to judge every image for each topic. So as is commonly done in IR research, we developed “pools” of images for each topic consisting of the top-ranking images in the runs submitted by participants {Buckley, 2005 #3538}. There were 13 research groups who took part in the task and submitted a total of 134 official runs. To create the pool for each topic, the top 40 images from each submitted run were combined, with duplicates omitted.

This resulted in pools with an average size of 892 images (range 470-1167). For the 25 topics, a total of 21,795 images were in the pools for relevance judgments.

The relevance assessments were performed by physicians who were also graduate students in OHSU biomedical informatics program. A simple interface was used from previous ImageCLEF relevance assessments. Nine judges, all medical doctors except for one image processing specialist with medical knowledge, performed the relevance judgments. All of the images for a given topic were assessed by a single judge. The number of topics assessed by each judge varied depending on how much time they had available, but varied from four to eight topics. Some judges also performed duplicate assessment of other topics. Half of the images for 20 of the 25 topics were judged in duplicate, 9,279 in all.

Once the relevance judgments were done, we could then calculate the results of the experimental runs submitted by ImageCLEF participants. We used the trec_eval evaluation package (available from trec.nist.gov), which takes the output from runs (a ranked list of retrieved items for each topic) and a list of relevance judgments for each run (called qrels) to calculate a variety of relevance-based measures on a per-topic basis that are then averaged over all the topics in a run. The trec_eval package includes MAP (our primary evaluation measure), binary preference (B-Pref) {Buckley, 2004 #3529}, precision at the number of relevant topics (R-Prec), and precision at various levels of output from 5 to 1000 images (e.g., precision at 5 images, 10 images, etc. up to at 1000 images). We also released the judgments so participants could perform additional runs and determine their results.

Although 134 runs were submitted for official scoring, many of these runs consisted of minor variations on the same technique, e.g., substitution of one term-weighting algorithm with another. We therefore limited our analysis of results to the best-performing run in a given run category from each group, for a total of 27 runs. Although this reduced our overall statistical power, it prevented groups that submitted multiple runs representing minor changes to algorithms from being over-represented in the statistical analysis.

Because our analysis was not hypothesis-driven, we limited our statistical analysis to an overall repeated measures analysis of variance (ANOVA) of MAP for the 27 runs as well as calculation of inter-rater relevance judgment agreement using the kappa statistic. Statistical analyses were performed using SPSS, version 12.0. Posthoc pairwise comparisons for the repeated measures ANOVA were done using the Sidak adjustment. For inter-rater agreement, the kappa statistic was calculated in two ways: with three categories (relevant, partially relevant, and not relevant) and with two categories (using the official category of relevance based on images judged as fully relevant).

Results

Run analysis

A total of 13 research groups submitted 134 runs. Table 3 lists the research groups, the number of runs submitted, and their general approaches. It also contains citations to each group's individual paper for more details. Table 4 shows the 27 best runs in each run category submitted by each group. Figure 3 shows the MAP for all 27 analyzed runs with 95% confidence intervals. The ANOVA analysis of MAP on the reduced set of 27 runs indicated that at least some runs were significantly different from others (p<0.001). Posthoc pair-wise comparison of MAP showed that significant difference from the top run IPALI2R_Tian started from I2Rfus.txt, about one-third down the rank. Figure 4 shows the rest of the performance measures for each run.

It can be seen that the best results came from the automatic-mixed run category. However, it can also be seen that some performance statistics do not follow the same trend as MAP. For example, the OHSUmanvis run outperforms all but the top few runs in precision at 10 and 30 images. Conversely, the SinaiEn_okapi_nofb_Topics run took a dip with those measures relative to others with comparable MAP.

*** Table 3

*** Table 4

*** Figure 3

*** Figure 4

Topic analysis

Our next analysis looked at differences by topic. Table 5 shows the results for each topic as well as averages for all topics and by topic categories. We again only used the best runs from each group for each run category to calculate these values in order to keep those completing larger numbers of runs within a run category from biasing the average. As seen in Table 5, a large diversity of results were obtained from the different topics. We do note that selecting which runs to use for this analysis could impact the results and, as such, note that this analysis should be used mainly to note the differences among the topics rather than the performance of systems on any particular one.

*** Table 5

Figure 5 plots the number of relevant images and MAP per topic on the same graph, showing a modest association between these measures. Figure 6 shows the best run in each run category plotted versus the various topic categories of visual, mixed, and semantic. It can be seen that visual retrieval techniques performed poorly compared to semantic queries, bringing down their overall performance.

*** Figure 5

*** Figure 6

Impact of variable relevance judgments

We also assessed the impact of variation in relevance judgments. Table 6 shows the overlap of judgments between the original and duplicate judges. Judges were more often in agreement at the ends (not relevant, relevant) than the middle (partially relevant) of the scale. For the 9279 duplicate judgments using three categories, the kappa score was 0.679 ($p<0.001$). The kappa statistic for strict relevance was 0.74, indicating "good" agreement.

*** Table 6

We also looked at how different relevance judgments impacted MAP. In addition to the official "strict" relevance, we also assessed "lenient" relevance, where partially relevant images were also considered relevant. We also combined the 9,279 duplicate judgments with the official ones using AND (both judgments had to be relevant for the image to be considered relevant) and OR (only one judgment had to be relevant for the image to be considered relevant) with both strict and lenient relevance. As shown in Figure 7, different judgments led to modest absolute changes in MAP but performance relative to other runs was largely unchanged.

*** Figure 7

Discussion

The ImageCLEF 2005 biomedical task developed a large test collection and attracted research groups who brought a diverse set of approaches to a common goal of efficacious image retrieval. Not only did these groups learn from their own experiments, but other researchers will subsequently be able to improve image retrieval by using the test collection that will now be available.

A variety of conclusions can be drawn from the experiments performed in ImageCLEF 2005. First, it was clear for most research groups that systems mixing visual and textual approaches performed better than those using either approach alone. In addition, our experiments also showed that systems employing textual approaches are more resilient to difficult visually oriented topics than visual systems are to difficult textually oriented topics. In other words, based on these results, image retrieval systems using that use visual techniques should also incorporate text retrieval capabilities for maximum performance.

A final conclusion was that MAP may not be the best measure for the image retrieval task. MAP measures the full range of retrieval results for a topic from low to high recall. In the image retrieval task, however, users may be more precision-oriented than recall-oriented. In other words, users may only want a small to moderate number of relevant images, and not every last relevant one. This is in distinction to, say, someone carrying out a systematic review who needs to retrieve every last relevant document in a text retrieval system. The problem with MAP versus other measures is exemplified in OHSUmanvis run. This run achieves very high precision at 10 and 30 images but much lower MAP than other runs with comparable precision at these levels. As such, this run may be desirable from the user's standpoint, even though the MAP is lower. Clearly further research is necessary to identify which measures are most important to the image retrieval tasks of real users.

This work has a number of limitations. First, like all test collections, the topics were artificial and may not be realistic or representative of how real users would employ an image retrieval system. Likewise, the annotation of the images may not be representative of how image annotation is done generally or represent best practice. And as with all test collections, the pools generated for relevance assessment only represent images retrieved by the techniques of the participating research groups. As such, there could have been other retrieval techniques that would retrieve other images that may be relevant.

We have a number of future plans, starting with ImageCLEF 2006. Because of the diversity of images and annotations, we plan to keep the same image collection and library structure for ImageCLEF 2006. We will, however, develop new topics. We plan to develop equal numbers of textual, visual, and mixed topics so we can better explore the differences among topic categories. Later on, we will enlarge the collection itself.

Additional future plans include carrying out user experiments on two fronts: one to see how users interact and perform with real systems using this collection and also to better elicit user information needs to develop even more realistic topics. With these experiments, we will also aim to assess performance measures to determine which are more representative for real tasks. This will be done by assessing which measures are best associated with the information needs of real users in specific searching situations.

We have created a large image retrieval test collection that will enable future research in this area of growing importance to biomedicine. We have also identified some observations that warrant further study to optimize the performance of such systems. The growing prevalence of images used for a variety of biomedical tasks makes imperative the development of better image retrieval systems and an analysis of how they are used by real users. The ImageCLEF test collections, with both system-oriented and user-oriented research around them, will contribute to further advances in this active research area.

Acknowledgements

This work was supported by a supplement to National Science Foundation (NSF) grant ITR-0325160. We also acknowledge the European Commission IST projects program in facilitating this work (through the SemanticMining Network of Excellence, grant 507505) and the Swiss National Funds (grant 205321-109304/1). Instructions for obtaining the data described in this paper can be obtained from the ImageCLEFmed Web site (http://ir.ohsu.edu/image/).

References

Table 1 - Collection origin and types for ImageCLEFmed 2005 library.

Collection Name	Image Type(s)	Annotation Type(s)	Original URL
Casimage {Rosset, 2004 #2570}	Radiology and pathology	Clinical case descriptions	http://www.casimage.com/
Mallinckrodt Institute of Radiology (MIR) {Wallis, 1995 #2998}	Nuclear medicine	Clinical case descriptions	http://gamma.wustl.edu/ home.html
Pathology Education Instructional Resource (PEIR) {Jones, 2002 #3520}	Pathology and radiology	Metadata records from HEAL database	http://peir.path.uab.edu/, http://www.healcentral.org/
PathoPIC {Glatz-Krieger, 2003 #2999}	Pathology	Image description - long in German, short in English	http://alf3.urz.unibas.ch/ pathopic/e/intro.htm

Table 2 - Items and sizes of collections in ImageCLEFmed 2005 library.

Collection Name	Cases	Images	Annotations	Annotations by Language	File Size (tar archive)
Casimage	2076	8725	2076	French - 1899 English - 177	1.28 GB
MIR	407	1177	407	English - 407	63.2 MB
PEIR	32319	32319	32319	English - 32319	2.50 GB
PathoPIC	7805	7805	15610	German - 7805 English 7805	879 MB

Table 3 - Research groups, runs submitted, general approaches, citation.

Institution	Group Code	Country	Runs	Brief description of runs submitted
CEA {Besancon, 2005 #3578}	CEA	France	5	All submitted runs were automatic with two visual and three mixed runs. Techniques used include the PIRIA visual retrieval system with texture, color and shape features and a simple word frequency-based text retrieval system.
U.Concordia - Computer Science {Rahman, 2005 #3572}	CINDI	Canada	1	One visual run containing a query only for the first image of every topic using visual features. The technique applied was an association model between low-level visual features and high-level semantic concepts mainly relying on texture, edge, and shape features.
U. and U. Hospitals Geneva {Müller, 2005 #3569}	GE	Switzerland	19	All submitted runs were automatic, including two textual, two visual, and 15 mixed runs. Retrieval relied mainly on the GIFT (visual) and easyIR (textual) retrieval systems. Gabor filters were used as texture descriptors and a multiscale color representation as layout features.
Inst. Infocomm Research	I2R	Singapore	7	All submitted runs were automatic and visual. First, visually similar images were selected manually to train the features. Then, a two-step approach for visual retrieval was used.
Institute for Infocomm Research {Xiong, 2005 #3571}	i2r	Singapore	3	All runs are visual with one automatic and two manual submissions. Main technique applied was the connection of medical terms and concepts to general visual appearance.
IPAL-CNRS (Institute for Infocomm Research) {Chevallet, 2005 #3570}	IPAL	Singapore	6	A total of 6 runs was submitted, all automatic with two being text only and the other a combination of textual and visual features. For textual retrieval, the text is mapped onto axes of the MeSH ontology (Pathology, Anatomy). Negatively weighted query expansion was used (remove unimportant anatomic regions and diseases from the results). Then, visual and textual results were combined for optimal results.
Daedalus &	MIRA	Spain	14	All runs submitted were automatic, with 4

Madrid U. {Martínez-Fernández, 2005 #3576}				visual and 10 mixed runs. As textual technique semantic word expansions with EuroWordNet were applied.
National Chiao-Tung U. {Cheng, 2005 #3575}	NCTU	Taiwan	16	All submitted runs were automatic, with 6 visual and 10 mixed runs. The system uses simple visual features (color histogram, coherence matrix, layout features) as well as text retrieval using a vector-space model with word expansion using Wordnet.
Oregon Health & Science U. Medical Informatics {Jensen, 2005 #3528}	OHSU	USA	3	Two manual and one automatic runs were submitted. One of the manual runs combined the output from a visual run using the GIFT with text. For text retrieval, the Lucene system was used.
RWTH Aaachen - Computer Science {Deselaers, 2005 #3567}	RWTHCS	Germany	10	Two visual runs with several visual features (downscaled image, Tamura texture) and classification methods of the IRMA project were submitted.
RWTH Aachen - Medical Informatics {Güld, 2005 #3568}	RWTHMI	Germany	2	Submitted runs include two manual mixed retrieval, two automatic textual retrieval, three automatic visual retrieval and three automatic mixed retrieval runs. The Fire image retrieval system was used with varied visual features (downscaled image, Gabor filters, Tamura textures) and a text search engine using English and mixed-language retrieval.
U. of Jaen - Intelligent Systems {Martin-Valdivia, 2005 #3574}	Sinai	Spain	42	All runs were automatic, with 6 textual and 36 mixed run. GIFT was used as a visual query system and the LEMUR system for text retrieval in a variety of configurations to achieve multilingual retrieval.
U. Buffalo SUNY - Informatics {Ruiz, 2005 #3573}	UB	USA	6	One visual and five mixed runs were submitted. GIFT was used as a visual retrieval system and SMART for text retrieval, with mapping of text to UMLS Metathesaurus terms.

Table 4 - Best runs from each group in each run category sorted by mean average precision (MAP). Also show are results from other evaluation measures, including R-Prec, binary preference (B-Pref), and precision at 10, 30, and 100 images (P10, P30, and P100 respectively).

Run identifier	Group	MAP	R-Prec	B-Pref	P10	P30	P100
Automated-Mixed							
IPALI2R_TIan	IPAL	0.2821	0.311	0.3848	0.616	0.5293	0.3152
nctu_visual+Text_auto_4	NCTU	0.2389	0.2829	0.3026	0.528	0.456	0.3116
UBimed_en-fr.TI.1	UB	0.2358	0.3055	0.3055	0.552	0.4507	0.2884
mirarf5.2fil.qtop	MIRA	0.1173	0.1692	0.1729	0.348	0.2773	0.1968
SinaiEn_kl_fb_ImgText2	Sinai	0.1033	0.1565	0.1745	0.28	0.2213	0.16
GE_M_10.txt	GE	0.0981	0.1499	0.1541	0.284	0.2133	0.1564
i6-3010210111.clef	RWTHCS	0.0667	0.1037	0.1108	0.216	0.1453	0.1212
ceamdItlTft	CEA	0.0538	0.0901	0.1033	0.248	0.1893	0.1052
Automated-Textual							
IPALI2R_Tn	IPAL	0.2084	0.2519	0.3288	0.448	0.376	0.2472
i6-En.clef	RWTHCS	0.2065	0.246	0.3115	0.4	0.3813	0.2288
UBimed_en-fr.T.Bl2	UB	0.1746	0.2117	0.2975	0.364	0.304	0.2276
SinaiEn_okapi_nofb_Topics	Sinai	0.091	0.1534	0.2238	0.14	0.16	0.128
OHSUauto.txt	OHSU	0.0366	0.0692	0.0746	0.132	0.116	0.0756
GE_M_TXT.txt	GE	0.0226	0.0536	0.0549	0.06	0.032	0.0524
Automated-Visual							
I2Rfus.txt	I2R	0.1455	0.2081	0.2183	0.36	0.3467	0.2368
mirabase.qtop	MIRA	0.0942	0.1343	0.146	0.304	0.22	0.1608
GE_M_4g.txt	GE	0.0941	0.1343	0.1461	0.304	0.22	0.1608
rwth_mi_all4.trec	RWTHMI	0.0751	0.1026	0.1335	0.288	0.2187	0.1248
i2r-vk-sim.txt	i2r	0.0721	0.115	0.1353	0.276	0.224	0.138
i6-vo-1010111.clef	RWTHCS	0.0713	0.1155	0.1162	0.26	0.192	0.1268
nctu_visual_auto_a8	NCTU	0.0672	0.1051	0.1185	0.28	0.2053	0.138
ceamdItl	CEA	0.0465	0.0825	0.0977	0.24	0.1627	0.0976
cindiSubmission.txt	CINDI	0.0072	0.0136	0.0855	0.008	0.0173	0.0124
Manual-Mixed							
OHSUmanvis.txt	OHSU	0.1574	0.2045	0.2066	0.488	0.4093	0.2204
i6-vistex-rfb1.clef	RWTHCS	0.0855	0.124	0.1349	0.332	0.2107	0.1392
Manual-Text							
OHSUmanual.txt	OHSU	0.214	0.2917	0.3372	0.464	0.3933	0.2596
Manual-Visual							
i2r-vk-avg.txt	i2r	0.0921	0.1472	0.1713	0.276	0.244	0.1612

Table 5 - Retrieval results for each topic (averaged across all runs) as well as topic categories (visual, mixed, and textual). (See Table 4 legend for definitions of result categories.)

Topic	Retrieved	Relevant	Relevant Retrieved	MAP	R-Prec	B-Pref	P10	P30	P100
1	976.3	201	84.9	0.1565	0.2053	0.3456	0.3333	0.3593	0.2748
2	950.7	160	58.8	0.0779	0.1496	0.1570	0.3185	0.2679	0.1693
3	991.4	232	72.7	0.0998	0.1517	0.2455	0.4519	0.3556	0.2389
4	954.1	165	70.3	0.1306	0.2061	0.2657	0.4852	0.3753	0.2478
5	939.4	155	98.1	0.3025	0.3802	0.4389	0.6444	0.5778	0.4526
6	968.1	301	80.4	0.0927	0.1481	0.2283	0.5370	0.4222	0.2633
7	864.0	37	10.9	0.1272	0.1562	0.1474	0.3000	0.1765	0.0744
8	854.8	32	10.1	0.1113	0.1481	0.1405	0.2667	0.1568	0.0593
9	907.9	148	41.1	0.1336	0.1797	0.2061	0.3519	0.3037	0.2115
10	985.2	69	44.4	0.2742	0.3306	0.3157	0.6037	0.4543	0.2815
11	971.9	90	23.0	0.1075	0.1461	0.1398	0.4185	0.3296	0.1363
12	984.2	24	15.6	0.0619	0.0756	0.0565	0.1074	0.0654	0.0485
13	985.7	411	175.4	0.1588	0.2525	0.3584	0.5000	0.4531	0.3711
14	963.7	138	33.2	0.0468	0.0902	0.1126	0.2778	0.1778	0.1074
15	916.7	103	34.8	0.1073	0.1546	0.1725	0.2481	0.2000	0.1563
16	942.0	8	1.6	0.0394	0.0509	0.0475	0.0407	0.0197	0.0059
17	943.2	21	11.6	0.0477	0.0653	0.0423	0.0667	0.0728	0.0522
18	934.9	28	15.7	0.0867	0.1124	0.0897	0.1333	0.1086	0.0633
19	855.4	48	14.4	0.1280	0.1674	0.1580	0.3148	0.2148	0.1004
20	925.6	26	9.8	0.0315	0.0755	0.0469	0.0593	0.0741	0.0441
21	967.9	295	107.4	0.1067	0.1871	0.2650	0.3185	0.3321	0.2581
22	966.9	81	23.0	0.0748	0.1203	0.1213	0.3630	0.2136	0.1070
23	919.4	144	43.3	0.1508	0.1672	0.2407	0.2704	0.2605	0.2026
24	972.6	3	1.5	0.0110	0.0000	0.0000	0.0037	0.0074	0.0059
25	925.1	124	60.1	0.2588	0.2915	0.3309	0.4519	0.4247	0.3181
Average	942.7	121.8	45.7	0.1170	0.1605	0.1869	0.3147	0.2561	0.1700
Visual	942.2	144.5	54.1	0.1467	0.2001	0.2391	0.4283	0.3435	0.2191
Mixed	944.2	107.5	40.2	0.0809	0.1229	0.1337	0.2209	0.1756	0.1195
Textual	939.1	90.3	35.0	0.1402	0.1529	0.1905	0.2420	0.2309	0.1756

Table 6 - Overlap of relevance judgments.

Duplicate Original	Relevant	Partially relevant	Not relevant	Total
Relevant	1022	94	102	1218
Partially relevant	157	83	153	393
Not relevant	236	199	7233	7668
Total	1415	376	7488	9279

Figure Legends

Figure 1 - Structure of test collection library.

Figure 2 - Example of visually (left) and semantically (right) oriented topics from the test collection.

Figure 3 - MAP for each run, sorted from highest to lowest, with 95% confidence intervals.

Figure 4 - All results from Table 4, sorted by MAP.

Figure 5 - Number of relevant images vs. MAP for the 25 topics based on results from each group's best run in each run category.

Figure 6 - MAP for the best performing run in each run category (denoted to the right of the graph) for each topic category. These results demonstrate that textual systems were more resilient for visual topics than visual systems were for textual topics.

Figure 7 - The impact of varying relevance judgments. The values of MAP are shown for each run with different sets of relevance judgments from the official Strict method to those using more lenient and/or incorporating duplicates judgments into the analysis, as described in the text.

Figure 1 - Structure of test collection library.

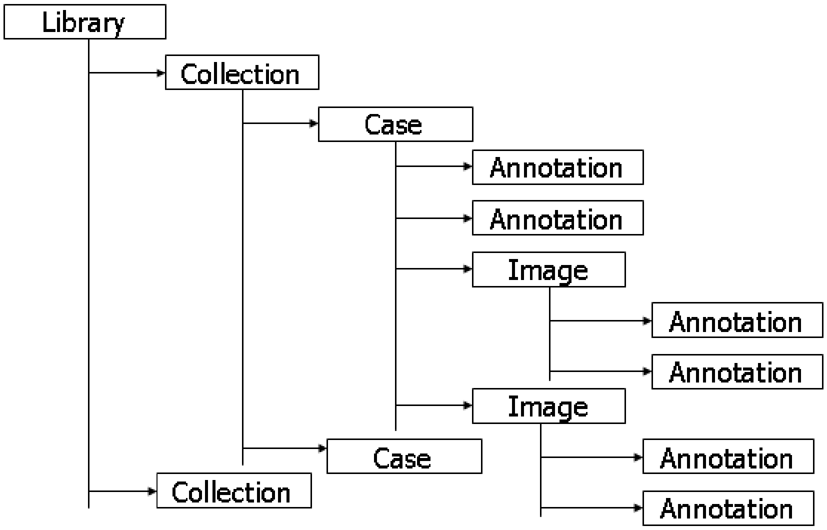
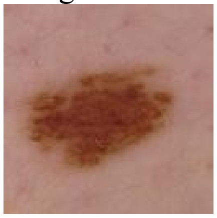


Figure 2 - Example of visually (left) and semantically (right) oriented topics from the test collection.

Show me photographs of benign or malignant skin lesions.
Zeige mir Fotos von gutartigen oder bösartigen Melanomen.
Montre-moi des images de lésions de la peau bénignes ou malignes.



Show me images of right middle lobe pneumonia.
Zeige mir Bilder einer Lungenentzündung des rechten mittleren Lungenlappens.
Montre-moi des images d’une pneumonie du lobe médial droit.

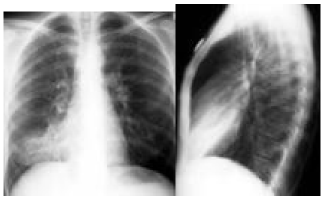


Figure 3 - MAP for each run, sorted from highest to lowest, with 95% confidence intervals.

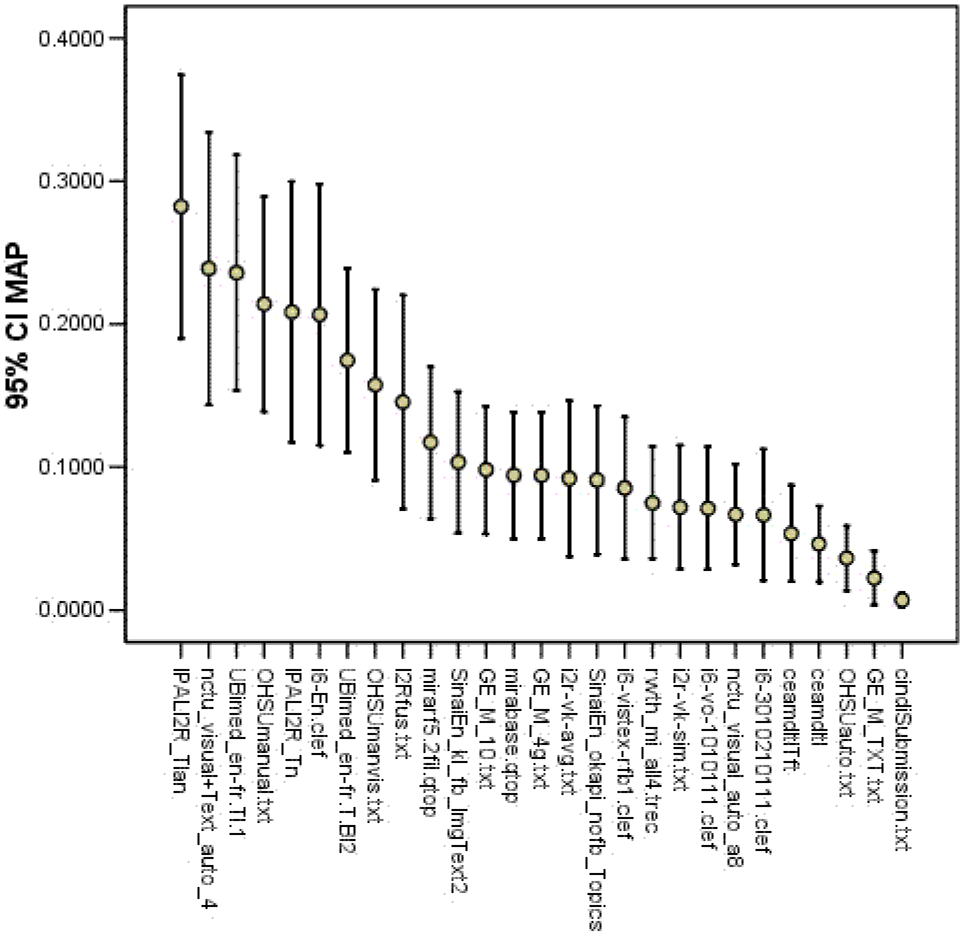


Figure 4 - All results from Table 4, sorted by MAP.

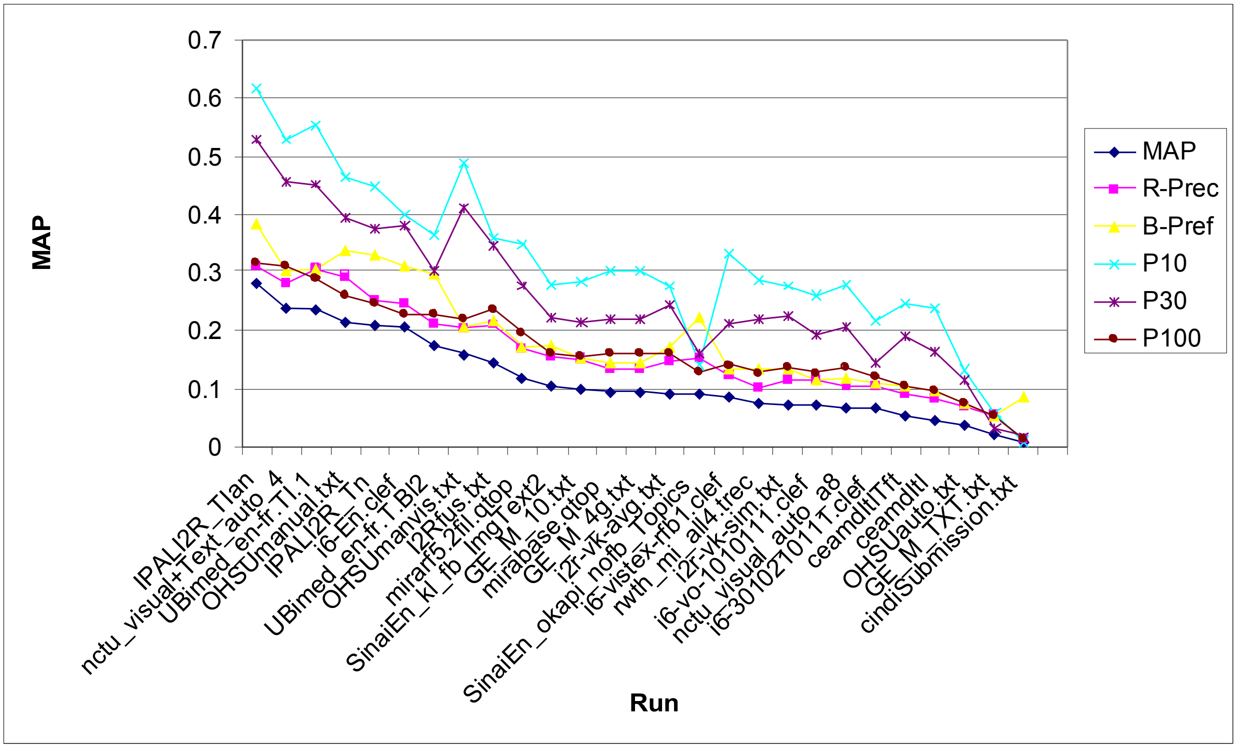


Figure 5 - Number of relevant images vs. MAP for the 25 topics based on results from each group's best run in each run category.

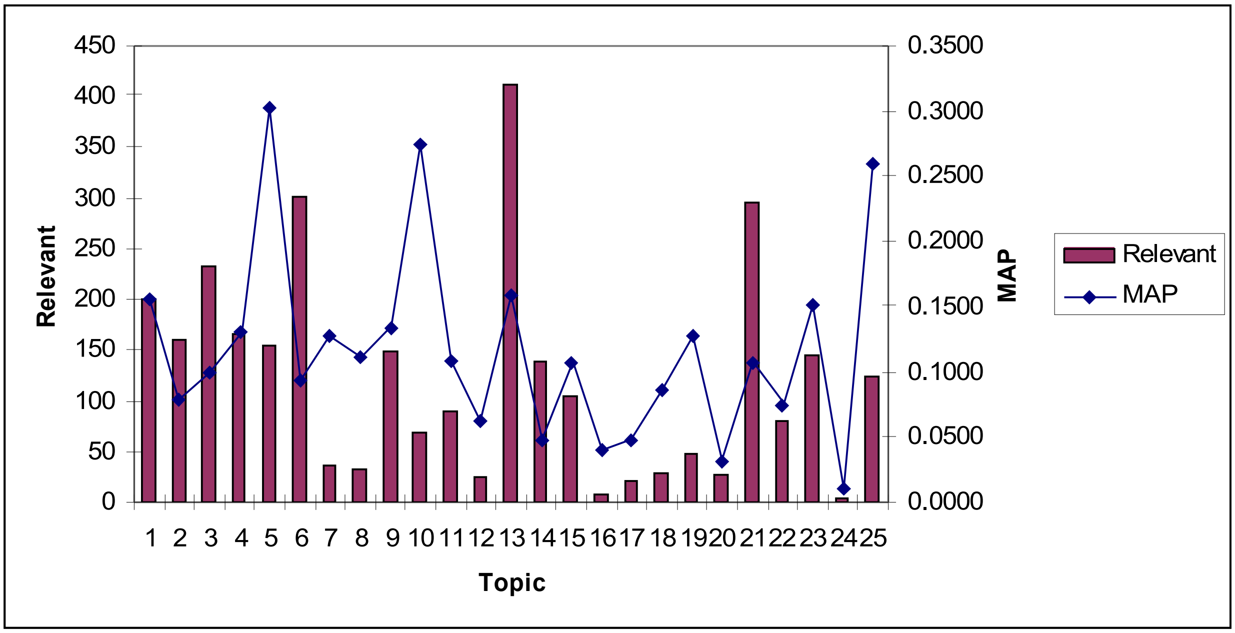


Figure 6 - MAP for the best performing run in each run category (denoted to the right of the graph) for each topic category. These results demonstrate that textual systems were more resilient for visual topics than visual systems were for textual topics.

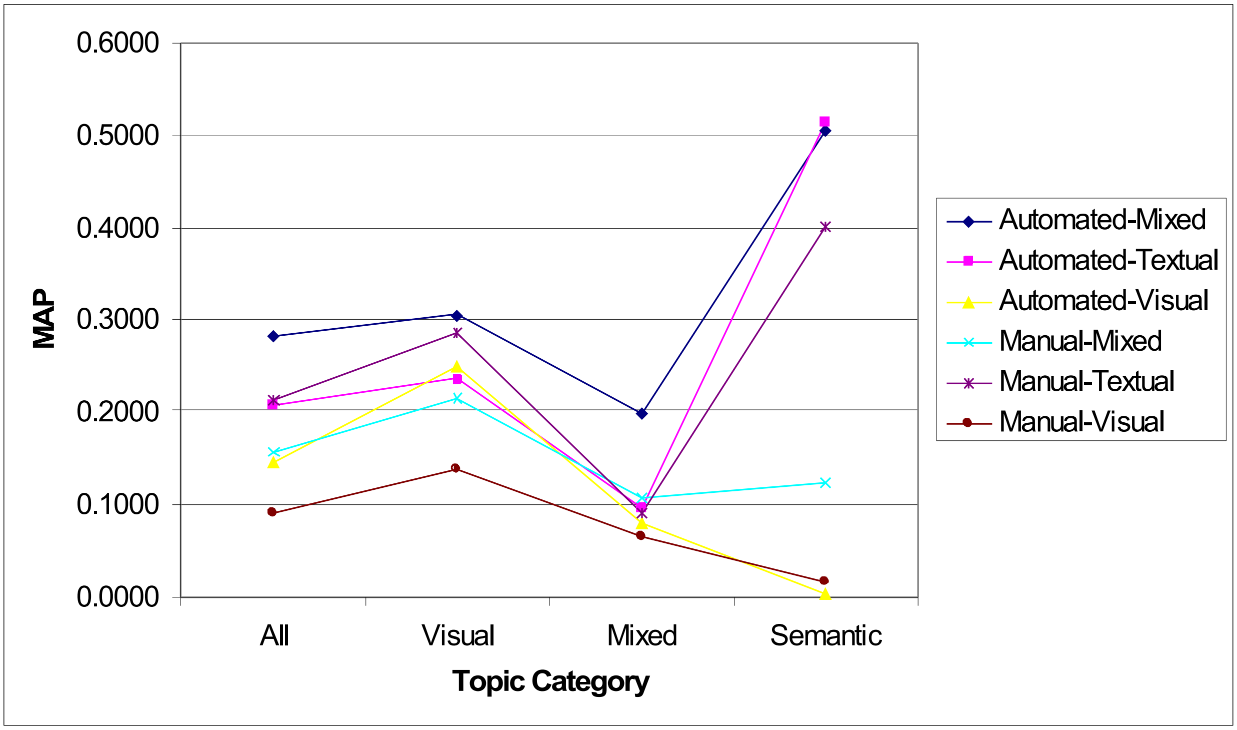
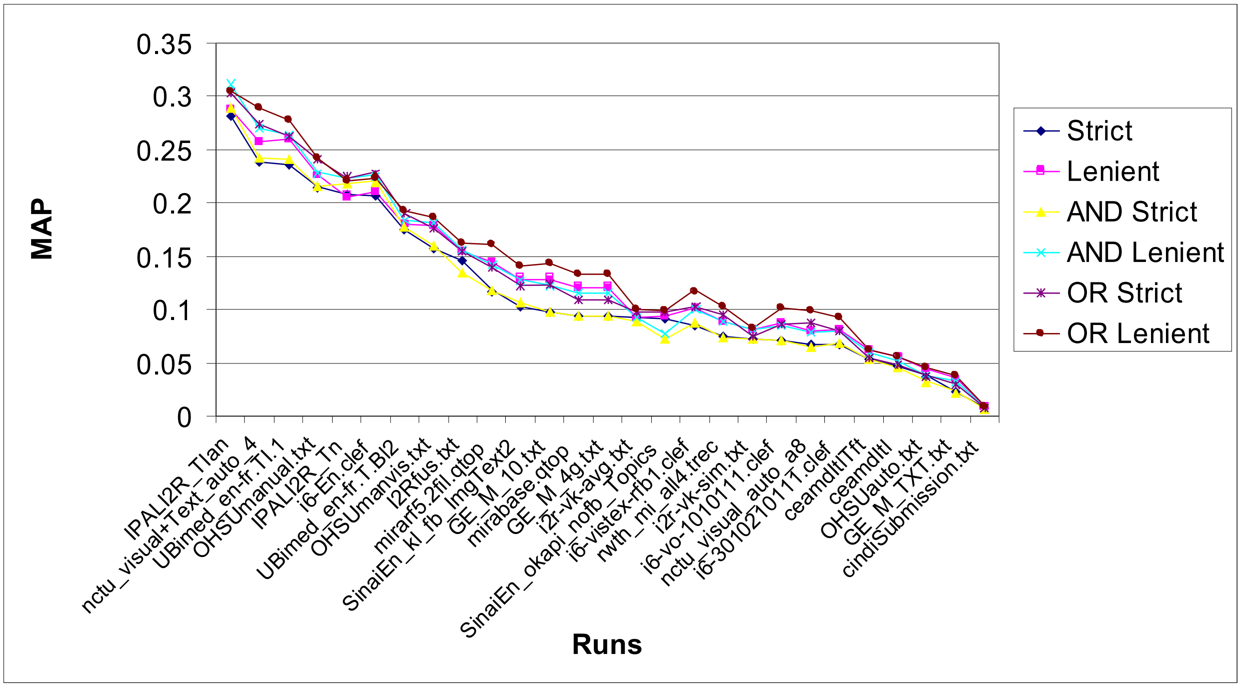


Figure 7 - The impact of varying relevance judgments. The values of MAP are shown for each run with different sets of relevance judgments from the official Strict method to those using more lenient and/or incorporating duplicates judgments into the analysis, as described in the text.


MAP
Strict
Lenient
AND Strict
AND Lenient
OR Strict
OR Lenient
IPALI2R_TIan
nctu_visual+Text_auto_4
UBimed_en-fr.TI.1
OHSUmanual.txt
IPALI2R_Tn
i6-En.clef
UBimed_en-fr.T.BI2
OHSUmanvis.txt
I2Rfus.txt
mirarf5.2fil.qtop
SinaiEn_kl_fb_ImgText2
GE_M_10.txt
mirabase.qtop
GE_M_4g.txt
i2r-vk-avg.txt
SinaiEn_okapi_nofb_Topics
i6-vistex-rfb1.clef
rwth_mi_all4.trec
i2r-vk-sim.txt
i6-vo-1010111.clef
nctu_visual_auto_a8
i6-3010210111.clef
ceamdItITft
ceamdItI
OHSUauto.txt
GE_M_TXT.txt
cindiSubmission.txt
Runs