Chapter 3

Analysis and Evaluation of Visual Information Retrieval

A fundamental part of the scientific method is the evaluation of research results: an objective comparison is essential to show scientific progress, because without a proper and critical evaluation, it is almost impossible to comment on the quality of research and to prove its progress. Hence, there is a need for standardised benchmarks that provide a platform for such objective evaluation.

There are many research domains in which such benchmarks have significantly advanced the fields. Examples include the *TPC benchmark*[1] for transaction processing in the database field and the *SPEC benchmark*[2] for computer systems performance; many other examples in which successful benchmarks proved vital to trigger significant improvements can be found in [142].

The field of visual information retrieval (VIR) is certainly no exception here. On the contrary, the lack of objective assessment of retrieval performance has hindered its research progress for a long time. Performance evaluation (benchmarking) for VIR systems (VIRS) thus presents a relatively young area of research and has only recently become a more active domain.

This chapter forms the second part of the literature review; it illustrates the aims and principles of performance evaluation in VIR, as well as its criticism and limitations, and also elaborates on the individual benchmark components.

[1] `http://www.tpc.org/`
[2] `http://www.spec.org/`

3.1 Aims and Principles of Benchmarking

This introductory section highlights the development of and motivation for VIR benchmarks, it deals with the major criticism of them and outlines the most commonly used methodology to carry out such evaluation.

3.1.1 Evolution and Motivation

Evaluation in information retrieval can look back on a long development phase, but while for text retrieval it is a very mature research field, in image retrieval it has just recently jump-started.

Evaluation in Information (Text) Retrieval

Research on performance evaluation in information retrieval began more than 40 years ago. The *Cranfield experiments* in 1962 [58] and 1966 [59] are generally considered to have been among the earliest benchmarking events for text retrieval. These experiments did not only emphasise the importance of creating test collections for comparative evaluations, they also provided the basis for the *Cranfield Collection* (1400 documents, 225 queries) that was created in the late 1960s and has heavily been used by researchers since then.

A lot of effort [122, 415, 416] was subsequently put into the creation of test collections in the following two decades, theoretically and practically. However, although these common collections were used by many researchers, there were problems with the consistency of the data and measures which made it difficult to compare the retrieval results. This lack of consolidation among groups resulted in broad generalisations and the unclear situation of not knowing whether systems actually had improved or not [414].

Such cooperation is more likely if groups can compare results across the same data, using the same evaluation method, and then meet at regular benchmarking events to openly discuss in a "friendly" environment how methods differ. The goal of such events should not be to show which systems are superior, but rather to allow

the comparison across a very wide variety of techniques, much wider than a single research group would normally be able to tackle on its own [165].

In order to address this missing element in information retrieval evaluation, the *Text REtrieval Conference*, short TREC[3], was initiated by the *National Institute of Standards of Technology* (NIST[4]) and strongly supported by the *Defense Advanced Research Projects Agency* (DARPA[5]) and the *US Department of Defence*[6]. The first TREC was held in 1992 and provided a very large test collection in order to encourage the interaction among research groups so that a new momentum in information retrieval could be generated [165, 166].

TREC turned out to be a huge success as it saw the participation of a rapidly growing number of research groups in the following years; it was running its 15th annual evaluation campaign in 2006 and is universally considered as the undisputed role model for information retrieval benchmarks.

First Evaluation Efforts (1990 - 2000)

In comparison to text retrieval, the development and evaluation in the much younger field of VIR lags far behind. The first reports of the performance of an image retrieval system in the early to mid 1990s consisted of simply displaying the screen-shots of the retrieval results for one or more sample queries [117]. This first approach lacked objective significance as one could always select certain queries that would return good results in order to highlight the own algorithm's benefits; it was thus neither an objective performance measure, nor could it be considered as a means of comparing different systems.

Consequently, a first comparison [481] of several systems was done (1997) in which three systems were compared and three sample results were printed out, leaving the judgment of the results up to the user. Again, no other performance measures (apart from the overall response time of the systems) were mentioned.

[3] `http://trec.nist.gov/`
[4] `http://www.nixt.gov/`
[5] `http://www.darpa.mil/`
[6] `http://www.dod.mil/`

This lack of objectiveness resulted in more and more researchers starting discussions on performance measures (or the lack of them) for VIR, like in the *Evaluation Framework for Interactive Multimedia Information Retrieval Applications* (MIRA), a project which was supported by the *European Union* [94].

In 1997, Narasimhalu *et al.* [312] presented the (arguably) first approach for a systematic evaluation of image retrieval systems, providing a good overview of VIR systems together with some guidelines on how objective measures for performance evaluation could be constructed. However, this paper neither proposed concrete performance measures nor did it carry out any evaluations so as to validate the effectiveness of these new guidelines. Further, some of the proposed measures that are based on the exact ranking of images within a result set might be hard to implement as it is very subjective to exactly rank items based on their similarity or relevance according to a sample image or a user request.

One year later in 1998, Smith [407] suggested to consider TREC as a role model and recommended to simply reuse the framework and techniques that were already well established for the much more advanced field of text retrieval: the establishment of a common image retrieval test-bed that comprises a standard image collection, benchmark queries, relevance assessments and evaluation methods. This work was, again, only theoretical and no validation with a system based on the proposed test-bed was done.

In the same year, an effort was taken by the University of Amsterdam to formulate a functional benchmark of image retrieval problems, the *Acoi Image Retrieval Benchmark* [315]: two image-sets (one of 1K images and one of 1M images) that were randomly retrieved from the Internet, and six queries and a baseline-run with their own implementation *Monet*[7] were provided. The only performance measure used, however, was the sum of all query execution times and no statement of image retrieval quality or comparisons to other systems are given. Later, this effort turned out to be more akin to a database performance benchmark rather than an image

[7] `http://monetdb.cwi.nl/Home/`

retrieval benchmark.

In 1999, Shapiro started the arguably first effort to create a database free of charge and copyright restrictions for research purposes: the *Database of the University of Washington*[8] (see also Section 3.2.8).

Theoretic Benchmark Designs (2000 - 2003)

The turn of the millennium saw an immense number of publications describing novel techniques and/or the development of innovative systems for VIR (an almost exhaustive overview describing the systems and their corresponding techniques in that phase was compiled in [99, 463]). Moreover, an exploding number of different performance measures [90, 212] to quantify image retrieval performance were proposed too.

With this ever increasing number of systems, techniques and measures, evaluation aspects became more and more crucial as it was practically impossible to compare any two systems. Consequently, several research groups recognised the urgent need for a standardised benchmark suite for VIR to be developed in order to allow the objective, profound and unbiased evaluation of image retrieval systems. It was commonly felt that only in doing so could the promising techniques be distinguished from the simply glossy ones and scientific progress be achieved. This led to several research groups proposing such benchmarks in parallel development.

Leung suggested constructing a benchmark suite for images with complex image contents in 2000 [234]. The proposed size of the image database was rather small (1000 images), and thus the relevance judgments for a number of selected query images were also kept quite small. The general query process was split into two parts, primary query processing (first page) and secondary query processing (relevance feedback), and several performance measures based on precision and recall were proposed for both of them. However, no example evaluation was done to verify the usefulness of the recommended measures.

In the same year, an automatic semantic-based benchmark for image browsing

[8] `http://www.cs.washington.edu/research/imagedatabase/`

systems based on a structured representation augmented by a thesaurus was proposed [308]. A baseline run with the `PicHunter` system [73] for eight queries on a relatively small set of 500 images was provided, and the "performance measures" were only short statements by the test users on the returned result sets like "many almost-relevant images" or "very bad user agreement with this annotation". Due to the massive annotation effort, this approach was only feasible for small image databases and has not been proceeded with any further [287].

Roughly at the same time, Müller *et al.* [302] presented a proposal for performance measures and means of developing a standard test suite for CBIR, similar to that used in information retrieval at TREC. Like [407] or [139], they claimed that many solutions from information retrieval could be adopted for VIR, despite the differences between the fields. Thus, rather than reinventing already existing techniques, a systematic review of evaluation methods in information retrieval and their suitability for VIR was undertaken.

Despite these increased efforts for a standardised platform for image retrieval evaluation, most of the researchers rather kept on using their own image sets and their own queries in order to highlight their own systems' benefits. For instance, more than 20 papers on image retrieval were presented at the *ACM Multimedia Conference 2001* [1], and all of them used different databases to show the performance of their algorithms [287].

The next promising effort was the *Benchathlon* initiative that held its first session in 2001 together with the BIRDS-I benchmark [154], which had been developed in parallel and had defined an evaluation methodology and performance measures. Quite a few researchers participated in theoretic discussions followed by several publications [47, 201, 302, 343] describing potential benchmark architectures, sample queries, relevance judgments and performance measures. In addition, a fully automatic benchmark [298] that was accessible via the Internet [297] was developed in order to make the *Benchathlon* even more attractive for researchers. The communication for query formulation and result transmission was handled by a com-

munication protocol based on a *Multimedia Retrieval Markup Language* (MRML [292, 303, 309, 310]) and a freely downloadable image database was provided to make results reproducible. In theory, this was a very promising and brilliant effort; unfortunately, the proposed architecture was not accepted by many research groups; not many participants could be attracted, and the goal to actually compare the performance of several different system was not reached [293].

The Breakthrough: Evaluation Events (2003 - present)

In fact, the VIR community was facing the same situation that the text retrieval community had faced in the 1980s two decades before [287, 414]: there was no effort by research groups to work with the same data, employ the same evaluation techniques and use these for comparative evaluation; and, in addition, there was also a lack of realistically sized image collections.

The great success of TREC in the text retrieval domain has shown that such cooperation is more likely if groups can compare results across the same data, using the same evaluation methodology, and then meet at regular benchmarking events to openly discuss in a "friendly" environment how methods differ [165, 166].

In 2001, TREC introduced a video track to provide an evaluation framework for video retrieval [403]. This track soon grew to an independent entity called TRECVID[9] in 2003 [401], with an increasing number of participants each year showing its importance in the field [328] (see also Section 3.6.1).

The *Cross Language Evaluation Forum* (CLEF[10]) is, like TRECVID, a spin-off from TREC and has been focussing on multilingual information retrieval since 2000 as an independent campaign. Following the successful examples of TREC and TRECVID, *ImageCLEF*[11] began as a part of CLEF in 2003 [342] and has been the first image benchmarking event to finally fulfil the calls for a TREC-style evaluation framework for image retrieval [139, 302, 407]. More information on *ImageCLEF* can be found in Section 3.6.2.

[9] `http://www.nlpir.nist.gov/projects/trecvid/`
[10] `http://clef-campaign.org/`
[11] `http://ir.shef.ac.uk/imageclef/`

Alternative evaluation events that have just recently started in the last two years (2005 and 2006) include *ImagEVAL*[12] (Section 3.6.3), the *PASCAL Visual Object Classes Challenge*[13] (Section 3.6.4), the *MUSCLE CIS Coin Competition*[14] (Section 3.6.5) and *INEX Multimedia*[15] (Section 3.6.6).

3.1.2 Criticism and Limitations

Although benchmarks are commonly perceived to be inevitable for progress in the field of VIR, they are not completely without criticism.

In 1998, Zobel had already criticised benchmarks in general because they would be counter-productive to (or even block) new innovations. He claimed that novel techniques would only be published if they outperformed the existing ones, and as a consequence researchers might often prefer to make small changes to existing techniques to fine-tune the performance rather than developing totally new techniques [510].

As far as VIR benchmarks are concerned, the main criticism was that current image retrieval benchmarks would study a somewhat artificial field, because: the state-of-the-art image retrieval systems would not be good enough yet to actually benchmark them; the search problem was overrated and a number of other issues like browsing, organising and image data mining were totally neglected; and the current solutions differed too widely from what a real user would need to be evaluated [120].

Recent criticism of evaluation events includes the fact that the data sets can both define and restrict the problems to be evaluated. In addition, although evaluation results and papers are usually made available publicly within evaluation campaigns, the original data can come with strings attached, generally because of copyright restrictions or the high cost of purchase from the original owners. Finally, there is the concern that the research directions of evaluation campaigns is over-influenced by the agencies who fund these evaluation campaigns [402].

[12] `http://www.imageval.org/`
[13] `http://www.pascal-network.org/challenges/VOC/`
[14] `http://muscle.prip.tuwien.ac.at/`
[15] `http://inex.is.informatik.uni-duisburg.de/2006/mmtrack.html`

Some of this criticism is not without reason. In fact, pure CBIR is still primarily working with low level image features such as colour, texture and shape; these rarely correspond to the concepts that users are actually looking for, and it is indeed essential to evaluate systems based on realistic tasks. Further, it is certainly true that current CBIR techniques do not really produce great results yet for concepts that real users are looking for [61].

However, Zobel's early concern that benchmarks would not leave any room for diversity, novelty or creativity can now be considered as a premature prediction. In fact, novelty and creativity have become even easier in evaluation environments in that they provide resources and collaboration, which has given rise to an increased range of novel approaches each year (see, for example, Section 7.3.2).

Furthermore, the general consensus is that system improvement can only be shown by systematic evaluation (and not evaluating at all would not advance any systems!), and that the benefits of evaluation events can not be outweighed by a few drawbacks [402].

3.1.3 Benchmark Components

As mentioned above, TREC has become the role-model for most successful benchmarking events in information retrieval. The main goal of TREC style benchmarks lies in the evaluation of an algorithm's ability to retrieve relevant information regarding a specific information need from a given document collection. Retrieval speed, response time or usability are thereby not of primary importance.

Figure 3.1[16] illustrates the annual cycle of events that is followed by TREC every year. The annual benchmark cycle, normally held within one calendar year, starts with a *call for participation* in which research groups are encouraged to participate. Then, the *tasks* are defined and the document collections are prepared and sent out to the participants that have registered for the track. These collections are the first essential component of a benchmark and will be further discussed in Section 3.2.

[16] taken from `http://trec.nist.gov/presentations/t2004.presentations.html`

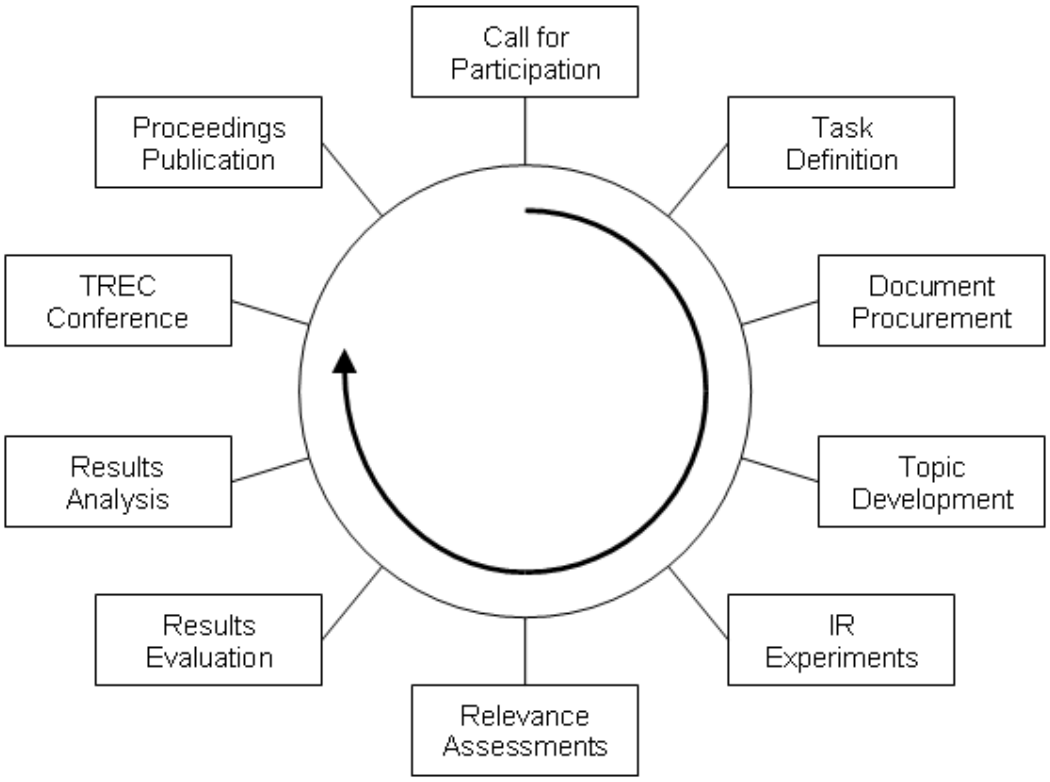


Figure 3.1: Annual cycle of a TREC-style benchmark.

Next, representative queries are defined (TREC called them *topics* and they will be referred to as such hereinafter) and they are then sent out to the participants. These representative topics are the second essential component of a benchmark and will be further discussed in Section 3.3.

The participants then perform their *information retrieval experiments* and are given a limited time to submit their retrieval results (*i.e.* a ranked list of documents) for the different topics and tasks. These results are then used to create the pools that serve as a basis for the *relevance assessments*, which are the third essential component of a benchmark and will further be discussed in Section 3.4.

Once these relevance judgments are completed, the performance of the systems can be *evaluated* in accordance with predefined performance measures and the results can be *analysed*. There is, in fact, an immense number of such performance measures, which are another vital component of the benchmark; the most significant of them are further discussed in Section 3.5.

The participants then use these results in their papers that are sent to the actual *conference*, where the systems are presented and compared on grounds of the evaluation results. Finally, all the techniques, new findings and evaluation results are printed in the *proceedings* of the conference, which complete the annual benchmarking event. These events are probably the most significant component of a benchmark as they ensure that the other four components do not just exist in theory but are also employed practically, and they are therefore further discussed in Section 3.6.

3.2 Document Collections

One of the main components of any benchmark is a collection of documents (*e.g.* images, texts, sounds, videos) that is representative of a particular domain [267]. Although there are hundreds of different document collections available, finding such resources for benchmarks has turned out to be quite hard, because visual resources especially are often quite expensive and copyrighted, which restricts both the large-scale distribution and the future access of the data for evaluation events. For evaluation purposes, an ideal image collection would be royalty-free, without copyright restrictions and accepted by the research community to be representative of its particular field.

This section first describes the by far most commonly used image database in VIR, the *Corel Photo CDs* (Section 3.2.1), and then mainly concentrates on collections that have actually been used in ad-hoc retrieval tasks and evaluation events (Sections 3.2.2 to 3.2.6), describing their visual contents and corresponding textual representations as well as their benefits and limitations.

The last two sections introduce collections used in alternative tasks (Section 3.2.7) and, finally, further collections that, albeit not used in evaluation events thus far, have either proved valuable for research in the past or are likely to be used in evaluation events in the future (Section 3.2.8).

3.2.1 The Corel Photo CDs

Commercial image collections are generally not used in large-scale evaluation events due to their high price and copyright restrictions. Yet, for a long time, the *Corel Photo CDs*[17] were the de-facto standard for image retrieval; they were the most commonly used image collection for VIR evaluation [193, 326, 367, 444, 445, 455, 456, 480, 507, 508] and therefore deserve to be examined a bit more closely.

Collection Content

This collection consists of more than 800 image CDs that cover many different aspects of life. Each CD can be purchased individually and contains 100 colour photographs from a certain category, including general themes like "sports", "winter" or "Europe", more specific ones like "boats", "dogs" or "mountains" and also quite abstract ones like "yellow" or "shapes" (see Figure 3.2 for sample images[18]).

Figure 3.2: Sample images from the Corel database.

All the photos on the CDs exhibit a very high resolution (3072x2048 pixels) and contain a few corresponding keywords. There is also the possibility of purchasing these images in differently compiled collections of smaller images, like the *Corel Gallery 1000000* that contains one million photos with a resolution of 384x256

[17] `http://www.corel.com/`
[18] Sample images were taken from CD 8 of Correl Gallery 1000000.

pixels [144, 239], or the *Corel Gallery 1300000* that contains 1.3 million photos with a resolution of 120x80 pixels [507, 508]. The photos in these collections do not have any captions other than their category name.

Benefits

The obvious advantage of the *Corel Photo CDs* is the high quality of the photos: they exhibit a high resolution, were taken by professional photographers and cover a wide range of different aspects of still natural images.

Limitations

Yet, in general, the *Corel Photo CDs* are quite problematic: first of all, these CDs are expensive to obtain because each of the image sets has to be bought individually; they are further protected by copyright and legal restrictions on use and therefore difficult to redistribute for large-scale evaluation events. Moreover, they are currently no longer available on the market [146] and therefore not accessible for researchers. Another problem is that the images only contain very limited written meta-data, making them less suitable for the evaluation of TBIR systems.

The biggest drawback, however, is that *Corel* does not offer one set of images but single sets on individual CDs or a collection of differently compiled sets. Groups that use the collection to evaluate the performance of their algorithms are therefore often using different subsets instead of the same images. This does not only impede the comparison of the performance of systems, it is even possible that subsets of these images can be tailored to make a system look better than it really is [294].

Despite all these disadvantages, the *Corel* images are still used in a few publications [39, 131, 170, 171, 239, 241, 244, 316, 351, 496, 506] to demonstrate retrieval performance; however, the trend is clearly going towards presenting retrieval results at evaluation events (see Section 3.6) where comparison is more objective.

Other Commercial Collections

In addition to the *Corel Photo CDs*, there are many other commercial image collections that would be quite useful for image retrieval evaluation, for example: the *Corbis Image Database*[19], *Getty Images*[20], *Photonica*[21], *Stock Photo*[22], *AccessStock*[23], and *AgeFotoStock*[24].

Yet, these large image collections are not suitable for large-scale evaluation events, mainly for two reasons: they are (1) expensive and (2) often copyrighted (or without a clear copyright statement) which restricts both the large-scale redistribution and future access of the data for evaluation purposes. Hence, these collections are not used in current VIR evaluation events.

3.2.2 The St Andrews Collection of Photographs

The *St Andrews Collection of Photographs (SAC)* is a subset of one of Scotland's most important archives of historic photography which was made available to the public via a web interface[25] in a large-scale digitalisation project by *St Andrews University Library* [66, 355]. This collection of 28,133 historic photographs from well-known Scottish photographers and photographic companies was a core component of the *ImageCLEF* ad-hoc retrieval task from 2003 to 2005 [62, 63, 64].

The digitised subset of the historic photo archive has recently been increased to more than 50,000 photos [66]; however, this section will only discuss the collection subset that was actually used at *ImageCLEF*.

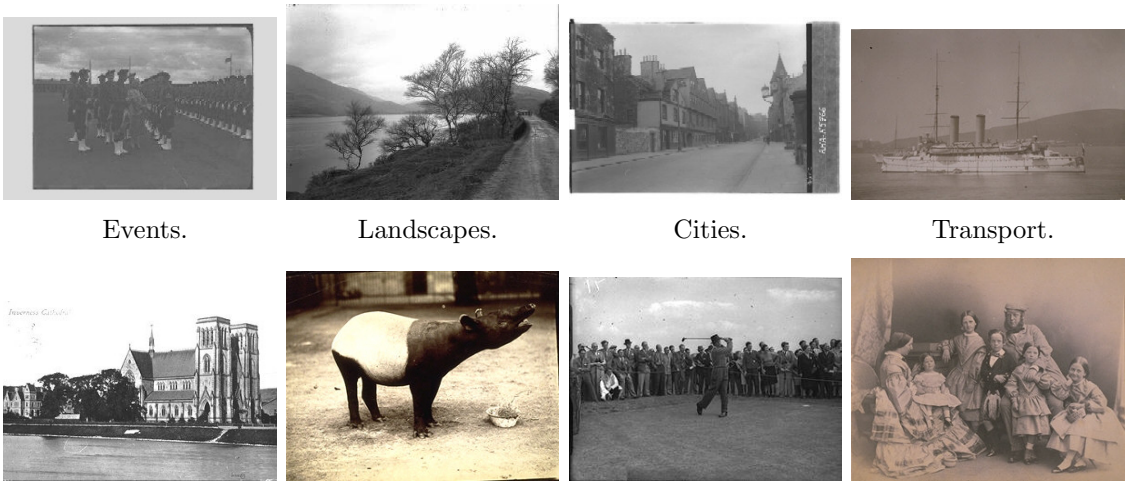
Collection Content

The majority of the photos in the SAC are monochrome or black and white (89.0%), due to the historic nature of the collection, and are specific to Scotland (67.1%) or

[19] `http://pro.corbis.com/`
[20] `http://www.gettyimages.com/`
[21] `http://www.photonica.com/`
[22] `http://www.stockphoto.net/`
[23] `http://www.accessstock.com/`
[24] `http://www.agefotostock.com/`
[25] `http://specialcollections.st-and.ac.uk/`

the UK (95.0%) and to life between 1840 and 1940 [66]; this includes photos and postcards of old towns and villages, nature (*e.g.* landscapes, animals), architecture (*e.g.* buildings, statues, monuments), events (*e.g.* war-related, royal visits), transport (*e.g.* ships, carriages, streets, bridges), family and individual portraits, and sports (especially golf).

Figure 3.3 displays sample images from these categories to illustrate the wide diversity of the collection.



Events. Landscapes. Cities. Transport.

Buildings. Animals. Golfers. Portraits.

Figure 3.3: Sample images from the SAC.

Not all the images in the SAC exhibit exactly the same size: the large versions of the images show an average resolution of 368x234 pixels; the corresponding thumbnails have 120x76 pixels (see [66] for detailed statistics).

Image Captions

Each photograph has an alphanumeric caption that consists of the following nine fields: (1) a unique record number, (2) a full title, (3) a short title, (4) a textual description of the image content, (5) the date when the photograph was taken, (6) the originator, (7) the location where the photograph was taken, (8) notes for additional information, and (9) its corresponding categories. Figure 3.4 provides a sample image and its respective caption.

Record ID:	JV-.094276
Title:	Frome, Somerset. Catherine Hill.
Short Title:	Catherine Hill, Frome.
Description:	Steep road lined with stone buildings with shops at ground level; awnings on shops windows right; goods in doorway, left.
Date:	Registered 1925
Originator:	J Valentine & Co
Location:	Somerset, England
Notes:	JV-94276 pc/mb/jf/mb DETAIL: Woman in long skirt and summer blouse and hat, carrying shopping basket. PCARD: Handwritten date and postmark 1912.
Categories:	[shops], [buildings - stone], [Somerset all views], [Collection - J Valentine & Co]

Figure 3.4: Sample image and its caption.

In addition to the plain text files, the captions were also encapsulated in a *Standard Generalised Markup Language* (SGML) format to be compatible with existing TREC collections (see Figure 3.5 for an example).

```
<DOC>
<DOCNO>stand03_1099/stand03_21287.txt</DOCNO>
<HEADLINE>Frome, Somerset. Catherine Hill.</HEADLINE>
<TEXT>
<RECORD_ID>JV-.094276</RECORD_ID>
Catherine Hill, Frome.
Steep road lined with stone buildings with shops at ground level; awnings on
shops windows right; goods in doorway, left.
Registered 1925
J Valentine & Co
Somerset, England
JV-94276 pc/mb/jf/mb DETAIL: Woman in long skirt and summer blouse and hat,
carrying shopping basket. PCARD: Handwritten date and postmark 1912.
<CATEGORIES>[shops], [buildings - stone], [Somerset all views], [Collection -
J Valentine & Co]</CATEGORIES>
<SMALL_IMG>stand03_1099/stand03_21287.jpg</SMALL_IMG>
<LARGE_IMG>stand03_1099/stand03_21287_big.jpg</LARGE_IMG>
</TEXT>
</DOC>
```

Figure 3.5: Sample SAC caption in SGML format.

The `<DOCNO>` tag contains the pathname of the image as a unique document identifier, and the title and categories are indicated by the `<HEADLINE>` and `<CATEGORIES>` tags respectively. The remaining caption fields are enclosed by the `<TEXT>` tag and are not structured. In addition, the `<SMALL_IMG>` and `<LARGE_IMG>` tags contain the path of the thumbnail and of the large version. Further examples and information about the SAC can be found in [65, 66] and the *St. Andrews University Library*[26].

[26] `http://www-library.st-andrews.ac.uk/`

Benefits

The SAC was used as the basis for *ImageCLEF* because of the following advantages: it (1) represents quite a "large" collection of images, (2) offers high quality, semi-structured image captions to support concept-based retrieval methods as well, and (3) permission was granted by St. Andrews Library to download and distribute the collection for use in *ImageCLEF* [64]. All these benefits facilitated the birth of *ImageCLEF*, as acquiring a suitable collection for large-scale evaluation events is, indeed, a non-trivial task.

Limitations

Although the SAC has provided a valuable contribution to *ImageCLEF*, there are a number of limitations to its use which include the following.

Firstly, the domain of the SAC is restricted to mainly photographs specific to life in Scotland and England from 100 years ago, which together with the excessive use of colloquial and domain-specific language affects both its use and effectiveness as a generic evaluation resource; it is hence questionable whether evaluation results using this collection are also transferable to other collections, which therefore limits the conclusions that can be drawn from research with the SAC.

Secondly, most of the images are monochrome or black and white photographs; they do not contain many clearly separated objects and a few are also of very poor quality (*e.g.* too dark, too blurry). This makes the SAC a very difficult collection for a purely visual analysis [291] as CBIRS predominately rely on information regarding colour, texture and shape.

Finally, the main problem with using the SAC for the comparative evaluation of image retrieval systems is, again, the restriction on copyright that hinders the redistribution and further use by researchers outside the CLEF campaign. The image collection, although free of charge, could only be legally used having officially registered for *ImageCLEF*; its replacement by the *IAPR TC-12 Image Benchmark* (see Chapter 7) makes the SAC again unavailable for research.

3.2.3 The Wikipedia Multimedia Corpus

The *Wikipedia Multimedia Corpus* is a subset of the *Wikipedia XML Corpus* [84], which comprises XML collections based on *Wikipedia*[27], an online encyclopedia that is collaboratively written by contributors from all over the world. This corpus was used in an image retrieval task of the *INEX 2006 Multimedia Track* [486].

Collection Content

The *Wikipedia Multimedia Corpus* comprises the following three collections:

- **The Wikipedia English Collection** contains 659,388 English XML files with filenames that are equivalent to the unique identifiers of that file (for instance: `15234.xml`); each file corresponds to exactly one article of *Wikipedia*.
- **The Wikipedia Image Collection** contains more than 300,000 JPEG images and presents a subset of the images referred to in the *Wikipedia English Collection*; not all the images referred to in the XML files are included because of copyright restrictions.
- **The Wikipedia Image XML Collection** contains exactly one meta-data file for each of the images of the *Wikipedia Image Collection*; these files contain very short image descriptions in XML format and correspond to the image information provided by *Wikipedia*.

The image collection does not only contain photographs, but also maps, satellite images, x-rays, graphs, drawings, sketches, illustrations, and figures (see Figure 3.6).

Figure 3.6: Sample images from the *Wikipedia Image Collection*.

[27] `http://www.wikipedia.org/`

These images come in all dimensions and sizes, from 30x30 pixels and 1 KB to 4800x3600 pixels and 4.7 MB; some even exhibit rather extreme dimensions, like 11880x1683 pixels.

```
<?xml version="1.0" ?>
- <article>
    <name id="205995">African_Buffalo.JPG</name>
    <image xmlns:xlink="http://www.w3.org/1999/xlink"
      xlink:type="simple" xlink:actuate="onLoad" xlink:show="embed"
      xlink:href="../pictures/African_Buffalo.JPG" id="205995"
      part="images-0">African_Buffalo.JPG</image>
  - <text>
      <h2>Summary</h2>
      An African Buffalo Bull. Photographed at Mabula Game
      Reserve, South Africa, 2004 by Paul M Rae.
      <p />
      <h2>Licensing</h2>
    - <wikitemplate parameters="1">
      - <wikiparameter number="0" last="1">
          <value>cc-by-2.5</value>
        </wikiparameter>
      </wikitemplate>
    </text>
  </article>
```

Figure 3.7: XML caption for the image `African_Buffalo.jpg`.

A sample image caption taken from the *Wikipedia Image XML Collection* is illustrated in Figure 3.7; this information corresponds with the one that is also available on the *Wikipedia Image* pages[28].

Benefits

The main benefit of the *Wikipedia Multimedia Corpus* lies in the large number of royalty-free images as well as in the extensive text that is associated with these, which can be downloaded after having registered at the *Wikipedia XML Corpus pages*[29]. This allows for close examination of both TBIR and CBIR. Further, the highly structured captions in XML format also provide the base for a testbed for XML retrieval evaluation.

Limitations

The varying dimensions, data types and copyright restrictions within the image collection can be seen as one of the few drawbacks of the collection; since anyone

[28] `http://en.wikipedia.org/wiki/Image:African_Buffalo.JPG`
[29] `http://www-connex.lip6.fr/~denoyer/wikipediaXML/`

can edit the text files, the quality of the image captions inherently varies within the collection as well.

3.2.4 The Lonely Planet XML Document Collection

The *Lonely Planet XML Document Collection* is based on the *Lonely Planet World Guide*[30] and was used in a combined text and image retrieval task at the *INEX 2005 Multimedia Track* [511].

Collection Content

Lonely Planet provided INEX with a collection of 462 XML documents containing information about destinations (*i.e.* countries, interesting regions and major cities), including an introduction, information about transport, culture, major events, facts and an image gallery of the local scenery. The collection comprises 1,787 low resolu-

Animals. Sport. People. Mountains.

Architecture. Maps. Landscapes. Cities.

Figure 3.8: Sample images of the *Lonely Planet* XML document collection.

tion JPEG photographs (400x300 pixels) showing a diverse range of topics including cities, architecture, landscapes, mountains, animals, people and sport scenes as well as 453 low-resolution maps (see Figure 3.8 for sample images).

[30] `http://www.lonelyplanet.com/worldguide/`

Image Captions

Unlike most of the other collections, the images in the *Lonely Planet XML document collection* are not further described by associated text files on a per image basis, but are embedded in one of the 462 XML documents (one for each travel destination) which themselves contain a number of images. For instance, the first image in the

```
<image filename="/images/BN836_13.jpg" object_type_id="68">
  <caption>The fins and scales of Sydney Opera House</caption>
  <photographer>Richard I'Anson</photographer>
</image>
```

Figure 3.9: XML captions in the Lonely Planet document collection.

second row of the examples shown in Figure 3.8 (the Sydney Opera House) belongs to the XML document that describes Sydney in the *Lonely Planet World Guide*[31]. Apart from a brief tag, which provides a short description of the image and further information on the photographer (see Figure 3.9), the XML document is generally concerned with the main destination (Sydney) and does not further elaborate on or describe the image directly.

Benefits

The major benefits of the *Lonely Planet XML document collection* are: it is a realistic images corpus which is also used in the real world, its content is very diverse, and it contains very exhaustive XML documents. All this allows for very interesting evaluation tasks, especially as far as multimedia retrieval from structured collections is concerned.

Limitations

Once again, the main limitations for this collection are very strict copyright regulations: the access to the *Lonely Planet* data is strictly limited to students and staff

[31] `http://www.lonelyplanet.com/worldguide/destinations/pacific/australia/new-south-wales/sydney/images/`

working directly on INEX[32]; an application to use the data has to be completed and signed before the permission for access is given. The agreement between INEX and Lonely Planet will expire on June 20, 2007; hence, the data will no longer be accessible for researchers after this point.

Other drawbacks can be found in the low resolution of the images which can limit the use of CBIR techniques, the low number of images in the collection, and the fact that the XML captions are rather general and do not directly describe the images but rather the destination they belong to.

3.2.5 The ImageCLEFmed Collection

The *ImageCLEFmed collection* is an archive of domain-specific photographs for the medical field which has been used in the medical ad-hoc retrieval tasks of *ImageCLEF*, called *ImageCLEFmed*, since 2004 [62, 63, 290].

Collection Content

This medical data set is, in fact, a composite of several medical sub-collections provided by independent medical institutions and hospitals that granted *ImageCLEF* the permission to use their data sets in its evaluation campaign.

The first collection used by *ImageCLEFmed* in 2004 [63] was the *Casimage Collection*[33]; most of its 8,725 images are radiology modalities (but it also contains photographs, presentation slides and illustrations) belonging to 2,075 medical cases (see Figure 3.10).

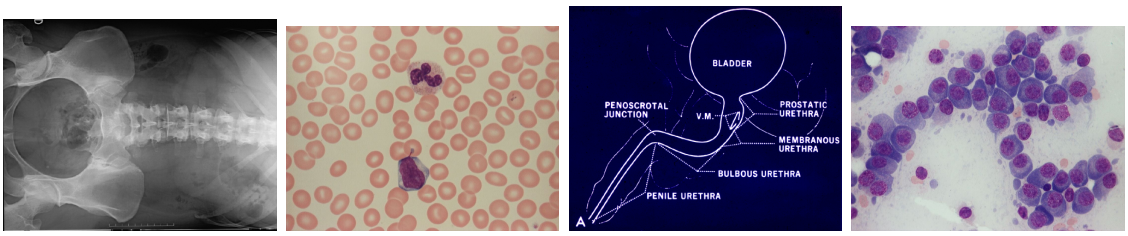


Figure 3.10: Sample images from the Casimage collection.

[32] A special consideration from the INEX organisers was sought to be allowed to access the collection in the frame of the literature review of this thesis.

[33] `http://www.casimage.com/`

In 2005, *ImageCLEFmed* [62] was also given the permission to use the *Pathology Educational Instructional Resource (PEIR)* data set[34]; this collection contains 33,000 mainly pathology images (see Figure 3.11).

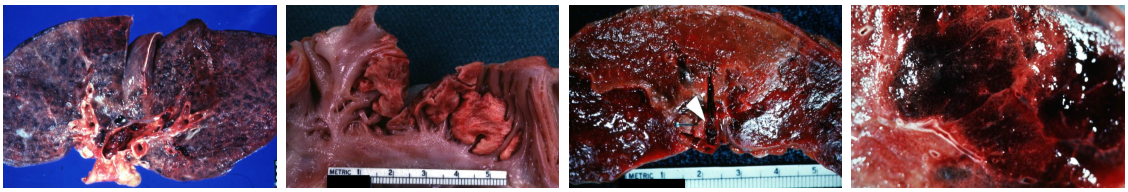


Figure 3.11: Sample images from the PEIR dataset.

Another dataset that was made available that year was the nuclear medicine database of the *Mallinckrodt Institute of Radiology*[35] (MIR) with more than 2,000 images mainly from the field of nuclear medicine [477]. Figure 3.12 shows sample images for the MIR dataset.

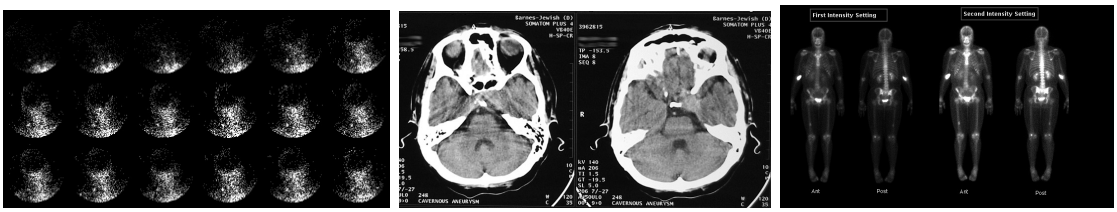


Figure 3.12: Sample images from the MIR database.

Likewise, the *PathoPic*[36] collection [134] was also included in 2005 and comprises 9,000 pathology images (see Figure 3.13 for examples).

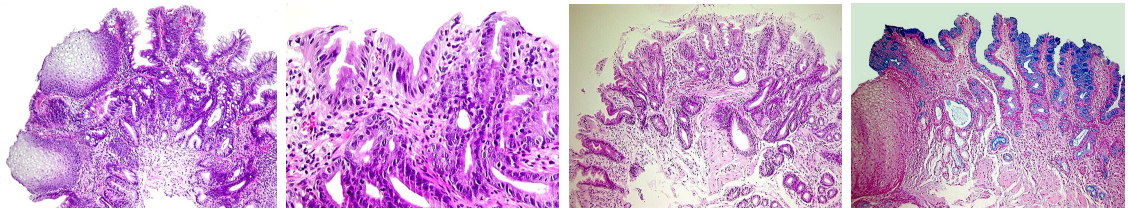


Figure 3.13: Sample images from the PathoPic collection.

[34] http://peir.path.uab.edu/
[35] http://www.mir.wustl.edu/
[36] http://alf3.urz.unibas.ch/pathopic/

Image Captions

The majority (95%) of the medical cases in the *Casimage database* have corresponding *case notes* which are written in XML, with 75% being annotated in French and 20% in English. These quite elaborate case notes can comprise several images and include, *inter alia*, a field for the title, diagnosis, free-text description, clinical presentation, hospital, department and keywords. Not all the cases, however, are described in such detail: 207 case notes are completely empty, and about 1,500 images in the collection are not attached to case notes at all [304, 365].

Similar to *Casimage*, the images in the *MIR database* are also assigned to medical cases which are described in an English XML file. These rather extensive descriptions are only encapsulated by one `<CASE>` tag. Yet, some sort of structured information still exists within the text as there are sections for, *e.g.*, diagnosis, findings, discussion and follow-up.

By contrast, each of the images of the PEIR dataset has a corresponding English caption based on the *Health Education Assets Library (HEAL) project*[37] and is thus not dependent on medical cases. The PEIR dataset also shows a very detailed representation structure, depicting information like the file name, a title, a description, a date of contribution, archiving and cataloguing, and the image source. More detailed information on the HEAL project can be found in [36]. Likewise, the *PathoPic collection* also comprises structured captions on a per image basis in German, however its captions are not as detailed as those of PEIR.

Sample captions for all four databases can be found in [289].

Benefits

The benefits of the *ImageCLEFmed collection* are quite obvious: the four medical datasets together build a relatively big set of images of a variety of medical fields. The high resolution and also the nature of the images are almost predestinated for CBIR, while the extensive, multilingual captions in English, German and French

[37] `http://www.healcentral.com/`

create a very realistic, comprehensive and versatile data set for the evaluation of concept-based image retrieval as well.

Limitations

Unfortunately, and analogue to the SAC (compare Section 3.2.2), most of these medical images are under copyright restrictions, and their redistribution to the participating research groups is only possible through a special agreement with the original copyright holders.

In the majority of cases, the captions do not describe the image content itself but rather the context in which the image was taken; they further include many abbreviations which are used in a non-standardised way and many terms which are very specific to the medical domain and unlikely to be found within most general-purpose dictionaries or normal stemmers.

Other medical databases

The four databases were chosen for the *ImageCLEFmed Collection* mainly because they are real-world collections and they were made available to *ImageCLEF* in a research context [289]. Although there are a number of medical databases available on the Internet (for example at the *Medical Image Resource Centre*, MIRC[38], or also [304] mentions several), they are not suitable for evaluation campaigns due to two reasons: 1) copyright restrictions hinder their large-scale distribution among participants, and 2) most of them lack query topics and/or a ground-truth (and the creation of such is very time-consuming and costly due to the necessary involvement of medical experts).

3.2.6 The ImagEVAL Corpora

The *ImagEVAL Corpora* is an image collection that was used as the development and test dataset for the *ImagEVAL 2006*[39] evaluation campaign [118]; the images

[38] `http://mirc.rsna.org/`
[39] `http://www.imageval.org/`

were provided by institutions such as the *French National Organisation of Museums* (Réunion des Musées Nationaux[40]), the French book and press group *Hachette*[41], the car manufacturer *Renault*[42] and the *French Ministry of Foreign Affairs*[43]; some images were also taken from *Wikipedia*.

Since all images in the corpora originate from governmental or commercial institutions, strict copyright regulations hinder the distribution of a complete package after the evaluation. The steering committee of *ImagEVAL* could only reach an agreement with the data providers that allows participants to use a part of the database for research purposes. The collection is therefore not available to researchers outside the *ImagEVAL* campaign[44].

3.2.7 Image Collections in Other Tasks

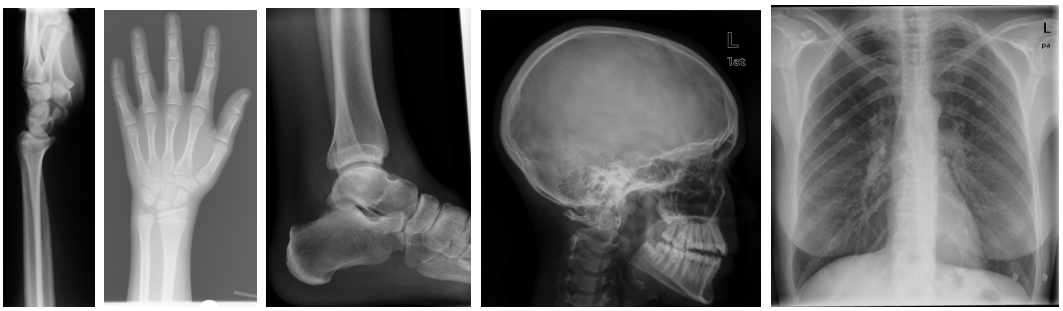
Although the following collections have not yet been used in ad-hoc retrieval tasks, they have made contributions to other evaluation set-ups such as object recognition, image classification and automatic annotation tasks as well as the evaluation of interactive retrieval (and usability), and are therefore briefly introduced hereinafter.

IRMA Database

The *IRMA database*[45] is a collection of more than 15,000 medical radiographs [230] that have arbitrarily been acquired from daily routine at the *Department of Diagnostic Radiology of the RWTH Aachen University*[46]. A subset of 10,000 images was used for an automatic image classification task at ImageCLEF 2005 [62, 86], and one of 11,000 images in 2006 [61] respectively.

All images in the IRMA database are provided as differently sized PNG files using 256 grey values as illustrated by the sample images in Figure 3.14. Each

[40] `http://www.rmn.fr/`
[41] `http://www.hachette.com/`
[42] `http://www.renault.com/`
[43] `http://www.diplomatie.gouv.fr/`
[44] The organisers of ImagEVAL were contacted and asked for access to the collections in the frame of this PhD thesis. This request could not be granted due to the strict copyright regulations.
[45] Image Retrieval in Medical Applications, `http://irma-project.org/`
[46] `http://www.rad.rwth-aachen.de/`



2227 2408 11226 12313 13447

Figure 3.14: Sample images from the IRMA database.

image is classified according to the so called *IRMA code*. More information on the IRMA database and code can be found in [229].

LTU Technologies

LTU (Look That Up) Technologies[47] provided their hand-collected dataset of mono-object images from 268 classes to an automatic object annotation task at *ImageCLEF* 2006 [61]. The image collection that was eventually used in that event had been reduced to 21 classes (and 81,211 images respectively) in order to facilitate the task for the participants in the first year.

Cups. Glasses. Phones. Books.

Figure 3.15: Sample LTU images from chosen categories.

The collection consists of two different types of images: first, there is the training data set in which each image contains only one object in a rather clean environment,

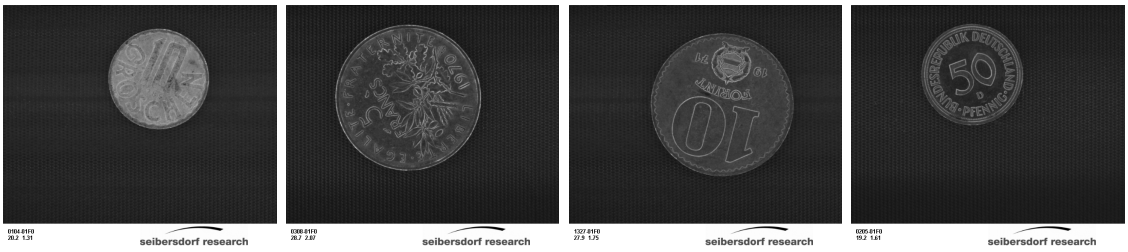
[47] `http://www.ltutech.com/`

i.e. the images show the object and some mostly homogeneous background; and second, there is the test data set in which these objects are in a more "natural setting", *i.e.* there is more background clutter than in the training images (see Figure 3.15). All the images are in PNG format, and most of them exhibit a resolution of 640x480 pixels.

Coin Images Seibersdorf (CIS) Benchmark

A somewhat unique image collection is the dataset of the *Coin Images Seibersdorf (CIS) Benchmark* [322] which was used at the *MUSCLE CIS Coin Competition 2006* [323], a coin recognition competition held by the *European Union Network of Excellence MUSCLE*[48].

This benchmark contains a large database of 60,000 black and white images of 30,000 test coins, which are further classified into 692 *coin classes* and 2,270 *coin-face classes*[49]. The competition data consists of a further 10,000 coins (20,000 images).



0104-01F0 0308-01F0 1327-01F0 0205-01F0

Figure 3.16: Sample CIS images with their identifiers.

Figure 3.16 shows four sample coin images of the CIS Benchmark (with front and reverse sides of each coin above/below each other). All images in the benchmark exhibit a resolution of 640x576 pixels and are stored in the PNG format. Each image is also represented by an eight-digit identifier which is coded into the first line of the image data. More information about the exact caption format can be found in [323].

[48] `http://muscle.prip.tuwien.ac.at/`

[49] Some coins changed their appearance over time, like a different image or a different text printed on the coin, thus there are more coin-face classes than coin-classes.

FlickR

In contrast to the aforementioned datasets, *FlickR*[50] is not a stand-alone, static image collection, but rather a photo-sharing web site which provides a platform for the online community to share, browse and tag images; it contains over 30 million freely accessible images[51] and was used in an interactive image retrieval task of CLEF, *iCLEF* [137].

FlickR is a large-scale, web-based image database that is based on a large social network of online users who upload, update and delete their images and describe them using freely chosen keywords (or so-called "tags"). These images comprise a broad variety of topics including people, places, landscapes, objects, animals, events, *etc.* (see Figure 3.17 for sample images).

Children. Cow. Snowboarding. City.

Figure 3.17: Sample *FlickR* images.

The advantages of using *FlickR* (or other online photo-sharing platforms such as *AlbumTown*[52], *Fotolog*[53], *Fotopic*[54], *MyPhotoAlbum*[55], *Webshots*[56], *Zooomr*[57], *etc.*) lie in the large amount of accessible, royalty-free images. Since such online photo-sharing repositories are used by the public, they represent a very realistic image collection of a broad range of image contents and have more or less excessive, multilingual captions - all of which would, in principle, make the collection a rich resource for image retrieval.

[50] `http://www.flickr.com/`
[51] As of August 2006.
[52] `http://www.albumtown.com/`
[53] `http://www.fotolog.com/`
[54] `http://fotopic.net/`
[55] `http://www.myphotoalbum.com/`
[56] `http://www.webshots.com/`
[57] `http://zooomr.com/`

However, despite the obvious benefits of these online collections, there are some drawbacks that hinder the use as a reliable dataset for evaluation events. One problem is that the images in these collections often exhibit different levels of copyright restrictions, and this unclear copyright situation often impedes the large-scale download and redistribution of the images. Having the collection locally stored, however, is crucial because using an online collection for evaluation would create further problems: CBIR is merely impossible since online collections hamper the necessary image analysis, captions are not controlled and are constantly changing over time, and collection frequencies become quite hard to calculate and are also constantly changing; the reproducibility of research results is therefore questionable.

The PASCAL Object Recognition Database Collection

The *PASCAL Object Recognition Database Collection*[58] is a compilation of image databases with the goal of providing a standardised collection of object recognition databases; it was created because the existing datasets did not provide a challenge for the current generation of object recognition algorithms anymore, with subsets of it being used at the 2005 and 2006 *PASCAL Video Object Challenges* [109, 110], an evaluation campaign for object recognition.

Bike (Training). Bike (Test). Car (Training). Car (Test).

Figure 3.18: Sample *Pascal* images.

The images in the collection are PNG images and show objects from a number of classes in mostly realistic scenes (*i.e.* no pre-segmented objects). Figure 3.18 illustrates sample images (both training and test images) for the categories "bikes" and "cars". The majority of the images in the compilation are annotated using

[58] `http://www.pascal-network.org/challenges/VOC/databases.html`

the class label provided by the original creators of the respective databases, and an additional PASCAL class label.

Most of the images and corresponding semantic descriptions are freely available for download without having to register for an evaluation event or any other form of agreement, while others do underly certain copyright regulations, yet the level of copyright restrictions can even vary within one collection itself. Furthermore, some images in the collections are of very poor quality.

3.2.8 Additional Royalty-Free Collections

The following image collections have not been used in image retrieval benchmarks, yet. However, some of them have either been of particular importance in the development of image retrieval evaluation (compare Section 3.1.1), have been cited by researchers using these collections or are likely to be used in the future; hence, they deserve to be briefly mentioned hereinafter.

University of Washington

The image database of the *University of Washington*[59] was the first collection of its kind to be made available free of charge and without copyright restrictions for image retrieval system evaluation. It was created in 1999 and later enlarged in collaboration with the *University of Geneva* [287].

Arbor greens. Campus in fall. Football. Spring flowers.

Figure 3.19: Sample images of the *University of Washington*.

The database is rather small and contains only about 1,100 images that are classified into 21 different groups (Figure 3.19 shows sample images together with

[59] `http://www.cs.washington.edu/research/imagedatabase/`

their group names). Most of the images are very similar within such a group, with the smallest group containing only 27 images and the largest 256. Keywords are provided in a separate text file, *e.g.* "arborgreens/Image01 trees bushes grass sidewalk" for the first image of Figure 3.19.

Benchathlon

The *Benchathlon* has a test data set[60] with images categorised according to their finishing status ("done" or "todo"), their resolution (from 96x64 pixels to 6144x4096 pixels) for finished images, and groups for unfinished ones. Figure 3.20 shows sample images for the status "todo" and its underlying group "0043".

Figure 3.20: Sample images of the test data set of the *Benchathlon*.

Image captions are provided for entire groups and not on a per image bases. For example, the first image of Figure 3.20 is annotated with: "Images 1-33: 1994 Soccer World Cup, Palo Alto, Stanford Stadium". The collection is available for free and without any copyright restrictions, but neither search tasks nor any ground-truth have been provided thus far [287, 293].

Amsterdam Library of Object Images

The *Amsterdam Library of Object Images*, short ALOI[61], is a collection of 1,000 objects under various imaging circumstances such as illumination and viewing angle as well as illumination colour for each image [127], creating a total of more than 110,000 images following the models of the *Colombia University Object Image Library* COIL-20 and COIL-100 databases [313, 314], the *Surrey Object Image*

[60] `http://www.benchathlon.net/resources/data.html`
[61] `http://staff.science.uva.nl/~aloi/`

Library (SOIL) [213] and work by Barnard [20], which recorded objects with varying viewing angles, illumination intensities and illumination sources respectively. Figure 3.21 illustrates a few sample images of ALOI.

Figure 3.21: Sample images of the ALOI.

All objects were recorded with a black background and in 24 different illumination angles, 12 illumination temperatures and 72 different object angles (with a rotation of 5 degrees between each of them), creating a total of 108 images for each of the objects. There is also a single text file which further describes the individual objects, stating the object name (*e.g.* "smiling duck"), the material of the object (*e.g.* "plastic"), how the object was stained (*e.g.* "pluriform"), and some additional surface properties (*e.g.* "shiny"). The collection is available copyright and royalty-free, but has not been used for evaluation events yet[62].

Alternative Collections

The *TRECVID video collections* have increasingly been used for image retrieval in the last couple of years as well [328]. The key frames can, indeed, be used for image retrieval and object recognition (although they exhibit a rather low resolution), and the tasks created correspond well to simple journalist search tasks. Since the videos also contain speech, multi-modal retrieval evaluation would be possible on these datasets as well.

Other interesting image data sets are the *CalPhotos*[63], a collection of 121,780 images of plants, animals, fossils, people, and landscapes, or the *Dataset of Annotated Animals* [160] that contains 59,795 images (8,114 showing animals). There

[62] As of 15 April 2007.
[63] `http://calphotos.berkeley.edu/`

are also databases available for computer vision research (*CVonline*[64] and the *Computer Vision Homepage*[65] at *Carnegie Mellon University* have long lists of test image collections), but they are rarely used for image retrieval as they do not represent realistic retrieval data.

3.3 Search Tasks

The second essential component of a TREC style benchmark is representative search requests (TREC calls these representative search queries *topics*, and they will also be referred to as such hereinafter). Topics are statements of information need (expressed as narrative text, a set of keywords, or images) which represent realistic user requests and ideally allow a good cross-section of the image contents to be searched [234, 469].

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 [224].

The selection of topics should not only be representative of the collection, but also be based on real-world queries. It is therefore crucial for any benchmark to define the evaluation's goal before the topic development process is being started ("What exactly do we want to evaluate?"), whereby such a goal is ideally based on real user information needs and not on 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 [293].

The definition of such an evaluation goal naturally influences the topic types as well as its development process and format; this section elaborates on each of these three aspects below.

[64] `http://homepages.inf.ed.ac.uk/rbf/CVonline/CVentry.htm`
[65] `http://www.cs.cmu.edu/afs/cs/project/cil/ftp/html/vision.html`

3.3.1 Topic Types

Topic types depend on the goal of an evaluation event and the corresponding tasks to achieve that goal. This section introduces the main topic and task types that have, in fact, been used in evaluation events thus far.

Ad-Hoc Retrieval

In TREC, the classic ad-hoc retrieval task is described as a 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.* the topics are not known to the system in advance). In general, the goal for an ad-hoc retrieval tasks is: given an alphanumeric statement (or an image) describing a user information need, find as many relevant documents (images) as possible from the given collection, without manual intervention or any further interaction [165].

The participants are required to produce a ranked list of images for each ad-hoc topic in the test collection, whereby this list is ordered by the decreasing likelihood that the image is relevant for that particular topic. Based on a function of the ranks of the relevant images within such a list, the effectiveness of the retrieval strategy can be computed for each single topic (see Section 3.5 for more information about these functions). The overall performance of a strategy can further be computed as the average score across the set of topics in the test collection [469].

Ad-hoc retrieval tasks are considered as the most common task type of standard evaluation queries: they are on the agenda in nearly all evaluation events that follow the TREC methodology (see also Section 3.6); they further resemble and simulate the mechanism of present day internet search engines such as *Google* or *Yahoo!*, which also provide a ranked list of relevant documents (images) on their first search page.

Target Search

The goal of target searches is to find exactly one particular image in a data collection to satisfy a specific user information need. An example of such a target search could

be: "Find that image of my friend Fernando with Marat Safin at the 2007 Australian Open". Since the required image is rarely found in an ad-hoc manner (*i.e.* on the first result page after only one search iteration), such evaluation efforts are often characterised by a more interactive component, in which aspects like relevance feedback or system usability play a significant role. The interactive tasks of *iCLEF* [137] and *ImageCLEF* [63] have carried out evaluation for these aspects.

In tasks in which only one image in the collection is relevant for a search request, the response time becomes increasingly important. In searches for an altered input image, search strategies also have to show a certain level of retrieval accuracy [287]. *ImagEVAL*, for example, offered such a task for searches of altered input images in 2006 [118].

Automatic Image Annotation and Classification

Another popular challenge in image retrieval evaluation are automatic image annotation and classification tasks in the field of object recognition. The goal is as follows: given an image collection and a certain number of predefined classes, identify objects that belong to these classes and label them accordingly (automatic annotation), or assign each of the images to one of these classes (automatic classification) respectively.

In general, participants would receive development (or training) data together with the predefined categories which can be used to optimise their algorithms. Then, the unclassified test data is sent out, and the participants are asked to automatically classify the images to the predefined groups handed out in the evaluation data. The most common measure for performance is the ratio of incorrectly classified images to the total number of images in the test data (error rate).

Automatic image annotation and classification tasks have been carried out at evaluation events such as *ImageCLEF* [60, 86], the *MUSCLE CIS coin competition* [323] and the *PASCAL Visual Object Classes Challenge* [109, 110].

Other Types

There exists a number of other functions that could be tested with a benchmark, but have not been offered in evaluation events thus far. For example, retrieval based on a sketch of an image would allow for an interesting evaluation challenge: retrieval strategies would have to deal with the additional challenge of incomplete information because the object of interest is not a photo itself but only a drawn sketch, often without texture, colour or background information. Other ideas for evaluation tasks are listed in [154, 287].

3.3.2 Dimensions of Topic Development

Deciding on which ad-hoc topics to include in the test collection 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 [64]. Thus, there are a number of factors [224] that should be taken into consideration when creating a topic for ad-hoc retrieval tasks. 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;
- 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" [338, page 1069]. This section presents a number of key considerations that are essential to guide the topic development process and to achieve the aforementioned goal.

User Need Analysis

The selection of topics should not only be representative of the collection, but also be based on real-world queries. Although some publications report on how real users query image databases in general (see Section 2.2.1 for further details), a preselection 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 specifications should subsequently be chosen for those topic candidates that seem appropriate to compare systems as well [293].

One approach to obtain a pool of candidate topics is to carry out a *log file analysis*. For example, the topics for the ad-hoc retrieval task from a historic photographic collection (SAC) at *ImageCLEF* 2004 [63] 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 [62] were based on a small survey administered to clinicians, researchers, educators, students and librarians at *Oregon Health & Science University* [289].

An alternative but innovative approach for the creation of topic candidates was taken by *INEX Multimedia* in 2006 [484]: 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 [224] in order to guarantee realistic and representative search requests (with the participants simulating the real users).

Number of Topics

Once a pool of candidate topics is established, the next question inherently arises: How many topics should be selected from this pool and be given to the participants? Retrieval system effectiveness is generally known to vary widely across topics, thus

the greater the number of topics, the more confident the experimenter can be in its conclusions [469]. 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 [234] while Spärck-Jones and Rijsbergen [416] found 250 usually acceptable, though little quantitative evidence exists to support these suggestions [473]. In 1998, Voorhees [466] 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 [475].

Estimated Size of Target Set

While many publications have discussed the appropriate number of topics, the number of expected relevant images for each of the topics has almost been ignored. Larsen [224] or Clough [63] 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 [234], 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.

Geographic Constraints

Several previous studies (for example [379, 503]) as as well as recent reports on current trends in online search engines (*e.g. Google Zeitgeist*[66]) have shown 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 [379]. 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.

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* [62, 63, 64], for example, particular topics were selected to deal with these challenges and further potential problems that are encountered during the translation of the topics as well. Examples of such (multilingual) text retrieval challenges include: dealing with proper names, compound words, abbreviations, morphological variants, idioms, acronyms and equivalent syntactic and semantic expressions [338, 492].

[66] `http://www.google.com/intl/en/press/zeitgeist.html`

Topic Difficulty

One of the key dimensions of the topic development process for VIR evaluation 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 decent results, which might prevent them from presenting their work and taking part in comparative evaluations. Providing a good variation in topic difficulty is therefore crucial 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 VIR (see also Section 5.1.2), little work has considered topic difficulty as a dimension for the topic development process: Eguchi *et al.* [102] investigated the topic difficulty for *NTCIR*[67] (*NII Test Collection for IR Systems*), which is the Asian counterpart of CLEF and conducts evaluation for cross-language information retrieval (CLIR) with a focus on Asian languages. However, no work has considered topic complexity for TBIR benchmarks, which is one of the major contributions described in this research and is further elaborated on in Chapter 5.

Other Factors

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

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

3.3.3 Topic Components

Most VIR evaluation events that offer ad-hoc retrieval tasks (such as [62, 118]) have adapted the TREC ad-hoc retrieval task to meet the needs of VIR. This also includes the TREC conception of topics: structured statements of user needs from which the queries are extracted [338]. Thus, topics in VIR events often comprise the following components: a title, a narrative description, and sample images. All these components explain the same information need, but for a different purpose.

Topic Title

The *topic title* is a short statement of a particular information need and corresponds to what a user would type into a concept-based search engine to satisfy that need. This can include one or several words, noun phrases or short sentences (for example: *small sailing boat*[68]). Topic titles are used in all concept-based retrieval evaluations.

Topic Narrative

The *topic narrative* is a more detailed, but clear and precise definition of the information need in order to unambiguously determine whether or not a particular image fulfils the given need. For example, at *ImageCLEF* 2005 [62], the following narrative description for the aforementioned topic title *small sailing boats* was used:

> *Relevant images will show one or more small sailing vessels (ships, boats). Sailing boats are exemplified by having masts and sails, and small typically means that the vessel will have no more than two sails. The boats can be at sea or docked on land, but must have visible sails to indicate the vessel is a sailing ship.*

Such narrative descriptions have been used in many evaluation events, including TREC, CLEF, ImageCLEF and INEX Multimedia [65, 167, 338, 486].

[68] Example taken from ImageCLEF 2005.

Sample Images

Sample images provide a visual description of the information need and are often included to attract a larger number of participants using visual (*e.g.* CBIR) approaches [64] or to support combined concept-based and content-based retrieval methods [62, 63]. Offering more than one sample image is hereby of utmost significance as QBE based on only one image lacks the necessary information to clearly identify the target-query concept and cannot achieve satisfactory query results, as shown in [30]. Figure 3.22 illustrates the sample images[69] used at *ImageCLEF* 2005 [62] for the topic *small sailing boat*.

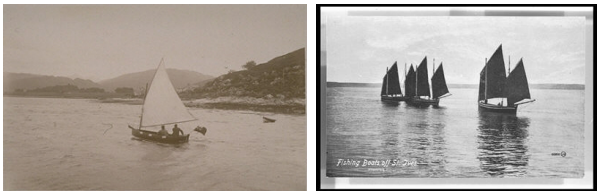


Figure 3.22: Sample images for topic "small sailing boat".

3.4 Relevance Assessments

Relevance assessments are a crucial part of a benchmark as they provide the link between a document (image) collection and the representative search tasks (topics); they are what turns a set of images and queries into a test collection by stating which images are relevant or not (and automatically creating the so called *ground-truth*) for each of the topics.

Two areas of concern can be identified with relevance assessments in general, which are both discussed in this section hereinafter: Section 3.4.1 is concerned with the quality and subjectivity of the judgments, while Section 3.4.2 deals with judgment completeness and introduces the state-of-the-art methods for relevance assessments.

[69] Taken from the SAC.

3.4.1 Assessment Quality and Subjectivity

Although "relevance" is a fundamental concept for information retrieval, it is not yet completely understood and often interpreted in different ways, also because of an inconsistently used terminology. An exhaustive review on the abundance of different definitions for relevance can be found in [277], while a classification of these definitions is attempted in [278].

Multi-grade Judgments

The majority of test collections have viewed relevance judgments as binary ("relevant" or "not relevant"), whereby documents were judged as relevant if any part of it is relevant, regardless of how small the proportion of that part is in relation to the entire document [165, 465]. Hence, most of the performance indicators for the evaluation of information retrieval were constructed based on these binary judgments (see Section 3.5).

While this simplification is certainly helpful for system designers and evaluators, multi-grade relevance judgments are generally believed to be more natural and closer to the judgments made in real life [419]. More often than not, some documents will appear to be more relevant than others, either because they contain more relevant information, or because the information they contain is highly relevant [202]. Moreover in VIR, there are often situations in which an image appears relevant, but one cannot be certain whether it fulfils the exact need described in the topic, because:

- only a part of the image is relevant;
- the required object is not recognisable or obscured;
- the relevant object is too small or in the background;
- the image appears relevant, but the caption is unable to confirm its contents.

Due to these different degrees or relevance and further potential uncertainties in the assessors' judgments, multi-grade relevance judgments are increasingly used in

current evaluation events. For example, ImageCLEF [62, 63, 64] has adopted a ternary judgment scheme: *relevant*, *partially relevant*, and *not relevant*; NTCIR [202, 209] uses four different degrees of relevance: highly relevant (S), relevant (A), partially relevant (B), and irrelevant (C).

Assessor Disagreement

Relevance assessment is an entirely subjective process, and this is especially true for the judgment of images. Different levels of prior knowledge and experience can influence the interpretation and understanding of an image. Furthermore, images can have different meanings or evoke different emotions for different people, or even mean different things to the same person under different circumstances or at different times [264, 396].

As a consequence, there will always exist a certain degree of disagreement among assessors about what constitutes a relevant image for a certain request. Yet, it is precisely because of the subjective nature of relevance assessments that IR evaluation has traditionally relied on a single assessor per query: in this way, the judgments are internally consistent, but admittedly represent only one single opinion [165].

However, more recent evaluation events have also adopted multiple judgments to reduce the influence of assessment subjectiveness. For instance, at TREC [466] and ImageCLEF [62, 63, 64], the topic author carries out the judgments as a primary assessor, with at least two additional assessors for each of the query topics, while NTCIR [202] avoids the involvement of the topic authors in the relevance assessment process and distributes the task over three different groups of users: information specialists, subject specialists, and ordinary users.

Creation of Relevance Sets

The use of multi-graded judgments and several assessors inherently leads to the creation of several sets of relevant documents (images) per query topic. Yet, the majority of IR evaluation metrics are constructed based on a single set of binary judgments (see Section 3.5); as a consequence, the individual multi-graded judg-

ments of the relevance assessors need to be combined to one set of relevant documents/images (so called *qrels*) for each query topic. One approach is thereby to combine the different sets of relevance judgments based on a particular level of overlap of relevant images between assessors. The resulting set of relevant images can thereby be established by either forming the *union* or *intersection* of the individual assessments [466].

In addition, the multi-graded relevance judgments also allow the creation of different relevance sets. At ImageCLEF, for example, *strict* sets contain images marked as relevant, and *relaxed* sets those marked relevant and partially relevant, and a total of four sets of relevance judgments is offered by the organizers: *union-strict*, *union-relaxed*, *intersection-strict*, *intersection-relaxed* [64].

NTCIR, by comparison, offers two sets of qrels for each topic: a *rigid relevant* set containing highly relevant and relevant images, and a *relaxed relevant* set containing highly relevant, relevant and partially relevant images [209]. In an alternative attempt to determine the rigid and relaxed sets, weights are assigned to each degree of relevance ($S = 3, A = 2, B = 1, C = 0$) and the average judgment of all assessors for each image and topic is scaled to the interval $(0, 1)$; the *rigid relevant* set would then contain images with an average relevance value between 0.67 and 1, and the *relaxed relevant* set the images with a value between 0.33 and 1 [202].

3.4.2 Assessment Coverage and Completeness

Ideally, every document in the collection would be judged for relevance for each of the topics. Such exhaustive examination of the collection was carried out for collections in the 1960s, 1970s and 1980s when the sizes were still small, *e.g. Cranfield* [58], until Spärck Jones and Van Rijsbergen pointed out that such a strategy would not work with larger collections [415].

For large-scale evaluation events, it is practically not feasible to judge the relevance of every image for each of the search topics due to limited resources and time. Consequently, a number of approaches have been taken to reduce the judg-

ment effort without compromising too much assessment coverage and completeness. This section provides an overview of the state-of-the-art approaches for relevance judgments that have proved successful in image retrieval evaluation events.

Pooling

One approach to reduce the cost in order to gain these judgments is the *pooling method* [416], whereby a set of candidate documents is created for each of the topics by computing the union of the top-ranked n documents from each of the independent retrieval efforts (runs) of the participants for that particular topic. Assessors (preferably experts in that domain) would then only judge the relevance of these candidate documents (also called *document pools*) for each query, rather than the entire collection.

This pooling approach is based on the assumption that highly ranked documents from each run will contain relevant documents, and hence nearly all relevant documents would consequently be found in these pools. Any unjudged document is assumed to be irrelevant by default [132]. The ranked lists should ideally originate from a diverse range of systems, because pooling can favour techniques that are variations of other systems whose results have been used in the document pool, and novel technologies with a large number of unjudged documents could thus be slightly disadvantaged [287]. Another question to be considered is the adequate choice for the size n: having n too small can lead to wrong recall values, while choosing n too high only yields in larger document pools and increases the assessment costs [165, 305].

Harman [167] and Voorhees [466] examined and approved the completeness of this approach. Further, Zobel [510] showed that, although pooling can affect results, these changes are rather small, not biased and below the level of user subjectivity. As a consequence, forming the ground-truth from the pooled output of multiple retrieval systems has become the standard process for many evaluation events [63, 165, 215, 217].

Several improvements to the pooling methods were also presented. For example,

Zobel [510] predicts the number of relevant documents in the lower ranks based on the number of relevant documents at the top of the ranking and implicitly forms a more shallow pool to save assessment time. In [68], *move-to-front* (MTF) pooling was proposed, which prioritises those systems that contribute the most relevant images to a document pool. In doing so, the assessment costs can be reduced considerably: the system ranks based on the examination of only 10% of the pool using MTF pooling and those based on an examination of the entire pool showed a correlation of 0.99, which can be considered equivalent. However, the use of MTF pooling runs the risk of biasing the relevance judgments in favour of systems that have high precision: if a run does not retrieve any relevant documents early in its ranking, it is going to be ignored even if it returns a lot of relevant documents lower in the ranking [467].

Interactive Search and Judge

The *interactive search and judge* (ISJ) strategy to create a set of relevance judgments was first introduced in 1998 [68] as an alternative to the pooling method. The main idea is to use a combination of interactive searching, judging and query re-formulation for each of the topics; the job of the assessor(s) is to search as many variations or refinements of the topic that one could think of, and only when no more relevant documents could be found, they would move on to the next topic.

Relevance judgments only based on ISJ can be quite time-consuming (Cormack [68] reports an average of more than two hours per topic for TREC-6), and the completeness of the judgments can be highly dependent on the assessor's level of motivation. ISJ is therefore often used to complement the pools as proposed by [217] and used at NTCIR and ImageCLEF [60, 62, 63, 64]. It was found that the combination of the pooling method and ISJ improves the coverage of relevant documents (in particular with queries that require a more general image), which consequently enhances recall [64, 217].

Predefined Ground-Truth

So far, all the approaches were concerned with building relevance assessments (and reducing the effort in building those) after the participants had submitted their results. However, it is also possible for some tasks to establish the ground-truth prior to result submission. For example, the test images in most object classification tasks such as [61, 86, 323] are pre-tagged prior to the evaluation or even come in already classified folders, which allows for a fast and efficient evaluation afterwards, but also provides a possibility for "cheating" (*e.g.* by knowing the ground-truth in advance, participants could alter their submissions manually).

The automatic pre-generation of relevance judgments based on the text captions in well annotated image collections was proposed in [201]. Although this would reduce the assessment efforts to a minimum, it would only shift the workload from the relevance assessment to the image annotation process, which is also expensive. Further, the quality and completeness of the generated ground-truth might be questionable unless such semantic image representations can model (1) user subjectiveness and (2) the fact that users might see the same image as a query for completely different goals, which is certainly not a trivial task [287]. The vantage of this approach, on the other hand, is that these alphanumeric image representations can be reused for the generation of relevance assessments in the future [298]. This approach has not been used in evaluation events thus far.

Some image collections such as the *Corel Photo CDs* offer predefined groups of similar images ("winter", "Colorado"), which would ease the relevance assessment process as these groups could directly be reused as search queries and would implicitly form the ground-truth as well. The limited number of such groups or terms supplied by a collection, however, might not be sufficient to create realistic search tasks. In other collections, it might be possible to create types of queries for which one can be certain that only a limited part of the collections will contain relevant images [287, 378].

Recent Development

Due to the high cost involved with the establishment of a ground-truth for topics in test collections, researchers have recently published many papers on reducing the relevance assessment effort. For example, Sanderson [378] presents different ways of building test collections without the use of system pooling. For instance, systems can be ranked reliably from a set of judgements obtained from a single system or from iterating relevance feedback runs. Aslam [11] proposes a technique based on random sampling, indicating that highly accurate estimates of standard performance measures can be obtained using a number of relevance judgments as small as 4% of the typical TREC-style judgment pool. Soboroff [413] randomly assigns relevance to documents in a pool and concludes that this could actually give a decent ranking of systems, while Carterette [44] proposes a method to construct a test collection with minimal relevance assessments.

3.5 Performance Measures

Once these relevance judgments are completed, the performance of the systems can be *evaluated* in accordance to a predefined performance measure and the results can be *analysed*. However, the question of which measures to use for the evaluation of retrieval effectiveness is not a trivial one to answer and has therefore received a lot of attention in the literature [302, 454].

Since VIR behaviour is, in general, too complex to be quantified in one single number, researchers have proposed a wide range of measures to quantify retrieval performance (for example, over 130 numbers based on more than 30 different measures are calculated by `trec_eval`[70], the evaluation software used at TREC). Different evaluation measures exhibit different properties with respect to:

- how closely related they are with user satisfaction criteria,
- how easy they are to interpret,

[70] `trec.nist.gov/trec_eval/`

- how meaningful aggregates such as average values are, and
- how much power they have to discriminate among retrieval results.

Consequently, since different performance measures evaluate different aspects of retrieval behaviour, these evaluation measures must be carefully selected in order to match the original objectives of an evaluation event [35]. This contradicts earlier publications such as [234, 287, 302] that had unsuccessfully attempted to establish a standardised set of performance measures for visual information evaluation in general.

Further, another fundamental factor that has to be considered for the selection of performance measures is the associated *error rate* - the likely error associated with the conclusion that "method A is better than method B". It has been shown that this error rate is inversely proportional to the number of topics used, and also that there are significant differences in error rates for various evaluation measures. This is not to say that measures with higher error rates should not be used, but rather that more topics or larger differences in score between retrieval strategies are required for evaluation experiments based on measures with higher error rates (in comparison with experiments based on measures with lower error rates) in order to be equally confident in the conclusion that one method is better than another [34].

This section describes the main evaluation measures for ad-hoc retrieval performance in the field of VIR. Most measures for ad-hoc retrieval tasks are based on ranked lists of images for each topic in a test collection: the higher the rank of a document, the more likely it is to be relevant for that particular topic (see also Section 3.3.1). In general, these measures use a function of the ranks of the relevant images in these lists to quantify the effectiveness of a retrieval method for each topic, which is then often aggregated across all topics in a test collection to describe the method's overall performance [469].

3.5.1 Precision and Recall-Based Measures

A large number of performance measures for ad-hoc tasks in VIR are based on the concept of *precision* and *recall*.

Precision and Recall

Let N denote the total number of images in a collection, n the number of images retrieved for a topic, n_r the number of relevant images retrieved, and r the total number of relevant images for that topic in the database. The *precision* P is defined as the ratio of relevant images retrieved to the total number of images retrieved,

$$P = \frac{n_r}{n}, \tag{3.1}$$

and thus measures the ability of a system to present *only* relevant items, whereas the *recall* R is defined as the ratio of relevant images retrieved to the total number of relevant images in the collection,

$$R = \frac{n_r}{r}, \tag{3.2}$$

hence, measuring the ability of a system to present *all* relevant items. Although both measures give a good indication of system performance, they are insufficient if they are just considered alone: one could always achieve $R = 1$ by just retrieving all images, or keep the precision high by retrieving only a few images. Precision and recall should therefore either be used together, for example in *precision-recall graphs*, or the number of images retrieved, a *cut-off value*, should be specified.

Precision and Recall at Cut-Off Values

Let $P(n)$ and $R(n)$ denote the precision and recall at n retrieved documents (cut-off value) hereinafter. The most significant cut-off values include:

P(10). The precision at 10 documents retrieved is easy to interpret and closely correlates with user satisfaction in tasks such as web searching [35]. Thus, $P(10)$ is

one of the most used performance measures and has been applied in many evaluation events such as *ImageCLEF* [62, 63, 64] or *INEX Multimedia* [486, 511].

However, the use of $P(10)$ as a measure for retrieval performance also comprises the two following problems: first, it lacks discrimination power among retrieval methods because the only change that affects $P(10)$ is a relevant image entering or leaving the top 10 ranks; and second, it averages very poorly because the constant cut-off at 10 images yields very different recall levels for different topics (*i.e.* the meaning of $P(10)$ is very different for a topic with 6 relevant images compared to a topic with 200 relevant images). Hence, due to both issues, $P(10)$ exhibits a substantial margin of error, which would require an increased number of topics (100) to allow for robust evaluation, according to a study by Buckley *et al.* [34].

P(20) and P(30). Buckley also showed that the error rate decreases with an increasing cut-off value, hence, in order to avoid such high numbers of topics, the precision at higher cut-off values such as $P(20)$ and $P(30)$ can be used, as was the case in, for instance, [62, 465]. While these measures show much more discriminating power and lower error rates in comparison with $P(10)$, high cut-off values automatically produce wrong and misleading values for topics with a small set of relevant images. For example, perfect retrieval for a topic with only 12 relevant images would only result in $P(30) = 0.4$.

P(r). TREC [465] bypassed this problem by using the so called *r-precision*, which is a special case for a precision value at a certain cut-off: r-precision is the precision after r documents have been retrieved, where r is the number of relevant images for the topic. Not only does this measure address the problems associated with precision at a constant cut-off level by evaluating each topic at the level where precision and recall are the same, it also allows for a more meaningful aggregation of the results across several topics and shows a much smaller margin for error than $P(10)$ [35]; r-precision was one of the measures of *ImageCLEF* in 2005 [62].

The precision at $r = 0.5$ is another special case of the r-precision as it is con-

cerned with the precision after half of the relevant images have been retrieved. This value corresponds to the value that would be plotted at $r = 0.5$ in a *recall-precision graph*; it is an interpolated value since $r = 0.5$ is undefined for topics that have an odd number of relevant images [454].

R(n). While most real-world applications are aiming towards high precision, there are also applications in which high recall is required. For example, in face recognition applications for crime prevention, it is vital to detect all potential terror threats, and one would prefer to browse through many result pages rather than miss out on the wanted target.

In retrieval evaluation, however, recall based performance measures have mostly been neglected so far. *ImageCLEF* used overall recall as one of its performance measures in 2005 [62], TREC has made use of $R(P(0.5))$ and $R(1000)$ [465], while it was also proposed to use $R(100)$ as a measure for the *Benchathlon* [299].

Graphical Representations

Precision versus recall graphs (PR graphs) are a standard evaluation method in IR [165] and have increasingly been used by the VIR community as well [428]. The benefits of PR graphs can be found in the abundance of information they carry, their quick and easy interpretation and the fact that they can distinguish well between different results (see Figure 3.23, taken from [302]).

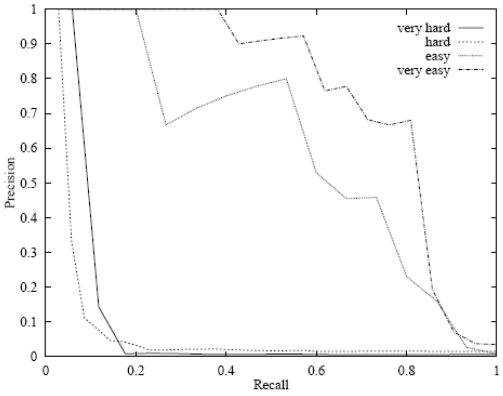
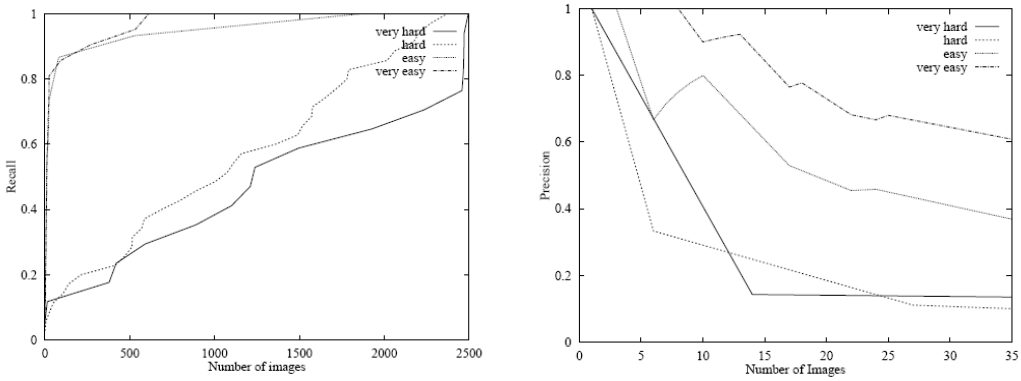


Figure 3.23: Precision vs. recall graph.

The limitations of a PR graph include its dependency on the number of relevant images for a given query and the fact that it is not possible to obtain practical information such as precision or recall after a given number of images have been retrieved. The latter drawback can be bypassed by separating the PR graph into a precision vs. number of images and a recall vs. number of images retrieved graph (see Figure 3.24, taken from [302]). Although *precision graphs* are very akin to PR



Recall graph. Precision graph.

Figure 3.24: Graphical representations.

graphs, they give a better indication of what might be a good number of images to retrieve, but are also more sensitive to the number of relevant images for a given topic. *Recall graphs*, on the other hand, look more positive than PR graphs and can distinguish well between hard and easy topics, but struggle to discriminate the easy from the very easy ones [287, 302].

Average Precision (AP)

The *average precision (AP)* is the non-interpolated, arithmetic mean of the precision values of all relevant images. Let P_i denote the precision of the relevant image i, and r the total number of relevant images for a query, then:

$$\mathrm{AP} = \frac{1}{r} \sum_{i=1}^{r} P_i \tag{3.3}$$

Hence, the precision value is calculated after each relevant image is retrieved (if a relevant image is not retrieved, its precision value is assumed to be 0.0) and then

averaged to get a single value for the performance of a query. This corresponds to the area underneath the recall-precision graph for the topic and reflects the performance over all relevant images, rewarding systems that retrieve relevant images quickly (*i.e.* highly ranked relevant images) [34].

AP is a very powerful and stable measure as it is based on much more information (and exhibits a much lower error rate) than precision values at a certain cut-off level like $P(20)$ or $P(r)$. The main drawback of AP is that it is not easily interpreted: an AP score of 0.3, for instance, can arise in a variety of ways, whereas a $P(20)$ score of 0.3 can only mean that 6 out of the top 20 images are relevant [35].

Mean Average Precision (MAP)

The *mean average precision (MAP)* is the non-interpolated, arithmetic mean of the average precision values of all individual topics. Let AP_i denote the average precision of topic number i, and Q the total number of query topics in a retrieval evaluation task, then:

$$\text{MAP} = \frac{1}{Q} \sum_{i=1}^{Q} AP_i \tag{3.4}$$

Since MAP inherits all the benefits of the AP as well, it has become one of the leading evaluation measures for ad-hoc retrieval evaluation (see Section 3.6).

However, one significant drawback of MAP lies in the fact that scores of better performing topics can mask changes in the scores of poorly performing topics; this can be overcome by using the geometric mean instead of the average mean [35].

Geometric Mean Average Precision (GMAP)

An essential feature of an image retrieval system is the ability to return at least passable results for any topic. System performance is usually reported as average performance, whereas an individual user only sees the effectiveness of a system on his or her requests and not the average performance of a system (and users who do not retrieve any relevant images for a request will hardly be consoled by the fact that other people's requests might be executed more successfully) [472].

Hence, the *geometric mean average precision (GMAP)* was introduced in the TREC 2004 robust track [470] in order to highlight poorly performing topics while remaining stable with as few as 50 topics; *GMAP* is the geometric mean of the average precision values across all topics.

Let AP_i denote the average precision of a query topic i, and Q the total number of query topics in a retrieval evaluation task:

$$\text{GMAP} = \left(\prod_{i=1}^{Q} AP_i \right)^{\frac{1}{Q}} \tag{3.5}$$

This measure emphasises topic scores close to 0.0 (the "bad results") while minimising differences between larger scores (the "good results") and therefore does not let better performing topics mask weaker ones. For example, if a run doubled the average precision score for topic A from 0.03 to 0.06 while decreasing topic B from 0.50 to 0.47, then MAP would be unchanged whereas GMAP would show an improvement.

Yet, one problem associated with this equation is the following: the average precision score for a single topic could also amount to 0, which would mean that *GMAP* for all topics would be 0 too (regardless of the precision values of all other topics). This can be avoided by transforming the Equation 3.5 using logarithmic identities, *i.e.* taking the log of the individual topic's average precision score, computing the arithmetic means of the logs, and exponentiating back for the final geometric MAP score:

$$\text{GMAP} = \exp \left[\frac{1}{Q} \sum_{i=1}^{Q} \ln AP(i) \right] \tag{3.6}$$

where

$$AP(i) = \begin{cases} \lambda & \text{if } AP_i = 0 \\ AP_i & \text{otherwise} \end{cases} \tag{3.7}$$

and λ is a value that is much lower that the least significant digit of the result generation script. Should the average precision value for a topic account to 0, then λ is added before taking the log in order to prevent single values from being 0. The `trec_eval` script, for example, reports scores to four significant digits, thus $\lambda = 0.00001$ is used for each topic in which no relevant images were found.

Excluding the research presented in this thesis, *GMAP* has only been used in text retrieval tasks so far [471, 472].

3.5.2 Rank-Based Measures

Apart from precision and/or recall based measures, a number of rank-based measures have been developed to quantify retrieval effectiveness as well. The main difference is that rank-based measures are neither influenced by the number of retrieved images nor by the total number of images in the data collection.

Rank of the First Relevant Image

The simplest measure based on rank is the *rank of the first relevant image* ($Rank_1$). As indicated by its name, the value of $Rank_1$ expresses the position of the first relevant image in an ordered result list.

Although this measure is very easy to interpret and has been used in TREC [474] as well as in CBIR [302], there are claims [465] that it is a poor measure of retrieval performance: first, it is unstable because a single topic can have an unreasonable effect on the average score; and second, large differences in a score might not correspond to how a user would perceive the same difference. For example, one system ranking the first relevant image of a difficult topic at rank 174 and another one ranking the first relevant image for the same topic at rank 892 will cause a large difference in the average score; to the user, on the other hand, this difference is virtually meaningless as both systems perform poorly.

Reciprocal Rank

Often, the reciprocal value of the rank of the first relevant image is used to scale the measure to the interval $[0, 1]$. Let $\mathcal{R}_R$ denote the set of images retrieved and $Rank_1$ the rank of the first relevant image r_1; the *Reciprocal Rank* $Rank_{rec}$ is defined as:

$$Rank_{rec} = \begin{cases} \frac{1}{Rank_1} & \text{if } r_1 \in \mathcal{R}_R \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

The reciprocal rank is more suitable than $Rank_1$ to compare averages and is therefore considered as a major measure in `trec_eval`, whereas $Rank_1$ is only a minor measure.

Average Rank of Relevant Images

Another measure based on rank is the *average rank of relevant images* [126]. Let $Rank_k$ denote the rank of the k^{th} relevant image in an ordered result set of images, and n_r the number of relevant images retrieved. The average rank of relevant images is defined as:

$$\overline{Rank} = \frac{1}{n_r} \sum_{k=1}^{n_r} Rank_k \tag{3.9}$$

Although this measure is a good indication for system performance, it is quite difficult to interpret since it depends on both the collection size and the number of relevant images for a query; it has proven to be quite unstable as just one relevant image with a very high rank can considerably affect its value. Hence, [154] introduced the *normalised average rank of relevant images* in order to make the measure less vulnerable to such potential outliers.

Mean Reciprocal Rank of Relevant Images

This single-valued measure is similar to the averaged rank measures mentioned before. Let $Rank_k$ denote the rank of the k^{th} relevant image in an ordered result set of images, and n_r the number of relevant images retrieved. The *mean reciprocal rank (MRR)* of relevant images is defined as:

$$\text{MRR} = \frac{1}{n_r} \sum_{k=1}^{n_r} \frac{1}{Rank_k} \tag{3.10}$$

MRR is both a good indicator for system performance and a stable measure and has therefore been used as a primary measure at *ImagEVAL* [279].

Binary Preference

Another single-valued measure that is formed on basis of only judged images is the *binary preference (bpref)* [35]. This measure is a function of the number of times

judged non-relevant images are ranked before relevant documents: let r denote the number of relevant images for a topic, r_k a relevant image and n_{rjn} a member of the first r judged non-relevant images as retrieved by a system:

$$bpref = \frac{1}{r}\sum_{k=1}^{r} 1 - \frac{|n_{rjn} \text{ ranked higher than } r_k|}{r} \tag{3.11}$$

This measure is very robust in the face of incomplete relevant judgments and has become more and more popular in large-scale information retrieval evaluation events such as [486].

Further, *bpref* is now also a major measure of `trec_eval`, although its implementation follows a slightly different definition compared to the one mentioned above:

$$bpref = \frac{1}{r}\sum_{k=1}^{r} 1 - \frac{|n_{rjn} \text{ ranked higher than } r_k|}{\min(r, n_{rjn})} \tag{3.12}$$

3.5.3 Measures for Other Tasks

The scope of this research concentrates on ad-hoc retrieval tasks. Thus, measures for other evaluation paradigms are only briefly introduced hereinafter.

Object Recognition Tasks

A common measure for object or face recognition is the *error rate (er)* [182], which is often also referred to as *fallout* [186, 233] and is defined as

$$er = \frac{n - r}{n} \tag{3.13}$$

where n is the total number of documents retrieved and r the number of relevant images. This measure is also used in object recognition tasks of *ImageCLEF* [61, 86].

Correct and *incorrect detection* were used in an object recognition context [330]. *Correct detection* refers to the number of correctly classified objects and *incorrect detection* to the number of incorrectly classified objects. These measures are equivalent to *precision* and *error rate* when they are divided by the number of retrieved images.

The *Wolf and Jolion metric (WJM)* is a measure for text area detection, which was proposed by Wolf and Jolion in [491] and was also used for a text area detection task at *ImagEVAL* 2006 [279].

Other performance measures for object classification are the *equal error rate (EER)* and the *area under curve (AUC)*, which are both based on the *receiver operator characteristic (ROC)* curve analysis and have been used as a primary measure for the object classification tasks at the *PASCAL Visual Object Classes Challenge* in 2005 and 2006 [109, 110].

Interactive Tasks

Interactive tasks are concerned with the study of VIR from a user-centred perspective. One of the crucial measures, as far as the usability and interactivity of a system is concerned, is the *response time* of a system: the interval between the instant at which a user submits an image query to a system and the instant at which the complete answer is received [319].

In interactive tasks, in which only one image in the collection is relevant for a search request, the performance can be measured by the *number of images* that a user has to look at before finding the target, the *time lapsed* or *links clicked* until the target is found [136].

Further, *aggregated precision* and *recall* values, which also consider relevance feedback for general multi-stage image retrieval, are proposed in [234], while [77, 237] provide further links to existing literature on relevance feedback.

3.6 Evaluation Events

The last four sections have dealt with the major components of a TREC-style evaluation: the development of document collections, the creation of representative search queries (topics), different forms of relevance assessments to establish a ground-truth, and a variety of performance measures in order to quantify that relevance and to rank submissions of participants.

However, all these benchmark components are not beneficial unless researchers can be motivated to make use of them – and the best motivation is evaluation events, in which researchers can use these aforementioned components to evaluate their systems, and then meet with other researchers in a "friendly" evaluation forum to present, discuss and compare their approaches and the corresponding results.

Evaluation events are probably the most significant component of a benchmark as they ensure that the other four components do not just exist in theory (which had been the case for a long time) but are also employed practically. Table 3.1 provides an overview of currently existing evaluation events (showing the number of participants and the number of tasks offered in parenthesis).

Event/Year	2001	2002	2003	2004	2005	2006
VidTREC, TRECVID	12 (2)	17 (3)	24 (4)	33 (4)	42 (5)	54 (4)
ImageCLEF			4 (2)	17 (3)	24 (4)	42 (5)
Pascal VOC					12 (4)	26 (4)
INEX MM					5 (1)	4 (2)
ImagEVAL						11 (5)
MUSCLE CIS						3 (1)

Table 3.1: Overview of VIR evaluation events.

This section describes such evaluation events[71] for VIR, which have only recently surfaced in the last few years.

3.6.1 TRECVID

In 2001, TREC [468] introduced a video track to provide an evaluation framework for video retrieval [403]. This track soon grew to an independent entity, the TREC Video Retrieval Evaluation (TRECVID[72]) in 2003 [401] and is generally considered as the first and (to date most successful) evaluation effort in the domain of VIR, as also indicated by the increasing number of participants each year (see Table 3.1).

[71] Evaluation events in very specific and application-dependent domains such as face recognition or fingerprint detection are not included.

[72] `http://www.nlpir.nist.gov/projects/trecvid/`

Evaluation Goals and Tasks

The main goal of TRECVID is to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. As a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations, the following challenges have been offered to its participants in the first six years [214, 215, 328, 401, 403, 404] of its existence:

- *Shot boundary detection.* This is an introductory task: identify the shot boundaries with their location and type (cut or gradual) in a given video. Shots are fundamental units of video, useful for higher-level processing.

- *High-level feature extraction.* Given a high-level concept such as "People", "Speech", "Indoor/Outdoor", "City/Landscape", "Day/Night", return a list of shots ranked by decreasing likelihood of detecting the presence of the feature.

- *Video search.* Given a multimedia statement of information need (topic), return a ranked list of at most 1000 common reference shots from the test collection which best satisfy the need. This task has included interactive, manually-assisted and/or fully automatic runs over the years.

- *Story segmentation.* Given a video test collection, identify the story boundaries with their location (time) and type (news or miscellaneous) in the given video clip(s). A story can be composed of different shots.

- *Low-level feature extraction.* This task is concerned with camera motion: given the video test collection, identify all shots for each of the three low-level features (left/right, up/down, zoom in/out) in which that feature (camera motion) is present.

- *Rushes exploitation.* Free exploration of rushes, which are the raw material (extra video, B-rolls footage) used to produce a video.

Participation

Table 3.2 provides an overview of the tasks that were run during the last six years, together with their corresponding number of participating groups (only groups that officially submitted their results are considered).

Task / Year	2001	2002	2003	2004	2005	2006
Shot boundary detection	9	8	14	17	21	25
High-level feature extraction		14	10	12	22	31
Video search	8	16	11	16	20	26
Story segmentation			8	8		
Low-level feature extraction					14	
Rushes exploitation					7	12
Total	12	17	24	33	42	54
Number of tasks	2	3	4	4	5	4

Table 3.2: Number of TRECVID participants per task and year.

Methodology

Most of the test data for TRECVID were broadcast news videos in MPEG-1 format from the USA (provided by NBC, CNN, MSNBC), but also Chinese (CCTV4, *Phoenix*, NTDTV) or Arabic video sources (LBC, *Al Hurra*) were used. The pooling method was applied for the relevance assessments in most search and feature extraction tasks [214, 215, 328, 401].

Further, apart from the standard precision and recall based measures for its search and feature extraction tasks, TRECVID used many other video-specific performance measures. For example, measures for the shot boundary detection tasks included *inserted transition count*, *deleted transition count*, *correction rate*, *deletion rate*, *insertion rate*, *error rate*, *quality index* and *correction probability* [403].

Analysis and Discussion

Although video retrieval is, in principle, different to image retrieval, TRECVID was briefly introduced in this section because of its similarities to image retrieval including the retrieval of key frames, the applied techniques for content-based retrieval, and the likeness of the search tasks which are often on a simple semantic level as

well. Other evaluation campaigns for video retrieval include ETISEO[73], PETS[74], AMI[75] and ARGOS[76].

3.6.2 ImageCLEF

The *Cross Language Evaluation Forum* (CLEF[77]) is, like TRECVID, a spin-off from TREC and has been focussing on multilingual information retrieval as an independent campaign since 2000. Following the successful examples of TREC and TRECVID, *ImageCLEF*[78] began as a part of CLEF in 2003 [342] and has arguably been the most significant and influential evaluation event for image retrieval thus far, because:

1. it was the first image benchmarking event to finally fulfil the calls for a TREC-style evaluation framework for image retrieval [139, 302, 407];

2. it has been attracting the largest number of research groups out of all image retrieval evaluation events: 24 in 2005 [63] and 42 in 2006 [62];

3. it has been focussing on realistic applications and has, similar to TREC, created some useful resources which allow the retrospective reproduction of the evaluation results.

Evaluation Goals and Tasks

The main goal of *ImageCLEF* is to investigate the effectiveness of combining text and image for retrieval, to collect and provide resources for benchmarking image retrieval systems, to promote the exchange of ideas which may help improve the performance of future image retrieval systems, and to evaluate these systems in a multilingual environment: the language used to express the associated texts or

[73] http://www.silogic.fr/etiseo/
[74] http://www.pets2006.net/
[75] http://www.amiproject.org/
[76] http://www.irit.fr/recherches/SAMOVA/MEMBERS/JOLY/argos/
[77] http://clef-campaign.org/
[78] http://ir.shef.ac.uk/imageclef/

textual queries should not affect retrieval, *i.e.* an image with a caption written in English should be searchable in languages other than English.

To achieve this goals, the following tasks have therefore been offered to its participants in the first four years [60, 62, 63, 64, 290] of its existence:

- *Ad-hoc retrieval.* Given an alphanumeric statement (and/or sample images) describing a user information need, find as many relevant images as possible from the given document collection (with the query language either being identical or different from that used to describe the images).

- *Object recognition.* Given an image collection and a certain number of predefined classes, assign each of the images to one of these classes (automatic *classification* tasks) or identify objects that belong to these classes and label the images accordingly (automatic *annotation* tasks).

- *Interactive evaluation.* This task is concerned with the study of cross-language image retrieval from a user-centred perspective.

ImageCLEF has provided these tasks within two main areas: retrieval of images from photographic collections and retrieval of images from medical collections. These domains offer realistic scenarios in which to test the performance of image retrieval systems, offering different challenges and problems to participating research groups.

Participation

The variety of tasks and domains have helped to attract participants from both academic and commercial research groups worldwide from communities including CLIR, CBIR and user interaction.

Table 3.3 provides an overview of the tasks that were run during the last four years, together with their corresponding number of participating groups (again, only groups that officially submitted their results are considered).

Year	2003	2004	2005	2006
Ad-hoc (historic photographs)	4	12	11	
Ad-hoc (medical photographs)		12	13	12
Ad-hoc (generic photographs)				12
Ad-hoc (visual only)				2
Automatic annotation (medical)			12	12
Automatic annotation (generic)				3
Interactive evaluation	1	2	3	(3)
Number of participants (total)	4	17	24	42
Number of tasks	2	3	4	5

Table 3.3: Number of ImageCLEF participants per task and year.

Methodology

The document collections for the ad-hoc retrieval tasks comprised historic photographs from 2003 to 2005 (*SAC*, see Section 3.2.2), medical images from 2004 to 2006 (*ImageCLEFmed*, see Section 3.2.5) and generic photographs in 2006 (*IAPR TC-12 Image Benchmark*, see Chapter 4). The object recognition tasks were carried out on medical collections (*IRMA*) in 2005 and 2006, and on a generic collection (*LTU*) in 2006. The interactive experiments also used the SAC from 2003 to 2005; in 2006, they were held as part of iCLEF [137] and used a general collection (*FlickR*). See Section 3.2.7 for a description of the *IRMA*, *LTU* and *FlickR* collections.

In 2003, 50 search queries (topics) in the form of short verbal descriptions, longer narratives and relevant sample images were provided for the only ad-hoc task from the historic collection; 2004 saw the creation of 25 such topics, while 28 of them were developed in 2005. For the medical ad-hoc tasks, 26 purely visual topics had been created in 2004, with 30 combined semantic and visual topics being developed in 2005 and 2006 each. In the medical annotation tasks, the participants were asked to categorise the images into 57 classes in 2005 and into 116 classes in 2006. In the general object annotation task, the images had to be assigned to 21 categories.

As for relevance assessments, the pooling technique combined with ISJ to complete the relevance assessments was used to establish the ground-truth for both the general and medical ad-hoc retrieval tasks. For the object annotation tasks, the ground-truth for the test data had been pre-determined by their image captions.

The primary performance measures for the ad-hoc tasks were *MAP*, *P(10)*, *P(r)*, the number of relevant images retrieved (n_r) and the total recall R, while the error rate (er) was used for the automatic object recognition and annotation tasks.

Analysis and Discussion

Further information on the *ImageCLEF* evaluation events can be found in the corresponding overview papers [60, 62, 63, 64, 86, 136, 137, 290]. Publications outside ImageCLEF include [65, 289, 288].

3.6.3 ImagEVAL

ImagEVAL [118, 279] conducts a slightly alternative approach to evaluation of multimedia information retrieval and concentrates on very specific search scenarios as required by their industry partners. After a first test campaign had locally been organised in 2005 to gain experience, the first official event was held in 2006 and also attracted participation from outside the borders of France.

Evaluation Goals and Tasks

The main goal of this campaign is to offer usage-oriented evaluation for multimedia retrieval (in comparison to the sometimes rather artificial laboratory style evaluations) in order to minimise the gap between technology evaluation and realistic industry needs. Thus, after several discussions with potential users of image retrieval systems, *ImagEVAL* offered the following five tasks to its participants:

1. *Copyright protection.* Given an original image, find all transformed images in an image collection (Task 1.1) and vice versa (Task 1.2). This resembles the search for illegally reused copyrighted images.

2. *Finding images to illustrate text.* Given a semantic concept (text), find a sample image to illustrate this concept.

3. *Text detection.* Given an image collection, find and interpret alphanumeric information (written text) in these images.

4. *Object identification.* Identify objects (or classes of objects) like cars, tanks, aircrafts, road signs, etc. in an image collection.

5. *Semantic feature extraction.* Given an image, differentiate whether it is black-and-white or colour, indoor or outdoor, night or day, town or countryside, *etc.* This concerns the recognition of general and context information.

Participation

More than 20 teams filed their registration for *ImagEVAL* 2006, out of which 11 (9 French, 2 international) also eventually submitted their results. Table 3.4 provides an overview of the participation for each individual task.

Task number	1.1	1.2	2	3	4	5
Participating groups	6	5	4	2	3	6
Runs submitted	21	12	22	5	9	12

Table 3.4: Participation at ImagEVAL 2006.

Methodology

ImagEVAL consequently built a diverse corpus covering the usage of their commercial partners and, at the same time, catering for each of the aforementioned tasks (see Section 3.2.6 and [344] for further information on the *ImagEVAL Corpora*). As for relevance assessments, the entire *ImagEVAL Corpora* had been pre-tagged by professional indexers to establish a ground-truth a priori. MAP was used as

Task number	1.1	1.2	2	3	4	5
Size of test set	45000	45000	700	500	1400	23572
Number of topics	50	60	25	500	10	13
Performance measure	MAP	MRR	MAP	WJM	MAP	MAP

Table 3.5: ImagEVAL 2006 task overview.

the primary performance measure for most tasks, MRR was used for Task 1.2 and the WJM for Task 3. Table 3.4 provides an overview of the methodology for each individual task.

Analysis and Discussion

The involvement of industry and real users in the preparation of the campaign and the creation of the image collection has certainly had a positive influence on the usage-oriented aspect of the tasks, but this also led to strict copyright regulations that limit the redistribution of the collections and thus also the reproducibility of the results. It might therefore be vital to the survival of *ImagEVAL* to shift to more accessible data collections in the future and to also offer these to non-participating researchers retrospectively. Further, rather than establishing a French counterpart to other retrieval evaluation events, it is felt that *ImagEVAL* should widen its horizon and make an effort to attract a more international audience as well.

3.6.4 PASCAL Visual Object Classes Challenge

The *PASCAL Visual Object Classes (VOC) Challenge* [109, 110] conducts evaluation of the state-of-the-art visual object recognition and localisation methods on "real-world" data. This yearly challenge is organised by the PASCAL Network of Excellence[79], it was first held in 2005 and has attracted participation across Europe and from the USA.

Evaluation Goals and Tasks

The main goal of the PASCAL VOC challenge is to recognise objects from a number of visual object classes in realistic scenes (*i.e.* objects are not pre-segmented); four classes (bicycle, car, motorbike, people) were selected in 2005, and ten (bicycle, bus, car, motorbike, cat, cow, dog, horse, sheep, person) in 2006. In both years, two separate tasks were offered to the participants:

1. *Classification task:* Given a certain number of predefined object classes, predict the presence (or absence) of at least one object for each of these classes in a test image.

[79] `http://www.pascal-network.org/`

2. *Detection task:* Given a certain number of predefined object classes, predict the bounding boxes of all objects corresponding to any of these predefined classes in a test image (if present).

Both tasks were further subdivided into two subtasks each, depending on the choice of training data. The *PASCAL Object Recognition Database Collection* formed the underlying image corpus for both tasks, with different subsets being used in 2005 and 2006 (see Section 3.2.7).

Participation

The VOC challenge in 2005 saw the participation of 12 groups from 9 institutions, out of which 9 groups from 8 institutions submitted 17 different methods to the classification tasks, and 5 groups from 4 institutions submitted 10 different methods to the detection tasks. In 2006, a total of 26 groups from 16 institutions participated, with 18 groups from 14 institutions submitting 23 runs (23 methods) to the classification tasks and 9 groups from 7 institutions submitting 10 runs (9 methods[80]) to the detection tasks.

Methodology

The participants were only allowed to submit one result set per method and were asked to provide a real-valued confidence of each object's presence for the classification tasks and/or of the detection of each object's bounding box for the detection tasks. The ground-truth was predefined by the existing image captions. *EER* and *AUC* based on a *ROC* curve analysis were used as performance measures for the classification tasks. The detection task was judged by *PR graphs*, which themselves were based on the area of overlap between the detected bounding boxes and the ground-truth bounding boxes. The principal quantitative measure is the *average precision for 11 thresholds on recall* as used by TREC.

[80] One method was used in both subtasks.

Analysis and Discussion

The evaluation of a wide range of methods for object classification and detection provided a valuable snapshot of the state-of-the-art technology in these tasks. In both years, the retrieval results were outstanding (both *EER* and *AUC* at around 0.90). This is not surprising as the difficulty of the tasks and the number of classes was tailored to the state-of-the-art object recognition techniques.

Issues that were raised in 2005 include the occurrence of errors in the data set, the difficulty level of the training data (which was too easy in comparison to the test data), the fact that it was unclear which performance measures would be applied, and the fact that the test data had been released together with its ground-truth, which allowed participants to fine-tune their results. The number of classes had been increased from four (2005) to ten (2006), but still seems very low to be representative of realistic object recognition in "real-world" environments. In both years, it was hard to motivate participants to submit results based on their own, unlimited training data.

3.6.5 MUSCLE CIS Coin Competition

The *MUSCLE Coin Images Seibersdorf (CIS) Competition* [322, 323] is another evaluation event for visual object classification, which was held by the *European Union Network of Excellence MUSCLE*[81].

Evaluation Goals and Tasks

The goal of the competition, which was held in 2006 for the first time, was to classify 20,000 previously unseen images into 362 different coin classes. This high number of classes is different to other object recognition campaigns (compare Section 3.6.4), which generally use a low number of classes. Further, participants had to correctly classify at least 70% of the coins in the competition, and their programs were not allowed to take more than eight hours to process a run of 5,000 coins.

[81] `http://muscle.prip.tuwien.ac.at/`

Participation

Seven groups registered for the competition, but only three eventually submitted their results (although the winner of the competition was promised an award which was endowed with prize money of 1,500 Euro).

Methodology

The main evaluation resource of the competition was the *CIS Benchmark*: a database of 80,000 coin images with 692 coin classes and over 2,200 coin types (see Section 3.2.7). The participants were given 60,000 images as training data and were then asked to assign the test data (20,000 previously unseen images) to 362 classes.

As for relevance assessments, due to the narrow domain of the images, the organisers were able to pre-determine the ground-truth with minimal uncertainty using the supplied image captions. The submissions of the participants were subsequently ranked according to the following formula:

$$S = 1n_{cc} + 0n_{cu} + 25n_{co} - 100n_{cw} \tag{3.14}$$

where n_{cc} denotes the number of correctly classified coins (including unknown coins correctly classified as unknown), n_{cu} the number of known coins classified as unknown, n_{co} the number of coin types in the training set that is correctly classified at least once, and n_{cw} the number of wrongly classified coins.

Analysis and Discussion

Two out of the three submissions passed the 70% criterion, which showed that the benchmark offers an environment that is tailored to the current state-of-the-art object classification methods. Ideas for future competitions include the classification (1) by coin type or (2) of partially occluded coins. A second coin classification competition is planned for 2007.

However, the question remains whether the results of such a specific task using a very narrow image domain can be transferred to other walks of life as well. Further,

it is questionable whether it is beneficial to run an evaluation event as a competition rather than in a more "friendly" environment in which researchers gather to present, share and discuss their potential research ideas.

3.6.6 INEX Multimedia

The *INEX Multimedia (INEX MM)* track conducts evaluation of retrieval strategies for XML based multimedia documents [486, 511]. This event was held as a track of the *INitiative for the Evaluation of XML Retrieval*[82] *(INEX)* for the first time in 2005 and is currently preparing for its third year running.

Evaluation Goals and Tasks

The goal of *INEX MM* is to provide an evaluation platform for multimedia retrieval in structured collections. In 2006, two tasks were provided for the participants:

1. *Multimedia Fragments.* Given a multimedia information need, find as many relevant XML fragments as possible. This represents a traditional ad-hoc task and was the only task offered in 2005.

2. *Multimedia Images.* Given an information need, find as many relevant images as possible. Since the type of the target element (an image) is defined, this ad-hoc task is concerned with image retrieval rather than element retrieval. This task was new in 2006.

Participation

A total of five research groups submitted their results in 2005, with only four groups doing so one year later in 2006.

Methodology

INEX MM used the *Lonely Planet collection* in 2005 and the *Wikipedia Multimedia Corpus* in 2006 (see Sections 3.2.3 and 3.2.4). In 2005, 25 topics were created

[82] `http://inex.is.informatik.uni-duisburg.de/`

whereby the participants had to provide the organisers with six candidate topics each (see [224] for a detailed description of the topic development process). The results had to be submitted in form of ranked lists of the relevant XML fragments for each topic. As for relevance assessments, the pooling technique was applied and the participants were involved in the assessment process (which was a requirement for participation). *MAP*, *P(10)* and *bpref* were used as the primary performance measure to report on the overall retrieval performance of the runs. Interpolated *recall-precision averages* and *MAP per topic* were also used for further analysis.

Analysis and Discussion

In 2005, INEX had created a solid basis for further instances of the multimedia track in the following years. By 2006, some of the major issues raised in 2005 could be solved, for example the heavily copyrighted *Lonely Planet* collection had been replaced by the more accessible *Wikipedia Multimedia Corpus*.

INEX will organise a multimedia track in 2007 again, whereby the major goals include to attract a larger number of participants (to allow for a more reliable evaluation of the retrieval results), and to motivate participants to make more use of the excellent visual resources provided to initiate the query process [485].

3.7 Summary

This chapter provided an overview of performance evaluation in the field of VIR. After an exploration of the young but active development phase together with some criticism of this research area, the major components of the most commonly used methodology for the evaluation of VIR systems, the TREC methodology, were described. These main benchmark components comprise a standardised image collection, representative search topics, relevance judgments to associate a ground-truth of relevant images for each of these topics, a set of performance measures according to which the results of the systems are evaluated, and evaluation events in which the other four components are used.

An analysis of each of these components that are currently used in retrieval evaluation events shows that most of them suffer from several limitations, which leave a lot of room for improvement: copyright restrictions hinder the redistribution of image collections and limit the reproducibility for other researchers; the search topics are often not representative of real-world information needs and are often either too difficult (and sometimes too easy) for the state-of-the-art image retrieval methods; most of the collections are static and cannot be easily adapted to different requirements or a change as far as evaluation goals are concerned; and there are no evaluation efforts that are concerned with the retrieval from generic collections of real-world photographs (such as pictures from holidays or sporting events).

The next four chapters of this dissertation will take on exactly these issues. Chapter 4 first presents the design of a representative collection of generic (real-world) photographs which is available free-of-charge and without copyright restrictions that would hinder its redistribution. Chapter 5 presents a framework for topic development and introduces a novel dimension to control the difficulty of these queries. Chapter 6 reports on an parametric benchmark administration architecture which allows for the quick adaptation to altered evaluation goals. Finally, Chapter 7 reports on the first evaluation event for multilingual ad-hoc retrieval from a generic photographic collection.