A Review of Benchmarking Content Based Image Retrieval

Gareth Loy* and Jan-Olof Eklundh
Royal Institute of Technology (KTH)
Stockholm, Sweden
tel: +46 8 790 6353 fax: +46 8 723 0302
e-mail: gareth@nada.kth.se

Abstract

Benchmarking Content Based Image Retrieval (CBIR) systems allows researchers and developers to compare the strengths of different approaches, and is an essential step towards establishing the credibility of CBIR systems for commercial applications. Here we introduce the problem of developing a benchmark, discuss some of the issues involved, and provide a review of current and recent benchmarking efforts. We also propose a solution to one of the key stumbling blocks hampering benchmarking campaigns to date: the availability of large royalty free test databases.

1 Introduction

A benchmark is a basis on which to compare performance. In the case of Content Based Image Retrieval (CBIR) systems [29] it is typically a common set of tasks, such as "detect all faces", together with a set of test images with accompanying ground truth, and some metrics for measuring performance. The aim of a benchmark is not just to evaluate the performance of an algorithm, but also to do so in a way that is directly comparable between different algorithms. Whilst this premise is simple enough, effectively implementing a benchmark is no simple task.

To be successful a benchmark must be embraced by the researchers and developers designing the algorithms. This requires that the benchmarking task, or tasks, be highly relevant to the intended tasks promoted for the CBIR systems, and it must be clear that improved performance on the benchmark correlates strongly with improved performance in the real world, i.e., improved end-user satisfaction.

Standardised performance metrics are also essential. Consider the face detection example above, the performance could be measured in many ways, e.g. via a simple binary response for each image: *Does this image contain a face?*; a numerical response: *How many faces are present?*; or a complex numerical response: *What are the locations of all faces in the image?*

*This work was carried out during tenure of a MUSCLE internal fellowship.

What are the locations of the eyes and mouth of all faces present? Should confidence values be associated with each response, and if so how should these be incorporated into the final performance measure?

In addition to broadly accepted and clearly understood tasks and performance metrics there is the question of test data. The purpose of CBIR systems is to efficiently search huge amounts of widely varying images for some specified content. It is therefore essential to benchmark the performance of these systems on large amounts of data, both in order to evaluate the performance in as realistic a manner as possible and obtain a statistically significant result. The problem then arises: where is this data obtained and how is the ground truth generated? It is desirable to distribute the test data freely, which introduces copyright issues, whilst the ground truth labelling of the data has the potential to be an extremely arduous task.

Difficulties aside, it is widely agreed that a well-designed benchmark is central to the advancement of the field. In 1893 Lord Kelvin [31] summarised the importance of benchmarking:

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the state of science.

Providing a clear and broadly understood performance measure both allows researchers to more fully understand the strengths and limitations of their systems and compare their results with other systems on a level playing field. It also provides industry, potential commercial partners, and potential buyers of CBIR systems with a universal measure of how good the systems really are, which is central to establishing the credibility of CBIR systems in the marketplace.

Benchmarking CBIR systems, and computer vision algorithms in general, has long been considered an important task [6] and has led to a number of extensive benchmarking efforts, e.g. [7, 8, 19, 17, 11, 10, 14, 15]. This paper reviews several of these, and then moves on to tackle one of the key issues hampering benchmarking at present, the availability of royalty free images and ground truth data. A proposed method is outlined for generating extensive benchmarking databases from the growing population of online images available from sites such as Flickr and licenced under the new Creative Commons copyright.

2 Benchmarking Projects

This section gives a brief overview of a number of current and recent benchmarking projects relevant to CBIR systems.

2.1 VIPER

The VIPER[1] (Visual Information Processing for Enhanced Retrieval) network based at the University of Geneva has been associated with several benchmarking efforts. These include, a web-

[1] http://viper.unige.ch/research/benchmarking/

based benchmark for Query By Example (QBE)-based CBIR systems, image browser benchmarking [20], and the Benchathlon[2] event. The group is also behind the development of the Multimedia Retrieval MarkUp Language[3] (MRML) that aims to provide a unified interface for multimedia retrieval and management software.

The web-accessible benchmark for QBE-based CBIR systems was designed to allow developers of image retrieval systems the opportunity to benchmark their systems online at any time. MRML was used to access the retrieval system, and queries were made based on image URLs and the results transmitted as URLs.

The Benchathlon began with BIRDS-I[4] (Benchmark for Image Retrieval using Distributed Systems over the Internet) and was an initial step towards a standardized benchmark for CBIR systems. BIRDS-I was an image retrieval benchmark that was presented as a contest during EI 2001. Participants were required to implement an image retrieval server to be tested against a ground-truth via a set of defined metrics. The Benchathlon aimed to develop a networked system benchmark for CBIR, along the lines of existing benchmarks for text retrieval and relational database management. Muller *et al.* [17] reported that while the Benchathlon initiated discussion amongst participants no systematic comparison between systems was started.

2.2 IAPR

The International Association for Pattern Recognition's (IAPR) Technical Committee 12 (TC-12) has worked towards a standardised benchmark for the comparison of multimedia retrieval systems [8, 23]. The aim was to identify and evaluate the relative strengths and weaknesses of different approaches by providing standard sets of data, queries, and performance metrics.

Leung and Ip [8] made several initial recommendations concerning the development of a CBIR benchmark. They proposed that an extensible suite of benchmarks should be developed to cater for the disparate requirements of different applications, and that benchmarking image collections must be made freely available to researchers and must be free from any conditions or restrictions of use. It was recommended that initially 1,000 images be used for a CBIR benchmark, and this number be increased over time. All images should be in JPEG format and should contain multiple objects, with diverse relationships between them. Twenty evaluation queries covering a representative cross-section of contents should be designed against these images. Queries should be based entirely on the visual contents of the images (meta-data must not be used) and each query should be correctly "answered" by a known ground truth set of images. The number of answer images per query should be less than a specified amount, as should the number of images the algorithm returns at any one time. The proposed measures for evaluating performance included: recall, precision, average number of stages for retrieving the relevant images, average rank of the relevant images, effectiveness of query language and model, and effectiveness of relevance feedback.

These specifications formed the basis of the IAPR TC-21 Benchmark [23]. The benchmark consisted of, a set of still natural images, a representative set or queries, ground truths

[2]http://www.benchathlon.net/
[3]mrml.net
[4]http://benchath.hpl.hp.com/

associated with these queries, and a set of recommended performance metrics. The benchmark concentrated on the retrieval of semantic content (e.g. image contains car, person, etc.).

2.3 Computer Vision Benchmarks

Computer vision algorithms offer a means for CBIR systems to extract semantic content from images. Two key areas of growing relevance for CBIR systems are segmentation and object recognition. Both of these have recently been the subjects of major benchmarking campaigns.

2.3.1 The Berkeley Segmentation Dataset and Benchmark

The Berkeley Segmentation Dataset and Benchmark[5] consist of 12,000 hand-labeled segmentations of 1,000 images from 30 human subjects [11]. The original images are from the Corel database. Half of the segmentations were done on colour images, and half on grayscale. The initial public release of this data consisted of all colour and grayscale segmentations for 300 images, and was divided into 200 training images and 100 test images. A method was also presented for measuring how well an automatically generated segmentation matched a ground-truth segmentation, and Matlab code was provided for running the benchmark. The motivation for the benchmark was to provide a basis for comparing different algorithms, and to track progress towards human-level performance.

The performance metrics developed for boundary detection [12, 13] are relevant for any boundary dataset – not just the hand-labelled segmentations in the Berkley benchmark. Overlaying all human segmented boundaries for a given image generated the ground truth. The benchmark then measured the performance of a given soft segmentation[6] against the ground truth. A threshold was applied to the soft boundary map at many different levels (e.g. 30). At each level a precision-recall curve was generated, where *precision* was the probability that a hypothesised boundary pixel was a true boundary pixel, and *recall* was the probability that a true boundary pixel was detected. Therefore, the curve showed the trade-off between misses and false positives as the detector threshold changes.

A summary statistic was also produced to provide a single number summarising the algorithm's performance. In the case where the precision-recall curves do not intersect, the curve furthest from the origin dominates. The harmonic mean of the precision and recall was calculated at each point on this curve and the maximum value over the curve was taken as the summary statistic encapsulating the algorithm's performance.

2.3.2 Evaluation of Feature Detectors and Descriptors

The past five years has seen a drastic increase in the success of feature-based methods in computer vision. These approaches have been used for CBIR on large databases [25, 30]

[5] http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

[6] a *soft* segmentation is not binary, instead it gives a continuous measure for each pixel describing the confidence that that pixel is a boundary.

as well as numerous other applications including, visual data mining [28], object retrieval in video [27, 26], model based recognition [5, 9, 21, 24], and object categorisation [2, 3, 4, 22].

These approaches involve three steps. Given a collection of images between which to find correspondences: detect a set of feature points in each image, construct a feature vector providing a local description of each feature point (typically invariant to orientation and scale), and match these feature vectors to find potential matching features over the collection of images. The success of these methods relies on detecting and describing feature points in a manner that is robust to some variations in lighting, viewpoint, scale and orientation, while providing feature vectors that are distinctive enough for good matching performance. Given two images containing the same object, these methods aim to extract a sufficient number of similar feature points from each image of the object to allow the objects to be matched between the two images.

Recently a concerted effort was made within the EU-funded project VIBES[7] to present a cohesive performance comparison of methods for detecting and describing local features[8]. This campaign focused on affine invariant methods that provide the most generic and useful means for feature-based recognition of objects or scenes. The treatment was presented in two papers, one describing methods for detecting local feature points [15] and the other describing methods for constructing feature vectors describing these points [14], software for executing the different methods and running the benchmark was also made available[9].

The first paper [15] reviews affine-invariant region detectors and considers performance under: blur, JPEG artifacts, light change, and scale change. Detectors considered are: Harris affine detector, Hessian affine detector, maximally stable extremal region detector, edge-based region detector, intensity extrema-based region detector, and entropy-based region detector.

The second paper [14] compares the performance of local descriptors. Descriptors considered are shape context, steerable filters, PCA-SIFT, differential invariants, spin images, SIFT, complex filters, moment invariants and cross-correlation.

2.4 ImageCLEF

ImageCLEF is the image retrieval track of the Cross Language Evaluation Forum (CLEF) [1, 18]. It is not strictly a CBIR benchmarking event as it allows the use of meta-data, i.e., text that appears around the image or in the image title. The primary purpose of the CLEF campaign is to promote the multi-lingual searching of such data, however, a secondary goal is to investigate combining text and CBIR.

ImageCLEF offers two main image retrieval tasks, one over collections of photographic images and one over collections of medical images. Since its inception in 2003, ImageCLEF has drawn interest from both academics and commercial research organisations from the areas of CBIR, Cross-Language Information Retrieval, and user interaction.

[7]http://www.nada.kth.se/ sullivan/VIBES/

[8]http://www.robots.ox.ac.uk/˜vgg/research/affine/

[9]http://www.robots.ox.ac.uk/˜vgg/research/affine/

2.5 Neural Networks Council Standards Committee

IEEE Neural Networks Council Standards Committee has a working group on pattern recognition benchmarks[10] with the goal of providing "easy access to data sets and algorithms which may be used for benchmarking and comparison purposes via the [internet]". There is a webpage with lists of datasets, algorithm code, and publications.

2.6 MUSCLE

The EU-sponsored Network of Excellence MUSCLE[11] (Multimedia Understanding through Semantics, Computation and Learning) is developing systems and methodologies for automatically extracting semantic information from multimedia data. The MUSCLE benchmarking workpackage[12] has been established to develop tools for evaluating and comparing these algorithms, and to promote the use of these tools amongst the MUSCLE members. To achieve this large test databases are being assembled and regular evaluation projects are planned.

In France the Million image CLIC (CEAList Image Collection) database[16] has been assembled by the Laboratoire d'ingnerie de la connaissance multimdia multilingue (LIC2M/CEA-LIST). The database contains 15,900 images that have undergone 69 different transformations. The group have also organised the ImagEVAL[13] competition for automatic recognition of pictures. The competition consists of several tasks including retrieval of transformed images (about 30,000 images generated from 2000 using various image transformations), combined text and image retrieval, text detection and recognition in images, object recognition, and attribute extraction (e.g. indoor, outdoor, people present, etc.). The Paris School of Mines has provided another image database containing images under different ground truth illumination conditions.

In Austria the massive CIS-Benchmark database of coin images, containing over 100,000 ground-truthed images and 1,500 object classes, has been assembled as part of Project Dagobert at the ARC Seibersdorf Research Centre. In addition, the Partial Planar Objects Database has been compiled at TU Graz and consists of 20 different objects seen from varying angles; ground truth includes both the object name and viewing angle.

Network members have also collect numerous databases containing video with ground truth labelling. The majority of databases compiled are available for public (if not always commercial) use.

2.7 TRECVID

A recent success story in an area closely related to CBIR benchmarking is TRECVID[14] that started in 2001 as a video retrieval track of the Text REtrieval Conference (TREC). In 2003 TRECVID grew into an independent evaluation forum for research in automatic segmentation,

[10] http://morden.csee.usf.edu/nnc/index1.html
[11] http://www.muscle-noe.org
[12] http://muscle.prip.tuwien.ac.at/index.php
[13] www.imageval.org
[14] http://www-nlpir.nist.gov/projects/trecvid/

indexing, and content-based retrieval of digital video. Its aim is to *promote progress in content-based retrieval from digital video via open, metrics-based evaluation*[15]. The evaluation is an annual event and provides a large test collection, uniform scoring procedures, and a two-day workshop for organizations interested in comparing the results of their video retrieval systems.

For 2005 the tasks considered for the TRECVID evaluation were: shot boundary detection, classification of types of camera motion, high-level feature extraction, a high-level search task which including query-based retrieval and browsing, and exploration of raw unedited BBC footage (rushes). High-level features were considered to be labels that were clear to humans, such as *people walking/running*: segment contains video of more than one person walking or running, or *US flag*: segment contains video of a US flag.

A request to establish a CBIR track at TREC was rejected on the grounds that there were no databases large enough that could be freely distributed [18].

2.8 PEIPA

The principal aim of the Pilot European Image Processing Archive[16] (PEIPA) is to provide information, datasets and software to measure and compare the effectiveness of image processing and computer vision algorithms. The archive is supported by the EU-funded project Performance Characterization in Computer Vision (PCCV) for Benchmarking Vision Systems[17], the University of Essex, and the British Machine Vision Association. It offers a comprehensive online resource covering many aspects of benchmarking and performance characterisation in computer vision, including tutorials on benchmarking and links to databases. PEIPA aims to distribute test datasets to researchers so they can quantify and compare the performance of their algorithms. A test harness called HATE has been prepared to automate much of this process, and results can optionally be uploaded to the PEIPA website and made available to other researchers interested in making comparisons.

3 The Challenge of Assembling Test Data

One of the greatest challenges facing the CBIR benchmarking community is the availability of test databases [18]. These databases not only need to contain thousands of images with associated ground truths, but they must be royalty-free so they can be freely distributed. This leads to two key problems: obtaining royalty free images, and generating ground truth.

Many databases do exist, but these have typically been limited to contain material that has been donated or specially captured by the creators of the database. This has restricted the development of very large databases and made database generation a labour intensive process. While some creative solutions have been found, such as collaboration with Project Dagobert to form the CIS-benchmark, and using image transformations to extend database size [16], it remains to find a *general* solution to generating diverse, realistic databases for CBIR benchmarking.

[15] Guidelines for the TRECVID 2005 Evaluation.

[16] http://peipa.essex.ac.uk/index.html

[17] http://peipa.essex.ac.uk/benchmark/

A solution to the copyright issue is offered by Creative Commons[18], a non-profit organisation that offers artists and authors an alternative to full copyright. Breaking away from the traditional “all rights reserved” copyright Creative Commons offers a “some rights reserved” copyright, allowing copyright holders to licence their material freely or with constraints such as *Noncommercial*, *No Derivative Works* or *Share Alike*. Importantly Creative Commons has been embraced by many online organisations including Flickr[19] — one of the most popular image sharing site on the internet — which is the home of millions images from users across the globe.

Flickr’s Creative Commons pool[20] provides a searchable database of publicly licensed photos. There are currently more than 4 million images available under Creative Commons licenses at Flickr and this number is steadily increasing. Many of these images also have “tags” describing their content and comments from users who have viewed the image.

Automatically building databases from publicly licensed images from sites like Flickr (or OpenPhoto[21]) simply requires writing a web-crawler to autonomously browse the site, downloading images and their associated tags, comments, and author and copyright information, and compiling this into a database. Tags and comments can provide an initial basis for generating ground truth, and author and copyright information can be used to ensure that copyright restrictions are adhered. As Flickr allows uses to “tag” and comment on images on an ongoing basis, it is possible to have the web-crawler automatically update the tags and comments associated with each image as additional information is posted. Thousands of users add images, tags and comments to Flickr on a daily basis. This is an invaluable opportunity to generate test data from the very environment that CBIR systems are built for, and to spur development towards the day when CBIR systems are helping the users of such sites find the content they require.

4 Conclusion

Today effective benchmarking of CBIR is within reach. Frameworks have been established and effective methodologies outlined. The challenge remains to tune performance measures to the needs of developers, potential users and buyers of CBIR systems, and assemble large realistic test databases. With the ever-increasing growing amount of image content available online under the Creative Commons flexible copyright, material for databases can be compiled rapidly and efficiently without traditional copyright restrictions. Combining this with forums like Flickr that allow multi-user posting, labelling and commenting of images, and it is becoming feasible to establish the realistic test databases required for a general “TREC-style” CBIR benchmark.

References

[1] Paul Clough, Mark Sanderson, and Henning Müller. The clef cross language image retrieval track (imageclef) 2004. In *CIVR*, pages 243–251, 2004. 5

[18] http://creativecommons.org/

[19] flickr.com

[20] http://www.flickr.com/creativecommons/

[21] http://openphoto.net/

[2] G. Csurka, C. Dance, C. Bray, and L. Fan. Visual categorization with bags of keypoints. In *Proceedings Workshop on Statistical Learning in Computer Vision*, 2004. 5

[3] Gy. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–640, 2003. 5

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, Madison, Wisconsin, June 2003. 5

[5] Vittorio Ferrari, Tinne Tuytelaars, and Luc J. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *ECCV (1)*, pages 40–54, 2004. 5

[6] O. Firschein, M. Fischler, and Takeo Kanade. Creating benchmarking problems in machine vision: Scientific challenge problems. In *Proceedings of the 1993 DARPA Image Understanding Workshop*, April 1993. 2

[7] N. Gunther and G. Beretta. Benchmark for image retrieval using distributed systems over the internet: Birds. Technical report, HP Labs, 2000. 2

[8] Clement H. C. Leung and Horace Ho-Shing Ip. Benchmarking for content-based visual information search. In *VISUAL '00: Proceedings of the 4th International Conference on Advances in Visual Information Systems*, pages 442–456, London, UK, 2000. Springer-Verlag. 2, 3

[9] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society. 5

[10] Stéphane Marchand-Maillet. Construction of a formal multimedia benchmark. In *Proceedings of the European Signal Processing Conference (EUSIPCO2002)*, Toulouse, France, September 2002. (Invited paper). 2

[11] D. Martin, C. Fowlkes, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *International Conference on Computer Vision*, July 2001. 2, 4

[12] David R. Martin, Charless Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using brightness and texture. In *NIPS*, pages 1255–1262, 2002. 4

[13] David R. Martin, Charless Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(5):530–549, 2004. 4

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans Pattern Recognition and Machine Intelligence*, pages 1615–1630, October 2005. 2, 5

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Accepted in International Journal of Computer Vision*, 2005. 2, 5

[16] P-A. Moëllic, P. Hède, G. Grefenstette, and C. Millet. Evaluating content based image retrieval techniques with the one million images clic testbed. In *Proc World Enformatika Congress*, pages 171–174, Istanbul, Turkey, February 25-27 2005. 6, 7

[17] H. Müller, P. Clough, A. Geissbuhler, and W. Hersh. Imageclef 2004-2005: results, experiences and new ideas for image retrieval evaluation. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI2005), to appear*, Riga, Latvia, 2005. 2, 3

[18] H. Müller, A. Geissbuhler, S. Marchand-Maillet, and P. Clough. Benchmarking image retrieval applications. In *Visual Information Systems*, 2004. 5, 7

[19] Henning Müller, Wolfgang Müller, Stéphane Marchand-Maillet, David McG. Squire, and Thierry Pun. A framework for benchmarking in visual information retrieval, 2003. 2

[20] Wolfgang Müller, Stéphane Marchand-Maillet, Henning Müller, and Thierry Pun. Towards a fair benchmark for image browsers. In *SPIE Photonics East, Voice, Video, and Data Communications*, Boston, MA, USA, 5–8 2000. 3

[21] Stepán Obdrzálek and Jiri Matas. Object recognition using local affine frames on distinguished regions. In *BMVC*, 2002. 5

[22] Andreas Opelt, Michael Fussenegger, Axel Pinz, and Peter Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV (2)*, pages 71–84, 2004. 5

[23] Paul Over, Clement H. C. Leung, Horace Ho-Shing Ip, and Micheal Grubinger. Multimedia retrieval benchmarks. *IEEE Multimedia*, 11(2):80–84, April-June 2004. 3

[24] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *CVPR (2)*, pages 272–280, 2003. 5

[25] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):530–535, 1997. 4

[26] Josef Sivic, Frederik Schaffalitzky, and Andrew Zisserman. Object level grouping for video shots. In *ECCV (2)*, pages 85–98, 2004. 5

[27] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003. 5

[28] Josef Sivic and Andrew Zisserman. Video data mining using configurations of viewpoint invariant regions. In *CVPR (1)*, pages 488–495, 2004. 5

[29] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000. 1

[30] Tinne Tuytelaars and Luc J. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems*, pages 493–500, London, UK, 1999. Springer-Verlag. 4

[31] Kelvin William Thompson. Popular lectures and addresses [1891-1894]. In *Bartlett's Familiar Quotations, Fourteenth Edition*, page 723a. 1968. 2