Special Issue on Image and Video Retrieval Evaluation

Allan Hanbury[a], Henning Müller[b], Paul Clough[c]

[a] *Information Retrieval Facility, Vienna, Austria*
[b] *University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland*
[c] *Dept. of Information Studies, University of Sheffield, UK*

Advances in technology, such as digital cameras, mobile phones and communications and networking are making visual media ubiquitous and readily accessible to a wide variety of consumers. To better manage this information, both description-based and content-based methods have been proposed [1, 2, 3, 4] for general as well as specialised domains [5]. However, although many techniques have been developed for image retrieval these are often hard to compare due to a disparity in datasets, performance measures and methodologies used to evaluate the techniques [6].

In recent years a multitude of benchmarks to evaluate multimedia retrieval systems have been created and a number of them used within comparative evaluation campaigns. Although proposals for the benchmarking of multimedia systems were made early on [7, 8, 6], Benchathlon[1] was the first large-scale event that provided evaluation resources and promoted discussions throughout the wider research community. Subsequent events then followed including TRECVid, ImageEVAL and ImageCLEF, which address different aspects of visual information retrieval evaluation.

TRECVid[2] started in 2001 as a task in the Text REtrieval Conference[3] (TREC), but in 2003 become an independent entity and has continually seen strong participation. TRECVid provides benchmarks to evaluate video retrieval systems [9], but is also important to image retrieval where evaluation of content-based algorithms can be performed on extracted video key frames. ImageEVAL[4], financed by the French research foundation, ran in 2006 with participants mainly from the French research community. The

[1] `http://www.benchathlon.net/`
[2] `http://www-nlpir.nist.gov/projects/trecvid/`
[3] `http://trec.nist.gov/`
[4] `http://www.imageval.org/`

event aimed to evaluate approaches for image filtering, content-based image retrieval (CBIR) and image classification. ImageCLEF[5] [10, 11] began in 2003 as a part of the Cross-Language Evaluation Forum[6] (CLEF), aiming to evaluate and compare multilingual information retrieval systems. ImageCLEF deals with retrieval of images from multilingual repositories, testing approaches that combine both visual and textual features for multi-modal retrieval. Strong participation in ImageCLEF over the past five years has shown the need for standardised system comparison and the importance of creating an infrastructure to support comparisons in this way.

The availability of benchmarks for evaluating image and video retrieval systems can dramatically reduce the effort required in evaluating new techniques. This instead allows researchers to work on developing novel approaches rather than issues associated with evaluation, such as defining the evaluation methodology and generating a benchmark.

1. Papers in the special issue

This special issue arose from a series of workshops on the evaluation of image and video retrieval held in conjunction with CLEF, and supported by the EU-funded MUSCLE Network of Excellence from 2005 to 2007[7]. A call for papers to contribute to this special issue was sent to workshop participants and published more widely throughout the research community. Submitted papers were peer-reviewed and five were selected to appear in this special issue.

The first paper by Smeaton, Over and Doherty [12] illustrates what can and cannot be learned from an established evaluation campaign track running over a long period of time. It presents the complete history (7 years) of the video shot boundary detection track of the TRECVid campaign. An overview of the TRECVid evaluation process and of the techniques submitted over the complete time period of the track are presented, followed by a detailed analysis of the 2005 results.

For all evaluation campaigns, data annotated with ground truth judgements is an important asset. This data can be used as both training data and the 'gold standard' against which to compare the results of different retrieval

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

[6] http://www.clef-campaign.org/

[7] http://muscle.prip.tuwien.ac.at/past_workshops.php

systems. Manually annotated data of high quality is essential in evaluation, but very time-consuming to produce. For a number of years, ImageCLEF made use of the IAPR TC12 (International Association for Pattern Recognition Technical Committee 12) dataset, consisting of 20,000 images annotated with textual descriptions written in three languages [13]. The paper by Escalante et al. [14] describes an extension of this dataset in which segments of the images are manually annotated. Experiments on this publicly-available data set and suggestions for further uses of the dataset are presented.

In cases where it is not possible to manually annotate the data (becoming more common as the size of datasets grow), semi-automated approaches to image annotation can be used. Such an approach is examined in the paper by Ulges et al. [15] which makes use of user-generated content downloaded from the video-sharing website YouTube[8] including video content and associated concepts/labels to be used as a ground truth for training automatic classifiers. The advantage of such an approach is that a large amount of training data covering a wide variety of topics can be quickly generated. However, a limitation is the resulting variability in quality and coverage of annotations generated.

The final two papers concentrate on the evaluation of low-level visual features for image retrieval: texture [16] and visual word codebooks [17]. In [16], the current methodology for evaluation of image classification by texture is analysed, tested and critiqued, and an improved evaluation methodology is proposed. This provides an example of useful but often neglected work on the analysis and improvement of evaluation methodology. The paper [17] targets mainly video retrieval evaluation. Several methods for generating 'codebooks' of visual features for retrieval are compared. The size (compactness) and retrieval quality are taken as the parameters to optimise, obtaining a trade-off between effectiveness and efficiency. Codebooks of visual features are currently the method of choice for many image and video retrieval techniques and the authors evaluate them using over 200 hours of video content.

2. Concluding remarks

The creation of standardised benchmarks for image and video retrieval is critical in comparing and improving systems. Evaluation campaigns, such

[8] `http://youtube.com`

as TREC and CLEF, have driven research in both academia and businesses alike. Bringing together researchers on common centrally-organised tasks and datasets has helped obtain a critical mass and develop successful approaches to image and video retrieval. However, most visual information retrieval evaluation has tended to focus on creating standardised benchmarks (or test/reference collections) for use in a laboratory-style setting. Saracevic [18] distinguishes six levels of evaluation for information systems that include image and video retrieval systems: (1) engineering level, (2) input level, (3) processing level, (4) output level, (5) use and user level and (6) social level. Much of the current research in evaluating retrieval systems tends to focus on levels 1–4, but levels 5 and 6 are important in producing effective operational systems. There have been attempts to evaluate video and image retrieval systems from a more user-centred perspective at both TRECVid (Interactive TRECVid) and ImageCLEF (iCLEF), but it is clear that further studies are required.

The link between producing image and video retrieval benchmarks and organising comparative evaluation events such as TREC and CLEF is clearly beneficial: it brings researchers from around the world together, testing a wide variety of systems and approaches on common tasks and using standard datasets. This enables comparison between various techniques and helps to stimulate progress in the field. Managed evaluation campaigns also help to obtain a critical mass and limit the administrative overhead with managing document collections, dealing with copyright issues and organising workshop events. Although these events usually follow an annual cycle of activities, having evaluations with a much shorter time scale where technologies can continuously be compared is something to consider in future events [19]. These could be based on distributed data sets and automatic evaluation of systems or components based on standardised query interfaces (e.g. Web services). Such evaluations would be particularly useful for quickly evaluating individual components, but also used in a combined way for evaluating retrieval systems as a whole.

Each of the papers in this special issue can be seen as dealing with parts of the considerations and technology needed in the implementation of such a continuous evaluation framework.

References

[1] P. G. B. Enser, Pictorial information retrieval, Journal of Documentation 51 (2) (1995) 126–170.

[2] A. Goodrum, Image information retrieval: An overview of current research, Journal of Information Science Research 3 (2) (2000) –.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.

[4] R. Datta, D. Joshi, J. Li, J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys 40 (2) (2008) 5:1–60.

[5] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, International Journal of Medical Informatics 73 (2004) 1–23.

[6] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, T. Pun, Performance evaluation in content-based image retrieval: Overview and proposals., Pattern Recognition Letters 22 (5) (2001) 593–601.

[7] J. R. Smith, Image retrieval evaluation, in: Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries, 1998, pp. 112–113.

[8] C. H. C. Leung, H. H.-S. Ip, Benchmarking for content-based visual information search, in: Proceedings of the 4th International Conference on Advances in Visual Information Systems, 2000, pp. 442–456.

[9] A. F. Smeaton, P. Over, W. Kraaij, Trecvid: Evaluating the effectiveness of information retrieval tasks on digital video, in: Proceedings of the international ACM conference on Multimedia, 2004, pp. 652–655.

[10] M. Grubinger, P. Clough, A. Hanbury, H. Müller, Overview of the imageclefphoto 2007 photographic retrieval task, in: Overview of the ImageCLEFmed 2007 Medical Retrieval and Medical Annotation Tasks, 2008, pp. 433–444.

[11] H. Müller, T. Deselaers, T. M. Deserno, J. Kalpathy-Cramer, E. Kim, W. Hersh, Overview of the imageclefmed 2007 medical retrieval and medical annotation tasks, in: Advances in Multilingual and Multimodal Information Retrieval: Proceedings of CLEF 2007, 2008, pp. 472–491.

[12] A. F. Smeaton, P. Over, A. R. Doherty, Video shot boundary detection: Seven years of trecvid activity, Computer Vision and Image Understanding Special Issue on Image and Video Retrieval Evaluation.

[13] M. Grubinger, P. Clough, H. Müller, T. Deselaers, The IAPR benchmark: a new evaluation resource for visual information systems, in: International Workshop OntoImage 2006, Genova, Italy, 2006, pp. 13–23.

[14] H. J. Escalante, C. A. Hernandez, J. A. Gonzales, L.-L. A., M. Montes, E. F. Morales, L. E. Sucar, L. Villasenor, M. Grubinger, The segmented and annotated iapr tc–tc12 benchmark, Computer Vision and Image Understanding Special Issue on Image and Video Retrieval Evaluation.

[15] A. Ulges, C. Schulze, M. Koch, T. Breuel, Learning automatic concept detectors from online video, Computer Vision and Image Understanding Special Issue on Image and Video Retrieval Evaluation.

[16] O. Drbohlav, A. Leonardis, Toward correct and informative evaluation methodology for texture classification under varying viewpoint and illumination, Computer Vision and Image Understanding Special Issue on Image and Video Retrieval Evaluation.

[17] J. C. van Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, J.-M. Geusebroek, Comparing compact codebooks for visual categorization, Computer Vision and Image Understanding Special Issue on Image and Video Retrieval Evaluation.

[18] T. Saracevic, Evaluation of evaluation in information retrieval, in: Proceedings of the 18th Annual International ACM SIGIR Conference, 1995, pp. 138–146.

[19] H. Müller, W. Müller, S. Marchand-Maillet, T. Pun, D. Squire, A web–based evaluation system for CBIR, in: Proc. ACM Multimedia, 2001, pp. 50–54.