

Chapter 8

Conclusion

This last chapter summarises the original work presented in this dissertation, recalls the scientific contributions and explains the limitations of this research.

8.1 Summary

This dissertation has investigated the system-centred evaluation of (multilingual) VIR from generic photographic collections and is composed of eight chapters, including the introduction and this concluding chapter; the remaining six main chapters are briefly summarised below.

Characteristics and Processes of Visual Information Retrieval

In recent years, there has been an increasing amount of literature on the main concepts and challenges of VIR. An unsolved problem to date is the so called semantic gap, which is the discrepancy between the information that one can automatically extract from visual data and the interpretation of the same data for a user in a given situation.

Research endeavours to bridge the semantic gap have thereby taken two contrary approaches: content-based image retrieval (CBIR) is based on purely visual features (such as colour, texture and shape) that can be directly extracted from images, while concept-based image retrieval (TBIR) relies on meta-data or additional alphanumeric representations associated with the images to express their semantics.

We provide an analysis and classification of visual information queries, similarity measures and the result generation process.

Analysis and Evaluation of Visual Information Retrieval

For the field of visual information search to advance, objective evaluation to identify, compare and validate the strengths and merits of different systems is essential. Uniform sets of data, queries, relevance judgments and measures of performance are therefore needed to provide a standardised platform (called benchmarks or test collections) to carry out such an evaluation, together with evaluation events to also attract researchers to make use of these components.

Such benchmarks have recently been developed (and evaluation events have been organised) for several domains of VIR, including the retrieval from historic or medical collections, object recognition and automatic annotation tasks for general collections as well as for specific ones like coin images or radiographs, user-centred evaluation of systems and also in related fields such as video retrieval, cross-language information retrieval and multimedia retrieval from structured (XML) collections. No efforts, however, had considered the evaluation of multilingual 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, *e.g.* pictures of holidays and events).

The goal of this research was therefore fill this gap by designing and implementing the required resources to carry out such an evaluation: the *IAPR TC-12 Benchmark*. These resources include (1) the design and development of a standardised image collection for this domain, (2) the creation of representative search topics and relevance judgments to associate a ground-truth of relevant images for each of these topics, (3) a set of performance measures to quantify, rank and evaluate the results, and (4) the organisation of an evaluation event to practically apply these components and provide them to the research community.

Data Design and Engineering

A core component of image retrieval benchmarks is a set of images that are representative of a particular domain. Finding such resources for general use is often difficult, not least because of copyright issues which restrict the distribution and future accessibility of data. This is especially true for visual resources that are often expensive to obtain and subject to limited availability and access for the research community.

We therefore report on the creation of an image collection called the *IAPR TC-12 image collection*, which we specifically designed and implemented to deal with the lack of resources for evaluation of VIR from generic photographic collections. The goal was to provide:

- a collection of general, real-world photographs suitable for a wider range of evaluation purposes;
- images with associated written information representing typical textual metadata to allow for the exploration of the semantic gap;
- semantic image descriptions in multiple languages as such real-life collections are inherently multilingual;
- a data set that is free of charge and copyright restrictions and therefore available to the general research community.

To achieve these goals, we first specified the requirements for the creation of such a collection, including the definition of rules for the *image selection* and *annotation* processes, which would subsequently allow for the strict control over the consistency and quality within all aspects of collection creation. We then acquired access to an image database of general photographs (photos of travel destinations, tourists and events) and, following the rules, we selected 20,000 images and annotated them in three languages: English, German and Spanish.

Task Creation and Visual Information Complexity

The second key component of the *IAPR TC-12 Image Benchmark* is a set of representative *search requests* (query topics). The specific goal was to develop a natural, balanced topic set accurately reflecting real world user statements of information needs for retrieval from the *IAPR TC-12 image collection*.

In general, such statements of user information needs are created against certain task parameters (dimensions) to allow for some control over the topic creation process. Thus, we first identified the dimensions specific to retrieval evaluation using the *IAPR TC-12 image collection*, which include the total number of topics provided and for each topic: the estimated number of relevant documents (images), the topic scope (*e.g.* broad or narrow, general or specific) and origin, the use of geographic constraints, the representation completeness of relevant images, the estimated difficulty, the likelihood of retrieval success using visual features only, and supplementary task creation parameters such as additional text retrieval challenges and feedback from participants.

To base the topic creation process on realistic user information needs, we first implemented a logging function for a web-based interface to the *IAPR TC-12 image collection* and subsequently analysed the search behaviour and query patterns specific to retrieval from this database. Based on the topic candidates following the results from the log file analysis, we then created a set of representative query topics against the aforementioned query dimensions.

No work had considered the topic difficulty for TBIR. To be able to also balance the query topics for difficulty, we designed a novel measure to quantify topic difficulty for TBIR based on both linguistic features of the topic and statistical information gained from the corresponding document collection. Experimental validation and a comparison with other approaches showed that the novel measure displays a strong negative correlation between topic difficulty and system effectiveness and gives an upper boundary of the correlation which can be achieved using a costly manual approach. We purport that having such an accurate measure en-

ables the creators of TBIR evaluation events to carefully select topics, making topic difficulty one of the most significant dimensions in the topic creation process.

Parametric Benchmark Design and Architecture

To facilitate the incremental development as well as the ongoing maintenance and administration of the benchmark collection (*i.e.* images and their corresponding semantic descriptions) and the creation and administration of the representative query topics, we designed and implemented a benchmark administration system.

The most significant benefit of this novel benchmark architecture can be found in its *parametric* nature, which allows for a fast adaptation to changed retrieval requirements or new evaluation needs. Collection parameters include the size of the collection, the contents and complexity of images and their geographic or temporal distribution. Examples for image representation parameters are their type, format, language, completeness and the quality level of orthography. The benchmark administration system thereby also supports this parametric benchmark paradigm and facilitates the quick reaction to such changes in research direction by simply altering the parameters and the subsequent regeneration of the required subsets.

Further merits of the benchmark administration system include the facilitation of the incremental collection development, the guidance of the creation, administration, translation and generation of representative search topics, and the efficient execution of relevance assessments.

System Evaluation and Analysis

The benchmark components summarised in the preceding sections certainly provide excellent resources to the information retrieval and computational vision communities to facilitate standardised laboratory-style testing of (predominately concept-based) image retrieval systems. However, such resources can only prove beneficial to research if they are actually used in evaluation events as well.

Hence, we have used the *IAPR TC-12 Image Benchmark* in a multilingual ad-hoc image retrieval task (called *ImageCLEFphoto 2006*) at the *ImageCLEF 2006*

evaluation campaign. Reasons for the choice of a multilingual environment as evaluation platform include:

- the task scenario offered by its ad-hoc retrieval task, which is very similar to that modelled by the *IAPR TC-12 Benchmark*;
- the broad range of audience and participation in prior *ImageCLEF* campaigns;
- the multilingual evaluation environment provided by *ImageCLEF*, which represents the most realistic model for evaluation of retrieval from general photographs since such real-life collections are inherently multilingual;
- the lack of the resources to organise an evaluation event on our own.

ImageCLEFphoto 2006 was the first evaluation event for (multilingual) ad-hoc retrieval from generic photographic collections, and we organised it following an adapted methodology that the Text REtrieval Conference (TREC) had successfully used in the text retrieval domain. The annual cycle of events thereby comprises (in chronological order): the call for participation, registration, document release, topic release, result submission, the creation of relevance assessments, result generation, the actual evaluation event, and the final publication of methods and results.

We highlight how the individual benchmark components were generated and used in the light of *ImageCLEFphoto*, including the image collection and the query topics as well as the relevance judgments and the choice for a particular set of performance measures. We analysed more than 150 system runs submitted by 12 participating groups from 10 different countries. Some of the findings include that:

- a combination of visual and textual features generally improves retrieval effectiveness;
- visual features often work well for more visual queries;
- multilingual image retrieval is as effective as monolingual retrieval;
- feedback and query expansion can help to improve retrieval effectiveness.

ImageCLEFphoto 2006 was the first large-scale evaluation event ever to actually investigate these findings for the domain of multilingual retrieval from a generic photographic collection.

We further analysed the test collection and the evaluation event itself, and, based on our results and feedback from participants, we claim that:

- the benchmark provides performance comparison of retrieval runs with high reliability and discrimination power (as quantified by the error rate and the proportion of ties);
- the difficulty of the retrieval tasks was appropriate (as quantified by the topic difficulty measure);
- the selection of performance measures was useful (as indicated by their correlation values);
- our methodology of parametric benchmarking for image collection and topic creation was validated and approved by the research community;
- we successfully addressed the barriers between research interests and real-world needs by organising an evaluation task modelled on a scenario found in multimedia use today.

Last but not least, and based on all the above, we purport that we successfully repaired the lack of evaluation for (multilingual) visual information retrieval from generic photographic collections.

8.2 Main Achievements

This section recalls the main scientific contributions of the research presented in this dissertation. These contributions have already been indicated in Section 1.3.3 and have been detailed in several chapters afterwards. The chapters on data design and engineering (4), task creation and visual information complexity (5), parametric

benchmark design and architecture (6) and on system evaluation and analysis (7), in particular, bear the content of these scientific contributions (see Figure 8.1).

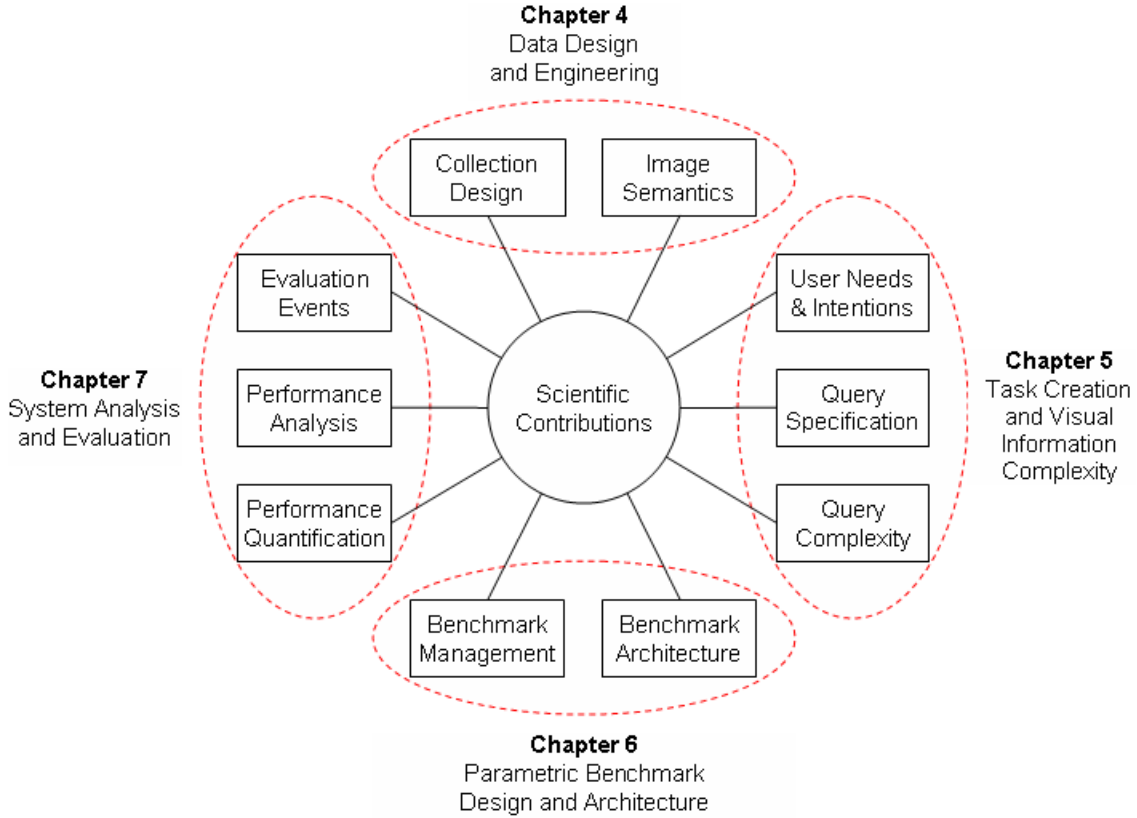


Figure 8.1: Scientific contributions.

We have studied and made contributions to the *design* of parametric test *collections*, the universality of *image semantics* and logical image representations across different languages and world views, the matching of *user intentions* and *query specifications*, *query complexity*, *benchmark management* and *architecture*, *performance quantification* and *analysis*, and the design of *evaluation events*. These contributions make possible a systematic calibration and comparison of system performance for (multilingual) VIR from generic photographic collections.

We have further shown that, with VIR, it is not just a matter of issuing queries against a database and obtaining results, but rather it requires the analysis of a multitude of variables and factors. The work presented in this dissertation therefore also enables a deeper understanding of the complex conditions and constraints

associated with visual information identification, the accurate capturing of user requirements, the correct expression of user queries, the complexity of queries, the execution of searches, and the reliability of performance indicators.

8.3 Limitations and Future Research

Although the topic creation process had been based on topic candidates derived from a log file analysis, and topics had been created against a number of dimensions to allow for additional control, there are still always negative voices that claim that topics were too contrived and not realistic at all. We therefore also recommend that further research be undertaken in the area of topic development and result generation.

More information on what types of searches users typically perform in the domains would, in general, help to establish a greater degree of accuracy in creating realistic topics for evaluation events. In the case of the *IAPR TC-12 Benchmark*, such investigation could be accomplished by re-analysing the log files from online access to the collection. While the original analysis was only based on 980 unique queries, the file has now accumulated more than 5,000 entries¹, representing a much more significant sample for investigation.

One drawback of the methodology for topic creation and management can be seen in the huge amount of work involved for the organisers of an evaluation event. Not only does the identification of topic candidates and the development of representative topics against several dimensions take up a considerable amount of time, but the translation of topics, the selection of sample images for query-by-visual-example approaches, and especially the carrying out of relevance assessments can also be very time-consuming and cumbersome tasks.

Solutions to ease the amount of work for organisers include (1) the idea to let participants choose their own sample images to start their visual queries or (2) to make it a requirement for participating groups at evaluation events to provide a

¹As of 27 April 2006.

number of topic candidates themselves (as practised at INEX Multimedia) and/or to also assist with relevance assessments. The question arises whether this would have any negative effects on the number of participants (*e.g.* INEX Multimedia could not attract more than five participants thus far).

It has further been suggested to save time and effort by replacing the proposed method for the difficulty estimation of topics, and using alternative automatic approaches instead. However, this would come at a cost of lowering correlation and ultimately being less successful at predicting system effectiveness, a compromise too severe to accept as we consider the quantification of topic difficulty as one of the key dimensions within the topic creation process.