Project no. FP6-507752

MUSCLE

Network of Excellence
Multimedia Understanding through Semantics, Computation and Learning

Proceedings of the third MUSCLE/ImageCLEF workshop on Image and Video Retrieval Evaluation

Due date of deliverable: 31.10.2007
Actual submission date: 31.10.2007

Start date of project: 1 March 2004 Duration: 48 months

Deliverable Type: R
Number: DN 2.1
Nature: Report
Task: WP 2

Name of responsible:
Allan Hanbury (hanbury@prip.tuwien.ac.at)

Revision 1.0

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Keyword List: Image retrieval evaluation, video retrieval evaluation

Proceedings of the Third

Workshop on Image and Video Retrieval Evaluation

September 18, 2007

Budapest, Hungary

Editors

Allan Hanbury, Vienna University of Technology

Henning Müller, University & Hospitals of Geneva

Paul Clough, University of Sheffield

Proceedings of the Third
Workshop on Image and Video Retrieval Evaluation

Budapest, Hungary
18 September 2007

Co-Chairmen
Allan Hanbury, Vienna University of Technology
Henning Müller, University & Hospitals of Geneva
Paul Clough, University of Sheffield

Supported by the MUSCLE Network of Excellence under the Sixth Framework Programme of the European Commission

CLEF is an activity of the DELOS Network of Excellence for Digital Libraries under the Sixth Framework Programme of the European Commission.

DELOS
Network of
Excellence on
Digital
Libraries

Preface

These are the proceedings of the Third Workshop on Image and Video Retrieval Evaluation, held on the 18th of September 2007 in Budapest, Hungary. The workshop was organised as a joint venture by the ImageCLEF campaign (part of the DELOS EU Network of Excellence) and the MUSCLE EU Network of Excellence.

The workshop consisted of three sessions, entitled:

- Evaluation Campaigns
- Evaluation Techniques
- Data, Topics and Ground Truth

The papers in these proceedings were presented by long-standing participants in the ImageCLEF evaluation campaign as well as by members of the MUSCLE consortium. We were honoured to have as keynote speakers at the workshop:

- Marcel Worring (University of Amsterdam), Chair of the IAPR Technical Committee 12 on Multimedia and Visual Information Systems.
- Thijs Westerveld (CWI), Co-organiser of the XML Multimedia Track at INEX (INitiative for the Evaluation of XML Retrieval).

We are grateful for financial support from the MUSCLE EU Network of Excellence, and to Dmitry Chetverikov of the Computer and Automation Research Institute of the Hungarian Academy of Sciences (MTA SZTAKI) for the use of their conference facilities.

Allan Hanbury
Henning Müller
Paul Clough

September 2007

Contents

Evaluation Campaigns

Evaluation Techniques

Data and Ground Truth Collection

Multimedia retrieval at INEX:
Finding images in structured context

Thijs Westerveld
CWI, Centrum voor Wiskunde en Informatica
INS1, Data Mining and Knowledge Discovery
Kruislaan 413
PO Box 94079
1090 GB Amsterdam
The Netherlands
e-mail: thijs@cwi.nl

Abstract

Images often do not appear in isolation. They are surrounded by text on webpages or come with tags and links to other images on photo sharing sites like flickr. Searching for media in such richly structured context is the object of study in the multimedia track at INEX. INEX, the initiative for the evaluation of XML retrieval, has studied the retrieval from structured document collections since 2002. It established an infrastructure and provided means, in the form of large test collections and appropriate scoring methods, for evaluating how effective content-oriented XML search systems are. The aim of the INEX multimedia track is to study the use of structural information to extract, relate and combine the relevance of different multimedia fragments. In this talk, I will present the test collections and evaluation procedures used for the INEX multimedia track and highlight some of our findings so far.

The MUSCLE Live Image Retrieval Evaluation Event [*]

Allan Hanbury, Branislav Mičušík, Julian Stöttinger
Pattern Recognition and Image Processing Group,
Institute of Computer Aided Automation,
Vienna University of Technology,
Favoritenstraße 9/1832,
A-1040 Vienna, Austria
e-mail: hanbury@prip.tuwien.ac.at

Abstract

This paper presents an overview of the live image retrieval evaluation event held at the CIVR 2007. It describes the organisation of the event and the dataset used, lists the queries used, and discusses the results.

1 Introduction

Image retrieval evaluation campaigns include ImageCLEF [1] and ImagEVAL [2]. However, these evaluation campaigns all run off-line. This means that the participants receive the queries usually a few weeks before the submission deadline. This gives them time to experiment with various system settings and does not impose a limit on the amount of time that the system has to process a query.

A more challenging task, which has received very little attention, is that of live evaluation. For such an evaluation, the participants set up their image retrieval systems in a single location and receive the queries as they should be entered into the system. This task introduces new challenges, including:

- the real-time performance of the system is important
- the user interface should be well designed to simplify interaction.

This paper describes a pilot event which was organised at the CIVR 2007 (Conference on Image and Video Retrieval) in Amsterdam. The focus was on retrieving images from a dataset of images largely annotated with text. Two types of query were used: visual queries and text queries. The aim was not to find the overall best system, rather to demonstrate what is possible with retrieval systems available today.

[*] This work was supported by the European Union MUSCLE Network of Excellence (FP6-507752).

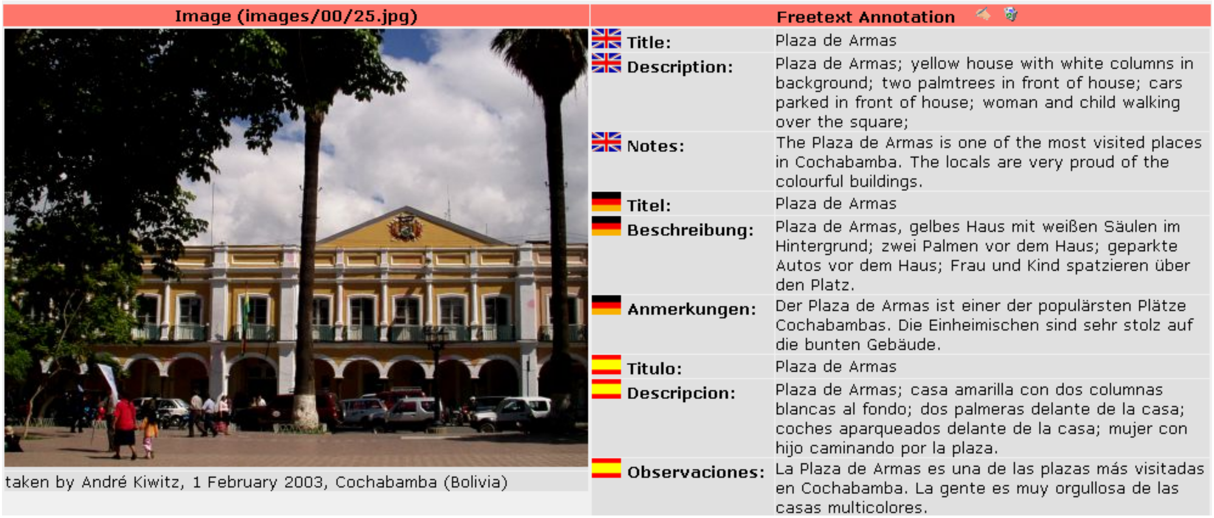


Image (images/00/25.jpg)	Freetext Annotation	
taken by André Kiwitz, 1 February 2003, Cochabamba (Bolivia)	Title:	Plaza de Armas
	Description:	Plaza de Armas; yellow house with white columns in background; two palmtrees in front of house; cars parked in front of house; woman and child walking over the square;
	Notes:	The Plaza de Armas is one of the most visited places in Cochabamba. The locals are very proud of the colourful buildings.
	Titel:	Plaza de Armas
	Beschreibung:	Plaza de Armas, gelbes Haus mit weißen Säulen im Hintergrund; zwei Palmen vor dem Haus; geparkte Autos vor dem Haus; Frau und Kind spatzieren über den Platz.
	Anmerkungen:	Der Plaza de Armas ist einer der populärsten Plätze Cochabambas. Die Einheimischen sind sehr stolz auf die bunten Gebäude.
	Titulo:	Plaza de Armas
	Descripcion:	Plaza de Armas; casa amarilla con dos columnas blancas al fondo; dos palmeras delante de la casa; coches aparqueados delante de la casa; mujer con hijo caminando por la plaza.
	Observaciones:	La Plaza de Armas es una de las plazas más visitadas en Cochabamba. La gente es muy orgullosa de las casas multicolores.

Figure 1: The annotation of one of the images in the IAPR-TC12 dataset (from [3]).

2 Dataset

For the dataset we used an extended version of the IAPR-TC12 dataset of 20 000 vacation images each annotated in English, German and Spanish [3]. An example of an image and its text annotations from this dataset are shown in Figure 1. We tried to make the task as realistic as possible. The version of the dataset used in the showcase event was therefore modified as follows:

- 1000 new images were added. All the text fields for these images were blank. This simulates the often rather bad photo annotation practice of many users.
- The *Description* field was removed from the image annotations (all other fields remained). The description field (a detailed description of the contents of an image) was found to favour text retrieval techniques in ImageCLEF 2006. It is also not realistic to expect users to annotate images in such a detailed way.
- Only the English annotations were kept. Very few people will annotate their image collections tri-lingually.

This dataset is a good simulation of a collection of photos that could be stored on a "standard" user's harddisk.

3 Event Organisation

Participants were asked to register to receive the dataset. The dataset was provided one month before the event, to allow participants time to install the dataset on their system. Some example queries were also provided at this stage (see Section 4.1).

Each participant was required to bring a laptop computer with their image retrieval system and the dataset installed, and set it up at the event. Internet access was not available. The queries were handed out one by one. The "searcher" designated by each team will used the system to find the images from the dataset matching the query (it was required that the same searcher do all the queries for a team). The searches could be as interactive as desired and go through as many iterations as needed. The searches on all the systems happened in parallel.

Two types of queries were provided:

Visual queries: these consist of a text query and one or more example images. Each system was to be used to find the solution to the question as fast as possible. The idea was to search for the images of the same object in the dataset which contain in the annotation file the desired information.

Text queries: these consist of only a text query. Each system was to be used to find as many images satisfying the query as possible.

A time limit of 5 minutes was imposed for each query. During this time, participants could interact with their systems as much as they wished. The query images were made available to each participant at the event on a memory stick. The queries were announced and handed out on slips of paper.

For the text queries, the ground truth was generated manually beforehand. This was in the form of a list of correct images for each query. The queries were ordered roughly based on the number of correct responses. For example, for query 1, there are 268 correct responses, while for query 6, there are only 7. For the visual queries, the ground truth was in the form of the correct answer along with some images leading to the answer.

4 Queries

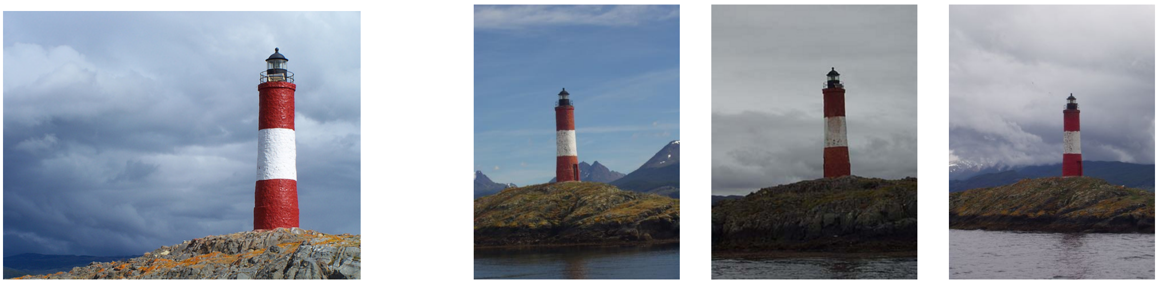
The example queries were sent to the participants one month before the event to provide a preview-view of the type of queries that would be provided. The event queries were kept secret and were only revealed one by one at the event.

4.1 Example queries

Visual queries

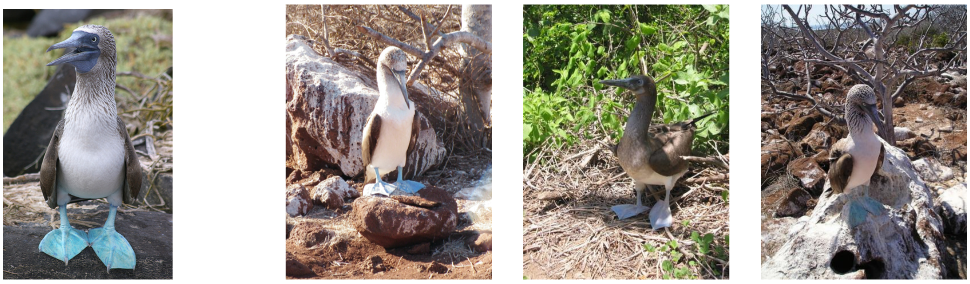
- **Where is the light house in the provided image, Fig. 2(a)?**
  Solution: Ushuaia, Argentina.

- **What is the name of the bird in the provided image, Fig. 3(a)?**
  Solution: Blue-footed Booby.



(a) (b) 10281.jpg (c) 10639.jpg (d) 19317.jpg

Figure 2: (a) The query image *not contained* in the dataset. The image was taken from FlickR. (b–d) Some of the similar dataset images containing the desired information in their annotation files.



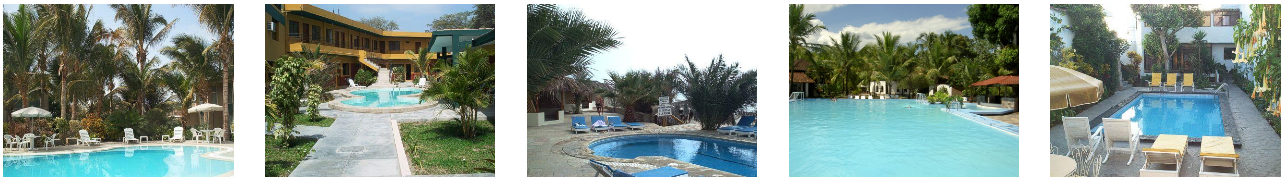
(a) (b) 16781.jpg (c) 5010.jpg (d) 4236.jpg

Figure 3: (a) The query image *not contained* in the dataset. The image was taken from FlickR. (b–d) Some of the similar dataset images containing the desired information in their annotation files.

Text queries

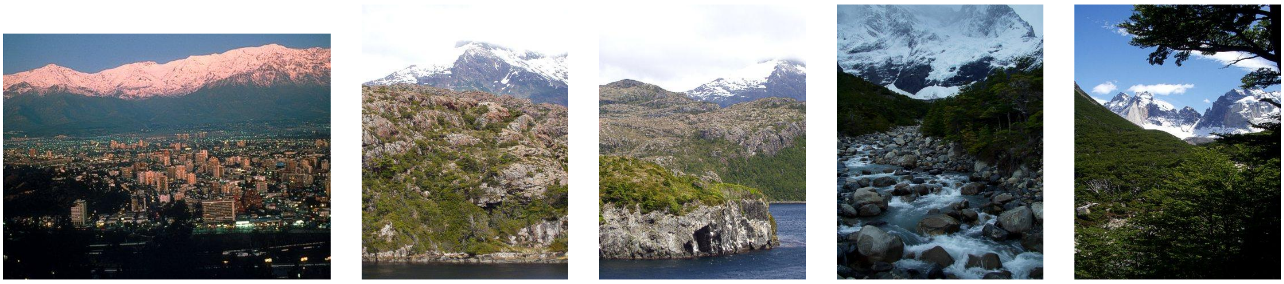
- **Find images with a swimming pool.**
  Some of the correct results are shown in Figure 4.

- **Find images with snowy mountains.**
  Some of the correct results are shown in Figure 5.



1228.jpg 1244.jpg 1254.jpg 12757.jpg 1311.jpg

Figure 4: Some of the correct results for the swimming pool query.



3381.jpg 3494.jpg 3499.jpg 3559.jpg 3564.jpg

Figure 5: Some of the correct results for the mountain query.

4.2 Event queries

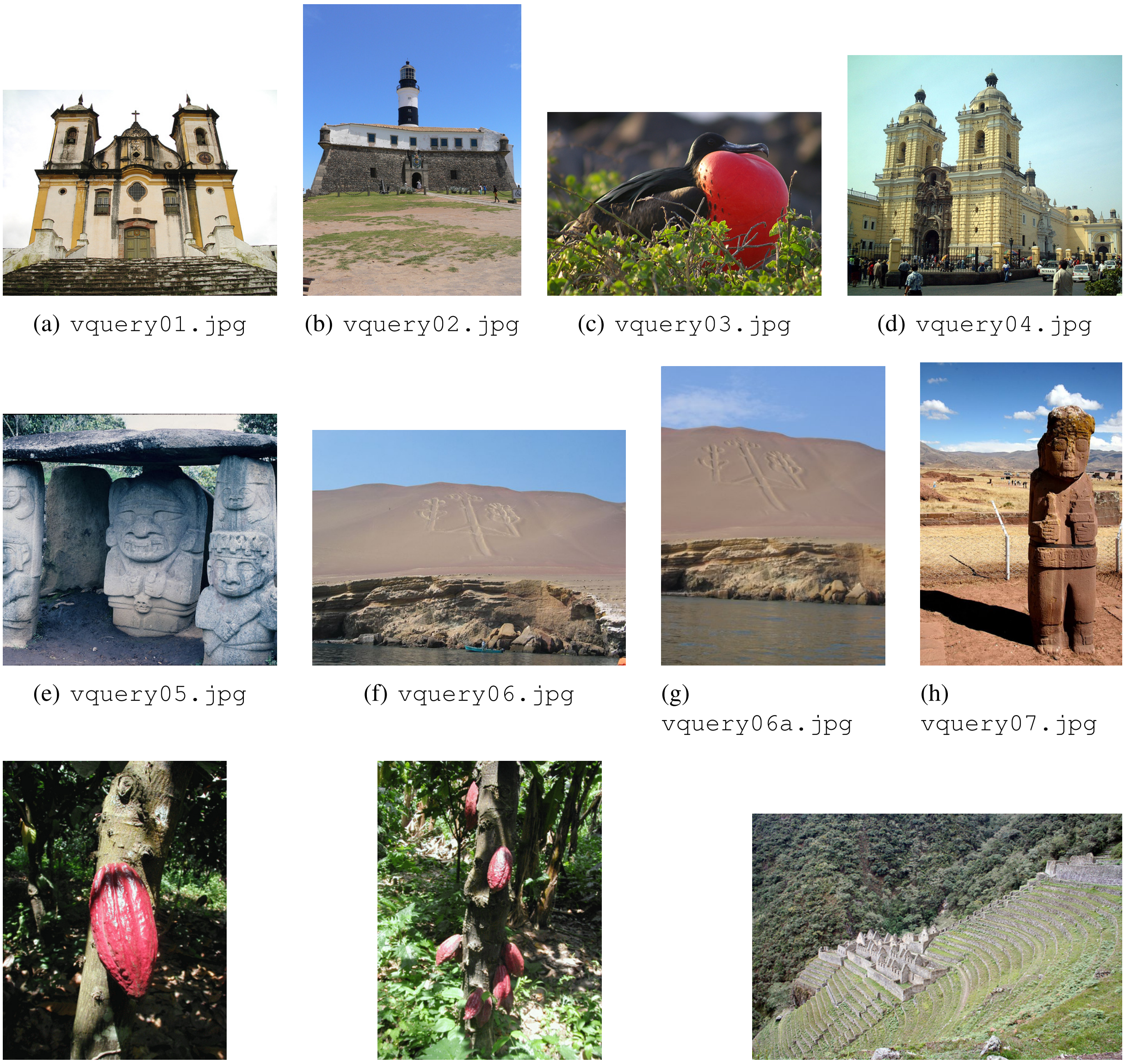
Visual Queries

The corresponding images are shown in Figure 6. They were all obtained from FlickR and hence are not in the dataset.

1. What is the name of the church in the `vquery01.jpg`?
2. What is the name of the building in the `vquery02.jpg`?
3. What type of bird is shown in `vquery03.jpg`?
4. Where (city, country) is the cathedral in `vquery04.jpg`?
5. Where (country) are the statues in `vquery05.jpg`?
6. Where (country) is the sand formation in `vquery06.jpg` and `vquery06a.jpg`?
7. Where (city, country) is the statue depicted in `vquery07.jpg`?
8. The seed in `vquery08.jpg` and `vquery08a.jpg` belongs to which plant?
9. What and where is it? `vquery09.jpg`

Text Queries

1. Find images of a waterfall
2. Find images of people riding or sitting on bicycles (bicycles without rider should not be found)
3. Find images showing a train (photographed from outside the train)
4. Find images showing one or more aeroplanes (photographed from outside the aeroplane)



(a) vquery01.jpg (b) vquery02.jpg (c) vquery03.jpg (d) vquery04.jpg

(e) vquery05.jpg (f) vquery06.jpg (g) vquery06a.jpg (h) vquery07.jpg

(i) vquery08.jpg (j) vquery08a.jpg (k) vquery09.jpg

Figure 6: The images provided with the visual queries at the event. All of these image were taken from FlickR and are hence not in the dataset.

5. Find images showing a Bolivian flag above a group of people [a schematic image of the Bolivian flag was provided]
6. Find images of the Red Footed Booby

5 Participation, Evaluation and Results

Three systems were entered into the event. They are:

- IKONA, content-based image search engine of INRIA IMEDIA (Nozha Boujemaa, Nicolas Hervé, Alexis Joly): http://www-rocq.inria.fr/imedia/cbir-demo.html
- RETIN Multimedia interactive retrieval system (Philippe H. Gosselin, Sylvie Philipp-Foliguet, Matthieu Cord, Julien Gony): http://dupont.ensea.fr/˜ruven/
- New-Phenix Server retrieval system (Jean-Yves Sage, Joël Huberson, Pierre-Alain Moëllic): http://www.new-phenix.com

For the visual queries, the amount of time taken for the first correct answer to be found was recorded. For the text queries, the ratio of correct to incorrect images within the first N images returned was calculated. The value of N was based on the number of correct images for each query in the ground truth.

For the text queries, the system with the best recall was RETIN and the system with the best precision was IKONA. For the visual queries, both the New-Phenix and IKONA systems performed equally well.

6 Conclusion

The event was appreciated by all participants.

The main criticism was that the evaluation criteria should be better defined. For example, one system had a window in which correct images found were placed. The other systems relied on a ranking of all images in the dataset based on how well they satisfied the query. For these latter systems, it was not clear after how many pages of results one should stop counting correct images.

A computer-supported evaluation as used by the VideOlympics should be examined to improve evaluation for future events. Here, the identifier of every correct image found was sent via a local network to a central server, which evaluated the performance of the systems.

References

[1] Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Mller. Overview of the imageclef 2006 photographic retrieval and object annotation tasks. In *CLEF Working Notes*, September 2006. 1

[2] Christian Fluhr, Pierre-Alain Moëllic, and Patrick Hede. Usage-oriented multimedia information retrieval technological evaluation. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 301–306, 2006. 1

[3] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark - a new evaluation resource for visual information systems. In *Proceedings of the International Workshop OntoImage'2006*, pages 13–23, 2006. 2

Problems with Running a Successful Multimedia Retrieval Benchmark [*]

Henning Müller[1], Thomas Deselaers[2], Michael Grubinger[3],
Paul Clough[4], Allan Hanbury[5], William Hersh[6]

[1]Medical Informatics, University and Hospitals of Geneva

[2]Computer Science, RWTH Aachen, Aachen Germany

[3]Victoria University, Melbourne, Australia

[4]Sheffield University, Sheffield, England

[5]Vienna University of Technology, Austria

[6]OHSU, Portland, Oregon, USA

e-mail: henning.mueller@sim.hcuge.ch

Abstract

Content-based image retrieval (CBIR) and multimedia retrieval are at the point where they are ready to leave the pure research status and become integrated into commercial prototypes and products. This requires techniques not only to be interesting as theoretical approaches but also to be comparable with respect to performance obtained. Similar to the text retrieval domain many years ago, several evaluation events or benchmarks have started these past years to compare multimedia retrieval techniques with a varying focus. TRECVID focuses on video, INEX Multimedia on structured data, ImageEval on visual retrieval and classification, and ImageCLEF on multimodal and multilingual data access. Running such a benchmark poses many problems and difficulties. This article summarises some of the major problems encountered in the organisation of ImageCLEF from 2003 to 2007 and tries to find solutions for at least part of the problems.

1 Introduction

Visual information is ubiquitous and the amount produced with cheap digital cameras is rising strongly. To better manage this information content-based images retrieval has been proposed proposed for general image retrieval [16, 6] as well as in specialised domains [11]. Many techniques have been developed for image retrieval but one of the problems is that most approaches

[*]This work was partially supported by the Swiss National Science Foundation (FNS) with grant 205321–109 and the American National Science Foundation (NSF) with grant ITR–0325160.

are very difficult to compare to each other as varying databases, performance measures, and methodologies are used [12].

In recent years several multimedia retrieval benchmarks with a varying focus have been created and run. Many ideas on benchmarking multimedia were already presented early [17, 8, 12] but the Benchathlon[1] was the first event generating a wider discussion in the community. TRECVID[2], the first real benchmark, started as a task in TREC but has since become an independent workshop on the evaluation of video retrieval systems [15]. The strong participation has also made this benchmark important for image retrieval where evaluation can be performed on extracted video key frames. Another initiative is ImagEval[3], financed by the French research foundation and with participants mainly from the French research community and mainly on visual retrieval of images and image classification. INEX[4] (INitiative for the Evaluation of XML retrieval) has also started a multimedia retrieval task in 2006. A fourth benchmarking event is ImageCLEF[5] [3, 2]. This event is part of the Cross-Language Evaluation Forum (CLEF) campaign to evaluate and compare multilingual information retrieval systems [14]. ImageCLEF concentrates on the retrieval of images from multilingual repositories and combining both visual and textual features for multimodal retrieval. A strong participation in ImageCLEF over the past three years has shown the need for standardised system comparison and the importance of creating an infrastructure to support the comparisons in this way. The connection of multimedia benchmarks with events such as TREC, CLEF, and INEX seems necessary to obtain a critical mass and limit the administration overhead concerning the management of document collections, copyright issues, organising a workshop, reserving rooms, registering participants, etc.

Such multimedia retrieval benchmarks can dramatically reduce the effort required by researchers to compare their approaches and start working on optimising the technical part of their work. They are able to concentrate on developing novel methods rather than issues associated with pure evaluation such as defining a methodology and obtaining a database.

This article describes experiences from running the ImageCLEF campaign over five years and tries to propose ways to limit these problems to a minimum.

2 Problems

This section summarises the largest problems encountered in five years of organising the ImageCLEF image retrieval benchmark [4, 5, 2, 10, 1].

2.1 Funding

To organise a benchmark with a yearly rhythm of:

[1] http://www.benchathlon.net/

[2] http://www-nlpir.nist.gov/projects/t01v/

[3] http://www.imageval.org/

[4] http://inex.is.informatik.uni-duisburg.de/2006/

[5] http://ir.shef.ac.uk/imageclef/

- participant registration,
- data release,
- results submission,
- topic definition,
- ground truthing,
- system evaluation,
- and workshop organisation.

takes much time of the organisers. Justifying the time spent on this benchmark is often hard for researchers as evaluation is often seen as part of technical work and not as research by itself. This means that these coordination efforts are sometimes tolerated but often not encouraged and much of the time spent is rather on the "free" time of researchers. Advantages for the organising researchers are the possibility to have an increased visibility in the community, a possibility to influence important research directions, and to publish on the obtained results. Real impact by an evaluation event can particularly be obtained through heavily funded initiatives such as TREC and TRECVID (both funded by NIST). Such funding guarantees a professional organisation, a sufficient marketing and a good evaluation of obtained data and results to optimise impact and to advance the research field. CLEF started as part of TREC and since its independence in 2000 has received minor funding, mainly in an indirect way through research projects funded by the European Union. This has allowed the running of the benchmark for several years but still required much personal work by the organisers and has the risk of creating a non-sustainable structure.

Lack of funding can result in the following problems:

- smaller participation of researchers in the benchmark due to a lack of marketing and confidence in the organisers;
- problems to get access to important data collections as high quality data can be expensive;
- problems with the ground truthing as specialists to judge the documents are usually expensive, and someone performing the ground truthing needs an incentive; bad data quality limits the validity of results;
- lack of analysing the results data obtained and thus a lack of creating new knowledge from available data;
- shortcomings in the organisation of the benchmark can frustrate participants and limit future participation;
- problems in creating a sustainable structure for benchmarking for long term support.

The only way to get out of this is to motivate funding agencies to finance research on evaluation in the same way as technical research. Benefits of benchmarking need to be taken into account as benchmarking is an infrastructure activity for research and can be a multiplier of research results in several domains.

2.2 Getting Access to Proper Data Sets

Image data in good quality is usually expensive and web sites selling images have sprung up over the last years (Corbis[6], Getty[7]). It is correct that image sharing sites such as FlickR[8] have proliferated as well and give access to many images with less restrictive licences. Still, many of these sites contain images that are not put there by the original copyright owners and using these for a benchmarking test collection can cause major problems. In the scientific domain it becomes clear that evaluations on small datasets are not an option and that it becomes important to share images among research groups to limit efforts [19, 13]. Some funding agencies such as the National Institutes of Health (NIH) in the US even require funded research to make their datasets available and similar initiatives exist in other domains.

Still, currently most visual datasets are copyrighted and their use for an evaluation campaign is often difficult. For ImageCLEF we have so far mainly taken image collections from institutions (medical teaching files), from libraries, personal photographic collections, and image collections available on the Internet through MIRC[9] (Medical Imaging Resource Centre). An access to a wider collection of images could strongly influence what exactly can be evaluated in benchmarks.

2.3 Advertise the Benchmark and Motivate Participants

One of the hardest problems in benchmarking is advertising such an event and motivating partners to participate as most researchers are always busy and do not like to read advertisement mails. In the multimedia domain this is particularly true as several research domains overlap in this field and it is hard to address all fields at the same time. Information retrieval, computer vision, machine learning, databases, information systems, digital libraries, e-learning and image processing have all their own methodologies for evaluation and their own conferences to present research results. All of them could address content-based image retrieval.

The main way to motivate research groups is through the reputation of researchers organising a benchmark and through personal contacts by inviting certain groups of researchers and then by mouth to mouth propaganda. Only reputation can create trust that the benchmark will be objective and unbiased. Through contacts between researchers the group of participants becomes larger and through presentations of the benchmark at conferences an even larger group can be informed and motivated to participate.

[6] http://www.corbis.com/

[7] http://www.gettyimages.com/

[8] http://www.flickr.com/

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

Once many groups register for benchmarks it still remains hard to motivate these groups to submit results and have their results compared with the other submissions. Without submissions the influence of these benchmarks will be very limited — the percentage of registrants submitting results in ImageCLEF is still unfortunately less than 50%. Incentives are the possibility to present results to a larger audience at the workshop and to publish on the data in proceedings of good quality (Springer Lecture Notes in Computer Science for CLEF). Still, several groups prefer using the data and task to see how their techniques work and then publish at other occasions. Organising a benchmark together with conferences in the field (such as ECDL for CLEF) can help to soften the problem of limited travel funds and time of participants. Only a high participation leads to important discussions among participants.

Another big problem is the fear of researchers to obtain poor results for their research and thus get problems with potential funding agencies that want to fund only the best technologies. Thus, benchmarking results cannot be taken out of a context and it does not have to be taken as a pure competition. Established techniques might obtain better short-term results but have less potential than some new approaches. This needs to be highlighted to participants to reduce the fear. Selection of oral presentation is at ImageCLEF not based on performance but rather based on interest and novelty of the technique.

2.4 Partners from Professional Companies

Partners from companies are important for multimedia retrieval benchmarks in two ways. Help with the organisation can professionalise research through indirect funding and publicity of companies, a field where they have more experience than most researchers do. They can also focus research towards real problems and realistic user models with realistic datasets through connections with their product development. The advantages can be on both sides: the companies get access to the newest technology and get ideas on how well these techniques work, while researchers get access to realistic tasks and maybe even commercial contacts to fund future projects.

Another problem are participants from companies at the events and a comparison of their techniques with those of other participants. Several companies cannot publish details on their algorithms as the algorithms are sometimes patented or should at least assure the advantage over competitors. As a consequence sometimes companies participate but give no details on the techniques they use, but instead broad descriptions (this is being practised by TRECVID). Some companies, mainly startups are even afraid that bad results would bring their products into discredit (or even reduce venture capital) and thus they would like to be able to remove their results from the final comparison if they turn out to be poor. Such an approach is currently being tested by ImageCLEF with the goal to improve the framework for commercial participation.

2.5 Realistic Tasks and User Models

The definition of tasks and topics depends mainly on the databases available and a very clear user model needs to be defined before tasks can be developed. This can help to tackle real problems and requirements but poorly defined topics can also limit the results of tasks completely.

Typically, realistic tasks can be gained from expert knowledge [5], from log files of system use [4, 1] or through interviews with experts [7].

When observing these information sources, one of the problems with multimedia retrieval is that most information needs are not formulated visually as only few running systems are currently employed. To develop visual tasks from text can be a hard task and requires time for selection. Another problem is the need to have an idea whether and how many relevant images for a certain topic exist in the database. Very broad topics can lead to an extremely large number of relevant items with the risk to miss some of them in the pooling process. If no relevant images exist, the task should also be omitted.

With a document collection and a source to define tasks, the user model can relatively easily be derived.

2.6 Ground Truthing or Gold Standards

Ground truthing for image retrieval evaluation is an expensive task and this is thus linked to the funding problems of many benchmarks. High quality annotations for many specialised tasks can only be performed by domain specialists, whereas some tasks such as image search tasks on personal collections can also be performed by the organisers themselves. INEX even lets participants judge documents of the pools, which limits the effort but creates a slight risk of a bias towards images one is sure the own system would find. To control the quality of relevance judgements several people can be asked to judge the same topics and then a kappa score on agreement between them can be calculated. Variations of judgements have been reported in several domains but they do not in general influence the evaluation results strongly [9, 20].

When using expert judges it is extremely important to define the topics well as human interpretation of seemingly clear information needs can vary strongly [9]! Variation among judges can be reduced through supplying a narrative with the topic explaining in more detail what is regarded as relevant and particularly what is not regarded as relevant. A description of non-relevance has to be highlighted to obtain high-quality results. In ImageCLEF a ternary judgement scheme is used: relevant, partially relevant and non-relevant. Despite the fact that we explain to judges to use partially relevant only when it is impossible to determine relevance, a significant proportion of judgements is in this category.

Another judgement problem concerns multilingual collections such as ImageCLEFmed. If judges are primarily familiar with one language and the judgement process requires to read the text, then a bias towards the native language can appear. A translation of the main terms or a mapping of multilingual text onto an ontology can help to limit the problem. It will only rarely be possible to have judges that are familiar with all languages.

In general, no complete judgement of a test collection is possible and thus a pooling technique has to be applied to judge the most important parts of the dataset based on the results submissions to not bias evaluation towards any system [20, 18]. There is a compromise to be made with respect to how many images to judge. The more images are judged the more time and money it takes and the better the results obtained can be, although it can also increase the fatigue of the judges.

2.7 Organisational Issues

Many benchmarks have a fairly similar model of organisation and a yearly cycle of events. To automate at least part of the process from registration, to document delivery, query submission relevance judgements and evaluation every benchmark seems to develop its own methodology depending on the domain. Within information retrieval TREC has helped massively to standardise at least part of the evaluation. Packages such as trec_eval[10] to evaluate runs based on a particular format of the participants' runs and the relevance judgements have helped to use the same measures and avoid calculation errors that can appear when developing software from scratch.

Even after several years of organisation of a benchmarking event, there are still many small errors happening in ImageCLEF. Among them are those in this short list of some of the main problems:

- errors or inconsistencies in the distributed data collections as no exhaustive tests were performed beforehand, and participants usually discover them at some point;
- incorrect submissions from participants that need to be corrected for correct evaluation, although formats were described and examples made available;
- incompletely or incorrect description of the techniques used for certain runs;
- incomplete descriptions for relevance judges due to time limitations resulting in lower quality judgements;
- delays due to other tasks of the organising researchers;
- problems with software for results submission or relevance judgements resulting in a loss of time for participants or judges.

2.8 Proving Advances and Benefits of Benchmarks

One of the most important parts in "selling" the utility of benchmarks is to show the improvement that they have brought to the domain. Again, this can be linked with funding and impact. When manpower is available it is much easier to prove the utility than when the manpower is lacking to analyse outcomes of benchmarks over time. TREC has shown that through detailed analysis of the results many important points can be shown such as the lack of a bias when using pooling techniques [20] or the fact that changing relevance judges generally does not change the ranking of performing systems significantly; measures such as B-Pref results also from TREC research. Benchmarks with less funding have a harder time doing these in depth analyses and will only be able to achieve minor impacts.

An easy way to prove performance is to measure the use of the created collections, topics and relevance judgements. Unfortunately, authors often reuse the resources for other publications but do not inform the organisers on this although it is requested from participants. Research on the web can bring up some of the publications but will always be incomplete. Through

[10] http://trec.nist.gov/trec_eval/

the number of reuses, the saved time of researchers can be estimated. Still, the most important part is the comparability of approaches and this is difficult to be measured: the comparability of techniques and focusing of researchers on promising techniques avoiding typical mistakes of the past.

Another way to show how a research field is improving is to run older techniques on new data and show where they are with respect to current techniques or to run newer techniques on older data. In ImageCLEF this shows well that the performance of participating systems has significantly improved over time.

3 Conclusions

Benchmarks in the multimedia field have enormously advanced the techniques developed in research labs. Re-creation of small datasets and the impossibility of comparing approaches have been reduced, and at several conferences approaches can now be compared on the same datasets making it possible to have a clear idea of advantages and disadvantages of various approaches. Instead of spending much time and money on the creation of datasets, research groups can start with standard datasets and participate in evaluation campaigns.

One of the main criticisms of benchmarks is a sort of standardisation of research and the tendency to reuse well-performing techniques and make minor modifications instead of developing completely new techniques that might have more potential for the future ("Do benchmarks kill innovation?"). Some of these criticisms are true and thus benchmarks cannot be used for completely new research domains but rather in domains where an established set of techniques has already been developed and that is at the point to be ready for a use in real prototype systems. Another point to soften this criticism is to attempt a quickly changing set of benchmarks to avoid running the same sort of tests every year. Similar to TREC where many tracks run between 2 and 5 years it is important to have changes in the types of tasks. Another important part is to include recommendations and new people from the community in organising new tasks to avoid the impression of an elitist organisation and to adapt running benchmarks towards real and up-to-date tasks of the users, which is the research community.

References

[1] Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In *CLEF 2006 Proceedings*, Springer Lecture Notes in Computer Science, pages 579–594, 2007. 2, 6

[2] Paul Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeffery Jensen, and William Hersh. The CLEF 2005 cross–language image retrieval track. In *Cross Language Evaluation Forum (CLEF 2005)*, Springer Lecture Notes in Computer Science, pages 535–557, September 2006. 2

[3] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul Clough, Julio Gonzalo, Michael Jones, Gareth J. F.and Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 597–613, Bath, UK, 2005. Springer. 2

[4] Paul Clough and Mark Sanderson. The CLEF 2003 cross language image retrieval task. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2003)*, 2004. 2, 6

[5] Paul Clough, Mark Sanderson, and Henning Müller. The clef cross language image retrieval track (ImageCLEF) 2004. In *The Challenge of Image and Video Retrieval (CIVR 2004)*, Springer Lecture Notes in Computer Science, pages 243–251, July 2004. 2, 6

[6] A. del Bimbo. *Visual Information Retrieval*. Academic Press, 1999. 1

[7] William Hersh, Jeffery Jensen, Henning Müller, Paul Gorman, and Patrick Ruch. A qualitative task analysis for developing an image retrieval test collection. In *ImageCLEF/MUSCLE workshop on image retrieval evaluation*, pages 11–16, Vienna, Austria, 2005. 6

[8] 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. 2

[9] Henning Müller, Paul Clough, William Hersh, and Antoine Geissbuhler. Variations of relevance assessments for medical image retrieval. In *Adaptive Multimedia Retrieval (AMR)*, volume 4398 of *Springer Lecture Notes in Computer Science (LNCS)*, pages 233–247, 2007. 6

[10] Henning Müller, Thomas Deselaers, Thomas Lehmann, Paul Clough, Eugene Kim, and William Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In *CLEF 2006 Proceedings*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, 2007 – to appear. Springer. 2

[11] 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. 1

[12] 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. Special Issue on Image and Video Indexing. 2

[13] Guiseppe Sasso, Hugo Raul Marsiglia, Francesca Pigatto, Antonio Basilicata, Mario Gargiulo, Andrea Francesco Abate, Michele Nappi, Jenny Pulley, and Francesco Silvano Sasso. A visual query–by–example image database from chest CT images: Potential role as a decision and educational support tool for radiologists. *Journal of Digital Imaging*, 18(1):78–84, March 2005. 4

[14] Jacques Savoy. Report on CLEF–2001 experiments. In *Report on the CLEF Conference 2001 (Cross Language Evaluation Forum)*, pages 27–43, Darmstadt, Germany, 2002. Springer LNCS 2406. 2

[15] 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. 2

[16] 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. 1

[17] 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. 2

[18] 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. 6

[19] Michael W. Vannier and Ronald M. Summers. Sharing images. *Radiology*, 228:23–25, 2003. 4

[20] Justin Zobel. How reliable are the results of large–scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York. 6, 7

Benchmarking the sensory and semantic gap

Marcel Worring, Cees Snoek, Jan-Mark Geusebroek, Arnold Smeulders
Intelligent Sensory Information Systems
Computer Science Institute
Faculty of Science, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
e-mail: worring@science.uva.nl

Abstract

Access to visual data is hampered by the sensory and semantic gap. The sensory gap is the gap between the object in the world and the information in an image recording of that scene. Due to variations in viewpoint, lighting and other circumstantial conditions all of these recordings will be different. To evaluate methods for bridging the sensory gap we have developed the ALOI dataset which are recordings of object with controlled variation of various conditions. The semantic gap is the lack of coincidence between what can be derived from the data and the user's interpretation of the same data. An influential benchmark for evaluation is TRECVID. Within the context of TRECVID we introduce the challenge problem for generic video indexing to gain insight in intermediate steps that affect performance of multimedia analysis methods, while at the same time fostering repeatability of experiments. To arrive at a challenge problem, we provide a general scheme for the systematic examination of automated concept detection methods, by decomposing the generic video indexing problem into 2 unimodal analysis experiments, 2 multimodal analysis experiments, and 1 combined analysis experiment. Finally, we consider the interacting user and will share with you our experiences in organizing the first Videolympics.

Thoughts on evaluation of image retrieval inspired by ImageCLEF 2007 object retrieval task [*]

Ville Viitaniemi and Jorma Laaksonen

Laboratory of Computer and Information Science, Helsinki University of Technology
02015 TKK, Finland
tel +358 9 4515973, fax +358 9 451 3277
e-mail: firstname.lastname@tkk.fi

Abstract

In this paper we discuss questions on information retrieval evaluation brought up by the analysis of the recent ImageCLEF 2007 object retrieval task. We question the adequacy of the pooling strategy used in this evaluation and discuss pooling in general. We also argue that use of MAP to rank the participants is not very meaningful in this case. Instead, the result of each query should be considered separately.

1 Introduction

Our group participated the ImageCLEF 2007 object retrieval task. This purely visual task—concerning the retrieval of objects from a collection of mostly touristic photographs—is a part of the Cross Language Evaluation Forum (CLEF), a benchmarking event for multilingual information retrieval. This paper was inspired by us being not fully confident of the statistical reliability of the reported ImageCLEF 2007 object retrieval task results. As a result, we posed a number questions and remarks towards the organisers of the task. Partly we hoped for definite answers, but partly we knew the questions to be open or very subjective, and wanted to hear the organisers' standpoint on these issues. However, it was decided to present the issues also to a larger audience in form of this paper in order to stimulate thoughts and discussion. Since then, the results have been re-evaluated due to a detected programming bug, but many of the issues still remain.

In addition to just posing questions, we went through some literature in order to seek answers to some of the questions. ImageCLEF 2007 object retrieval task results were obtained using pooling, a commonly used information retrieval evaluation technique to reduce the number

[*] Supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *Finnish Centre of Excellence in Adaptive Informatics Research*. Special thanks to the ImageCLEF 2007 object retrieval task organisers Thomas Deselaers and Allan Hanbury for being helpful concerning the details of the evaluation procedure.

of manual document relevance judgements needed. In standard TREC[1] settings, the round-robin pooling, used also in ImageCLEF, has proven to work well, but in those settings the general performance level has been sufficient to include in the judging pools many or most of the relevant documents present in the databases [7, 13]. Here, as we shall see later, this is not the case. We suspected the pooling process to be one of the main factors causing the possible unreliability of the evaluation results. In particular, we suspected the pools to be too shallow in the specific case of this image collection and these queries, where the general level of performance seems to be very modest.

To somewhat clarify the matters and provide a reference, we manually annotated the whole test image collection, as the collection was of a still rather manageable size. For the current concrete need of evaluating the performance in this particular task, this naturally makes the issue of pooling irrelevant. However, we still devote the matter some thought. For fairness' sake, we have to admit that there are also other reasons for us wanting to re-evaluate the results of the task. The reason is that in our own submissions we made a mistake that excluded our runs completely or partly from the pool selection procedure for some queries, which seems to degrade our results essentially. If this was actually the case, it would testify for the pools being too shallow to reach *reusability* —the possibility to fairly evaluate new methods without need to redo the pooling — that can be taken also as a criterion of adequate pooling depth. On the other hand, the pooling process might favour our runs, as we submitted half of the total runs. The matter deserves investigation in order to be able to get an unbiased picture of the capabilities of the image analysis techniques that generated the participating runs.

2 Image collections and the retrieval task

For reference, we summarise here the retrieval task description and later the results from the ImageCLEF 2007 object retrieval task overview paper [4]. In the task the goal was to find all bicycles, buses, cars, motorbikes, cats, cows, dogs, horses, sheep, and persons from the IAPR TC-12 database [6], using the manually-annotated images of PASCAL NoE VOC 2006 Challenge [5] for training. Both the approximately 2600 training and 20000 test images are realistic photographs of real-life situations, where the objects of interest appear as parts of complex scenes. However, in the training images the number of appearances of each object type is roughly equal, whereas in the test set the difference can be more than three orders of magnitude. The test images also exhibit a wider variety of scene types, large portion of which containing objects not present in the training data. For example, many test images contain llamas or alpacas that were not present in the training data, but very closely resemble the training category "sheep".

One might very well ask whether trying to learn from training data different from the test data makes any sense. On the other hand, the goal rarely is to learn to exactly replicate the training data. Instead, one tries to learn a model that fits the training data and generalises well to test data. The key in this process is to evaluate which of the training examples are similar to the test images. In the present case, the bit patterns of the images and even the low-level image

[1] http://trec.nist.gov/

features might not be similar at all, and so it is justifiable to talk about differently distributed training and test data. However, if we make a transformation to a semantic space, where each dimension is the presence of some object class, the data sets might no longer appear so different.

Average precision (AP) is used as a single figure measure to measure retrieval quality for each of the ten queries. The overall retrieval quality is measured—as in many other evaluations—by mean average precision (MAP) of all queries. For the ground truth, the actual relevance of all of the images was not originally judged by humans. Instead, the commonly-used round-robin pooling technique was applied, where 100 of the top-ranked retrievals of each run were collected to form a pool for each query topic. Each pool was then manually judged for that specific query topic. In addition, each pool was augmented with part of the images of the remaining pools as it was noticed that judging all ten query topics simultaneously was not much more laborious than judging just one topic. The decision to collect such additional relevance judgements, however, was made only during the judging process and the exact number of additionally judged images remains unfortunately unknown. Thus the results with additional relevance judgements are to be seen only as suggestive examples, although in practice they might give more truthful picture then the normal annotation pools.

Table 1 shows the statistics of the relevance judgement, along with the statistics of our manual annotation of the whole test database that was completed afterwards. For comparison of the annotations made by the original judges for the pools and us for the whole database, we list in the fifth column of the table the number of pooled images that would have been relevant according to our complete annotation. We notice there to be some differences due to the subjectivity of the annotations, but generally the differences are minor and the retrieval results with and without pooling are rather well comparable in this sense.

Table 1: Judging statistics.

query	query name	size of pool	relev. in pool (judges)	relev. in pool (complete annot.)	additional relev.	relev. in database
1	bicycle	1437	66 (4.6%)	73 (5.0%)	254	655 (3.3%)
2	bus	1449	23 (1.6%)	32 (2.2%)	69	218 (1.1%)
3	car	1578	200 (13%)	196 (13%)	522	1268 (6.3%)
4	motorbike	1478	7 (0.47%)	13 (0.74%)	28	86 (0.43%)
5	cat	1502	5 (0.33%)	3 (0.20%)	18	7 (0.04%)
6	cow	1504	7 (0.47%)	8 (0.53%)	23	49 (0.25%)
7	dog	1488	9 (0.60%)	9 (0.60%)	22	72 (0.36%)
8	horse	1513	13 (0.86%)	14 (0.93%)	94	175 (0.88%)
9	sheep	1693	5 (0.30%)	0 (0.0%)	42	6 (0.03%)
10	person	1437	554 (39%)	621 (43%)	3939	11248 (56%)

The annotation of the whole test database was performed by a single person looking systematically through all the images. An average time spent annotating a single image was a couple of seconds. The images tagged as relevant were double-checked more carefully for all but the two most numerous query topics. The process revealed the subjectivity of some of the queries, in particular as the queries were not tailored for this image collection. For example, it had to be decided where to put the borders of query “car” among sedans, estate wagons, suburban vehicles, land cruisers, pick-ups and vans. It had to be chosen, of how many pixels must a "person" consist in order to be tagged as such, even if the classification was very probable due to contextual information. Table 1 reveals also the fact that none of the animals in the pools that were judged to be sheep actually was a sheep (with possibly one exception), but llamas or similar animals. In contrast, some queries, such as "horse", were straightforward to annotate. We believe the annotations to be reasonably accurate but probably not completely error-free. One positive property of the annotations, however, is the internal consistency the single annotator made and interpreted all the necessary subjective decisions.

3 Comparison of retrieval results

Tables 2, 3 and 4 show the AP results of the evaluation using the normal pooling, additional relevance information and complete annotations of the test database, respectively. The tables list the AP values for all the 26 runs submitted to the evaluation. The method identifiers are the same as in [4], but for the purposes of the discussion in this paper the actual identity of runs does not matter. Due to our own mistake, our runs with prefix HUTCIS missed the pool selection process either completely (the normal pool) or partly (the additional pool) for query topics 4–10, inserting random images in the pool instead of top-ranking images. Therefore, only results for queries 1–3 can be considered representative for HUTCIS in the pooled evaluation. This has to be kept in mind also when assessing the effectiveness of the chosen pooling strategy in light of the retrieval results.

The runs have been ranked according to MAP, which we argue not to be very meaningful in this case. Actually, we think the queries and the runs are so different that there might be little sense in making any overall ranking at all, the results of interest lying revealed by query-wise statistics. Besides the queries being very different in terms of number of relevant images and visual difficulty, another factor is brought by the number of relevant images in the database exceeding that of the maximum number of retrieval results (1000) the runs were allowed to return for the queries "car" and "person". This makes the queries qualitatively easier than the other runs as the runs are not required to retrieve all of the relevant, but only the preferred 79% or 8.9% of the relevant images, for "car" and "person" respectively. On the other hand, the AP numbers in the tables do not have maximum of 1.0 for these queries, but 0.79 ("car" with full annotations), 0.25 ("person" with the additional pooling) and 0.89 ("person" with full annotations). Keeping this in mind, we notice that the best "person" results are actually very close (97%) to this maximum. In TRECVid, the AP measure is normalised to maximum of unity in such cases, but here such normalisation is not done, keeping the results easily reproducible by the `trec_eval` tool of the TREC campaign.

Still another argument against the usefulness of MAP in this case is the differences in the submitted runs in terms of which of the query topics their MAP originates from. If we take, for example, the runs budapest-acad-budapest-acad314 and HUTCIS_SVM_FULLIMG_ALL and compare their performance with the complete relevance information, we find that although quite similar MAP, essentially all of Budapest MAP comes from very good performance in a single query, "bicycle", whereas HUTCIS gathers its MAP from several sources. Is it really meaningful to compare whether it is desirable to be really good in one query topic or rather good in several? How about the scaling of the "person" query AP? Additionally, the number of queries is quite small and thus the effect of single query to the MAP can be decisive. The ranking of MAP values could change essentially if just one query, say "motorbike", was replaced by some other query.

Table 2: AP results of the task with normal relevance information. The numbers have been multiplied by 100 to improve readability.

rank	run id	1	2	3	4	5	6	7	8	9	10	MAP
1	HUTCIS_SVM_FULLIMG_ALL	**21.3**	1.5	**28.1**	0.3	0.0	0.2	0.2	0.8	21.0	0.8	**7.4**
2	HUTCIS_SVM_FULLIMG_IP+SC	10.2	1.2	25.8	0.2	0.0	0.2	0.1	1.4	20.3	1.6	6.1
3	HUTCIS_SVM_FULLIMG+BB	13.0	1.5	11.4	0.1	0.0	0.4	0.1	0.6	**22.4**	1.1	5.1
4	HUTCIS_SVM_FULLIMG_IP	9.3	1.3	23.6	0.1	0.0	0.1	0.1	**2.6**	3.0	1.2	4.1
5	MSRA-MSRA_RuiSp	2.5	1.9	7.9	**3.5**	0.9	0.0	0.3	0.7	2.1	**13.7**	3.4
6	HUTCIS_SVM_BB_BAL_IP+SC	4.9	1.3	2.4	0.0	0.0	**1.5**	0.0	0.1	10.4	0.4	2.1
7	HUTCIS_PICSOM1	3.2	1.3	13.5	0.1	0.0	0.1	0.2	0.0	0.4	0.5	1.9
8	HUTCIS_PICSOM2	1.7	1.4	13.2	0.1	0.0	0.0	**0.4**	0.0	0.6	0.5	1.8
9	HUTCIS_SVM_BB_ALL	5.8	1.6	2.4	0.0	0.0	1.4	0.0	0.1	4.4	0.3	1.6
10	HUTCIS_SVM_BB_BB_IP+SC	5.2	1.3	2.8	0.1	0.0	0.8	0.0	0.1	4.4	0.4	1.5
11	HUTCIS_SVM_BB_FULL_IP+SC	8.1	1.7	1.5	0.1	0.0	1.0	0.0	0.1	2.4	0.3	1.5
12	RWTHi6-HISTO-PASCAL	0.3	**2.7**	1.7	0.1	0.0	0.1	0.2	0.4	0.1	9.4	1.5
13	HUTCIS_SVM_BB_BB_IP	4.0	0.7	1.2	0.1	0.0	0.1	0.0	0.2	2.1	0.3	0.9
14	MSRA-MSRA-VLM_8_8_640_ful	0.6	0.2	1.2	0.3	0.0	0.0	0.0	2.0	0.1	3.6	0.8
15	HUTCIS_SVM_BB_BAL_IP	3.9	1.6	1.2	0.1	0.0	0.2	0.0	0.1	0.5	0.3	0.8
16	MSRA-MSRA-VLM-8-8-800-HT	0.9	0.1	1.1	0.4	0.0	0.0	0.0	0.6	0.3	3.2	0.7
17	PRIP-PRIP_HSI_ScIvHarris	0.2	0.0	0.6	0.2	**2.6**	0.2	0.1	0.0	0.0	2.2	0.6
18	budapest-acad-budapest-acad314	2.3	0.1	0.0	0.5	1.1	0.0	0.0	0.0	0.0	0.1	0.4
19	HUTCIS_SVM_BB_FULL_IP	0.2	1.2	1.0	0.1	0.0	0.1	0.0	0.1	1.2	0.2	0.4
20	NTU_SCE_HOI-NTU_SCE_HOI_1	2.0	0.1	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4
21	PRIP-PRIP_cbOCS_ScIvHarr2	0.1	0.0	0.3	0.0	0.8	0.0	0.0	0.0	1.3	1.0	0.4
22	budapest-acad-budapest-acad315	0.3	0.0	0.0	1.5	0.0	0.1	0.3	0.0	0.1	0.6	0.3
23	INAOE-TIA-INAOE_SSAssemble	0.5	0.1	0.1	0.1	0.0	0.3	0.0	0.0	0.0	0.6	0.2
24	INAOE-TIA-INAOE-RB-KNN+MRFI_ok	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.1
25	INAOE-TIA-INAOE-RB-KNN+MRFI	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.1
26	INAOE-TIA-INAOE-RB-KNN	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.1

4 Reliability of the retrieval results

Our prior belief was that using the HUTCIS methods, the results of this type of task with differently distributed training and test images would not be very much better than random. Performance levels close to the noise level are challenging with respect to statistical reliability. There are at least two kinds of statistical unreliability that could be present in the results. First of all, the finite number of relevant and non-relevant images cause purely random orderings of the images to exhibit non-vanishing AP performance. We demonstrated this component of the unreliability—determined completely by the numbers of relevant and non-relevant images—by tabulating percentiles of the AP distributions of 20000 purely random runs in Table 4. We notice that in most queries, most of the participant runs—but not all—rise clearly above the level of random performance. We also confirm that different queries are somewhat different in terms of random performance. Another observation is that for the judgement pools, this kind of fluctuation does not seem to be a larger factor than for the results using complete annotations.

Table 3: AP results of the task with additional relevance information. The numbers have been multiplied by 100 to improve readability.

rank	run id	1	2	3	4	5	6	7	8	9	10	MAP
1	HUTCIS_SVM_FULLIMG_ALL	7.2	1.3	**14.9**	0.4	0.0	0.4	0.1	**4.1**	2.7	5.8	**3.7**
2	HUTCIS_SVM_FULLIMG_IP+SC	3.7	0.7	14.1	0.5	0.0	1.1	0.0	3.5	3.0	**6.3**	3.3
3	HUTCIS_SVM_FULLIMG_IP	3.3	0.7	13.0	0.8	0.0	**1.4**	0.0	4.0	0.8	5.0	2.9
4	HUTCIS_SVM_FULLIMG+BB	4.7	1.6	6.5	0.7	0.0	0.5	0.0	2.0	**3.1**	4.2	2.4
5	HUTCIS_PICSOM1	1.5	1.2	7.5	0.2	0.0	0.3	0.1	0.4	0.6	5.1	1.7
6	HUTCIS_PICSOM2	1.0	1.1	7.4	0.3	0.0	0.2	0.2	0.3	0.6	4.6	1.6
7	MSRA-MSRA_RuiSp	1.4	0.7	5.4	1.0	0.3	0.1	0.1	0.3	0.3	5.1	1.5
8	HUTCIS_SVM_BB_BAL_IP+SC	1.9	**2.3**	1.1	0.3	0.0	0.7	0.0	0.5	1.8	3.2	1.2
9	HUTCIS_SVM_BB_BB_IP+SC	1.9	2.2	1.3	0.5	0.1	0.5	0.0	0.6	1.0	3.2	1.1
10	HUTCIS_SVM_BB_BB_IP	1.5	0.5	0.9	3.8	0.2	0.2	0.0	0.7	0.4	2.9	1.1
11	HUTCIS_SVM_BB_ALL	2.2	1.3	1.1	0.5	0.0	0.7	0.0	0.7	1.1	3.1	1.1
12	budapest-acad-budapest-acad314	**9.1**	0.2	0.0	0.7	0.3	0.1	0.0	0.1	0.0	0.0	1.1
13	HUTCIS_SVM_BB_FULL_IP+SC	3.0	1.4	0.8	0.2	0.1	0.6	0.0	0.6	1.0	3.0	1.1
14	RWTHi6-HISTO-PASCAL	0.2	1.0	1.5	0.3	0.0	0.0	0.1	0.2	0.1	5.0	0.8
15	HUTCIS_SVM_BB_BAL_IP	1.5	0.8	0.7	0.6	0.3	0.2	0.0	0.6	0.2	2.9	0.8
16	budapest-acad-budapest-acad315	0.2	0.0	0.0	**6.2**	0.0	0.0	0.1	0.0	0.2	0.1	0.7
17	HUTCIS_SVM_BB_FULL_IP	0.3	0.8	0.6	0.4	0.1	0.1	0.0	0.8	0.4	2.7	0.6
18	MSRA-MSRA-VLM_8_8_640_ful	0.3	0.3	1.3	0.1	0.0	0.0	0.0	0.3	0.1	3.2	0.6
19	MSRA-MSRA-VLM-8-8-800-HT	0.4	0.1	1.1	0.1	0.0	0.0	0.0	0.3	0.1	2.5	0.5
20	PRIP-PRIP_HSI_ScIvHarris	0.1	0.0	0.4	0.1	**1.5**	0.1	**0.3**	0.1	0.1	1.8	0.5
21	NTU_SCE_HOI-NTU_SCE_HOI_1	1.2	0.2	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4
22	PRIP-PRIP_cbOCS_ScIvHarr2	0.1	0.0	0.2	0.8	0.4	0.0	0.0	0.1	0.2	1.1	0.3
23	INAOE-TIA-INAOE_SSAssemble	0.1	0.1	0.2	0.0	0.0	0.3	0.0	0.1	0.1	1.1	0.2
24	INAOE-TIA-INAOE-RB-KNN	0.1	0.0	0.3	0.1	0.0	0.0	0.0	0.0	0.0	1.0	0.2
25	INAOE-TIA-INAOE-RB-KNN+MRFI_ok	0.3	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.8	0.1
26	INAOE-TIA-INAOE-RB-KNN+MRFI	0.3	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.8	0.1

In contrast to the first source of unreliability, the second source affects only the pooled results. It originates from just a subset of relevant images ending to a judgement pool. This causes both random-like statistical fluctuations and also systematic bias towards systems that work well for images for which also many other systems find likely to be relevant. As an example of another kind of bias, one can easily think of a situation where the pooled evaluation favours methods performing extremely well among the easy cases that appear in the beginning of retrieval results and much worse in difficult case, in comparison to methods performing more evenly among both easy and difficult cases. At this time, we could not evaluate unreliability caused by pooling in any systematic manner. Anecdotal evidence of systematic bias can be clearly observed in the result tables, however. Most prominent is the Budapest results for “bicycles” and “motorbikes” being completely overshadowed by other runs with normal judgement pools, alleviated somewhat by inclusion of additional relevance information. However, the complete relevance information reveals the results actually to be overwhelmigly better than the other runs, and perhaps the most interesting results of this whole evaluation in general. The relevance pooling thus appears somehow inadequate.

Comparing the fourth and last columns of Table 1, one notices that for most queries, the pooling works to some degree in the sense that the fraction of relevant images in the pools is larger than in the database in general. However, the reached relevant image fractions are still quite small, contrary to the case of many other retrieval evaluations. It seems to be that in the sense of pooling, most of the submitted runs perform only modestly better than random. There-

Table 4: AP results of the task using complete annotation of the database. The numbers have been multiplied by 100 to improve readability.

rank	run id	1	2	3	4	5	6	7	8	9	10	MAP
1	budapest-acad-budapest-acad314	**28.3**	0.3	0.1	1.8	0.2	0.2	0.0	0.0	0.0	0.0	**3.1**
2	HUTCIS_SVM_FULLIMG_ALL	4.1	**1.2**	**10.6**	0.4	0.0	0.6	0.1	**3.8**	0.0	8.3	2.9
3	HUTCIS_SVM_FULLIMG_IP+SC	2.6	1.0	11.1	1.0	0.0	1.0	0.1	3.2	0.0	8.2	2.8
4	HUTCIS_SVM_FULLIMG_IP	2.4	1.1	10.3	1.8	0.0	**1.1**	0.1	3.0	0.0	8.1	2.8
5	HUTCIS_SVM_FULLIMG+BB	3.0	1.1	4.2	0.6	0.0	0.7	0.1	2.5	0.0	**8.6**	2.1
6	budapest-acad-budapest-acad315	0.6	0.0	0.1	**18.5**	0.0	0.1	0.1	0.0	0.6	0.1	2.0
7	HUTCIS_SVM_BB_ALL	1.6	0.9	0.5	0.3	0.0	0.6	0.1	1.5	0.0	8.3	1.4
8	HUTCIS_SVM_BB_BB_IP+SC	1.4	1.0	0.7	0.3	0.0	0.5	0.1	1.1	0.0	8.4	1.4
9	HUTCIS_SVM_BB_FULL_IP+SC	2.0	0.8	0.4	0.2	0.0	0.8	0.1	1.1	0.0	8.2	1.3
10	HUTCIS_PICSOM1	0.9	0.7	4.5	0.6	0.0	0.3	0.1	0.7	0.0	5.6	1.3
11	MSRA-MSRA_RuiSp	0.9	0.5	3.6	0.6	0.7	0.1	0.1	0.4	0.0	6.0	1.3
12	HUTCIS_SVM_BB_BAL_IP+SC	1.3	0.8	0.5	0.2	0.0	0.5	0.1	0.8	0.0	8.4	1.3
13	HUTCIS_PICSOM2	0.8	0.6	4.2	0.5	0.0	0.3	0.1	0.4	0.0	5.4	1.2
14	HUTCIS_SVM_BB_BB_IP	1.1	0.7	0.4	1.4	0.0	0.3	0.0	1.0	0.0	7.2	1.2
15	HUTCIS_SVM_BB_BAL_IP	1.1	0.8	0.3	0.3	0.0	0.4	0.1	0.9	0.0	6.9	1.1
16	HUTCIS_SVM_BB_FULL_IP	0.3	0.9	0.3	0.3	0.0	0.3	0.0	1.1	0.0	6.6	1.0
17	RWTHi6-HISTO-PASCAL	0.4	0.2	1.4	0.2	0.0	0.1	0.0	0.2	0.0	5.5	0.8
18	NTU_SCE_HOI-NTU_SCE_HOI_1	1.2	0.7	2.4	0.0	0.0	0.0	0.1	0.1	0.0	0.8	0.5
19	MSRA-MSRA-VLM_8_8_640_ful	0.4	0.3	0.7	0.1	0.1	0.0	0.0	0.3	0.0	2.5	0.4
20	MSRA-MSRA-VLM-8-8-800-HT	0.3	0.2	0.5	0.0	0.0	0.1	0.0	0.2	0.1	2.5	0.4
21	INAOE-TIA-INAOE_SSAssemble	0.1	0.0	0.1	0.0	0.0	0.2	0.0	0.2	0.0	3.2	0.4
22	PRIP-PRIP_HSI_ScIvHarris	0.1	0.0	0.3	0.1	**1.4**	0.1	0.0	0.0	0.0	1.5	0.4
23	INAOE-TIA-INAOE-RB-KNN+MRFI_ok	0.5	0.1	0.6	0.0	0.0	0.2	0.0	0.0	0.0	2.2	0.4
24	INAOE-TIA-INAOE-RB-KNN+MRFI	0.5	0.1	0.6	0.0	0.0	0.2	0.0	0.0	0.0	2.2	0.4
25	INAOE-TIA-INAOE-RB-KNN	0.3	0.0	0.5	0.1	0.0	0.1	0.0	0.0	0.0	2.2	0.3
26	PRIP-PRIP_cbOCS_ScIvHarr2	0.1	0.0	0.1	0.5	0.1	0.0	0.1	0.1	0.0	0.8	0.2

fore, it might be worthwhile to try to improve the pooling performance by favouring the better of the methods. This could be achieved, for example, with algorithms like RankBoost [11]. The idea of concentrating the judging effort non-uniformly to the queries based on need estimated progressively during judging [13] also appears to have potential, as observed also by [7]. Another possible way to increase the number of relevant images in the pools without increasing manual judging effort is to make the pools deeper, but then manually judge only a sample of the images. When properly done, this sacrifices only a very small fraction of accuracy in return for a significant savings in judging effort. However, the pools have to be sampled carefully and the performance measure has to be adapted accordingly. An example of such adaptation is the inferred average precision (infAP) [12], adopted, for instance, by the TRECVid 2006 evaluation in the high-level feature detection task [8]. Literature also mentions methods to evaluate the retrieval accuracy without making any manual relevance judgements at all, but such methods tend to measure the popularity of documents among participants instead of their genuine relevance [1]. If the overall retrieval accuracy is good among all participants, such techniques might be useful, but this seems not to be the case here.

In general, at least two requirements can be posed to a pooling scheme: *rank preservation* and *reusability*. Preservation of method ranks is typically evaluated by comparing the pooled ranks with a reference ranking, in this case it could be the ranking using the complete annotation. Kendall's τ and a pre-determined threshold as reliability criterion is a common means of evaluating rank preservation.

Table 5: Percentiles of AP performance distribution of random runs. The numbers have been multiplied by 100 to improve readability.

Pooling	Percentile	1	2	3	4	5	6	7	8	9	10	MAP
Normal	50	0.03	0.01	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.04
	75	0.05	0.03	0.11	0.02	0.00	0.02	0.02	0.02	0.00	0.21	0.06
	90	0.11	0.08	0.16	0.06	0.05	0.06	0.06	0.07	0.05	0.27	0.09
	95	0.17	0.14	0.22	0.12	0.11	0.13	0.13	0.13	0.11	0.33	0.13
	99	0.52	0.54	0.52	0.51	0.56	0.60	0.54	0.52	0.48	0.44	0.36
Additional	50	0.08	0.03	0.15	0.01	0.01	0.01	0.01	0.04	0.02	1.01	0.15
	75	0.12	0.06	0.20	0.04	0.03	0.03	0.03	0.07	0.04	1.11	0.17
	90	0.18	0.11	0.27	0.08	0.08	0.08	0.08	0.12	0.10	1.20	0.20
	95	0.24	0.16	0.32	0.14	0.14	0.15	0.14	0.18	0.15	1.26	0.23
	99	0.46	0.51	0.43	0.51	0.56	0.55	0.57	0.52	0.51	1.37	0.37
Full database	50	0.19	0.07	0.34	0.03	0.00	0.02	0.03	0.06	0.00	2.82	0.37
	75	0.24	0.11	0.41	0.06	0.02	0.05	0.06	0.10	0.02	2.94	0.39
	90	0.30	0.17	0.48	0.12	0.06	0.10	0.11	0.15	0.06	3.05	0.42
	95	0.35	0.22	0.53	0.17	0.12	0.15	0.17	0.21	0.12	3.11	0.44
	99	0.46	0.50	0.62	0.50	0.55	0.52	0.53	0.48	0.48	3.23	0.54

The other requirement for pooling is reusability. This means that the judging should be comprehensive enough so that a method that does not participate in the selecting the judging pools would not significantly suffer from this. This is a common and useful criterion for benchmark tasks as it facilitates the incremental evaluation of novel methods without the need to redo the pool selection and thus affect also the results of earlier methods. Actually, Ian Soboroff [9] claims that the success of whole Cranfield paradigm, paradigm of evaluating ranked retrieval runs based on relevance judgements, is entirely due to reusability: only reusability facilitates evaluating experiments in an easily reproducible and extendable manner. As an example of application of the reusability criterion, the databases and tasks of CLEF 2001 [2] were required and found to be reusable.

For the present case, reusability could be obtained by using sufficiently deep judging pools. In the present case this could be practically impossible goal as some query topics are so small that they might require the annotation of almost the whole database. There exists also an alternative route to reusability: abandon round-robin pooling and do random sampling instead. This would immediately lead to reusability. As the pooling does not seem to be operating too efficiently anyway, it could we weighed whether the small performance advantage might be traded away for reusability.

5 On combining query-wise results

Information retrieval evaluations always have to deal with the matter of selecting a suitable performance measure. Average precision (AP) has become one of the standard measures of a performance of a single query. To combine the performances of multiple queries, one often takes simply the mean average precision (MAP) over all queries. We discuss some of the properties of the MAP measure in this section in general level, and ask ourselves how some of its shortcomings could be improved. We do this in spite of already earlier arguing that it is not too meaningful to perform an overall ranking of the rather different runs on basis of the very heterogeneous queries of the present evaluation.

Ease of comparison of performance against random level is one of the qualities the AP measure lacks. As Table 4 demonstrates, random behaviour may manifest itself in different magnitudes of AP performance for different query topics. In contrast to AP, this kind of ease of interpretation is one of the desirable properties of the ROC AUC measure that was in use for example in PASCAL VOC Challenge [5] until 2006. Could this somehow be incorporated into AP and consequently into MAP? One possibility would be to normalise the AP values against such shifting of zero-level performance, for example by subtraction.

Another property of the MAP measure is that it treats all query topics uniformly, which may not be desirable. For example, in the present case the overall MAP performance is often dominated by the performance in few of the topics where the absolute AP values are large, such as “bicycle” or “person”. One could imagine a situation where percentually large improvement, say 100% in a query topic with lower general performance would be overshadowed by an improvement of 10% in a more dominant topic. It is matter of taste whether this is actually what is wanted, but if a fairer weighting of query topics is sought after, information theory could be looked at. Another approach is taken by GMAP measure adopted by TREC [10] that replaces arithmetic mean with geometric mean, effectively giving more weight to proportionally large changes in small AP values than same absolute AP change in larger values.

The presented ways to modify the MAP combining do not deal with the varying difficulty of different topics. The similar AP (or some normalised version thereof) values could result either from a system performing well in a visually challenging query or fairly poorly in an easy query. The different difficulty levels of queries could be taken into account for example by normalising the results to the mean, maximum or some other statistic of the performance of other runs participating the evaluation.

An important observation to make that the required level of statistical reliability depends on whether one wants to be able to rank the participants of the evaluation separately in terms of each query or just to obtain an overall ranking of the participants over all queries. In principle, this present evaluation is about evaluating the performance of various techniques, not ranking different research groups. In principle, a group is not required to use the same algorithm for resolving all queries but might use different, manually specified algorithms for different queries. In practice, however, same algorithm seems often to be used throughout. Another argument for the need of the ability to evaluate the queries separately is the qualitative heterogeneity of queries. Different runs may be effective in different queries. To identify the kind of method to be used for each query, the evaluation should be statistically reliable enough to rank the runs in terms of each query separately.

The same statistical reliability is often easier to obtain for a composite measure such as MAP. When summarising the results into such a metric, the fluctuations in individual queries tend to cancel and the result gets more reliable as concepts are added. The property can be taken advantage of in IR evaluations, where a certain level of reliability in MAP measure is more efficiently (in terms of judging effort) achieved by using a large number of shallowly pooled queries, compared to a few deeply pooled (and thus more reliable) queries [3]. However, if some component queries of a composite metric are very unreliable, incorporating them might degrade the overall reliability. In the present database, the queries “cat” and “sheep” might be examples of such harmful queries due to their small proportion of true positives in the database. A more reliable composite metric could then be obtained by leaving these queries out of the metric and combining all the other queries. This idea could be taken further by including all the queries into summarisation process but assigning them different weights. The weight would be a function of a suitable reliability indicator, such as a measure inversely proportional to the variance of the query performance in random trials.

6 Conclusions and future work

We believe to have demonstrated that there are reasons to doubt the reliability of the results of the ImageCLEF 2007 object retrieval task, at least the results obtained using relevance pooling. In the light of the retrieval result tables, the presently used judgement pooling seems to fulfill neither the rank preservation nor the reusability criterion. Partly to demonstrate the shortcomings of the used pooling approach and partly to make the retrieval task reusable for novel methods, we performed complete annotation of the database.

It seems that the database and queries for the ImageCLEF 2007 object retrieval task were chosen rather ad hoc, perhaps to satisfy the information needs of a hypothetical user in one situation. If this kind of evaluations are organised in the future, it should be asked whether certain technical requirements posed by the need to evaluate the retrieval results should be taken into account already when the retrieval task is designed. The idea of taking an image collection from an existing evaluation as training data and evaluating it in another collection whose details were unknown seems not to have been a total success. The queries, of which there were a balanced number of training images, turned out to have very different number of relevant images in the test collection, thus making the results of the different queries difficult to compare. In addition, the evaluation procedure seems not to be perfectly suited for the test database, resulting in some technical problems. However, the evaluation was not a total failure, either. The query-wise results themselves are very interesting. This is because the task is able to partially separate the detection of actual object itself from use of contextual information that has turned out to be very beneficial in other evaluations such as the Pascal NoE VOC Challenge.

In order to be able to fully trust, understand and utilise the evaluation results a careful reliability study would have to be performed, for example using the randomisation methodology in similar way as in analysis of TRECVid 2006 high-level feature detection task results. In TRECVid, the randomisation experiments were performed on MAP, regarding single-query APs as constituents of the map that could randomly be swapped among a pair of runs. In contrast, here the MAP does not seem very useful, and randomisation would have to be performed within the individual queries. This could mean partitioning the relevant images into several partitions, which would seem to be hopeless for the smaller queries. It is completely possible that one would have to abandon the hope of provably reliable results for some of the weakest individual queries.

The design and execution of the outlined randomisation experiments remains a future task. The question to ask is, however, is this benchmark task of sufficiently high quality and the results so interesting one one hand, and questionable on the other that such experiments would be worth performing. The results for “bicycle” and “motorbike” would seem to be clear enough even without further analysis: the Budapest approaches seem to be the only ones that really work. For the other queries, the results seem to be either bad or, in the case of “person”, the task is probably too easy as the achieved performance is almost perfect.

References

[1] J. A. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of ACM SIGIR ’03*, pages 361–362, July 2003. 7

[2] Martin Braschler. CLEF 2001 - overview of the reults. In *CLEF 2001*, number 2406 in Lecture Notes in Computer Science, pages 9–26. Springer, Berlin-Heidelberg, 2002. 8

[3] Gordon V. Cormack and Thomas R. Lynn. Power and bias of subset pooling strategies. In *Proceedings of ACM SIGIR ’07*, Amsterdam, July 2007. 9

[4] Thomas Deselars and Allan Hanbury et al. Overview of the ImageCLEF 2007 object retrieval task. In *CLEF 2007 Working Notes*, 2007. 2, 4

[5] Mark Everingham, Andrew Zisserman, Chris Williams, and Luc Van Gool. The Pascal Visual Object Classes Challenge 2006 (VOC2006) results. Technical report, 2006. Available on-line at `http://www.pascal-network.org/`. 2, 9

[6] Michael Grubinger, Paul Clough, and Clement Leung. The IAPR TC-12 benchmark for visual information search. *IAPR Newsletter*, 28(2):10–12, April 2006. 2

[7] Sabrina Keenan, Alan F. Smeaton, and Gary Keogh. The effect of pool depth on system evaluation in TREC. *Journal of the American Society for Information Science and Technology*, 52(7):570–574, 2001. 2, 7

[8] Paul Over, Tzveta Ianeva, Wessel Kraaij, and Alan F. Smeaton. TRECVID 2006 - an overview. In *TRECVID Online Proceedings*. TRECVID, March 2007. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html. 7

[9] Ian Soboroff. A comparison of pooled and sampled relevance judgments in the TREC 2006 Terabyte track. In *1st International Workshop on Evaluating Information Access (EVIA)*, pages 22–31, Tokyo, Japan, May 2007. 8

[10] TREC 2006 common evaluation measures, 2006. Online proceedings of 15th Text Retrieval Conference (TREC 2006) , available via http://trec.nist.gov/. 9

[11] Huyen-Trang Vu and Patrick Gallinari. Using RankBoost to compare retrieval systems. In *Proceedings of ACM CIKM ’05*, pages 309–310, Bremen, Germany, 2005. 7

[12] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of ACM CIKM ’06*, pages 102–111, Arlington, VA, USA, November 2006. 7

[13] Justin Zobel. How reliable are the results of large-scale information retrieval experiments. In *Proceedings of ACM SIGIR’98*, pages 307–414, Melbourne, Australia, August 1998. 2, 7

Consolidating the ImageCLEF Medical Task Test Collection: 2005-2007*

William Hersh, MD[1], Henning Müller, PhD[2],
Jayashree Kalpathy-Cramer, PhD[1], Eugene Kim[1]

[1]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University

[2]Section of Medical Informatics, University & Hospitals of Geneva, Geneva, Switzerland

3181 SW Sam Jackson Park Rd., BICC

Portland, OR   97239

+1-503-494-4563

e-mail: hersh@ohsu.edu

Abstract

The goal of the ImageCLEF medical image retrieval task (ImageCLEFmed) has been to improve understanding and system capability in search for medical images. This has been done by developing a test collection that allows system-oriented evaluation of medical image retrieval systems. From 2005-2007, test collections were developed and used for ImageCLEFmed. This paper describes our recent work consolidating the test collections into a single unified collection of 66,662 images and their annotations; 85 topics classified by amenability to visual, textual, or mixed retrieval methods; and relevance judgments. This will provide a comprehensive test collection for further testing of systems and algorithms in medical image retrieval..

1 Introduction

Images play a variety of uses in health care and biomedical research. Despite their widespread use, however, we know little about how those who use them find and manage them. Two small analyses have found that the image use tends to be related to the "role" of the user, such as clinician, educator, researcher, etc. [1, 2]. As there are growing numbers of image collections and search interfaces proliferating on the World Wide Web as well as closed

* 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/).

networks, we believe it is important to understand users' needs as well as provide systems that meet those needs.

The goal of the ImageCLEF medical image retrieval task (ImageCLEFmed) is to improve understanding and system capability in search for medical images [3]. This has been done by developing a test collection that allows system-oriented evaluation of medical image retrieval systems. As with most collections, we have strived to make the content and search topics for this collection as realistic as possible. For three years running, ImageCLEF has featured a medical retrieval task based around ad hoc retrieval. The collection of images came from four sources initially, with two additional ones added in the third year. Each collection is used "as is," i.e., its annotations are used from the original source. This paper describes the recent effort by the project to consolidate the three years of test collections into a single collection that aims to provide a test bed for evaluating systems and algorithms that perform medical image retrieval.

2 Background

ImageCLEF is a part of the Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org), a challenge evaluation for information retrieval from diverse languages [4]. CLEF itself is an outgrowth of the Text Retrieval Conference (TREC, trec.nist.gov), a forum for evaluation of text retrieval systems [5]. TREC and CLEF 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 [6]. 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) [7]. 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 [8].

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 [9, 10]. More recently, focus has shifted to the biomedical research domain in the TREC Genomics Track [11]. Test collections are also used increasingly for image retrieval outside of medicine [12].

In this paper, we describe our efforts to create a single consolidated test collection. In the remaining sections, we describe the content, topics, relevance judgments, and future plans for the merged collection.

3 Content

The conceptual structure of the content of the ImageCLEFmed test collection is as follows. The entire *library* consists of multiple collections. Each *collection* is organized into cases that represent a group of 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 (e.g., HEAL tagging), and/or a textual annotation. All of the images and annotations are stored in separate files. An XML file contains the connections between the collections, cases, images, and annotations. Figure 1 shows a graphical depiction of the library, while Figure 2 shows the XML metadata format.

The image library for ImageCLEFmed 2005 and 2006 consisted of the first four collections listed in Tables 1 and 2 (Casimage, MIR, PEIR, and PathoPIC). In 2007, we added the latter two collections listed in those tables (myPACS and CORI). Table 1 describes the image collections, their image and annotation types, and their origins, while Table 2 lists the numbers of images and annotations (including amounts in each language) as well as the archived file size. Figure 3 shows an example case from the Casimage collection, demonstrating how multiple different images and image types can be part of a case. However, note that the largest collection, PEIR, is not organized into cases per se (or, using our framework, has one image per case). The image library for the consolidated test collection will be the entire library, which is the same as that used for ImageCLEFmed 2007.

4 Topics

A total of 85 topics have been developed over 2005-2007 for ImageCLEFmed. Each topic has been provided with an information statement in English, French, and German, as well as an index image of a relevant image for use by visual retrieval systems. Because we discovered early on that results on different tasks varied by whether the topic was amenable to visual or textual retrieval, we classified each topic as visual, textual, or mixed.

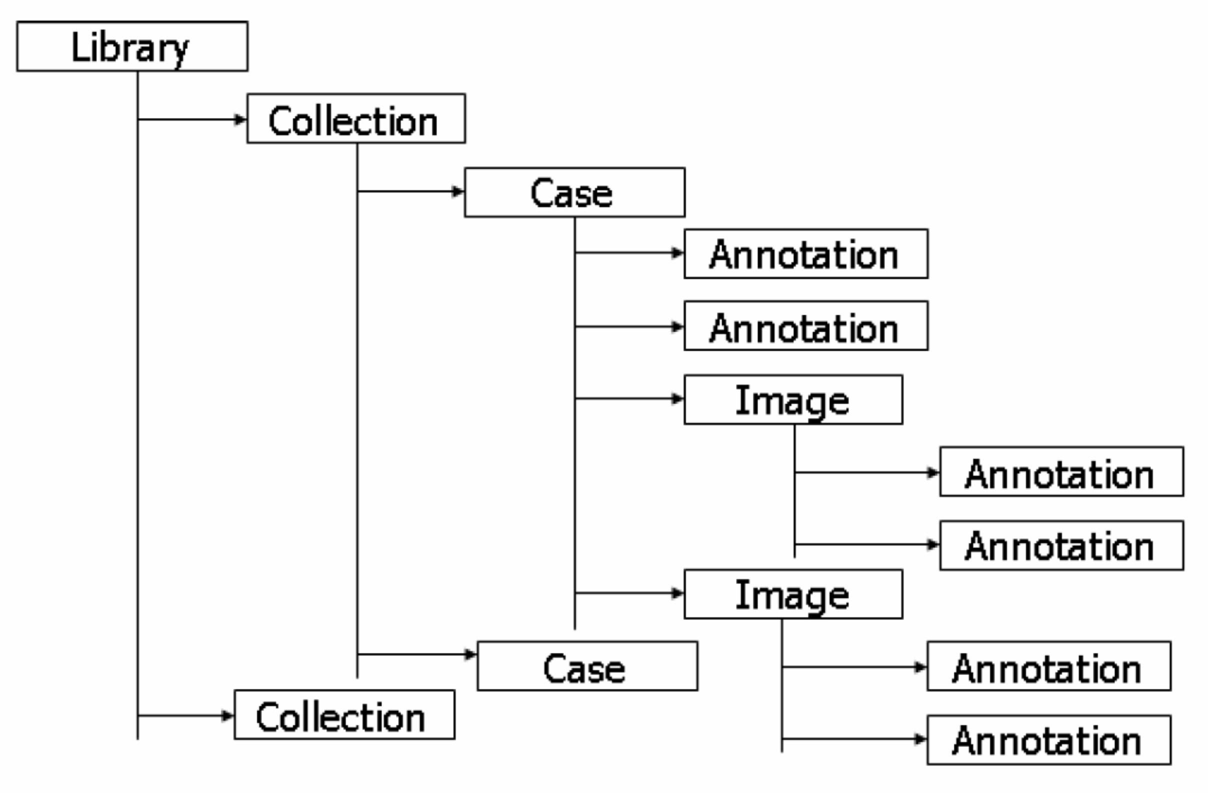


Figure 1 - Structure of ImageCLEF medical image retrieval task (ImageCLEFmed) test collection content.

```
<library>
    <collection>
        <name>name-text</name>
        <cases>
            <case>
                <id>identifier-text</id>
                <images>
                    <image>
                        <id>identifier-text</id>
                        <imagefile>file-name-text</imagefile>
                        <annotation lang=” “>file-name-text</annotation>
                        <annotation lang=” “>file-name-text</annotation>
                    </image>
                <images>
                <annotation lang=” “>file-name-text</annotation>
                <annotation lang=” “>file-name-text</annotation>
            </case>
        </cases>
    </collection>
</library>
```

Figure 2 - Structure of ImageCLEF medical image retrieval task (ImageCLEFmed) XML metadata format for the content.

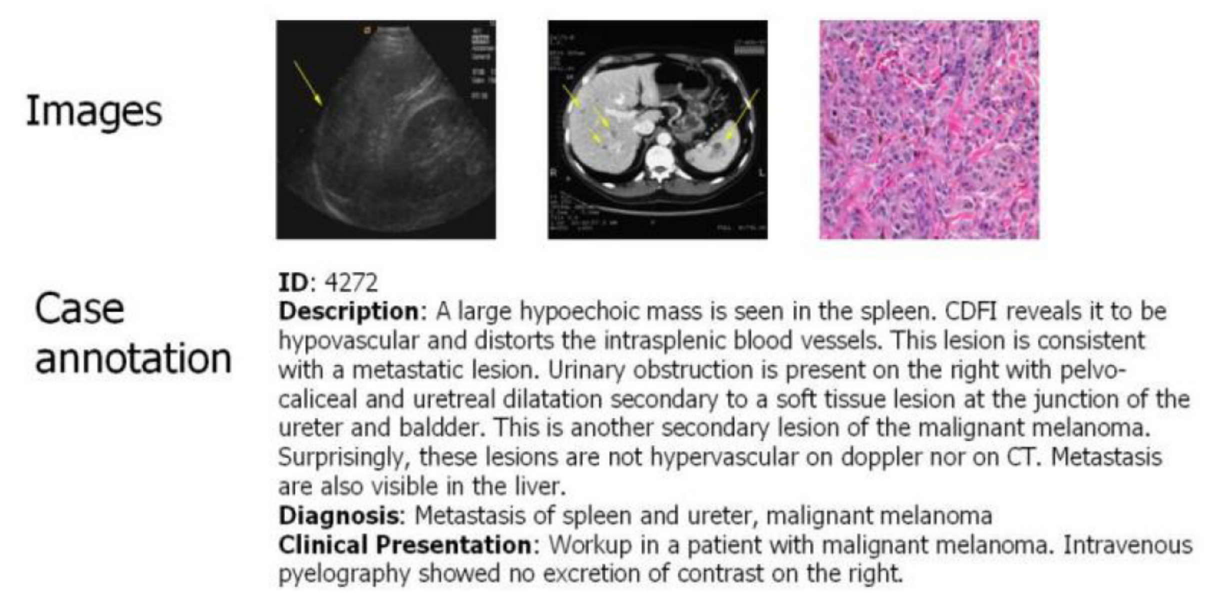
There were 25 topics in 2005 and 30 each in 2006 and 2007. In each year, each topic was numbered from 1, i.e., 1-25 in 2005 and 1-30 in 2006 and 2007. In the consolidated test collection, the topics from 2005 are numbered 1-25, those from 2006 are numbered 26-55, and those from 2007 are numbered 56-85. A sample topic from the consolidated collection is shown in Figure 4.

Table 1 - ImageCLEF medical image retrieval task (ImageCLEFmed) image collections, image and annotation types, and their origins.

Collection Name	**Image Type(s)**	**Annotation Type(s)**	**Original URL**
Casimage	Radiology and pathology	Clinical case descriptions	http://www.casimage.com/
Mallinckrodt Institute of Radiology (MIR)	Nuclear medicine	Clinical case descriptions	http://gamma.wustl.edu/home.html
Pathology Education Instructional Resource (PEIR)	Pathology and radiology	Metadata records from HEAL database	http://peir.path.uab.edu/
PathoPIC	Pathology	Image description - long in German, short in English	http://alf3.urz.unibas.ch/pathopic/e/intro.htm
MyPACS	Radiology	Clinical case descriptions	http://www.mypacs.net/
Clinical Outcomes Research Initiative (CORI) Endoscopic Images	Endoscopy	Clinical case descriptions	http://www.cori.org/

Table 2 - ImageCLEF medical image retrieval task (ImageCLEFmed) numbers of images and annotations (including amounts in each language) as well as the archived file size

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
myPACS	3577	15140	3577	English - 3577	390 MB
Endoscopic	1496	1496	1496	English - 1496	34 MB
Total	47680	66662	55485	French - 1899 English - 45781 German - 7805	5.15 GB



ID: 4272
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.

Figure 3 - An example ImageCLEF medical image retrieval task (ImageCLEFmed) case from the Casimage collection.

```
<topic>
  <number>55</number>
  <EN-description>Show me images of findings with Alzheimer's Disease.
      </EN-description>
  <DE-description>Zeige mir Bilder von Fällen mit einer Alzheimer Diagnose.
      </DE-description>
  <FR-description>Montre-moi des images d'observations avec la maladie
      d'Alzeimer.</FR-description>
  <year>2006</year>
  <query-images>
    <image>images2006/3-10a.jpg</image>
    <image>images2006/3-10b.jpg</image>
  </query-images>
  <query-type>semantic</query-type>
</topic>
```

Figure 4 - Topic 55 from the consolidated ImageCLEF medical image retrieval task (ImageCLEFmed) test collection.

5 Relevance Judgments

Relevance judgments in ImageCLEFmed have been performed by physicians who are also students in the OHSU biomedical informatics graduate program. They have been paid an hourly rate for their work. The pools for relevance judging have been created by selecting the top ranking images from all submitted runs. The actual number selected from each run varied by year, but was usually about 30-40, with the goal of having pools of about 800-1200 images in size for judging. Judges have been instructed to rate images in the pools are definitely relevant (DR), partially relevant (PR), or not relevant (NR). In ImageCLEFmed 2005 we used only DR images for the gold standard, but in 2006 and 2007 we used DR and PR images.

For the consolidated test collection, we need to perform relevance judgments for the new 2007 images applied to the 2005 and 2006 topics. This process is currently underway. We are also judging some images whose names were erroneous in the 2007 pools due to their being incorrect in the submitted runs. Relevance judging will take place in August-September, 2007.

Although the reliability of judging has been slightly better than that obtained from relevance judgments of textual documents in clinical [9] and genomics [11, 13] tasks, we have found instances of incorrectly judged images, especially with regards to the modality of the image, which is vitally important in image retrieval. To that end, we plan to nominate images for rejudgment, and all future judges will be asked to adhere to the following instructions:

1. Note that a topic can refer to one or more of the following: (a) an imaging modality, (b) an anatomical location, (c) a view and/or (d) a disease or finding. An image should only be considered relevant if it meets all the terms mentioned explicitly in the topic (i.e., should be an AND, not an OR). For instance, in the topic “CT liver abscess,”

only CT scans showing a liver abscess should be considered relevant. Pathology or MRI images of liver abscesses should not be considered relevant. Images of other abscesses should not be considered relevant. An x-ray image associated with an annotation that refers to a need for a CT scan in the future should not be considered relevant.

2. When a photograph is the desired imaging modality, i.e., it says “image of” or picture of,” only photographic images should be considered relevant. Although, technically, microscopic images of histology/pathology may be considered to be photographs, in this context, they should not be considered relevant.
3. Pathology in the query refers to pathological images (microscopic/gross pathology), not the state of being abnormal.
4. Refer to the sample images provided with each topic for a better understanding of desired imaging modalities.
5. Synonyms of terms should be considered relevant in the topic. For instance, any MeSH synonyms of the search terms should be considered relevant. As an example, cholangiocarcinoma is a synonym of bile duct cancer. But on the other hand, the liver/biliary system/pancreas should not be considered synonymous with the entire gastrointestinal system.

6 Future Work

When the work described in sections 2-4 is complete, we will have a medical image retrieval collection with 66,662 images and their annotations; 85 topics categorized by amenability to visual, textual, or mixed retrieval; and about 800-1200 relevance judgments per topic. Our goal is to contact the dozen or so major participants in ImageCLEFmed from 2005-2007 submit baseline runs so that baseline levels of performance can be ascertained. Our hope is that additional researchers will use the collection to evaluate new approaches to image retrieval in the future.

We do plan to continue ImageCLEFmed in 2008 but hope to look at new types of tasks beyond the ad hoc retrieval used in 2005-2007. We aim to expand our previous work in user assessment to develop use cases for new tasks.

References

1. Hersh WR, et al. *A qualitative task analysis of biomedical image use and retrieval. MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*. 2005. Vienna, Austria. http://muscle.prip.tuwien.ac.at/workshop2005_proceedings/hersh.pdf.
2. Müller H, et al. *Health care professionals’ image use and search behaviour. Proceedings of Medical Informatics Europe 2006*. 2006. Maastricht, Netherlands. 24-32. http://www.sim.hcuge.ch/medgift/publications/MIE2006_Mueller.pdf.

3. Hersh WR, et al., *Advancing biomedical image retrieval: development and analysis of a test collection.* Journal of the American Medical Informatics Association, 2006. 13: 488-496.
4. Braschler M and Peters C, *Cross-language evaluation forum: objectives, results, achievements.* Information Retrieval, 2004. 7: 7-31.
5. Voorhees EM and Harman DK, eds. *TREC: Experiment and Evaluation in Information Retrieval*. 2005, MIT Press: Cambridge, MA.
6. Sparck-Jones K, *Reflections on TREC.* Information Processing and Management, 1995. 31: 291-314.
7. Buckley C and Voorhees EM, *Retrieval System Evaluation*, in *TREC: Experiment and Evaluation in Information Retrieval,* Voorhees EM and Harman DK, Editors. 2005, MIT Press: Cambridge, MA. 53-75.
8. Buckley C and Voorhees E. *Evaluating evaluation measure stability*. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000. Athens, Greece: ACM Press. 33-40.
9. Hersh WR, et al. *OHSUMED: an interactive retrieval evaluation and new large test collection for research*. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1994. Dublin, Ireland: Springer-Verlag. 192-201.
10. Hersh WR, *Interactivity at the Text Retrieval Conference (TREC).* Information Processing and Management, 2001. 37: 365-366.
11. Hersh WR, et al., *Enhancing access to the bibliome: the TREC 2004 Genomics Track.* Journal of Biomedical Discovery and Collaboration, 2006. 1: 3. http://www.j-biomed-discovery.com/content/1/1/3.
12. Clough P, et al. *Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks*. *Evaluation of Multilingual and Multi-modal Information Retrieval - Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*. 2006. Alicante, Spain: Springer Lecture Notes in Computer Science. in press. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/cloughOCLEF2006.pdf.
13. Hersh W, et al. *TREC 2005 Genomics Track overview*. *The Fourteenth Text Retrieval Conference - TREC 2005*. 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/GEO.OVERVIEW.pdf.

Content-based Video Tagging for Online Video Portals*

Adrian Ulges[1], Christian Schulze[2], Daniel Keysers[2], Thomas M. Breuel[1]

[1]University of Kaiserslautern, Germany

[2]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern

{a_ulges,tmb}@informatik.uni-kl.de,

{christian.schulze,daniel.keysers}@dfki.de

Abstract

Despite the increasing economic impact of the online video market, search in commercial video databases is still mostly based on user-generated meta-data. To complement this manual labeling, recent research efforts have investigated the interpretation of the visual content of a video to automatically annotate it. A key problem with such methods is the costly acquisition of a manually annotated training set.

In this paper, we study whether content-based tagging can be learned from user-tagged online video, a vast, public data source. We present an extensive benchmark using a database of real-world videos from the video portal *youtube.com*. We show that a combination of several visual features improves performance over our baseline system by about 30%.

1 Introduction

Due to the rapid spread of the web and growth of its bandwidth, millions of users have discovered online video as a source of information and entertainment. A market of significant economic impact has evolved that is often seen as a serious competitor for traditional TV broadcast. However, accessing the desired pieces of information in an efficient manner is a difficult problem due to the enormous quantity and diversity of video material published. Most commercial systems organize video access and search via meta-data like the video title or user-generated tags (e.g., *youtube*, *myspace*, *clipfish*) – an indexing method that requires manual work and is time-consuming, incomplete, and subjective.

While commercial systems neglect another valuable source of information, namely the content of a video, research in *content-based video retrieval* strives to automatically annotate (or 'tag') videos. Such systems learn connections between low-level visual features and high-level semantic concepts from a training set of annotated videos. Acquiring such a training set manually is costly and poses a key limitation to these content-based systems.

*Work supported partially by the Stiftung Rheinland-Pfalz für Innovation, project InViRe (961-386261/791)

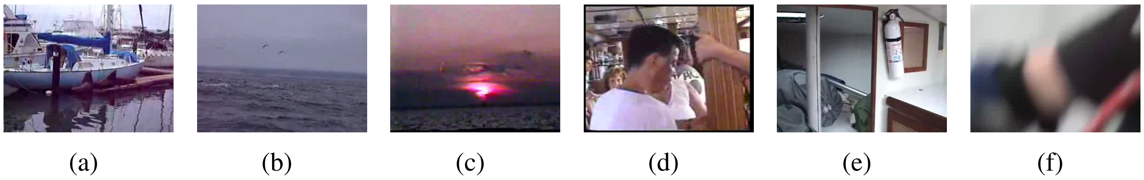


Figure 1: Some sample keyframes extracted from a video with the tag ’sailing’. Tagging such material is aggravated by the complexity of concepts (a,b), varying appearance (b,c), shots not directly visually linked to sailing (d,e), and low production quality (f).

In this paper, we study a different kind of training set, namely videos downloaded from online video portals (a similar idea has been published for images before, learning from Google Image Search [3]). Online videos are publicly available and come in a quantity that is unmatched by any dataset annotated for research purposes, providing a rich amount of tagged video content for a large number of concepts.

On the backside, online video content is extraordinarily difficult to interpret automatically due to several of its characteristics. First, its diversity is enormous: online video is produced world-wide and under various conditions, ranging from private holiday snapshots to commercial TV shows. Second, semantic concepts are often only linked indirectly to the actual visual content. These phenomena are illustrated in Figure 1, which shows keyframes from online videos tagged with the concept ‘sailing’. The visual appearance of frames varies due to several reasons: first of all, the concept itself is so complex that it can be linked to multiple views, like shots of the boat or ocean views (1(a), 1(b)). Second, the appearance varies greatly among shots of the same kind (1(b), 1(c)). Third, there are shots not directly linked to sailing in a visual sense (1(d),1(e)) and garbage frames that occur frequently in home video (1(e)).

In this paper, we study whether - despite these difficulties - automatic tagging can be learned from online video. Our key contributions are: first, the description of an extensive, publicly available database of real-world online videos associated with 22 user-defined tags. Second, a benchmark of online video tagging that presents quantitative results for several combinations of visual features and strategies of fusing evidence over the shots of a video.

2 Related Work

Though to our knowledge there exists no prior work on video tagging with the focus on online material, we briefly discuss related work in content-based image and video retrieval that forms the basis for our study.

Content-based image retrieval [16] provides a useful pool of techniques and is strongly related to video annotation via the use of keyframes. Particularly, image annotation has been dealt with by modelling latent visual concepts in images [4, 14] or joint occurrences of local descriptors and tags [11]. Also, multiple instance learning methods have been used to detect local features associated with a concept [21]. However, only few prototypes perform this task online and at large scale, e.g. the (ALIPR) server [8].

When it comes to video content, annotation often follows a keyframe extraction for efficiency reasons, leading to an image retrieval problem [11, 20]. However, some approaches employ the video-specific concept of *motion*, for example in form of activity descriptors [19] or spatio-temporal features [2]. Generally, significantly less work can be found for video. One fundamental reason for this are copyright issues that cause a lack of large, publicly available datasets and hence aggravate quantitative comparisons.

A remarkable exception is TRECVID[1], an annually video retrieval contest with the goal of creating a stock of best practice for video retrieval. The contest includes quantitative evaluations on an extensive corpus of news video. In its ”high-level features” task, the automatic annotation of shots is addressed. To boost standardization and comparability of results, groups share large sets of manual annotations , low-level features, and baseline results [17].

3 Database

Our evaluation is done on a database of real-world online videos we downloaded from the video portal *youtube.com*, and is therefore publicly available via the video URLs. These videos are tagged and grouped into semantic categories (like *sports* or *travel & places*) by users during upload. We selected 22 frequent tags and associate a canonic category with each of them, which helps to remove ambiguities that may be present if using tags only: e.g., a search for ‘beach’ returns beach scenes as well as music videos by the Beach Boys. Categories offer a straightforward way to avoid such problems in our analysis.

The youtube API was used to download 75 videos for each tag+category combination, obtaining a database of 1650 videos (total: 145 hrs.). The whole set was separated into a training set (50 videos per tag) and a test set (25 videos per tag). You can find more details, including the complete list of tags and the database itself (all URLs) at *demo.iupr.org/videotagging*.

3.1 Duplicate Removal

Duplicate videos uploaded multiple times by different users pose a problem for our evaluation. We identify them by extracting a signature from each downloaded video. This signature consists of the combined information of the change of color and the amount of motion of the bright and dark centroids between adjacent frames of a video [6]. In contrast to the $YCbCr$-converted frames used in [6], we represent the change of color by the MPEG-7 Color Layout Descriptor (CLD) [10], which proved to be more robust to color gradation. If the edit distance between the signatures of two videos from the same category is below a certain threshold, one of them is considered a duplicate and removed. Most of the duplicate videos could be eliminated using the combined signatures. Videos that were modified by adding additional content to the beginning or end of the video, however, could not be reliably identified as duplicate. Those near-duplicate videos caused suspiciously good results in the tagging experiments, which we used to detect and eliminate them manually.

[1] `http://www-nlpir.nist.gov/projects/t01v/`

3.2 Keyframe Extraction

To cope with the large amount of 145 hrs. of video data, we first reduce each video to a set of representative keyframes (though we enrich our representations with shot-level motion-based descriptors as well). In practice, often the first frame or center frame of a shot is chosen, which causes information loss in case of long shots containing considerable zooming and panning. This is why unsupervised approaches have been suggested that provide multiple keyframes per shot [5, 12]. Since for online video the structure varies strongly, we use a two-step approach that delivers multiple keyframes per shot in an efficient way by following a divide-and-conquer strategy: shot boundary detection – for which reliable standard techniques exist – is used to divide keyframe extraction into shot-level subproblems that are solved separately.

shot boundary detection: shot boundaries are detected using an adaptive thresholding [9] of differences of MPEG-7 color layout features [10].
intra-shot clustering: to determine the keyframes for a shot, we follow an unsupervised approach similar to [12]. We fit a Gaussian mixture model to the descriptors of all frames within a shot using K-Means, and for each mixture component the frame nearest to the center is extracted as a keyframe. The number of components is determined using the Bayesian Information criterion (BIC), which balances the number of keyframes explaining the shot versus the fitting error.

The tradeoff of this simplification is the loss of inter-shot reasoning: for example, in a dialog scene that alternates between two actors, the shot produces many similar keyframes. An additional grouping step might overcome this problem, but we omit it here. The keyframe extraction gives us a set of 75.000 frames for the whole database.

4 Tagging System

The purpose of our system is – given a video X – to return a *score* for each tag t that corresponds to the posterior $P(t|X)$. Figure 2 gives an overview of our setup: in a preprocessing stage, visual features are extracted (see Section 4.1). These features can be *global* frame descriptors based on color, texture, or motion, or based on *local* patches. For both classes, different statistical models are used (Section 4.2).

Since a video may contain many shots, each of them with several keyframes, the evidence gathered from all shots and keyframes must be fused to a global decision, a problem that we tackle using *direct voting* [13]. Optionally, the results for different features can be combined in a late fusion step.

4.1 Visual Features

In the feature extraction step, we scale videos to the same format (320×240) and extract several visual features. The first category contains descriptors for the whole frame and are thus referred to as *global* features:

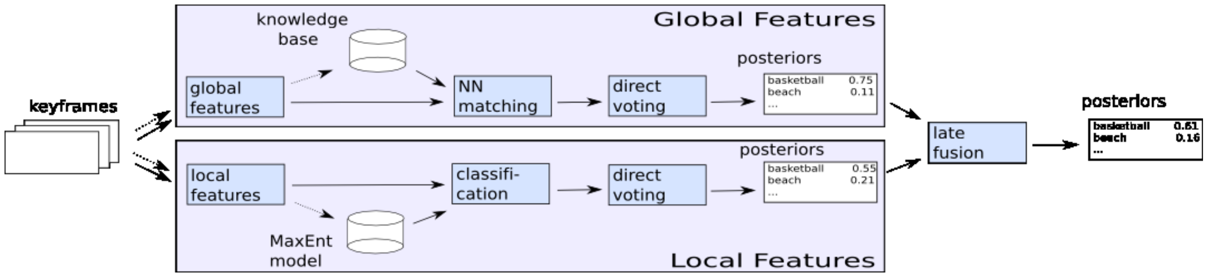
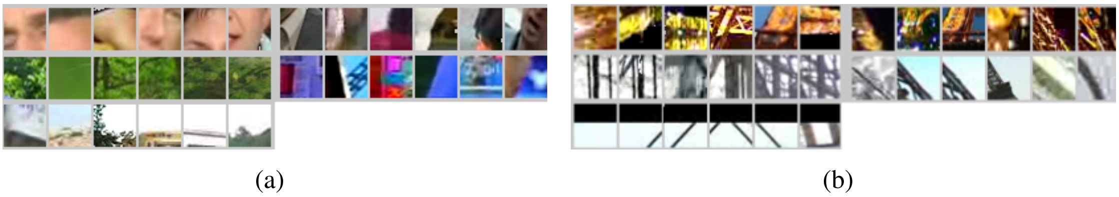


Figure 2: An overview of our tagging system: during offline processing (dashed line) features extracted from a training set are used to train tag models. When tagging a new video (solid line), its features are classified, and evidence from all keyframes is combined using direct voting. Results for local and global features can be combined in a late fusion.

color: RGB color histograms with $8 \times 8 \times 8$ bins
texture: Tamura texture features [18]
motion: some semantic features can be characterized better by their motion pattern than by color or texture. Here, we use a simple compressed domain feature of motion vectors extracted by the MPEG-4 codec XViD[2] to describe *what* motion occurs as well as *where* it occurs. For this, the spatial domain is divided into 4×3 regular tiles, and for each tile a regular 7×7 histogram of all motion vectors in the associated shot is stored (clipped to $[-20, 20] \times [-20, 20]$). The resulting 588-dimensional descriptor is the same for all keyframes in a shot.

While these global features are frequently used in practice, modern recognition systems are often based on collections of local image regions to make them more robust against partial occlusion, deformations, and clutter. Many of these *patch-based* methods have been published recently, among them the 'bag-of-visual-words' model [1, 3, 15, 17]. Here, visual features are clustered according to their appearance, and histograms over the clusters indicate what 'kinds' of features appear with what frequency, an analogy to the 'bag-of-words' model from textual information retrieval. This approach is referred to as *local* in the following:
visual words: a visual vocabulary is learned by extracting and clustering features of overlapping 32×32 patches regularly sampled in steps of 16. To extract the features, patches are transformed to YUV color space, and the discrete cosine transform (DCT) is applied to each channel. From the resulting DCT coefficients, 36 low-frequency components are extracted in a zigzag-pattern for the intensity, and 21 coefficients for each chroma component, obtaining a 78-dimensional descriptor. K-Means clustering in Euclidean space is performed with 500 clusters, five of which are illustrated by a few sample patches in Figure 3(a). Interestingly, it can be seen that some of those visual words can be related to certain semantics, like 'plants' or 'face parts'. During recognition, patches are extracted and assigned to the nearest cluster. A histogram of cluster frequencies is stored as a 500-dimensional feature.

[2] www.xvid.org


(a)

(b)

Figure 3: Left: samples from our visual codebook: 6 patches belong to the same cluster (or 'visual word', respectively). Right: sample patches from 5 of the most discriminative visual words for the tag 'eiffeltower' (patches sampled from eiffeltower images only).

4.2 Statistical Modelling

Given a new video X associated with features $x_1, .., x_n$ each extracted from one of its keyframes, our tagging relies on the estimation of the posterior $P(t|x_1, .., x_n)$. This is done in two steps: first, for each frame, a posterior $P(t|x_i)$ is estimated for each feature x_i, and second, the single estimates are fused to a global posterior $P(t|X)$. For the first step, we use separate strategies for global and local features.

Global Features - NN Matching: for each global feature x_i, a nearest neighbor x' is found among all keyframes in the knowledge base (a kd-tree with Euclidean distance is used for fast matching [13]), giving an estimate $P(t|x_i) \approx \delta(t, t(x'))$.

Local Features - Maximum Entropy: For histograms over visual words, we adapt a discriminative approach based on maximum-entropy that has successfully been applied to object recognition before [1]. The posterior is modeled in a log-linear fashion:

$$P(t|x_i) \propto \exp\left(\alpha_t + \sum_{c=1}^{C} \lambda_{tc} x_i^c\right), \tag{1}$$

where x_i^c is the cth entry in the visual word histogram for frame x_i. The parameters $\{\alpha_t, \lambda_{tc}\}$ are estimated from our training set using an iterative scaling algorithm [1].

Fusing Key Frames - Direct Voting: to tag the whole video, the evidence from all keyframes must be fused to a global posterior $P(t|x_1, .., x_n)$. For this purpose, we propose a simple *direct voting* strategy:

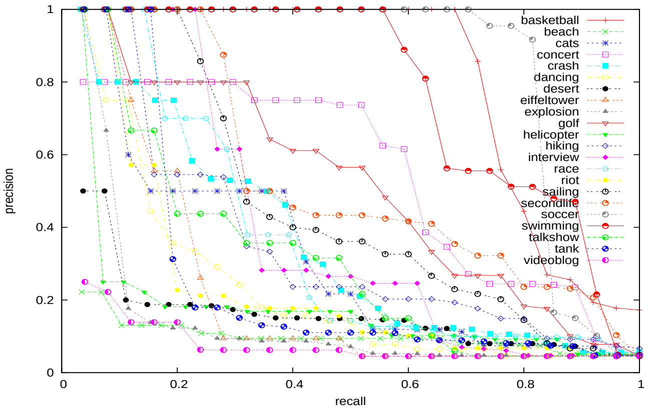
$$P(t|X) = \frac{1}{n} \sum_{1}^{n} P(t|x_i) \tag{2}$$

Direct voting can be seen as an equivalence of the sum rule in classifier combination. Statistical motivations for the approach can be found in [7, 13]. Particularly, in [7] theoretical reasons for the excellent robustness against errors in weak local estimates are described and validated in experimental studies – a property that is important in our context, since many keyframes in $\{x_1, .., x_n\}$ may not be visually related to the true tag and thus give erroneous matches.

Late Fusion - Sum Rule: To combine posteriors obtained from local and global features, we use a *late fusion* by applying the sum rule.

method	MAP
(1) color+texture, EF	0.267
(2) motion	0.170
(3) col+tex+motion, EF	0.290
(4) (col+tex, EF) + mot, LF	0.274
(5) visual words	0.281
(6) (col+tex+mot, EF) + vis. words, LF	0.345
(7) (col+tex, EF) + mot + vis. words, LF	0.345

(a)



(b)

Figure 4: Left: Experimental results in terms of Mean Average Precision (MAP). EF=early fusion, LF=late fusion. Right: the recall-precision curves for our best system (7). The average precision per concept varies between 0.840 (soccer) and 0.117 (beach)

5 Experiments

In our benchmark, we apply several combinations of features and fusion strategies to our test set of 550 videos. We obtain a posterior (*score*) for each combination of test video and tag. By sorting all test videos according to this score, we obtain a ranked list for each tag in which relevant videos should come first. We measure quality in terms of the *mean average precision* (MAP). Our results are subsumed in Figure 4(a) and outlined in the following.

(1) Baseline - Color and Texture: Our baseline system uses color histograms and Tamura features in an *early fusion* (EF), i.e. NN matching is performed on concatenated feature vectors. The nearest neighbors for some frames are illustrated in Figure 5.

(2) Motion: While the baseline gives good results for concepts with a fixed global composition of frames (e.g., soccer, basketball), it performs poorly when the action that takes place is critical. Such action can be captured well using the motion descriptor as outlined in Section 4.1. Though we obtain a worse overall result, the motion descriptor gives a strong improvement for some concepts that come with typical motion patterns like 'interview' (40%) or 'riot' (44%).

(3)+(4) Fusing Baseline and Motion: We combine the baseline features and motion features using early fusion (EF) or late fusion (LF). First improvements relative to the baseline can be achieved (MAP=0.290 for early fusion, 0.274 for late fusion).

(5) Visual Words: We train a discriminative maximum entropy model on all visual word histograms from the training set and use the posterior from Equation (1) as the score. Our results give a slight improvement compared to the baseline (MAP=0.281).

To visualize which patches the model uses to discriminate between classes, we use the fact that coefficients from the model are related to the discriminative power of a visual word: if $\lambda_{tc} \gg \lambda_{t'c} \quad \forall t' \neq t$, the feature c is a strong discriminator for class t. Figure 3(b) illustrates 5

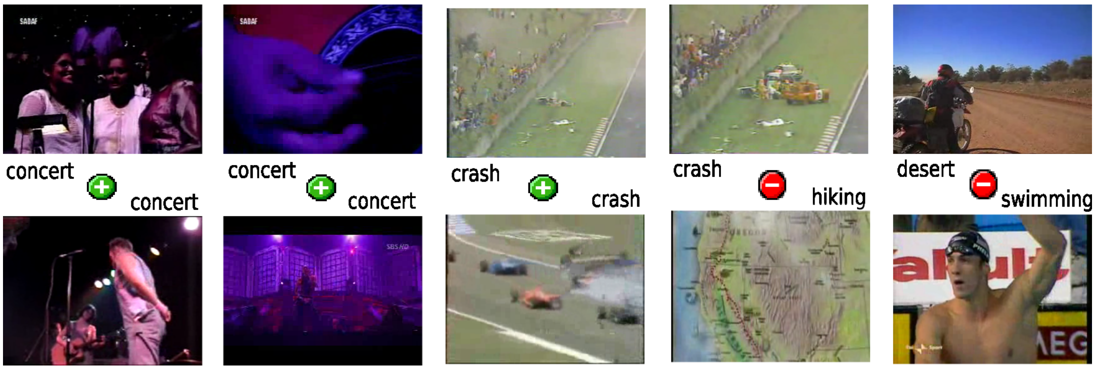


Figure 5: Some sample votes given by our baseline system (1). The upper row shows frames from the test set, below each of them its nearest neighbor in the training set. The three left columns show correct votes, erroneous ones are on the right.

of the 20 visual words that maximize the discriminative power

$$disc_t(c) = \lambda_{tc} - \max_{c'} \lambda_{tc'} \tag{3}$$

for 'eiffeltower', the tag with the strongest improvement (148%) relative to the baseline. These patches indicate how the system learns the visual structure of the tower.

(6)+(7) Best System - All Features: finally, we integrate global features with visual words using a late fusion. We obtain the best results with an MAP of 0.345 (an improvement of about 30% relative to the baseline). The associated recall-precision curves for all concepts are illustrated in Figure 4(b). It can be seen that the MAP varies strongly between concepts: the best results are achieved for sports videos (soccer - MAP=0.83, basketball - 0.80, swimming - 0.77), which often come with a well-defined color layout of the screen. Rather difficult concepts for our approach are associated ill-defined classes of actions (like 'dancing' - 0.21) or widely varying appearance (like 'beach' - 0.10).

Figure 6 illustrates some of the results for the category 'golf' (MAP=0.50), with representative key frames from the top 5 videos of the ranked retrieval list (first row) and from the golf videos with lowest posterior are plotted. While the top 5 show 4 correct samples (2 × golf, 2 × frisbee golf) and one false positive (6(c)), the 5 false negatives show only one typical golf video (6(j)). The others - though tagged with 'golf' - turn out to be difficult special cases: an indoor golf scene (6(f)), a VW Golf (6(g)), a hockey match (6(h)), and a comic (6(i)). It is obvious that finding the correct tag from such content is very difficult using visual features only.

6 Discussion

We have proposed online videos as a source for training models for content-based video retrieval. Our experimental results on a publicly available database of online videos suggest that such training is possible, though a difficult problem, and that the performance varies strongly between concepts.

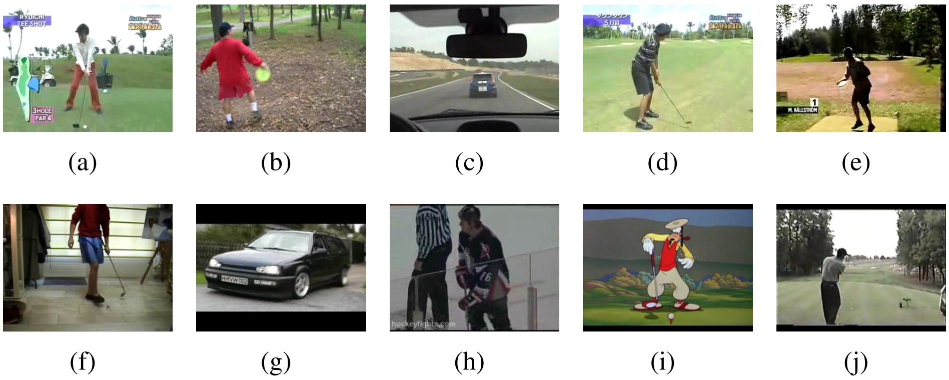


Figure 6: Sample frames from the 5 videos with the highest score for the tag 'golf' (first row) and from the 5 golf videos with lowest score (second row).

Another problem that should be mentioned is a certain redundancy of visual content in the database whose influence is difficult to quantify. Though we remove duplicate videos, still different levels of near-duplicates exist: first, popular scenes are reused and recomposed by different users (e.g., the 'best trick plays in soccer'). Second, *series* of videos often play at the same location or share a similar production style. Our system makes use of such redundancy, but quantifying its influence on tagging performance is difficult, and we have not tackled it yet.

References

[1] Deselaers T. and Keysers D. and Ney H., 'Discriminative Training for Object Recognition Using Image Patches', *CVPR*, pp.157-162, Washington, DC, 2005.

[2] DeMenthon D. and Doermann D., 'Video Retrieval using Spatio-Temporal Descriptors', *ACM Intern. Conf. on Multimedia*, pp.508-517, Berkeley, CA, 2003.

[3] Fergus R. and Fei-Fei L. and Perona P. and Zisserman, A., 'Learning Object Categories from Google's Image Search', *Computer Vision*, Vol. 2, pp.1816-1823, 2005.

[4] Barnard K. and Duygulu P. and Forsyth D. and de Freitas N. and Bleib D. and Jordan M., 'Matching Words and Pictures', *J. Mach. Learn. Res.*, Vol. 3, pp.1107-1135, 2003.

[5] Hanjalic A. and Zhang H., 'An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster-Validity Analysis', *IEEE Trans. Circuits Syst. for Video Tech.*, Vol. 9, No. 8, pp.1280-1289, 1999.

[6] Hoad T.C. and Zobel J., 'Detection of Video Sequences using Compact Signatures', *ACM Trans. Inf. Systems*, Vol. 24, No. 1, 2006.

[7] Kittler J. and Hatef M. and Duin R. and Matas J., 'On Combining Classifiers', *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 20, No. 3, pp.226-239, 1998.

[8] Li J. and Wang J., ‘Real-time Computerized Annotation of Pictures’, *Intern. Conf. on Multimedia*, pp.911-920, Santa Barbara, CA, 2006.

[9] Lienhart R., ‘Reliable Transition Detection in Videos: A Survey and Practitioner’s Guide’, *International Journal of Image and Graphics*, Vol. 1, No. 3, pp.469-286, 2001.

[10] Manjunath B.S. and Ohm J.-R. and Vasudevan V.V. and Yamada A., ‘Color and Texture Descriptors’, *IEEE Trans. on Circuits Syst. for Video Techn.*, Vol. 11, No. 6, 2001.

[11] Feng S.L. and Manmatha R. and Lavrenko V., ‘Multiple Bernoulli Relevance Models for Image and Video Annotation’, *CVPR*, pp.1002-1009, Washington, DC, 2004.

[12] Hammoud R. and Mohr R., ‘A Probabilistic Framework of Selecting Effective Key Frames for Video Browsing and Indexing’, *Intern. Worksh. on Real-Time Img. Seq. Anal.*, pp.79-88, Oulu, Finland, 2000.

[13] Paredes R. and Perez-Cortes A., ‘Local Representations and a Direct Voting Scheme for Face Recognition’, *Workshop on Pattern Rec. and Inf. Systems*, pp.71-79, 2001.

[14] Fei-Fei L. and Perona P., ‘A Bayesian Hierarchical Model for Learning Natural Scene Categories’, *CVPR*, pp.524-531, San Diego, CA, 2005.

[15] Sivic J. and Zisserman A., ‘Video Google: A Text Retrieval Approach to Object Matching in Videos’, *ICCV*, pp.1470-1477, Washington, DC, 2003.

[16] Smeulders A. and Worring M. and Santini S. and Gupta A. and Jain R., ‘Content-Based Image Retrieval at the End of the Early Years’, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, No. 12, pp.1349-1380, 2000.

[17] Snoek C. et al., ‘The MediaMill TRECVID 2006 Semantic Video Search Engine’, *TRECVID Workshop* (unreviewed workshop paper), Gaithersburg, MD, 2006.

[18] Tamura H. and Mori S. and Yamawaki T., ‘Textural Features Corresponding to Visual Perception’, *IEEE Trans. on Systems, Man, and Cybern.*, No. 6, Vol. 8, pp.460-472, 1978.

[19] Vasconcelos N. and Lippman A., ‘Statistical Models of Video Structure for Content Analysis and Characterization’, *IEEE Trans. Image Process.*, Vol. 9, No. 1, pp.3-19, 2000.

[20] Snoek C. and Worring M. and van Gemert J. and Geusebroek J.-M. and Smeulders A., ‘The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia’, *Intern. Conf. on Multimedia*, pp.421-430, Santa Barbara, CA, 2006.

[21] Yang C. and Lozano-Perez T., ‘Image Database Retrieval with Multiple-Instance Learning Techniques’, *Int. Conf. on Data Eng.*, pp.233-243, San Diego, CA, 2000.

On the Creation of Query Topics for ImageCLEFphoto

Michael Grubinger
School of Computer Science and Mathematics, Victoria University
PO Box 14428, Melbourne VIC 8001, Australia
Phone: +61 3 9919 4577 Fax: +61 3 9919 4050
email: michael.grubinger@research.vu.edu.au

Paul D. Clough
Department of Information Studies, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
Phone: +44 114 2222664 Fax: +44 114 2780300
email: p.d.clough@sheffield.ac.uk

Abstract

The selection of realistic and representative search requests (or topics) presents one of the most crucial challenges of benchmark creation: not only should these request be representative for the document collection used, but they should also reflect real user information needs, so that the effectiveness measured with the benchmark will correspond to that one might expect to obtain in a practical setting as well.

In this paper, we present the methodology we used to develop the query topics for *ImageCLEFphoto*, a benchmark event for the evaluation of visual information retrieval from generic photographic collections: first, we carry out a log file analysis to establish a pool of realistic and representative topic candidates for the document collection in question. Based on these topic candidates, we then create a set of representative topics against a number of dimensions to provide an element of control of the topic selection and development processes, and we finally discuss the generation and translation of these topics as well as their distribution to the participants of *ImageCLEFphoto*.

1 Introduction

ImageCLEFphoto is the general photographic ad-hoc retrieval task of the *ImageCLEF* evaluation campaign and provides both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information retrieval from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections as well). The evaluation scenario is thereby similar to the classic TREC[1] ad-hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (*i.e.* topics are not known to the system in advance) [19]. The goal of the simulation is: given an alphanumeric statement and/or sample images describing a user information need, find as many relevant images as possible from the given collection (with the query language either being identical or different from that used to describe the images) [2].

The benchmark resources to carry out such an evaluation typically comprise several components: (1) a document collection, (2) a set of representative search requests (topics), (3) a list of relevance judgements for each topic, and (4) a set of performance indicators to compare the retrieval results. One of the most significant and difficult challenges of creating a benchmark is selecting suitable query topics - structured statements of user needs which consist of a title (a short sentence or phrase describing the search request in a few words), a narrative (a description of what constitutes a relevant or non-relevant image for each request) and three sample images that are relevant to that search request. Deciding on which topics to include in the benchmark is crucial because if they are not representative of the collection, or they differ from real user requests, the effectiveness measured with the test collection will not correspond to that which one might expect to obtain in a practical setting [5].

Thus, there are a number of factors [12] that should be taken into consideration when creating topics for ad-hoc retrieval tasks (such as *ImageCLEFphoto*). For example, topics should:

- reflect real needs of operational systems;
- represent the type of service an operational system might provide;
- be authored by an expert in (or someone familiar with) the subject areas covered by the collection;
- be diverse and allow a good cross-section of the image contents to be covered;
- differ in their coverage, for example broad or narrow topic queries;
- be assessed by the topic author.

The ultimate goal is to "achieve a natural, balanced topic set accurately reflecting real world user statements of information needs" [16].

To achieve this goal for *ImageCLEFphoto*, we used the following methodology to facilitate the topic development process: we first carried out a log file analysis to form a pool of realistic and representative candidate topics for the document collection used (Section 2). Based on these topic candidates, we then created a set of representative search requests against a number of dimensions to provide an element of control over the development and selection process (Section 3), and we finally distributed the completed and translated query topic files to the participants of *ImageCLEFphoto 2006* and *2007* (see Section 4).

[1] Text REtrieval Conference, `http://trec.nist.gov/`

2 User Need Analysis

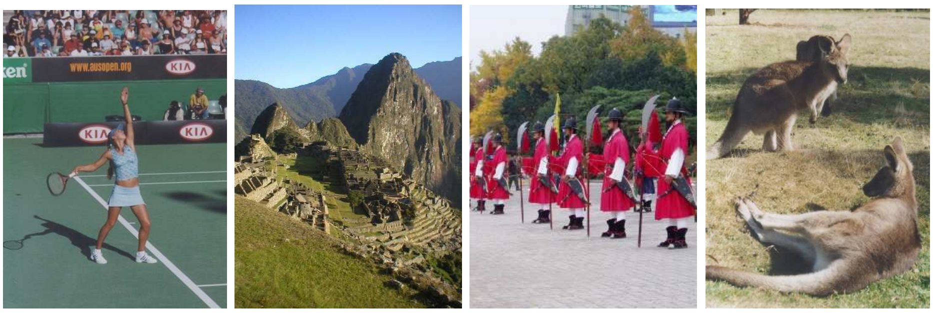
The selection of topics should not only be representative for the collection, but also be based on real-world queries. Although some publications report on how real users query image databases in general (*e.g.* [1, 8], a pre-selection of topics candidates should be based on realistic queries for that particular database in question (which can rarely be provided by such general studies), and images and textual formations should subsequently be chosen for those topic candidates that seem appropriate to compare systems as well [15].

One approach is to carry out a log file analysis to obtain a pool of topic candidates. For example, the topics for the ad-hoc retrieval task from a historic photographic collection at *ImageCLEF 2004* [4] were based on an analysis of a log file from online-access to that collection. Another possibility to create a pool of topic suggestions is to request such possible search queries from searchers or experts familiar with the domain of the document collection. For instance, the topics for the medical ad-hoc retrieval task at *ImageCLEF 2005* [3] were based on a small survey administered to clinicians, researchers, educators, students and librarians at *Oregon Health & Science University* [14].

An alternative but innovative approach for the creation of topic candidates was taken by *INEX Multimedia* in 2006 [21]: the participants were given the image collection and, once familiar with the contents of the collection, they were asked to submit at least six topic candidates following a guide for topic development [12] in order to guarantee realistic and representative search requests (with the participants simulating the real users).

2.1 Document Collection

The document collection of the *IAPR TC-12 Benchmark* [11] provided the resources for *ImageCLEFphoto 2006* and *2007*. This archive contains 20,000 colour photographs taken from locations around the world and comprises a varying cross-section of still natural images. Figure 1 illustrates a number of sample images from a selection of categories.



Sports. Landscapes. People. Animals.

Figure 1: Sample images from the *IAPR TC-12 collection*.

Each image in the collection has a corresponding semi-structured caption in English, German and Spanish, consisting of the following seven fields: (1) a unique identifier, (2) a title, (3) a free-text description of the semantic and visual contents of the image, (4) notes for additional information, (5) the provider of the photo, and fields describing (6) the location and (7) date the photo was taken.

The majority of images have been provided by *viventura*[2], an independent travel company that organises adventure and language trips to South America. Travel guides accompany the tourists and maintain a daily online diary including photographs of trips made and general pictures of each location including accommodation, facilities and ongoing social projects. The collection is publicly available for research purposes and, unlike many existing photographic collections used to evaluate image retrieval systems, this collection is very general in content with many different images of similar visual content, but varying illumination, viewing angle and background. This makes it a challenge for the successful application of techniques involving visual analysis. More information on the design and implementation of the *IAPR TC-12 Benchmark* can be found in [11].

2.2 Search Characteristics

It is important for any benchmark to define the goal of the evaluation before the topic development process is being started ("What exactly do we want to evaluate?"), whereby such goal is ideally based on real user information needs and not a computer vision expert's interest. Only by this means will the evaluation event deliver results that correspond to what a user would expect from a system and systems consequently be optimised for these goals [15]. We therefore decided to model the evaluation scenario of *ImageCLEFphoto* such that it closely corresponds to that of customers and employees of *viventura* requesting images from the *IAPR TC-12 photographic collection*.

Due to the lack of literature to report on search behaviour for this particular scenario, we first set up a logging function in order to monitor the user information needs specific to that collection and to further create a pool of potential topic candidates. The data was collected from 1 February to 15 April 2006, with the log file containing 980 unique queries. Most of the queries were performed in German or Spanish and later translated to English. The average query length for English was 2.45 words, with a standard deviation of 1.61 words; the longest query comprised 12 words and the shortest was one word. According to the log file, the following search characteristics could be identified for retrieval from the *IAPR TC-12 photographic collection*:

- *Query Types:* most of the queries are short noun phrases, often with a place adjunct.
- *Length:* the majority of English queries (59%) is between 2 and 5 words. 37% are single word queries, the rest (4%) is 6 words or longer. The German queries are slightly shorter.
- *Nouns:* people search for both general nouns and proper names.
- *Adjectives:* only a few queries use adjectives, mostly with colour information. Adjectives are mainly used in queries including solely one or two objects, but not in longer ones.
- *Verbs:* only a few queries use verbs to indicate an action.

[2] `http://www.viventura.de/`

- *Geographic constraints:* many queries involve additional geographic information, some with specific descriptions ("*in* La Paz"), while others make use of spatial operators ("*near* Lake Titicaca", "*around* Quito").
- *Prepositions:* used irregularly, some people make use of them ("churches *in* Ecuador"), others do not ("churches Ecuador").
- *Temporal constraints:* people generally do not (yet) look for pictures restricted to certain periods or years.

2.3 Search Patterns

The main search requests are for general and specific tourist destinations, people, landscapes, regions, accommodations, animals, social projects, actions, and specific objects as well as abstract terms. The majority of requests in one of these search areas thereby follows a specific pattern as illustrated in Table 1.

Search Pattern	Example
LOCATION	Rio de Janeiro
COUNTRY	Brazil
REGION	Patagonia
LOCATION - COUNTRY	Rio de Janeiro, Brazil
TOURIST DESTINATION	Mitad del mundo
TOURIST DESTINATION - LOCATION	Mitad del mundo, Quito
TOURIST DESTINATION - COUNTRY	Mitad del mundo, Ecuador
ACCOMMODATION	Host families
ACCOMMODATION - SPECIFICATION	Host families with swimming pool
ACCOMMODATION - LOCATION	Host families near Lake Titicaca
ANIMAL	Boobies
ANIMAL - LOCATION	Boobies in Ecuador
ANIMAL - SPECIFICATION	Blue-footed boobies
ANIMAL - SPECIFICATION - LOCATION	Blue-footed boobies in Ecuador
PEOPLE	Surf instructor
PEOPLE - SPECIFICATION	Godchildren with red cap
PEOPLE - LOCATION	Godchildren in Peru
PEOPLE - PROPER NAMES	André Kiwitz
PEOPLE - PROPER NAMES - LOCATION	André Kiwitz in Botogá
OBJECT	Church
OBJECT - SPECIFICATION	Church with one tower
OBJECT - LOCATION	Church in Ecuador
ACTION	Surfing
ACTION - LOCATION	Surfing in Brazil
SOCIAL PROJECT	Kindergarten project
SOCIAL PROJECT - LOCATION	Kindergarten project in Quito
ABSTRACT TERM	Guerilla
ABSTRACT TERM - LOCATION	Agriculture in Ecuador
LANDSCAPE	Mountain scenery
LANDSCAPE - LOCATION	Mountain scenery in Patagonia

Table 1: Search patterns specific to the *IAPR TC-12 collection*.

Many search patterns thereby exhibit some kind of geographic constraint, which concurs with previous studies for retrieval from general photographic collections [17, 23].

3 Topic Development and Dimensions

The log file analysis did not only offer direct insight into request characteristics and search patterns specific to the *IAPR TC-12 image collection*, it also provided a pool of 980 unique topic candidates which formed the foundation for the topic development process.

To provide an element of control over the selection from these topics candidates, we considered the following dimensions for the selection of the final set of topics that was eventually distributed to the participants of *ImageCLEFphoto 2006*: topic quantity, the estimated number of relevant images, geographical constraints, additional challenges for both concept-based and content-based retrieval, the difficulty of the topic, and feedback from previously held evaluations.

3.1 Topic Quantity

How many topics should be selected from this pool of topic candidates and be given to the participants? The performance of retrieval systems usually varies largely between different topics, and since this variation is in general greater than the performance variation of different retrieval approaches on the same topic, the retrieval performance must be averaged over a large number of versatile topics in order to judge whether one retrieval strategy is (in general) more effective than another [12].

The greater the number of topics, the more confident the experimenter can be in his conclusions. Yet, it is not practical either to include an arbitrarily large number of topics in a retrieval experiment, as each topic requires relevance judgments, which are costly to produce. Even if a large source of topics is available, a compromise between result robustness and assessment effort has to be found to allow for a feasible evaluation.

Many experienced researchers have made suggestions regarding how many topics are sufficient. For example Leung mentioned 20 topics [13] while Spärck-Jones and Rijsbergen [18] found 250 usually acceptable, though little quantitative evidence exists to support these suggestions. In 1998, Voorhees showed that system rankings based on results from less than 25 topics are relatively unstable. TREC has therefore defined 25 as the minimum number for topics, with 50 topics being the preferred default [20].

Consequently, we decided to select 60 topics to represent typical search requests for the *IAPR TC-12 Benchmark*. This number is slightly higher than the preferred default used in TREC in order to further increase the reliability of the retrieval results. To make the task realistic, we took 40 topics directly from the log file (semantically equivalent but perhaps with slight syntactic modification, *e.g.* “lighthouse sea” to “lighthouses at the sea”) and derived 10 further topics from entries in the log file (*e.g.* “straight roads in Argentina” changed to “straight roads in the USA”). The remaining 10 topics were not taken directly from the log file, but based on domain knowledge of the topic authors and created to test various aspects of text and image retrieval (*e.g.* “black and white photos of Russia”).

3.2 Estimated Size of Target Set

While many publications have discussed the appropriate number of topics, little work has covered the number of expected relevant images for each of the topics. Larsen [12] or Clough [4] claim that topic concepts should cover both narrow and broad aspects as well as general and specific queries, which automatically results in a variation of target set sizes as well. In [13], a maximum target size of 15 relevant images for queries on a database with 1000 images is proposed, which seems quite small.

In general, the number of relevant images for a topic should not be too high in order to limit the retrieval of relevant images by chance and to keep the relevance judgment pools to a manageable size. Having the number of relevant images too low, on the other hand, might restrict the use of performance measures; for example, *P(20)* is not very meaningful for a topic with less than 20 relevant images.

Hence, for *ImageCLEFphoto*, we aimed for a target set size between 20 and 100 relevant images and therefore had to further modify some of the topics (broadening or narrowing the concepts). The minimum was chosen in order to be able to use *P(20)* as a performance measure, whereas the upper limit of relevant images should limit the retrieval of relevant images by chance and keep the relevance judgment pools to a manageable size.

3.3 Geographic Constraints

Several previous studies (for example [17, 23]) show that search requests on general real-world databases exhibit a considerable percentage of geographic constraints. Such geographic constraints include:

- place names (*e.g.* Melbourne, Sydney);
- other locators (*e.g.* post code, ZIP);
- adjectives of place (*e.g.* Australian, European);
- terms descriptive of location (*e.g.* state, county, city);
- geographic features (*e.g.* island, lake);
- directions (*e.g.* south, north).

A *geographic query* was consequently defined as a query that includes at least one of these geographic constraints [17]. Therefore, in order to be representative of realistic search requests on real-life collections, evaluation topics should also contain a certain percentage of geographic queries.

Similar to these previous analyses of search log files, we also found many search requests to exhibit some kind of geographical constraint (*e.g.* specifying a location) within the topic candidates for *ImageCLEFphoto*. Therefore, we selected 24 topics with a geographic constraint (*e.g.* "tourist accommodation *near Lake Titicaca*" specifies a location and spatial operator *near*), 20 topics with a geographic feature or a permanent man-made object (*e.g.* "group standing in *salt pan*") and 16 topics with no geography (*e.g.* "photos of female guides").

3.4 Text Retrieval Challenges

Evaluation in concept-based image retrieval should also cover text retrieval challenges in addition to user needs of image archives. In *ImageCLEF*, for instance, particular topics are typically selected to deal with these challenges and further potential problems that are encountered during the translation of the topics as well [3]. Examples for such (multilingual) concept-based retrieval challenges include: dealing with proper names, compound words, abbreviations, morphological variants, idioms, acronyms and equivalent syntactic and semantic expressions [16, 22].

For many of the *ImageCLEFphoto* topics, successful retrieval using concept-based methods will therefore require the use of query analysis (*e.g.* expansion of query terms or logical inference). These reflect examples found in the log files, *e.g.* for the query "group pictures on a beach", many of the annotations will not use the term "group" but rather terms such as "men" and "women" or the names of individuals.

Similarly for the query "accommodation with swimming pool" (also from the log file), the query will result in limited effectiveness unless "accommodation" is expanded to terms such as "hotel" and "B&B". Queries such as "images of typical Australian animals" require a higher level of inference and access to world knowledge (this query is not found in the log file but could be a feasible request by users of an image retrieval system).

Apart from the aforementioned investigation of general versus specific concepts and the additional challenge of vocabulary mismatches between query topics and image annotations, we also offered various other challenges for concept-based image retrieval such as the inclusion of ambiguous terms like "San Francisco" in the topic "people in San Francisco" (which can either refer to the Californian city but also to South American churches consecrated to Francis of Assisi) and the use of abbreviations such as "USA" in the topic "straight roads in the USA".

3.5 Visual Retrieval Challenges

We also classified all topics regarding how "visual" they were considered to be. An average rating between 1 and 5 was obtained, which we based on the retrieval score from a baseline retrieval system (FIRE, see [6] for more information) and on the opinion of three experts in the field of image analysis, who we had asked to rate these topics according to whether content-based image retrieval techniques would produce:

- (1) very bad or random results;
- (2) bad results;
- (3) average results;
- (4) good results;
- (5) very good results.

Based on these findings, we then classified a total of 30 topics as "semantic" (levels 1 and 2) for which visual approaches would be highly unlikely to improve results (*e.g.* "cathedrals in

Ecuador"), 20 topics as "neutral" (level 3) for which visual approaches may or may not improve results (*e.g.* "group pictures on a beach"), and 10 topics as "visual" for which content-based approaches would be most likely to improve retrieval results (*e.g.* "sunset over water").

3.6 Topic Difficulty

One of the most important dimensions of the topic development process for evaluation of visual information retrieval is the appropriate choice of topic difficulty (*i.e.* the difficulty for retrieval systems to return relevant images).

As image retrieval algorithms improve, it is necessary to increase the average difficulty level of topics each year in order to maintain the challenge for returning participants. However, if topics are too difficult for current techniques, the results are not particularly meaningful either. Moreover, it may prove difficult for new participants to obtain good results and prevent them from presenting results and taking part in comparative evaluations. Providing a good variation in topic difficulty is therefore very important as it allows both the organisers and participants to observe retrieval effectiveness with respect to topic difficulty levels.

While quantifying task difficulty is not a totally new concept in the field of visual information retrieval, little work has considered topic difficulty as a dimension for the topic development process: Eguchi *et al.* investigated the topic difficulty for *NTCIR*[3] (*NII Test Collection for IR Systems*) [7].

We also examined the difficulty of the topics and categorised them with respect to another new measure defined by Grubinger [9]: 4 topics were thereby classified as "easy" (*e.g.* "bird flying"), 21 as "medium" (*e.g.* "pictures taken on Ayers Rock"), 31 as "hard" (*e.g.* "winter landscape in South America") and 4 as "very hard" (*e.g.* "tourist accommodation near Lake Titicaca"). See [9] for the exact definition of these topic difficulty levels.

3.7 Feedback of Participants

Another important factor for topic development is the integration of feedback from participants in prior evaluation events. The success of these events is often compared by their number of participants, thus it seems sensible to always develop search topics based on ongoing consultation with past (and potential future) participants.

For example, *ImageCLEF* participants had suggested in prior events that we provided groups of similar topics in order to facilitate the analysis of weak performing queries. We also considered this input in the topic development process for *ImageCLEFphoto 2006* and clustered the topics in groups of up to five topics. An example for topics in one cluster is: "people in bad weather", "destinations in bad weather", "Machu Picchu in bad weather".

[3] `http://research.nii.ac.jp/ntcir/`

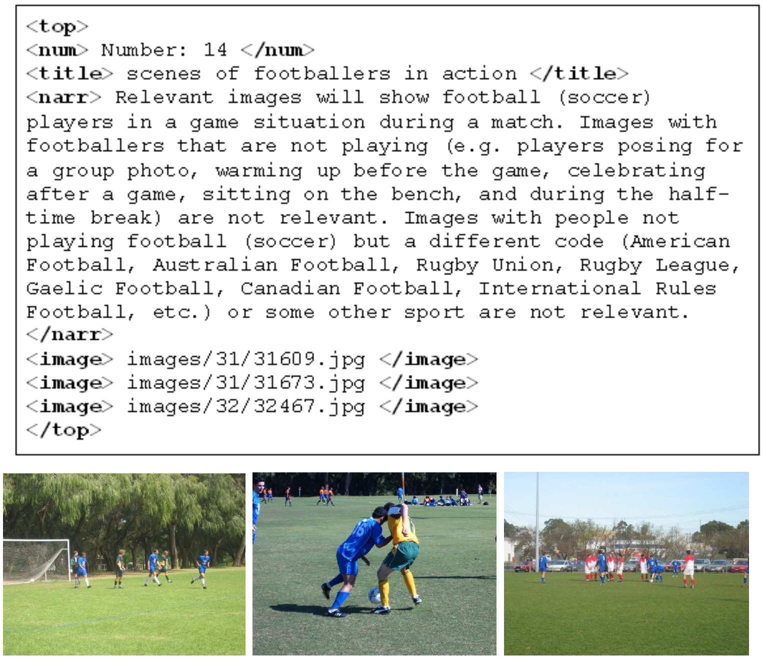
4 Topic Overview and Distribution

Table 2 provides an overview of the 60 query topics representative for the *IAPR TC-12 image collection*, which were eventually released to the participants of *ImageCLEFphoto 2006*. Detailed information on each of these topics as well as the exact distribution of the topics across these dimensions can be found in the Appendix of [9].

ID	Topic Title	ID	Topic Title
1	accommodation with swimming pool	31	volcanos around Quito
2	church with more than two towers	32	photos of female guides
3	religious statue in the foreground	33	people on surfboards
4	group standing in front of mountain	34	group pictures on a beach
	landscape in Patagonia	35	bird flying
5	animal swimming	36	photos with Machu Picchu in
6	straight road in the USA		the background
7	group standing in salt pan	37	sights along the Inka-Trail
8	host families posing for a photo	38	Machu Picchu and Huayna Picchu
9	tourist accommodation near		in bad weather
	Lake Titicaca	39	people in bad weather
10	destinations in Venezuela	40	tourist destinations in bad weather
11	black and white photos of Russia	41	winter landscape in South America
12	people observing football match	42	pictures taken on Ayers Rock
13	exterior view of school building	43	sunset over water
14	scenes of footballers in action	44	mountains on mainland Australia
15	night shots of cathedrals	45	South American meat dishes
16	people in San Francisco	46	Asian women and/or girls
17	lighthouses at the sea	47	photos of heavy traffic in Asia
18	sport stadium outside Australia	48	vehicle in South Korea
19	exterior view of sport stadia	49	images of typical Australian animals
20	close-up photograph of an animal	50	indoor photos of churches or cathedrals
21	accommodation provided by host families	51	photos of goddaughters from Brazil
22	tennis player during rally	52	sports people with prizes
23	sport photos from California	53	views of walls with unsymmetric stones
24	snowcapped buildings in Europe	54	famous television (and
25	people with a flag		telecommunication) towers
26	godson with baseball cap	55	drawings in Peruvian deserts
27	motorcyclists racing at the	56	photos of oxidised vehicles
	Australian Motorcycle Grand Prix	57	photos of radio telescopes
28	cathedrals in Ecuador	58	seals near water
29	views of Sydney's world-famous landmarks	59	creative group pictures in Uyuni
30	room with more than two beds	60	salt heaps in salt pan

Table 2: *ImageCLEFphoto 2006* topics.

Retrieval from multilingual real-life collections such as the *IAPR TC-12 image collection* is inherently multilingual, thus one key part of evaluation in *ImageCLEFphoto* was to provide queries in a language different from that used to describe the images. As a consequence, we translated the topic titles into 15 languages: German, Spanish, French, Italian, Portuguese, Dutch, Russian, Polish, Danish, Swedish, Finnish, Norwegian, Japanese, and Simplified and Traditional Chinese. The topics were translated by native speakers and verified by at least another native speaker. The choice of languages was thereby based on previous submissions to *ImageCLEF* (these 15 languages were exactly the ones that had actually been used in *ImageCLEF* 2005) and on participants' feedback.



```
<top>
<num> Number: 14 </num>
<title> scenes of footballers in action </title>
<narr> Relevant images will show football (soccer)
players in a game situation during a match. Images with
footballers that are not playing (e.g. players posing for
a group photo, warming up before the game, celebrating
after a game, sitting on the bench, and during the half-
time break) are not relevant. Images with people not
playing football (soccer) but a different code (American
Football, Australian Football, Rugby Union, Rugby League,
Gaelic Football, Canadian Football, International Rules
Football, etc.) or some other sport are not relevant.
</narr>
<image> images/31/31609.jpg </image>
<image> images/31/31673.jpg </image>
<image> images/32/32467.jpg </image>
</top>
```

Figure 2: Topic with three sample images.

Figure 2 displays an example for a generated English topic file as eventually distributed to the participants of *ImageCLEFphoto 2006*. The corresponding relevance assessments, retrieval results and topic performance analyses can be found in the *ImageCLEF* overview papers [2, 10].

5 Summary and Future Prospects

This paper reported on a methodology we used to develop the query topics for *ImageCLEFphoto 2006*, a benchmark event for the evaluation of visual information retrieval from generic photographic collections. First, we carried out a log file analysis to establish a pool of realistic and representative topic candidates for the document collection in question. Based on these topic candidates, we created a set of representative topics against a number of dimensions to provide an element of control of the topic selection and development process, and we finally discussed the generation and translation of these topics as well as their distribution to the participants of *ImageCLEFphoto*.

At the *ImageCLEF 2006* breakout session, the vast majority of the participating groups considered the number and difficulty of the topics as appropriate and agreed with the topic creation process being based on several query dimensions. Only two participants pointed out that they found the topics a bit too contrived, while two other participants would have liked to see more than 60 topics for evaluation. Due to this very positive feedback, and to facilitate the comparison of retrieval techniques, we reused the topics created in 2006 also for *ImageCLEFphoto 2007*.

For *ImageCLEFphoto 2008*, however, we are planning to create a new set of query topics. These topics will be based on an updated *viventura* log file (to develop realistic topics) and will again be created against the dimensions introduced in this paper.

References

[1] Linda H. Armitage and Peter G. B. Enser. Information Need in the Visual Document Domain. Technical Report Research and Innovation Report 27, British Library Research and Innovation Centre, London, UK, 1996.

[2] Paul David Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006)*, Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).

[3] Paul David Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeffery Jensen, and William Hersh. The CLEF 2005 Cross–Language Image Retrieval Track. In *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 535–557, Vienna, Austria, September 20–22 2005. Springer.

[4] Paul David Clough, Henning Müller, and Mark Sanderson. Overview of the CLEF Cross–Language Image Retrieval Track (ImageCLEF) 2004. In *Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, number 3491 in Lecture Notes in Computer Science (LNCS), pages 597–613, Bath, UK, September 15–17 2004. Springer.

[5] Paul David Clough and Mark Sanderson. The CLEF 2003 Cross Language Image Retrieval Track. In *Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, number 3237 in Lecture Notes in Computer Science (LNCS), pages 581–593, Trondheim, Norway, August 21–22 2003. Springer.

[6] Thomas Deselaers, Tobias Weyand, Daniel Keysers, Wolfgang Macherey, and Hermann Ney. FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval. In *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum (CLEF 2005)*, volume 4022 of *Lecture Notes in Computer Science (LNCS)*, pages 652–661, Vienna, Austria, September 20–22 2005. Springer.

[7] Koji Eguchi, Kazuko Kuriyama, and Noriko Kando. Sensitivity of IR Systems Evaluation to Topic Difficulty. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume 2, pages 585–589, Las Palmas de Gran Canaria, Spain, May 29–31 2002.

[8] Peter G. B. Enser. Pictorial Information Retrieval. *Journal of Documentation*, 51(2):126–170, 1995.

[9] Michael Grubinger. *Analysis and Evaluation of Visual Information Systems Performance*. PhD thesis, School of Computer Science and Mathematics, Victoria University, Melbourne, Australia, April 2007.

[10] Michael Grubinger, Paul David Clough, Allan Hanbury, and Henning Müller. Overview of the ImageCLEFphoto 2007 photographic retrieval task. In *CLEF Working Notes*, Budapest, Hungary, September 2007.

[11] Michael Grubinger, Paul David Clough, Henning Müller, and Thomas Deselears. The IAPR–TC12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06*, pages 13–23, Genoa, Italy, May 22 2006.

[12] Birger Larsen, Andrew Trotman, Börkur Sigurbjörnsson, Shlomo Geva, Mounia Lalmas, and Saadia Malik. INEX 2006 Guidelines for Topic Development. In *INEX 2006 Workshop Pre-Proceedings*, pages 373–380, Schloss Dagstuhl, Germany, December 18–20 2006. DELOS: Network of Excellence on Digital Libraries.

[13] Clement H. C. Leung and Horace Ip. Benchmarking for Content-Based Visual Information Search. In *Fourth International Conference On Visual Information Systems (VISUAL'2000)*, number 1929 in Lecture Notes in Computer Science (LNCS), pages 442–456, Lyon, France, November 2–4 2000. Springer.

[14] Henning Müller, Paul Clough, William Hersh, Thomas Deselaers, Thomas Lehmann, and Antoine Geissbuhler. Using heterogeneous annotation and visual information for the benchmarking of image retrieval systems. In *Internet Imaging VII*, volume 6061 of *SPIE Proceedings*, page 606105, San Jose, CA, USA, January 16 2006.

[15] Henning Müller, Antoine Geissbuhler, Stéphane Marchand-Maillet, and Paul David Clough. Benchmarking Image Retrieval Applications. In *The Seventh International Conference on Visual Information Systems (VIS'2004)*, pages 334–337, San Francisco, CA, USA, September 2004. Knowledge Systems Institute.

[16] Carol Peters and Martin Braschler. Cross Language System Evaluation: The CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, 22(12):1067–1072, 2001.

[17] Mark Sanderson and Janet Kohler. Analyzing Geographic Queries. In *Online Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004 (GIR '04)*, Sheffield, UK, July 25–29 2004.

[18] Karen Spärck Jones and Cornelis Joost van Rijsbergen. Information Retrieval Test Collections. *Journal of Documentation*, 32:59–75, 1976.

[19] Ellen M. Voorhees. Overview of the Seventh Text REtrieval Conference (TREC–7). In *Proceedings of the Seventh Text REtrieval Conference (TREC–7)*, pages 1–24, Gaithersburg, MD, USA, November 9–11 1998. Department of Commerce, National Institute of Standards and Technology.

[20] Ellen M. Voorhees and Donna Harman. Overview of the Sixth Text REtrieval Conference (TREC–6). *Information Processing and Management*, 36(1):3–35, 2000.

[21] Thijs Westerveld and Roelof van Zwol. Benchmarking Multimedia Search in Structured Collections. In *INEX 2006 Workshop Pre-Proceedings*, pages 313–320, Schloss Dagstuhl, Germany, December 18–20 2006. DELOS: Network of Excellence on Digital Libraries.

[22] Christa Womser-Hacker. Multilingual Topic Generation within the CLEF 2001 Experiments. In *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001)*, number 2406 in Lecture Notes in Computer Science (LNCS), pages 389–393, Darmstadt, Germany, September 3–4 2002. Springer.

[23] Vivian Zhang, Benjamin Rey, Eugene Stipp, and Rosie Jones. Geomodification in Query Rewriting. In *Online Proceedings of the Third Workshop on Geographic Information Retrieval at SIGIR 2006 (GIR'06)*, Seattle, WA, USA, August 10 2006.

Towards a Region-Level Automatic Image Annotation Benchmark[*]

H. Jair Escalante, Manuel Montes and L. Enrique Sucar

Computer Science Department

National Institute on Astrophysics, Optics and Electronics,

Luis Enrique Erro # 1, Puebla, México,72840

e-mail: {hugojair, mmontesg, esucar}@ccc.inaoep.mx

Michael Grubinger

School of Computer Science and Mathematics,

Victoria University, Australia

PO Box 14428, Melbourne VIC 8001, Australia

email: michael.grubinger@research.vu.edu.au

Abstract

Automatic image annotation at region-level consists of the task of assigning labels to regions within segmented images. This is a very important task for the development and improvement of image retrieval methods. Many image annotation methods have been proposed so far, reporting relatively good results on this task. However the lack of benchmark collections have caused that region-level methods are often evaluated using image-level collections. On the other hand, a few region-level collections are available, although these are small and composed of unrealistic images; which difficult the evaluation of the expected annotation performance on real collections. In this paper we propose the creation of a benchmark collection for the evaluation of region-level automatic image annotation methods, in order to provide a valuable resource to the image annotation and retrieval communities. We describe a methodology for creating such a benchmark and present some work we have performed towards building it; work in progress and future directions are also discussed. The main goal of this paper is to obtain feedback from the benchmarking community as well as to establish collaborations in order to carry out this exhaustive, but very important, task.

[*] We would like to thank Allan Hanbury because of the facilities and support for the publication of this paper.

1 Introduction

The task of assigning semantic labels (words) to images is known as image annotation. This is a very important step towards developing more effective image retrieval systems. For text-based image retrieval, annotations are indispensable features; while for content-based image retrieval methods, annotations can provide them with semantic information for improving their performance. There are two ways of facing this problem, at image-level and at region-level. In the first case, labels are assigned to the entire image as an unit, not specifying which words are related to which objects within the image. In the second approach, which can be conceived as a high-level object recognition task, the assignment of labels is at region level; providing a one-to-one correspondence between words and regions. The last approach can provide more semantic information for the retrieval task, although it is more challenging than the former. Image annotation, however, is not an easy task; manual annotation is both infeasible for large collections and subjective due to annotator's criteria. In consequence, there is an increasing interest on developing automatic methods for image labeling.

In the last few years many interesting methods have been proposed for automatic image annotation (*AIA*) [7, 16, 23, 2, 8, 1, 4, 22, 6, 5, 19, 17, 21, 18]. Most of these approaches have reported relatively good results on the *AIA* task, however the evaluation criterion used in most of these works make their results not very reliable. Usually, evaluation of *AIA* methods is carried out using small collections of unrealistic images. Furthermore, the performance assessment of most of the region-level methods has been done using collections and evaluation protocols designed for image-level *AIA*, which can not provide a reliable estimation of the methods' accuracy. These sort of evaluations are done because of the lack of a reliable benchmark for the task of *AIA*. Until now, only a few region-level *AIA* collections are available. However collection's size, images' type and even copyrights make these collections not completely accessible and reliable for benchmarking. In this paper we propose the creation of a region-level *AIA* benchmark, which can also be used for evaluating image-level methods. Specifically, we propose the manual annotation at region-level of the *IAPR-TC12 benchmark*, a recently released collection of photographs manually annotated at image-level [12]. We describe some work already carried out for images' segmentation and feature extraction, as well as available tools for manual annotation. Furthermore, we describe work in progress and research issues that should be considered for the development of a reliable benchmark. The main goal of this paper is obtaining feedback from the benchmarking community that can help us for enhancing the reliability of the proposed benchmark; also we are looking for collaborations because this is a very exhaustive task.

The rest of this paper is organized as follows. In the next section we describe the motivation for building a benchmark on *AIA* at region-level. Then in Section 3 we describe the proposed methodology for the full annotation of the *IAPR-TC12 benchmark*. Next in Section 4 research advances are presented. Finally in Section 5 we summarize the proposal and discuss important issues that should be considered when creating this benchmark.

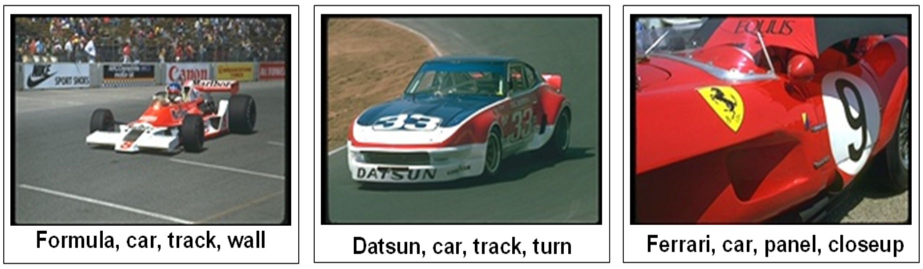


Figure 1: Sample images belonging to the concept *Auto Racing* are shown.

2 Motivation

AIA is a relatively new research area with the goal of providing visual-semantics into the image retrieval task. Many interesting approaches have been proposed [23, 2, 8, 1, 4, 22, 6, 5, 19, 17, 21, 18], based on statistical and probabilistic models [23, 2, 8, 1, 4, 5, 19], information retrieval models [22, 17, 21, 18, 15] and supervised learning [6, 9]. Most of these methods have been evaluated using different subsets of the Corel^R image collection (annotated at image-level) and adopted a protocol proposed by Duygulu et al [8]. However, while this protocol may be some reliable for image-level annotation, region-level methods can not be reliably evaluated by using it.

2.1 Current image collections

Seminal papers on *AIA* are due to Mori and Duygulu et al, with the introduction of the co-occurrence and machine translation models, respectively [23, 8]. The image collection used by Duygulu et al in their reference work as well as the evaluation protocol proposed became an standard for the evaluation of *AIA* methods [16]. For their experiments, subsets of the Corel image collection were used [8]. The Corel collection consists of around 800 CD's, each containing 100 images related to a common semantic concept. Each image is accompanied by a few keywords describing the semantic or visual content of the image. For example, in Figure 1 sample images belonging to a common concept are shown.

Besides the Corel collection is large enough for obtaining significant results. There are several problems with this collection because of its commercial nature that make it not a reliable collection. First, images are unrealistic because most of them were taken by professional photographers in difficult poses and under controlled situations. Second, it contains the same number of images related to each of the semantic concepts, as a result it is a balanced collection. Third, Corel images are annotated at image level, limiting its applicability to image-level methods. Finally, given that the collection is commercial it is copyright protected and as a result images can not be distributed among researchers. Furthermore, the collection is no longer available hindering the evaluation for some methods.

Recently, a few efforts have been carried out towards developing a benchmark collection for the evaluation of *AIA* methods [14, 3, 5, 13]. Hanbury et al *re-annotated*, at image-level, a large subset of almost 60,000 images as containing animals or not, in most of these images the name of the animal is also provided [14]. For region-level annotation Hanbury et al provided a collection of 1289 manually segmented images of animals in which each segment was annotated. Barnard et al also presents a dataset of 1041 manually segmented images annotated at region-level too [3]. Images correspond to a broader concept domain than that considered by Hanbury et al; furthermore, Barnard et al considered WordNet, a semantic network, and a established methodology for the annotation process [3, 13]. Each region is therefore annotated according to a set of rules based on concepts and their synonyms as defined in WordNet. Carbonetto et al provides smaller subsets of Corel images annotated at region-level [5]. Other benchmarks from the object recognition community are useless for evaluating *AIA* algorithms because images in such collections are also unrealistic and are limited to only a few objects[1]. Current segmentation algorithms are far away of providing accurate segmentations, and therefore performance evaluation on manually segmented images is not a reliable estimator of the performance of *AIA* methods on real scenarios. Furthermore the size of the above described collections is very small and all of the images are still unrealistic. In consequence a large collection of automatically segmented realistic images is needed.

2.2 Evaluation of AIA methods

The evaluation protocol introduced by Duygulu et al has been used by most of *AIA* methods proposed [7, 16, 23, 8, 1, 4, 22, 19, 17, 21, 18]. It is designed to evaluate annotation performance by assessing image retrieval efficacy using automatically generated annotations. Intuitively, this protocol provides a measure of the labels overlap between the generated image-level annotations and the original image-level labels of the images. The protocol consist of splitting the image collection into training and test sets. The first set is used by the *AIA* methods for learning and training, then images in the test set are annotated using the trained method. Next, queries are defined by considering the set of labels in the test-set vocabulary. Queries consist of combinations of one, two, three and four keywords in such vocabulary. These queries are used for retrieving images, from the test collection, by considering the automatically generated labels. An image is said to be relevant to a query if the retrieved image includes the query in the original (manual) annotation of such image. Standard evaluation measures like precision and recall are used for evaluating the retrieval performance. As we can see annotation accuracy at region level can not be effectively evaluated under this protocol. Because we can never know if the annotation method is accurately assigning annotations to regions or, instead, if the method performed well by chance. This evaluation methodology can be useful for evaluating *AIA* methods at image-level; but even when it can be useful for giving an idea of the performance of region-level *AIA* methods, it can not be considered for reliable evaluations.

A more reliable methodology has been considered for the evaluation of annotation methods by Carbonetto et al [5]. They considered subsets of the Corel collection annotated at region-level. Under Carbonetto's methodology a measure of the percentage of correctly annotated regions is considered. This is a reliable estimate because we can know how many regions were correctly annotated. Instead of just measuring accuracy at image-level. The problem with this

[1] See for example the **ALOI** collection and the **Caltech** repository.

approach is that image collections annotated at region-level are required. Peter Carbonetto have made available small subsets of the Corel collection annotated at region-level. However these datasets have the same drawbacks of any Corel subset, namely the unrealistic nature of images.

3 Towards and AIA benchmark

Our proposal for creating a benchmark on *AIA* consists of annotating at region-level a large collection of realistic images. Specifically we consider the *IAPR-TC12* benchmark [12] an actual standard for the evaluation of image retrieval methods. We selected this collection because it has already two desirable properties for an *AIA* benchmark, namely its size (large enough for obtaining significant results) and the images source (realistic photographs). The *IAPR-TC12* collection consist of 20,000 images annotated at image-level. Images consist of photographs taken by tourist around several places in the world. Annotations are available in English, German and Spanish. These were generated following a strict methodology and annotation rules [12]. The *IAPR-TC12* benchmark is copyright protected and it is available only for academic and research purposes.

The proposed methodology for creating the *AIA* benchmark consist of the following steps.

1. Segmentation
2. Feature extraction
3. Defining vocabulary
4. Defining annotation rules
5. Manual annotation of images
6. Publication of the benchmark

Since our objective is to provide a benchmark that can be used for evaluating both region-level and image-level *AIA* methods, we should obtain regions for each image in the collection. For this purpose we propose to use automatic segmentation methods because of the size of the collection, and because manual segmented collections can not give a reliable estimate of method's accuracy in real collections. We have considered two segmentation approaches, the first one is based on the normalized cuts algorithm [24], and the second one consists of splitting images into squared patches (grid segmentation). The first method is the most used algorithm in the *AIA* literature [8, 2, 1, 17, 21, 18, 19, 4], while with the second, better results have reported by *AIA* methods [5, 9].

Once the collection have been completely segmented visual attributes should be extracted from each region. For this step we propose extracting a large number of attributes from the regions, including color, texture, shape and orientation information. A large number of features will allow benchmark's users to perform feature selection in order to obtain the best set of features for their annotation methods.

A crucial step towards creating the benchmark consists of defining the annotation vocabulary for the collection, that is the set of labels to choose from for assigning them to regions. We may consider one of the annotation approaches identified by Hanbury: *free text*, *keywords* or *classification based on ontologies* [13]. *Free text* descriptions is one of the most easiest ways of annotation, because the user can annotate regions according to its own knowledge and vocabulary. However under this approach the same object can be labeled with different annotations and other inconsistencies may arose. The *keyword approach* is the most used in *AIA* collections, Corel for example uses it. Under this approach a predefined vocabulary of keywords (or even arbitrary keywords) are used for annotating images. This strategy is a good candidate for defining the benchmark vocabulary. The last approach consists of using *semantic networks* for assigning labels, this approach is similar to that of using keywords with the difference that annotations are arranged into a hierarchy of concepts, resulting in a semantically annotated collection.

The selection of a keyword vocabulary is ongoing work. Actually we are seriously thinking on using the list of keywords identified by Hanbury in a recent study [13]. This study comprises the analysis of the labels' vocabulary used in several *AIA* collections. Hanbury selected a set of keywords and arranged them according to an ontology [13]. The next step consists of defining a consistent methodology for the annotation process in order that several annotators could participate in the annotation task. This is another important step because of it depends the objective annotation of the collection and the consistency of the benchmark. We intend to establish several annotation rules, like those proposed by Barnard et al [3].

The main reason for creating a benchmark like this is to provide researchers with the segmented collection as a tool for the evaluation of their systems. Therefore the benchmark will be made publicly available for academic and research purposes, of course, under the agreement of the actual copyright owners of the collection. A future work extension could be the proposal of a track within the *ImageCLEF* forum for the evaluation of region-level image annotation systems.

4 Research status

At the moment we have segmented the *IAPR-TC12* collection into regions using both the normalized cuts algorithm and the grid approach. A total of 100,000 regions resulted from the segmentation with normalized cuts and around 480,000 regions are available under the grid segmentation approach (considering 24 patches per image, of course the size of the patches in the grid approach can be modified for obtaining smaller or bigger regions). Segmentation is not computational expensive, it may took almost 2 days for segmenting the entire collection. Sample segmented images from *IAPR-TC12* collection are shown in Figure 2. As we can see good image partitions can be obtained with the normalized cuts method, while other segmentations can be poor with this approach. The grid method always shows the same performance.

Simultaneously, when images were segmented visual attributes were also extracted from each of the regions. The set of attributes considered consisted of color, texture, shape and orientation information. Therefore, each of the regions is described by a vector of attributes. The vectors' size varies because we used different tools for the distinct segmentation algorithms. Though in the near future we will extract the same patterns from both segmentations.

For the process of segmentation and feature extraction we have used two recently developed software tools. The first one is an interface developed at our institution that uses the normalized cuts algorithm for segmenting a given image collection [20]. This tool also includes methods

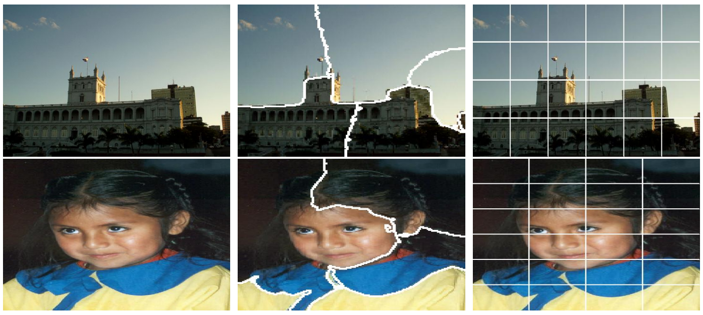


Figure 2: Sample images from the *IAPR TC-12* collection. From left to right, original image, image segmented with normalized cuts and image segmented with the grid approach.

for feature extraction and manual annotation of regions. Additionally the interface provides options for the re-segmentation of images and for joining adjacent regions annotated with the same label. A second interface for segmentation and manual annotation is that provided by Perter Carbonetto, this tool also includes options for segmentation with both normalized cuts and the grid approach. Further, this tool provides methods for soft-annotation of images, that is several labels can be assigned to each region.

Regarding the manual annotation step we have performed a first attempt for the annotation of the *IAPR-TC12* collection, however some issues were encountered. Four our participation in the photographic retrieval task at *ImageCLEF2007* we decided to make use of *AIA* methods for improving accuracy of retrieval [10]. For this purpose we created a training set of manually annotated regions. Randomly, a small subset of the collection segmented with normalized cuts was selected and manually annotated according to a defined vocabulary a keywords. For defining this vocabulary we looked at the textual part of the task's topics [11]. The keyword-annotations were defined according to a handmade ontology, built according to the nouns appearing in the topics. In Table 1 the label's vocabulary proposed for annotating images is shown, as well as the keywords related to each of the labels. As we can see some keywords like *building* and *person* comprise many concepts, though several other are not associated to any other keyword. This fact together with poor segmentation made difficult the process of annotation. Using this pseudo-ontology a small subset of images was annotated and used for training and *AIA* method in order to annotate the rest of the segmented collection. The automatically generated labels were used for expanding queries and documents in the ad hoc retrieval task at *ImageCLEF2007* [10].

Our experience at *ImageCLEF2007* give evidence that poor segmentation as well as the wrong definition of the annotation vocabulary can difficult the annotation process. The segmentation problem can be alleviated by using the grid segmentation approach. Because, even when segmentation is poor, it is consistent through any type of images. While segmentation algorithms like normalized cuts have irregular performance depending on the images. The vocabulary definition problem can de addressed by defining a consistent vocabulary of keywords, based on semantic knowledge (just as the keyword list proposed by Hanbury [13]). On the other hand, results obtained in the retrieval task give evidence that the use of annotation for image retrieval can help improving retrieval performance, although some issues should be addressed.

Keyword	**Related keywords**
animal	fish, bird, reptile, kangaroo, . . ., seals, sea lions (any animal)
boat	-
building	accommodation, tourist-accommodation, hotel, hostel, cities, stadium, school, bridge, grandstand, ruin, wall
church	mosque, cathedral
clouds	fog
flag	-
furniture	bed, tv, room
grass	football, ground, sports-field
mountain	landscape, volcanoes, sights
other	-
person	group of persons, people, footballers, players, families, god daughter, tennis player, god children, god son, guide, woman, girls
plate	meat dish, dish
prize	medals, trophies, cups
road	straight road, highway, square, street
sand	salt pan, beach, desert, salt heap, salt pile
sky	-
snow	winter
statue	monument
sun	sunset
swimming-pool	-
tower	telescope, lighthouse
trees	-
vehicle	motorcycle, car, buses, bicycles, forklifts, trains
water	sea, lake, waterfall, river

Table 1: Annotation vocabulary defined for the creation of a training set of annotated regions for *INAOE-TIA's* participation at ImageCLEF2007 [10].

Work in progress consists on the definition of the annotation vocabulary and the methodology to follow for the annotation process. We are requesting feedback and collaborations for both tasks[2].

5 Conclusions

We have proposed the creation of a benchmark for the evaluation of region-level image annotation methods. Our proposal comprises the segmentation, feature extraction and annotation at region-level of images in the *IAPR-TC12* collection. The main goal of this paper is to obtain feedback from the benchmarking community in order to create a robust standard collection. We have already segmented the collection using normalized cuts and a grid approach, and visual features have been extracted from each of the resulting regions. Furthermore, software tools available for manual annotation have been briefly described. Currently we are in the process of defining the vocabulary of keywords allowed for annotation. The next stage of the project will consist of annotating the generated regions, a slow and time-consuming process that should be done some day and start as soon as possible.

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003. 2, 3, 4, 5

[2] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, volume 2, pages 408–415. IEEE, IEEE, 2001. 2, 3, 5

[3] Kobus Barnard, Quanfu Fan, Ranjini Swaminathan, Anthony Hoogs, Roderic Collins, Pascale Rondot, and John Kaufhold. Evaluation of localized semantics: Data, methodology and experiments. *International Journal of Computer Vision (to appear)*, 2007. 3, 4, 6

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127 – 134. ACM press, 2003. 2, 3, 4, 5

[5] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general context object recognition. In *Proceedings of the 8th European Conference on Computer Vision*, pages 350–362, 2005. 2, 3, 4, 5

[6] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. on PAMI*, 29(3):394–410, 2007. 2, 3

[7] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval - approaches and trends of the new age. In *Proceedings ACM International Workshop on Multimedia Information Retrieval*, Singapore, 2005. ACM Multimedia. 2, 4

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision*, volume IV of *LNCS*, pages 97–112. Springer, 2002. 2, 3, 4, 5

[9] H. J. Escalante, M. Montes, and L. E. Sucar. Word co-occurrence and mrf's for improving automatic image annotation. In *In Proceedings of the 18th British Machine Vision Conference (BMVC 2007) To appear*, Warwick, UK, September, 2007. 3, 5

[2] We will appreciate any comment, suggestion and support offer, if interested please contact the first author of this paper via e-mail.

[10] H. Jair Escalante, C. A. Hernandez, A. Lopez H. Marin-Castro, E. Morales, L. E. Sucar, M. Montes, and L. Villasenor. Inaoe-tia participation at imageclef2007. In *Working Notes of the CLEF (to appear)*, Budapest, Hungary, 2007. CLEF. 7, 8

[11] M. Grubinger, P. Clough, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007. 7

[12] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. 2005. 2, 5

[13] Allan Hanbury. Review of image annotation for the evaluation of computer vision algorithms. Technical Report 102, PRIP, Vienna University of Technology, 2006. 3, 4, 5, 6, 7

[14] Allan Hanbury and Alireza Tavakoli Targhi. A dataset of annotated animals. In *Proceedings of the Second MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation*, Czech Repulic, 2006. 3, 4

[15] J. S. Hare, P. H. Lewis, P. G.B. Enser, and C. J. Sandom. A linear-algebraic technique with an application in semantic image retrieval. In H. Sundaram, editor, *ACM International Conference on Image and Video Retrieval, CIVR*, volume 4071 of *LNCS*, pages 31–40. ACM, Springer-Verlag, 2006. 3

[16] J. S. Hare, P. H. Lewis, P. G.B. Enser, and C. J. Sandom. Mind the gap: Another look at the problem of the semantic gap in image retrieval. In Hanjalic A. Chang, E. Y. and Eds. Sebe, N., editors, *Proceedings of Multimedia Content Analysis, Management and Retrieval*, volume 6073, San Jose, California, USA, 2006. SPIE. 2, 3, 4

[17] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, New York, NY, USA, 2003. ACM Press. 2, 3, 4, 5

[18] J. Jiwoon and R. Manmatha. Using maximum entropy for automatic image annotation. In P. Enser, editor, *Procc. international conference on image and video retrieval (CIVR 2004)*, volume 3115 of *LNCS*, pages 24–32, Dublin IR., 2004. Springer. 2, 3, 4, 5

[19] V. Lavrenko, R. Manmatha, and J.Jeon. A model for learning the semantics of pictures. In Sebastian Thrun, Lawrence Saul, and Bernhard Scholkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004. 2, 3, 4, 5

[20] H. Marin-Castro, L. E. Sucar, and E. F. Morales. Automatic image annotation using a semi-supervised ensemble of classifiers. In *In Proceedings of the 12th Iberoamerican Congress on Pattern Recognition (CIARP 2007), To appear*, 2007. 6

[21] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proceedings of International Conference Image and Video Retrieval*, volume 3115 of *LNCS*, pages 42–50. Springer, 2004. 2, 3, 4, 5

[22] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear. 2, 3, 4

[23] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management (in conjunction with ACM Multimedia Conference 1999)*, 1999. 2, 3, 4

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 5