Chapter 1
Seven Years of Image Retrieval Evaluation

Paul Clough, Henning Müller, and Mark Sanderson

Abstract In this chapter we discuss evaluation of Information Retrieval (IR) systems and in particular ImageCLEF, a large–scale evaluation campaign that has produced several publicly–accessible resources required for evaluating visual information retrieval systems and is the focus of this book. This chapter sets the scene for the book by describing the purpose of system and user–centred evaluation, the purpose of test collections, the role of evaluation campaigns such as TREC and CLEF, our motivations for starting ImageCLEF and then a summary of the tracks run over the seven years (data, tasks and participants). The chapter will also provide an insight into lessons learned and experiences gained over the years spent organising ImageCLEF, and a summary of the main highlights.

1.1 Introduction

The contents of this book describe ImageCLEF, an initiative for evaluating cross–language image retrieval systems in a standardised manner thereby allowing comparison between the various approaches. ImageCLEF ran for the first time in 2003 as a part of the Cross–Language Evaluation Forum (CLEF), leading to seven years of activities which are summarised in this book. As of 2010, however, the ImageCLEF evaluation campaign is still running evaluation tasks. A major outcome of ImageCLEF has been the creation of a number of publicly–accessible evaluation

Paul Clough
University of Sheffield, Sheffield, United Kingdom e-mail: p.d.clough@sheffield.ac.uk

Henning Müller
University of Applied Sciences Western Switzerland (HES–SO), TechnoArk 3, 3960 Sierre, Switzerland e-mail: henning.mueller@sim.hcuge.ch

Mark Sanderson
University of Sheffield, Sheffield, United Kingdom e-mail: m.sanderson@sheffield.ac.uk

resources. These benchmarks have helped researchers develop new approaches to visual information retrieval and automatic annotation by enabling the performance of various approaches to be assessed. A further outcome, arguably less tangible but just as important, has been to encourage collaboration and interaction between members of various research communities, including image retrieval, computer vision, Cross–Language Information Retrieval (CLIR) and user interaction.

The possibility of creating a publicly available benchmark or test collection for evaluating cross–lingual image retrieval systems was a key objective of the Eurovision project[1]. This included dissemination through an international body, such as CLEF, and in 2002 a new multimedia evaluation task for CLEF was proposed (Sanderson and Clough, 2002). At the same time the CLEF community were looking for new avenues of research to complement the existing multi–lingual document retrieval tasks being offered to participants. Image retrieval was seen as a natural extension to existing CLEF tasks given the language neutrality of visual media, and motivated by wanting to enable multi–lingual users from a global community access to a growing body of multimedia information.

In addition the image retrieval community was calling for a standardised benchmark. Despite the many advances in areas such as visual information retrieval, computer vision, image analysis and pattern recognition over 20 or so years, far less effort has been placed on comparing and evaluating system performance (Müller et al, 2004). Although evaluation was conducted by some researchers, the availability of often only small and copyrighted databases made it hard to compare between systems and provide conclusive results. Calls for a systematic evaluation for image retrieval systems were suggested as a way to make further advances in the field and generate publicly–accessible evaluation resources (Smith, 1998; Goodrum, 2000; Müller et al, 2001), similar to evaluation exercises being carried out in text retrieval such as the U.S. Text REtrieval Conference or TREC[2] (Voorhees and Harman, 2005).

Although Forsyth (2002) argued that such an evaluation of content–based retrieval systems was not productive because the performance of such techniques was too low, the impact of having evaluation resources available for comparative evaluation could clearly be seen in events such as TREC in the text retrieval community and could equally be assumed to advance visual retrieval systems in a similar manner. Over the years, evaluation events such as Benchathlon[3], TRECVID[4], ImagEval[5] and ImageCLEF have helped to foster collaboration between members of the visual retrieval community and provide the frameworks and resources required for systematic and standardised evaluation of image and video retrieval systems. Chapter 27 discusses in more detail various evaluation campaigns for multimedia retrieval.

---

[1] The Eurovision project was funded by the UK Engineering and Physical Sciences Research Council (http://www.epsrc.ac.uk) grant number GR/R56778/01

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

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

[4] http://trecvid.nist.gov/

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

1.2 Evaluation of IR Systems

Evaluation is the process of assessing the 'worth' of something and evaluating the performance of IR systems is an important part of the development process (Saracevic, 1995; Robertson, 2008). For example, it is necessary to establish to what extent the system being developed meets the needs of the end user, to show the effects of changing the underlying system or its functionality on system performance, and enable quantitative comparison between different systems and approaches. However, although most agree that evaluation is important in IR, much debate exists on exactly how this evaluation should be carried out. Evaluation of retrieval systems tends to focus on either the system or the user. Saracevic (1995) distinguishes six levels of evaluation objectives, not mutually exclusive, for information systems, including IR systems:

1. The *engineering level* deals with aspects of technology, such as computer hardware and networks to assess issues such as reliability, errors, failures and faults.
2. The *input level* deals with assessing the inputs and contents of the system to assess aspects such as coverage of the document collection.
3. The *processing level* deals with how the inputs are processed to assess aspects such as the performance of algorithms for indexing and retrieval.
4. The *output level* deals with interactions with the system and output(s) obtained to assess aspects such as search interactions, feedback and outputs. This could include assessing usability for example.
5. The *use and user level* assesses how well the IR system supports people with their searching tasks in the wider context of information seeking behaviour (e.g. the user's specific seeking and work tasks). This could include, for example, assessing the quality of the information returned from the IR system for work tasks.
6. The *social level* deals with issues of impact on the environment (e.g. within an organisation) and could include assessing aspects such as productivity, effects on decision–making and socio–cognitive relevance.

The first three levels (1–3) are typically considered part of system–centred evaluation; the latter three (4–6) part of user–centred evaluation. For many years evaluation in IR has tended to focus on the first three levels, predominately through the use of standardised benchmarks (or test/reference collections) in a laboratory–style setting. The design of a standardised resource for IR evaluation was first proposed over 50 years ago by Cleverdon (1959) and has since been used in major information retrieval evaluation campaigns, such as TREC (Voorhees and Harman, 2005), CLEF (Peters and Braschler, 2001) and the NII Test Collection for IR Systems or NTCIR (Kando, 2003).

Over the years the creation of a standard test environment has proven invaluable for the design and evaluation of practical retrieval systems by enabling researchers to assess in an objective and systematic way the ability of retrieval systems to locate documents relevant to a specific user need. Although this type of evaluation has met with criticism, such as whether the performance of a system on a benchmark reflects

how a system will perform in an operational setting and the limited involvement of end users in evaluating systems, it cannot be denied that this kind of organised large–scale evaluation has done the field tremendous good, both within and outside the environment of evaluation campaigns (Chapter 27 describes the strengths and weaknesses of evaluation campaigns). However, it is important to acknowledge that IR systems are increasingly used in an interactive way and within social contexts. This has motivated evaluation from a user–centred evaluation perspective to assess performance at the latter three levels: output, use and user, and social (Borland, 2000; Dunlop, 2000; Ingwersen and Järvelin, 2005; Petrelli, 2008; Kelly, 2010). Projects such as MIRA (an evaluation framework for interactive and multimedia information retrieval applications) started to address this for visual information from 1996 (Dunlop, 2000).

The contents of this book are mainly related to system–centred evaluation of visual information retrieval systems: the resources generated to support evaluation and advances in image retrieval and annotation that have resulted from experiments within ImageCLEF. This is not to imply that user–centred evaluation has been ignored. In fact, from the very beginning ImageCLEF ran an interactive image retrieval task (described in Chapter 7) that was later subsumed by the interactive CLEF track (iCLEF). In addition, where possible, evaluation resources that are described in the following chapters, were designed with realistic operational settings in mind. However, our primary aim has been to first create the necessary resources and framework in which researchers could develop and compare underlying techniques for visual retrieval across multiple domains and tasks.

1.2.1 IR Test Collections

A core activity of evaluation campaigns such as TREC and CLEF has been to create reusable benchmarks for various tasks and domains in IR (Robertson, 2008; Sanderson, 2010 – to appear). Similar to other fields in science a benchmark provides a standard by which something can be measured. The design of a standardised resource for evaluation of document retrieval systems (a *test collection* was first proposed in the late 1950s in the Cranfield I and II projects (Cleverdon, 1959, 1991), and has since become the standard model for comparative evaluation of IR systems. In this approach to testing IR systems, commonly referred to as the Cranfield paradigm, the focus is on assessing the performance of how well a system can find documents of interest given a specification of the user's information need in a way that is abstracted from an operational environment. Laboratory–based evaluation is popular because user–based evaluation is costly and complex and it is often difficult to interpret results obtained with end users.

The main components of a typical IR test collection are:

1. A *collection of documents* representative of a given domain (each document is given a unique identifier *docid*). Collections created for and used in ImageCLEF are discussed in Chapter 2.

2. A set of *topics* or *queries* (each given a unique identifier *qid*) describing a user's information needs expressed as narrative text or sets of keywords. For image retrieval, topics may also include example relevant images. Topic creation within ImageCLEF is discussed further in Chapter 3.
3. A set of *relevance judgments (qrels)*, or ground truths, provide a representative sample of which documents in the collection are relevant to each topic (a list of qid/docid pairs). Although relevance judgments are commonly binary (relevant/not relevant) the use of *graded* relevance judgments is also commonly utilised in IR evaluation (e.g. highly relevant/partially relevant/not relevant). This has implications for which performance measures can be used to evaluate IR systems. The topic of gathering relevance assessments for ImageCLEF is discussed in Chapter 4.

Performance measures, such as precision and recall, are used to provide absolute measures of retrieval effectiveness, e.g. what proportion of relevant documents are returned by the IR system (see Chapter 5 for further details on IR evaluation measures). Together, the test collection and evaluation measures simulate the users of a search system in an operational setting. In evaluations such as CLEF, the focus is not on absolute values but on relative performance: system outputs can be compared and systems ranked according to scores obtained with the evaluation measures (i.e. comparative testing). Although test collections were originally used to evaluate ad hoc[6] retrieval, evaluation campaigns, such as TREC and CLEF, have extended the use of test collections to other tasks (e.g. document filtering and routing, document classification and automatic annotation).

Evaluation campaigns, such as TREC and CLEF, are founded upon the Cranfield paradigm and make use of test collections to evaluate various aspects of information access. However, a 'TREC–style' evaluation not only includes producing evaluation resources, such as test collections, but also community building through holding organised annual workshops to present and discuss findings with other researchers. Figure 1.1 shows activities commonly undertaken in the evaluation 'cycle' of TREC (although applicable to other campaigns such as CLEF and NTCIR). For TREC and CLEF this cycle operates runs during one year; some evaluation campaigns operate over a longer period (e.g. NTCIR runs the cycle over 18 months). The cycle begins with a call for participation followed by an expression of interest from participating groups and registration. Evaluation tasks are centred on tracks (e.g. ImageCLEF is a track of CLEF) that may involve one or many tasks. The track organisers must define their tasks for prospective participants in addition to preparing the document collection and topics. This may also involve preparing and releasing training data beforehand. The participants run their IR experiments according to a variety of parameters to produce system outputs in standard format (called *runs*) and will submit what they consider their *n* best runs to the evaluation campaign. Typically the runs

[6] Ad hoc retrieval as defined by TREC simulates the situation in which a system knows the set of documents to be searched, but the search topics are not known to the system in advance. It is also characterised by a detailed specification of the user's query (title, narrative description and keywords) and searches are required to achieve high recall.

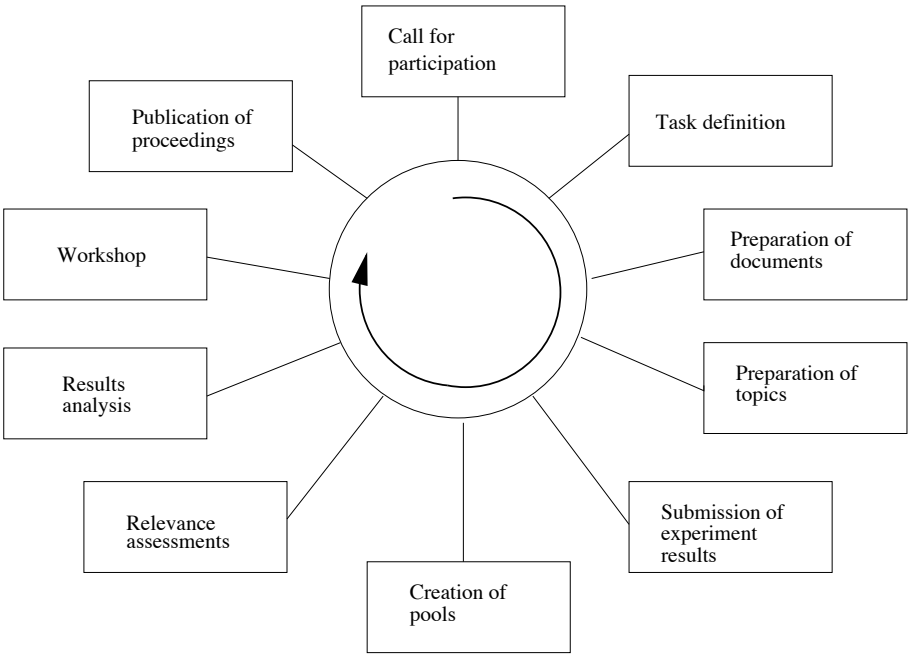

Fig. 1.1: Annual cycle of activities in a TREC–style evaluation (adapted from `http://trec.nist.gov/presentations/TREC2004/04intro.pdf`).

will be based on varying search parameters such as the use of relevance feedback or various combinations of visual and textual modalities.

A sub–set of runs, chosen by the organisers, is used to create *document pools*, one for each topic (Kuriyama et al, 2002). Domain experts (the assessors) are then asked to judge which documents in the pool are relevant or not. Document pools are created because in large collections it is infeasible to judge every single document for relevance. These assessments (qrels) are then used to assess the performance of submitted runs. Evaluation measures are used to assess run performance based on the number of relevant documents found. Although relevance is subjective and can vary between assessors, investigations have shown that relevance assessments can provide consistent evaluation results when ranking runs relative to one another (Voorhees, 2000). Results are released and analysed prior to holding a workshop event to share and discuss findings. Finally, the activities and results are written up in some kind of formal publication, such as workshop proceedings.

1.2.2 Cross–Language Evaluation Forum (CLEF)

CLEF began in 2000 to promote the development of multi–lingual information access systems (Peters and Braschler, 2001). CLEF grew out of the Cross–Language IR track of TREC that ran from 1997–1999. The aims of CLEF are[7] (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross–language contexts, and (ii) creating test–suites of reusable data which can be employed by system developers for benchmarking purposes. In the 2009 CLEF campaign the following main tracks were run:

- Ad hoc track, which deals with multi–lingual textual document retrieval;
- ImageCLEF track, which concerns cross–language retrieval in image collections;
- iCLEF track, which addresses interactive cross–language retrieval;
- QA@CLEF track, which covers multiple language question answering;
- INFILE track, which concentrates on multi–lingual information filtering;
- LogCLEF track, which copes with log analysis from search engine and digital library logs;
- CLEF–IP track, which studies multi–lingual access and retrieval in the area of patent retrieval;
- Grid@CLEF track, which performs systematic experiments on individual components of multi–lingual IR systems.

In total there have been 10 CLEF campaigns to date, involving around 200 different participating groups from around the world. Several hundred different research papers have been generated by CLEF participants over the years describing their evaluation experiments and the state of the art contributions to multi–lingual information access.

1.3 ImageCLEF

1.3.1 Aim and Objectives

ImageCLEF first ran in 2003 with the aim of investigating cross–language image retrieval in multiple domains. Retrieval from an image collection offers distinct characteristics and challenges with respect to one in which the document to be retrieved is text (Clough and Sanderson, 2006). For example, the way in which a query is formulated, the methods used for retrieval (e.g. based on low–level features derived from an image, or based on associated textual information such as a caption), the types of query, how relevance is assessed, the involvement of the user during the search process, and fundamental cognitive differences between the interpretation of visual versus textual media. For cross–lingual IR the problem is further complicated

[7] These aims have been taken from the CLEF website: `http://www.clef-campaign.org/`

by user queries being expressed in a language different to that of the document collection or by multi–lingual collections. This requires crossing the language barrier by translating the collection, the queries, or both into the same language. Although the tasks and data sets used in ImageCLEF changed over the years the objectives broadly remained the same:

- To investigate the effectiveness of combining textual and visual features for cross–lingual image retrieval. The combination of modalities is the subject of Chapter 6.
- To collect and provide resources for benchmarking image retrieval systems. These resources include data sets, topics and relevance assessments, which are discussed in Chapters 2–4 and in the track overviews (Chapters 7–12).
- To promote the exchange of ideas to help improve the performance of future image retrieval systems. Work from selected participants from ImageCLEF 2009 is found in Chapters 14–24.

To meet these objectives a number of tasks have been organised by ImageCLEF within two main domains: (1) medical image retrieval and (2) non–medical image retrieval, including historical archives, news photographic collections and Wikipedia pages. Broadly speaking the tasks fell within the following categories: ad hoc retrieval, object and concept recognition, and interactive image retrieval.

Ad hoc retrieval. This simulates a classic document retrieval task: given a statement describing a user's information need, find as many relevant documents as possible and rank the results by relevance. In the case of cross–lingual retrieval the language of the query is different from the language of the metadata used to describe the image. Ad hoc tasks have been run by ImageCLEF from 2003 to 2009 for medical retrieval and non–medical retrieval scenarios, see Chapters 7 and 12 respectively.

Object and concept recognition. Although ad hoc retrieval is a core image retrieval task, a common precursor is to identify whether certain objects from a pre–defined set of classes are contained in an image (object class recognition), assign textual labels or descriptions to an image (automatic image annotation) or classify images into one or many classes (automatic image classification). Chapters 11 and 12 summarise the ImageCLEF object and concept recognition tasks, including medical image classification.

Interactive image retrieval. Image retrieval systems are commonly used by people interacting with them. From 2003 a user–centred task was run as a part of ImageCLEF and eventually subsumed by the interactive CLEF (iCLEF) track in 2005. Interaction in image retrieval can be studied with respect to how effectively the system supports users with query formulation, query translation (in the case of cross–lingual IR), document selection and document examination. See Chapter 7 for further details on the interactive image retrieval tasks of CLEF.

Table 1.1: Participation in the ImageCLEF tasks 2002–2009, distinct number of participants by year and chapter references for further details.

Task	2003	2004	2005	2006	2007	2008	2009	See Chapter
General images								
Photographic retrieval	4	12	11	12	20	24	19	8
Interactive image retrieval	1	2	*2*	*3*	–	*6*	*6*	7
Object and concept recognition				4	7	11	19	11
Wikipedia image retrieval						12	8	9
Robot vision task							7	10
Medical images								
Medical image retrieval		12	13	12	13	15	17	13
Medical image classification			12	12	10	6	7	12
Total (distinct)	4	17	24	30	35	45	65	

1.3.2 Tasks and Participants

Table 1.1 summarise the tasks run during ImageCLEF between 2003 and 2009 and shows the number of participants for each task along with the distinct number of participants in each year. The number of participants and tasks offered by ImageCLEF has continued to grow steadily throughout the years from four participants and two tasks in 2003 to 65 participants and seven tasks in 2009. Participants have come from around the world to participate in ImageCLEF from both academic and commercial institutions. It is difficult to summarise all of the ImageCLEF activities between 2003 and 2009 and we have not provided an exhaustive account, but in brief these are some of the key events year by year:

- In *2003* the first ImageCLEF task was run at the 4th CLEF workshop by Mark Sanderson and Paul Clough involving two tasks and four participants.
- For *2004* a medical image retrieval task organised by Henning Müller was added to ImageCLEF giving a total of three different tasks. This attracted submissions from 17 participating groups and began the focus for us on medical images.
- In *2005* a new medical image annotation task was introduced bringing the total number of tasks offered to four. William Hersh, Thomas Deserno, Michael Grubinger and Thomas Deselaers joined the organisers and we received approximately 300 runs from 24 participants. The interactive task moved to iCLEF in collaboration with Julio Gonzalo and Jussi Karlgren.
- In *2006* 30 participants submitted runs to four tasks that included a new non–medical object annotation task organised by Allan Hanbury and Thomas Deselaers. A new data set (IAPR–TC12) was also developed for the ad hoc retrieval task (referred to as ImageCLEFphoto).
- In *2007* a total of 35 participants submitted runs to four tasks: multi–lingual ad hoc retrieval, medical image retrieval, hierarchical automatic image annotation for medical images and photographic annotation through detection of objects, a purely visual task. Jayashree Kalpathy–Cramer joined the organising team.

- In *2008* we included a new task for cross–lingual image retrieval from Wikipedia (called WikipediaMM) where participants could exploit the structure of Wikipedia for retrieval. This attracted submissions from 12 participants and overall a total of 45 groups submitted over 1,000 runs to ImageCLEF tasks. The photographic retrieval task experimented with promoting diversity in image retrieval and the interactive task, now a part of iCLEF, created a novel evaluation utilising data from Flickr and undertaking log analysis. Thomas Arni, Theodora Tsikrika and Jana Kludas joined the organisers.
- The *2009* ImageCLEF track was run at the 10th and final CLEF workshop. We had the largest number of participants to ImageCLEF (65 groups) across six tasks which included a new robot vision task organised by Andrzej Pronobis and Barbara Caputo that attracted seven participants. Monica Lestari Paramita also joined the organising team of the ImageCLEFphoto task that used a new data set from Belga, a news agency from Belgium, containing over 500,000 images.

1.3.3 Data sets

A major contribution of ImageCLEF has been to collect a variety of data sets for use in different tasks. Table 1.2 shows all 16 data sets used in ImageCLEF over the seven years, which are further discussed in Chapter 2. The table shows the data set, year added to the ImageCLEF campaign, the total number of images contained in the data set and languages used to annotate the image metadata. For data sets where the same data set has been used but added to in subsequent years, such as the Radiological Society of North America (RSNA), the final number of images has been reported in the table. Clearly noticeable is that many collections are annotated in English. As a cross–language track of CLEF the focus has been primarily on translating user's queries (*query translation*) for bilingual retrieval from a query in a non–English language into English. Other CLEF tracks have focused on other cross–language issues such as bilingual retrieval between other language pairs and multi–lingual retrieval: searching document collections that contain texts in multiple languages.

1.3.4 Contributions

Each of the overview chapters in this book (Chapters 7–13) provides a description of activities conducted in ImageCLEF and summarises contributions made in each of the areas covered. This includes a summary of test collections and ground truths produced for each task that have been used within various research communities. It is clear from the participant's reports (Chapters 14–24) that many novel and interesting techniques have been developed as a part of the experiments carried out for ImageCLEF. This highlights the benefits of TREC–style evaluation for IR sys-

Table 1.2: A summary of data sets used in ImageCLEF 2003–2009.

Data set	Year Added	#Images	Annotation Languages
General images			
St Andrews (SAC)	2003	28,133	English
IAPR–TC12	2006	20,000	English, Spanish, German
Belga	2009	498,920	English
LTU	2006	1,100	–
PASCAL VOC	2007	2,600	–
Flickr MIR	2009	25,000	–
INEX MM	2008	150,000	English
KTH–IDOL2	2009		
Medical images			
IRMA	2005	14,410	–
Casimage	2004	8,725	English, French
MIR	2005	1,177	English
PEIR	2005	32,319	English
PathoPIC	2005	7,805	English, German
MyPACS	2007	15,140	English
CORI	2007	1,496	English
RSNA	2008	75,000	English

tems. Chapter 27 highlights the benefits (and limitations) of evaluation campaigns for multimedia retrieval researchers, but overall we believe that ImageCLEF has made a number of contributions including the following:

Reuseable benchmarks: one of the largest obstacles in creating a test collection for public use is securing a suitable collection of images for which copyright permission is agreed. This has been a major factor influencing the data sets used in the ImageCLEF campaigns. The ImageCLEF test collections provide a unique contribution to publicly available test collections and complement existing evaluation resources for a range of retrieval tasks and scenarios. These resources include the IAPR–TC12 photographic collection (Grubinger et al, 2006), a segmented version of the IAPR–TC12 data set (Escalante et al, 2010) and Casimage (Müller et al, 2004).

Evaluation measures: a range of performance measures have been experimented with or developed for ImageCLEF including Geometric Mean Average Precision (GMAP), Cluster Recall (for assessing diversity) and a new evaluation metric based on ontology scoring for the 2009 image annotation task (Nowak et al, 2010).

Open forum for exchange of research: ImageCLEF has actively promoted discussion at the CLEF workshops about approaches to ImageCLEF tasks. In addition, a number of activities[8] have been organised in conjunction with the CLEF workshop and a number of European projects: the First, Second and Third MUSCLE/ImageCLEF Workshops on Image and Video Retrieval Evaluation in

[8] See `http://www.imageclef.org/events/` for further details and access to workshop proceedings.

2005–2007, the QUAERO/ImageCLEF Workshop on Multimedia Information Retrieval Evaluation in 2008 and the Theseus/ImageCLEF Workshop on Multimedia Information Retrieval Evaluation.

Publications: the CLEF workshop proceedings provide a published set of formal papers that describe ImageCLEF activities over the years. In addition, the organisers of ImageCLEF co–ordinated a Special Issue on Image and Video Retrieval Evaluation (Hanbury et al, 2010) in the journal Computer Vision and Image Understanding (CVIU) and a Special Issue on Medical Image Annotation in ImageCLEF 2007 (Deselaers et al, 2009) for Pattern Recognition Letters (PRL).

Advances in state of the art: ImageCLEF has run various tasks in different image retrieval settings. For example the medical image retrieval task has provided a set of resources for assessing the performance of medical retrieval systems based upon realistic tasks and topics. The organisers have involved medical professionals in creating realistic tasks and carrying out relevance assessments. Chapter 6 on fusion techniques for combining textual and visual information demonstrates a positive contribution in exploring the use of multiple modalities for image retrieval.

1.3.5 Organisational Challenges

Based on our experiences with ImageCLEF over the past seven years we have encountered a number of challenges with running a TREC–style multimedia retrieval evaluation benchmark. The main organisational challenges are detailed below with suggested solutions (adapted from Müller et al (2007)).

One of the greatest challenges facing the organisation of ImageCLEF has been *funding*. Organising a successful event requires a certain level of commitment from the organisers and their host institutions, e.g. to create suitable data sets, organise and pay for relevance assessments, to maintain regular communication with participants and assist with producing publications from the evaluation event (e.g. workshop proceedings). The ImageCLEF organisers have relied on the support of national and international funding bodies in addition to voluntary effort. Running an evaluation campaign over several years requires thinking about funding beyond the lifetime of a single research project. A strength of ImageCLEF has been to involve several different people to distribute the workload and costs.

To produce reusable evaluation resources for multimedia retrieval systems requires *obtaining access* to data sets and *permission* from the owners to distribute the content to participating groups. This is a significant challenge for high–quality multimedia data sets that are often copyrighted and subject to limited distribution. ImageCLEF has been able to gain access to a number of data sets, some with little or no copyright restrictions. Availability of data sets has a direct impact on what can be evaluated in the evaluation campaign and on reusability of the data set after the lifetime of the evaluation campaign.

A difficult task is often *advertising* the evaluation campaign and *motivating participation*. This is particularly relevant to multimedia retrieval where it is often time–consuming to develop systems for specific tasks and submit runs. This is clearly seen by comparing the number of groups that register for the task (to obtain the data sets) compared to the number who eventually submit results: commonly lower than 50%. ImageCLEF has also had to actively advertise the event across multiple domains because of the cross–disciplinary nature of the tasks. ImageCLEF has benefitted from being part of CLEF that already had a following of participants, was well–known in the IR field and offered participants the chance to publish their results in a good quality publication: the Springer Lecture Notes in Computer Science, after the workshop.

An often difficult task has been to encourage *input from commercial organisations*: both collaborating with organisers (e.g. to suggest suitable search tasks) and participating in the evaluation event itself. Ideally having commercial input enables participants to tackle current real–world challenges and offer businesses an opportunity to investigate what state of the art approaches can achieve on their data sets. The 2010 CLEF campaign has been organised around themes that both academics and businesses have identified as important areas of research requiring investigation.

Creating *realistic tasks and user models* is important in estimating the effectiveness of systems in an operational setting based on results obtained in a laboratory–setting using the benchmarks provided. In ImageCLEF, for example, we have developed realistic search tasks and queries based on the knowledge of experts (e.g. discussions with medical professionals in the case of the medical image retrieval tasks) and analysing query logs generated by existing search systems.

A further challenge in ImageCLEF has been to *efficiently create the ground truths*. This is linked with funding as it is often an extensive and time–consuming task. Approaches such as pooling and interactive search and judge are often used to reduce the amount of assessor time required for judging the relevance of documents, but completeness of relevance judgments and variations amongst assessors must be taken into account. A further issue is that criteria for assessing relevance in multimedia retrieval is often different from assessing the results of text retrieval systems, particularly for medical images (Sedghi et al, 2009). This may require the use of domain experts to make the judgments which relies on access to such people and their availability to make judgments.

1.4 Conclusions

To improve multimedia retrieval systems we need to have appropriate evaluation resources, such as test collections, that offer researchers access to visual data sets, example queries and relevance judgments. Over the past seven years ImageCLEF has provided such resources, together with providing a forum in which researchers have been able to interact and discuss their findings. ImageCLEF has provided mainly resources for system–centred evaluation of image retrieval systems, but has also

maintained a relationship with user–centred evaluation of image retrieval systems, mainly through its relationship with the CLEF interactive track (iCLEF).

However, there are still many issues to address with regards to evaluation and the results of ImageCLEF by no means provide a 'silver bullet' solution to evaluating image retrieval systems. There is still a tension between running system–centred and user–centred evaluation on a large scale for image retrieval (e.g. (Forsyth, 2002)). Most image retrieval in practice is interactive and should be seen as a priority for future image retrieval evaluation campaigns. Attempts have been made to run interactive tasks, but participation continued to be low across the years. This is not just a problem with image retrieval but an issue with IR evaluation in general.

Specific areas that are still ripe for exploration include: investigating which performance measures best reflect user's satisfaction with image retrieval systems and incorporating measures such as system response time; further investigation of the information seeking behaviours of users searching for images, such as their goals and motivations, search contexts, the queries issued and their reformulation strategies, and especially criteria shaping a user's notion of relevance; assessing user behaviours such as browsing, an important search strategy for image retrieval; continuing to develop publicly–accessible data sets covering multiple domains, tasks and varying in size; investigating the utility of test collections in image retrieval evaluation, especially with respect to the user to generate realistic test resources. Only by doing this can we start to address some of the concerns expressed by researchers such as Saracevic (1995), Forsyth (2002) and Smith (1998).

Acknowledgements We gratefully acknowledge the help of Norman Reid from St. Andrews University Library in Scotland for providing us access to the historic set of photographs for the first ImageCLEF evaluation campaign. Our thanks also extend to all other data providers who have enabled us to distribute and use their content. We thank the national and international funding bodies who have supported ImageCLEF over the years in a variety of ways. In particular we thank the European Union for support through the funding of various EU projects that have supported task organisers. Thanks go to all those involved in carrying out relevance assessments across the tasks without which whom we would have no gold standard to benchmark against. Finally a huge thank you goes to Carol Peters, the co–ordinator of CLEF, who supported ImageCLEF from the very beginning and through seven years of activities.

References

Borland P (2000) Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation 56(1):71–90

Cleverdon C (1959) The evaluation of systems used in information retrieval. In: Proceedings of the International Conference on Scientific Information — Two Volumes, Washington : National Academy of Sciences, National Research Council, pp 687–698

Cleverdon CW (1991) The significance of the Cranfield tests on index languages. In: 14th annual international ACM SIGIR conference on research and development in information retrieval, ACM, Chicago, IL, USA, pp 3–12

Clough PD, Sanderson M (2006) User experiments with the eurovision cross–language image retrieval system. Journal of the American Society for Information Science and Technology

57(5):679–708
Deselaers T, Müller H, Deserno TM (2009) Editorial to the special issue on medical image annotation in ImageCLEF 2007. Pattern Recognition Letters 29(15):1987
Dunlop M (2000) Reflections on MIRA: Interactive evaluation in information retrieval. Journal of the American Society for Information Science 51(14):1269–1274
Escalante HJ, Hernández CA, Gonzalez JA, López-López A, Montes M, Morales EF, Sucar LE, Villaseñor L, Grubinger M (2010) The segmented and annotated IAPR TC–12 benchmark. Computer Vision and Image Understanding 114(4):419–428
Forsyth D (2002) Benchmarks for storage and retrieval in multimedia databases. In: Proceedings of storage and retrieval for media databases, SPIE Photonics West Conference, pp 240–247
Goodrum A (2000) Image information retrieval: An overview of current research. Informing Science 3(2):63–66
Grubinger M, Clough PD, Müller H, Deselaers T (2006) The IAPR TC–12 benchmark — a new evaluation resource for visual information systems. In: Proceedings of the International Workshop OntoImage 2006 Language Resources for Content–Based Image Retrieval, held in conjunction with LREC 2006, pp 13–23
Hanbury A, Clough PD, Müller H (2010) Special issue on image and video retrieval evaluation. Computer Vision and Image Understanding 114:409–410
Ingwersen P, Järvelin K (2005) The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series). Springer, Secaucus, NJ, USA
Kando N (2003) Evaluation of information access technologies at the NTCIR workshop. In: Peters C, Gonzalo J, Braschler M, Kluck M (eds) Comparative Evaluation of Multilingual Information Access Systems Fourth Workshop of the Cross–Language Evaluation Forum, CLEF 2003, Trondheim, Norway, Lecture Notes in Computer Science, vol 3237, pp 29–43
Kelly D (2010) Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval 3(1–2):1–224
Kuriyama K, Kando N, Nozue T, Eguchi K (2002) Pooling for a large–scale test collection: An analysis of the search results from the first NTCIR workshop. Information Retrieval 5(1):41–59
Müller H, Müller W, McG Squire D, Marchand-Mailet S, Pun T (2001) Performance evaluation in content–based image retrieval: Overview and proposals. Pattern Recognition Letters 22(5):593–601
Müller H, Geissbuhler G, Marchand-Maillet S, Clough PD (2004) Benchmarking image retrieval applications. In: Proceedings of the tenth international conference on distributed multimedia systems (DMS'2004), workshop on visual information systems (VIS 2004), pp 334–337
Müller H, Deselaers T, Grubinger M, Clough PD, Hanbury A, Hersh W (2007) Problems with running a successful multimedia retrieval benchmark. In: Proceedings of the third MUSCLE / ImageCLEF workshop on image and video retrieval evaluation
Nowak S, Lukashevich H, Dunker P, Rüger S (2010) Performance measures for multilabel classification — a case study in the area of image classification. In: ACM SIGMM International conference on multimedia information retrieval (ACM MIR), ACM press, Philadelphia, Pennsylvania
Peters C, Braschler M (2001) Cross–language system evaluation: The CLEF campaigns. Journal of the American Society for Information Science and Technology 52(12):1067–1072
Petrelli D (2008) On the role of user–centred evaluation in the advancement of interactive information retrieval. Information Processing and Management 44(1):22–38
Robertson S (2008) On the history of evaluation in ir. Journal of Information Science 34:439–456
Sanderson M (2010 – to appear) Test Collection Evaluation of Ad–hoc Retrieval Systems. Foundations and Trends in Information Retrieval
Sanderson M, Clough PD (2002) Eurovision — an image–based CLIR system. In: Workshop held at the 25th annual international ACM SIGIR conference on research and development in information retrieval, Workshop 1: Cross–Language Information Retrieval: A Research Roadmap, ACM press, Philadelphia, Pennsylvania, pp 56–59

Saracevic T (1995) Evaluation of evaluation in information retrieval. In: 18th annual international ACM SIGIR conference on research and development in information retrieval, ACM, Seattle, OR, USA, pp 138–146

Sedghi S, Sanderson M, Clough PD (2009) A study on the relevance criteria for medical images. Pattern Recognition Letters 29(15):2046–2057

Smith JR (1998) Image retrieval evaluation. In: Proceedings of the IEEE Workshop on Content–Based Access of Image and Video Libraries (CBAIVL 1998), IEEE Computer Society, Washington, DC, USA, pp 112–113

Voorhees EM (2000) Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management 36(5):697–716

Voorhees EM, Harman DKe (2005) TREC: Experiments and Evaluation in Information Retrieval. MIT Press, Cambridge, MA