A web-based evaluation system for content-based image retrieval *

Henning Müller, Wolfgang Müller,
Stephane Marchand-Maillet, Thierry Pun
Computer Vision Group, University of Geneva
24 Rue du Général Dufour,
CH-1211 Genève 4, Switzerland
Henning.Mueller@cui.unige.ch

David Squire
Computer Science and Software Engineering
Monash University
David.Squire@csse.monash.edu.au

ABSTRACT
This papers describes a benchmark test for content-based image retrieval systems (CBIRSs) with the query by example (QBE) query paradigm. This benchmark is accessible via the Internet and thus allows to evaluate any image retrieval system which is compliant with the Multimedia Markup Language (MRML) for query formulation and result transmission. Thus it allows a quick and easy comparison between different features and algorithms for CBIRSs. The benchmark is not only based on a standardized communication protocol to do the communication between the benchmark server and the benchmarked system, but it also uses a freely downloadable image database for the evaluation to make the results reproducible. A CBIR system that uses MRML and other components to develop MRML-based applications can be downloaded free of charge as well.

The evaluation is based on several queries and known relevance sets for these queries. Several answer sets for the same query image are possible if user judgments of several users exist, thus almost any sort of user judgment can be incorporated into the system. The final results are averaged over all the queries.

The evaluation of several steps of relevance feedback based on the collected relevance judgments is also included into the benchmark. The performance of relevance feedback is often regarded to be even more important than the performance in the first query step because only with relevance feedback the adaptation of the system to the users subjective goal can be measured. For the evaluation of a system with relevance feedback, the same evaluation measures are used on the query results as for the first query step.

*(Produces the permission block, copyright information and page numbering). For use with ACM_PROC_ARTICLE-SP.CLS V2.0. Supported by ACM.

Keywords
web-based evaluation, content-based image retrieval, benchmark

1. INTRODUCTION
Proper evaluation has always been regarded as very important for content-based image retrieval (CBIR). Despite this importance of standardized evaluation, most systems were evaluated with a variety of non-standardized performance measures and with image databases not accessible to reproduce and compare the results. Many systems were only evaluated with one example query result, which gives no objective impression of the quality of a system at all. System evaluation was a widely neglected topic in image retrieval.

In text retrieval, a closely related field, standardized performance comparisons were proposed as early as the 60s with the SMART system in 1961 [16] and the Cranfield test in 1966 [1]. With the Text REtrieval Conference (TREC) [19] (`http://trec.nist.gov/`) a clearly defined and accepted benchmark started in 1992 and has been done every year since. This benchmark really helped the research field because good techniques could be distinguished form poor techniques.

The MIRA framework (Evaluation Frameworks for Interactive Multimedia Information Retrieval Applications) (see `http://www.dcs.gla.ac.uk/mira/`) was the first iniative to take a more formal approach to evaluate Multimedia Retrieval systems. A number of conferences and workshops were held within this framework.

In 1997 Narasimhalu et al [14] give a formal comparison of different sorts of retrieval systems and how the systems can be evaluated based on users giving ranked relevance sets for a number of query images. Concrete performance measures or image databases to use were not proposed and there is also no example of an evaluation given.

In recent years appeared more publications on the subject. Starting from discussion at conferences such as Visual99 and SPIE Photonics West 2000 the need for a standardized performance evaluation, standardized image databases, and especially an event where systems can be compared, became apparent.

In 1999 Dimai [3] described a rank-based measure to compare two different feature sets or retrieval systems to overcome the shortcomings of precision and recall. To compare two systems this might work, but for a benchmark many systems need to be compared. It is also important not to compare the systems based on a single performance measure, but on several measures because for different application areas different characteristics might be important. Koskela et al. [7] describe performance measures to quantify how close together clusters of images are in feature space based on their ranks. This can be used to compare different features and techniques. Leung [8] gives a detailed proposal for a benchmark with stating performance measures and the approximate sizes of the databases. He proposes to have a database of roughly 1000 images for a start and have a number of categories with not more than $15-20$ relevant images for a query. An example evaluation with the measures is not given in the article. In [10, 11] an approach similar to TREC is used for retrieval system evaluation. Measures are proposed and a fully automatic benchmark is implemented based on these measures, with an example evaluation based on one CBIR system. None of these papers discusses in detail the hard question of how to obtain a large freely available image database for performance comparison and how to obtain relevance judgments for this database.

By far the most promising approach to a CBIR benchmark so far is the Benchathlon (see `http://www.benchathlon.net/`). It started from discussions at SPIE Photonics West 2000 and had its first prototypical, running system at the Photonics West conference in 2001. The techniques of the benchmark are described in [4]. For the Photonics West conference in 2002 a larger database and a more sophisticated benchmark is planned. Several researchers from different fields are currently working on this benchmark.

Most of the proposals are evaluating the performance of Query by Example (QBE) systems. For browsing systems (or target search) there are measures proposed in [2], where a real user has to interact with a database and a more elaborated approach in [12], where extensive annotation is used to simulate the users behavior in selecting images.

To compare systems relying on different paradigms such as QBE, image browsing or search by keyword is a very hard task if feasible at all. This aspect is not addressed here.

This paper describes a solution based on the measures and techniques used by TREC for text retrieval, but transferred to the field of image retrieval. The system can use different kind of relevance judgments and several image databases to allow a maximum flexibility with respect to the images and judgments that are available. Performance measures proposed by other researchers are implemented as well to be able to compare the different measures.

2. THE TECHNOLOGICAL BASIS FOR A FULLY AUTOMATED BENCHMARK

For a fully automatic evaluation of image retrieval systems there has to be a common access method to the retrieval system, so that the benchmark server can automatically perform all the queries, including feedback queries, and receive the results for a performance evaluation. Other problems for any evaluation are a freely available image databases to make the obtained results comparable and reproducible. The hardest and most work-consuming task is most likely to obtain ground truth data for the images, especially for a large image database.

2.1 MRML

The Multimedia Retrieval Markup Language (MRML, see `http://mrml.net/`) is an XML-based communication protocol for CBIR, which was developed to separate the query interface from the actual query engine. It was developed for QBE and thus contains tags for query by positive and negative examples. A detailed technical description can be found in [13].

The client can open a session on the server, and configure it according to the needs of its user (interactive client) or its own needs (eg. benchmark test).

A basic query consists of a list of images and their corresponding relevance levels, assigned by the user. In the following example, the user has marked two images: `1.jpg` positive and `2.jpg` negative. All images are referred to by their URLs.

```
<mrml session-id="1" transaction-id="44">
<query-step session-id="1"
 resultsize="30"
 <user-relevance-list>
  <user-relevance-element
   image-location="http://viper.unige.ch/1.jpg"
   user-relevance="1"/>
  <user-relevance-element
   image-location="http://viper.unige.ch/2.jpg"
   user-relevance="-1"/>
 </user-relevance-list>
</query-step>
</mrml>
```

The server will return the retrieval result as a list of image URLs, ordered by their relevance to the query.

2.2 Image databases

A general approach to a CBIR benchmark has to have a maximum flexibility with respect to the image databases used. Because of this, we allow the benchmark to use any kind of database where the images can be distributed on the Internet and thus can be accessed by URL.

Many existing CBIRSs use the Corel (`http://www.corel.com/`) image collections for an evaluation, which contains groups of 100 images each with roughly the same subject. Still, the images are rather expensive and copyrighted and the choice of groups determines the difficulty of the query task. The images of MPEG-7 [9] are as well copyrighted and might not be used in publications or on the Internet which makes them unusable for a performance comparison between systems. Another possibility is the image collection of the Department of Water Resources (DWR) in California that is available without charge for non-commercial use from the University of California in Berkeley (`http://elib.cs.berkeley.edu/`

photos/tarlist.txt). This database is relatively large (more than 25,000 images) but has only a limited number of different contents. No relevance judgments are currently available for this database.

For a first test we decided to use the database of the University of Washington in Seattle (http://www.cs.washington.edu/research/imagedatabase/groundtruth/) for the evaluation explained in this paper because it offers ground truth in form of image clusters and annotations for most of the images. It is also available free of charge and without any copyright. Unfortunately it is still very small with only 922 images in 14 image clusters, but we hope to enlarge the set with the help of other research groups. We also tested the benchmark on the database of the Télévision Suisse Romande (TSR) where user judgments of five users are available. Unfortunately this database is copyrighted.

We plan to include other databases that we can get with relevance judgments. A standard database of, for example, the Benchathlon will be very useful for the context of a web-based evaluation.

2.3 Relevance Judgments

One of the hardest tasks for the evaluation of CBIR systems is the process of obtaining relevance judgments. Often image databases contain clusters of images with the same objects ("cars", "airplanes") like the Corel collection or images of different regions ("mountains", "cities") like the database of the University of Washington. In this case the clusters can be regarded as ground truth and one image of the cluster can be taken as example image for a QBE query. Still it is often the case that an image from a cluster has more similarities with images from other clusters than with the same cluster. Visual similarity within a cluster can vary in a large span. For these reasons predefined clusters are not always a very good choice as relevance judgments.

These fixed image clusters also neglect the subjectiveness of the users of a CBIRS. With the same query image users can look for a completely different answer set [18]. To model this user subjectivity real user tests have to be performed with several users as in [17]. There is also the possibility to use textual annotations of images for the generation of groundtruth. More about the classification of images can be read in [6].

The evaluation stated in this paper are based on the clusters of the image database of the University of Washington, because this database is freely available and thus the results are reproducible. Tests have also been performed with real user judgments on databases of the TSR. The TSR database can also be evaluated via the web interface.

2.4 Performance measures

In this paper we mainly use the performance measures described in [11] to have a set of measures similar to those used in TREC. These measures can be used for all the different image databases and also all the different kinds of relevance judgments mentioned. The measures are:

- $Rank_1$, $\bar{Rank}$ and $\widetilde{Rank}$: rank at which first relevant image is retrieved, average rank and normalized average rank of relevant images (see Eq. 1);
- $P(20)$, $P(50)$ and $P(N_R)$: *precision* after 20, 50 and N_R images are retrieved;
- $R_P(.5)$ and $R(100)$: *recall* at *precision* .5 and after 100 images are retrieved;
- PR graph.

A simple average rank is difficult to interpret, since it depends on both the collection size N and the number of relevant images N_R for a given query. Consequently, we normalize the average rank by these numbers and propose the *normalized average rank*, $\widetilde{Rank}$:

$$\widetilde{Rank} = \frac{1}{NN_R}\left(\sum_{i=1}^{N_R} R_i - \frac{N_R(N_R-1)}{2}\right) \qquad (1)$$

where R_i is the rank at which the ith relevant image is retrieved. This measure is 0 for perfect performance, and approaches 1 as performance worsens. For random retrieval the result would be 0.5.

We use several measures based on precision and recall despite criticism on precision and recall already in the 60s [16] and as well in [3, 7] for the use in CBIR. Precision and recall, especially in form of the precision/recall graph and in form of precision or recall at important cutoff point, are still the standard in text retrieval and are easy to interpret.

For a user who is in general looking at $20-50$ images on screen it is very important how many relevant images he can actually see as a result. For a more user-centered evaluation it does not really make a difference whether a relevant image is retrieved at position 1000 or 2000, whereas position 50 or 51 can make a larger difference if the user cannot see the image in the response set displayed on screen.

We are also integrating the normalized average rank measure proposed in [4, 15], that basically proposes a penalization for images that are not retrieved at all. The measures proposed in [8] will also be integrated to show how easily adaptable the system is to different and new performance measures.

3. THE WEB-BASED BENCHMARK

A description of the web-based benchmark is available from http://viper.unige.ch/evaluation/. This page contains the prerequisites for the execution of the benchmark and links to a number of benchmark resources. An example system using MRML can be downloaded at http://www.gnu.org/software/gift/

3.1 Overview

Figure 1 shows the general structure of the benchmark. The communication between the benchmark server and the benchmarked systems is done in MRML. The benchmarked systems basically only need to know the URLs of the images in the database. The performance measures are openly visible as well.

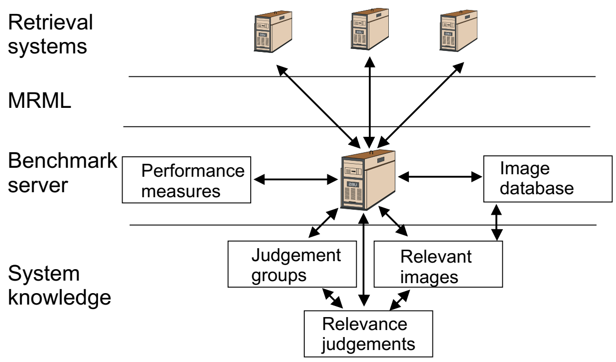

Figure 1: Structure of the automated benchmark.

The ground truth data for the images and even the images chosen as query images should not be known by the benchmarked systems as they can try to cheat when this information is available. If a system knows the image classes, it can of course always return a perfect response. Normally this phase of getting the ground truth should be done after all the systems have returned the results.

For the web-based benchmark the image groups are for now not hidden. A system could thus cheat and get good results. However, this web-based benchmark is so far more a research tool than an official benchmark, so there is no need to hide the ground truth data. For an official version this should be different.

3.2 Communication framework

We can see in Figure 2 that only the first step in the communication, the configuration of the benchmark and the last step of the communication are not done in MRML. The configuration is done via a CGI interface, and the results are displayed on any web browser after the execution of the benchmark.

With the first step of communication in MRML the benchmark server contacts the system to benchmark and verifies that the chosen database is available and that the system speaks MRML. If this is the case, all queries are done as single image queries (QBE). After the first query step, all performance measures are calculated for the available relevance judgments and they are averaged over all the relevance judgments and queries. Now positive and negative relevance feedback is generated for every relevance judgment group (respectively every user) and every query. Based on this generated feedback, a new set of queries is executed and again the performance measures are calculated based on the relevance judgments, and the performance measures are averaged in the end. This step can be repeated as often as necessary. In the very end, all the averaged performance measures are displayed on the web browser.

3.3 Configuring the benchmark

The CGI Interface shown in Figure 3 allows the user to enter a number of parameters that the system needs to execute the benchmark.

The *system name* is only meant as an identification of the

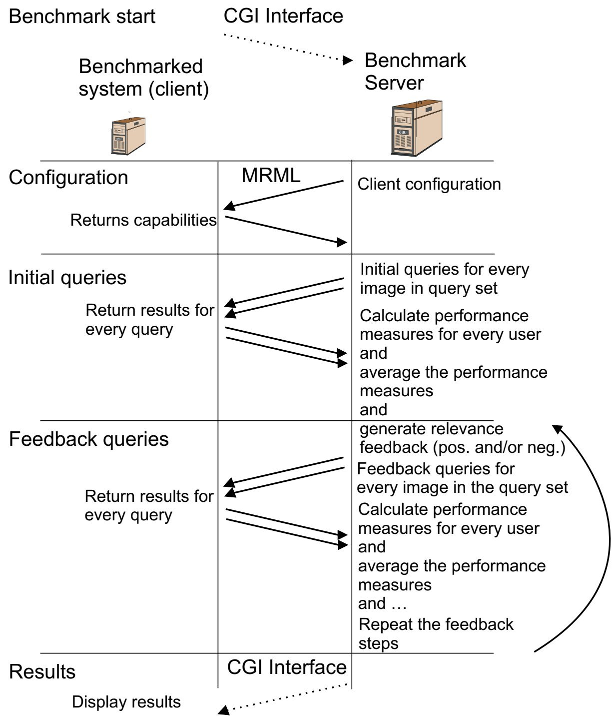

Figure 2: Structure of the automated Benchmark steps for the communication.

benchmarked system to the server, it can be left at anonymous if the developers want their system to stay unknown. Important for the communication are the *host name* and the *port number* of the system to benchmark. These two parameters are absolutely needed to start the MRML communication on this socket. The choice of a *database* determines the queries and the relevance judgments the web-based benchmark will use. The *database ID* is important for the benchmark server to chose the database via MRML. For flexibility reason we allow to use any database ID and do not link it directly with a database name. The number of *feedback steps* finally determines the number of query steps that are done with the system. The first step is in this context the step with only one query image and no feedback.

4. RESULTS OF AN EXAMPLE RUN

To demonstrate that the benchmark works, we use the GIFT (GNU Image Finding Tool) `http://www.gnu.org/software/gift/` to have a client with an MRML interface. The evaluation with this client was executed for the database of the University of Washington and for the database of the TSR as well. As an example we used four steps of relevance feedback for the evaluation, to show the possibility to evaluate relevance feedback with our benchmark. The results are put into tables for readability. Normally the results are directly shown on screen.

4.1 Image database of the University of Washington

The database of the University of Washington consists of 922 images that are in 14 different categories, normally geographical areas like “Cannon Beach” or “Mountains”. We

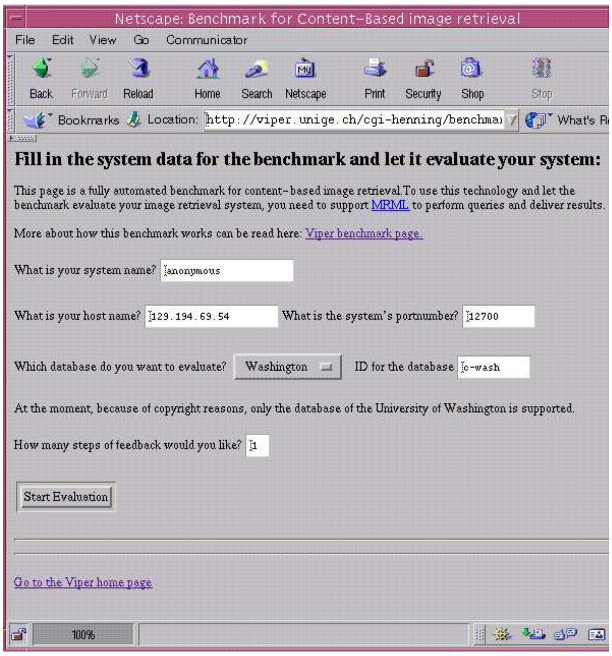

Figure 3: A screenshot of the web-based benchmark

always took the first image of a group as a query image and all the images of a group as the relevance set, no matter how visually similar or different they are.

Measure	no RF	RF 1	RF 2	RF 3	RF 4
N_R	65.14	65.14	65.14	65.14	65.14
t	1.23 s.	2.18 s.	2.49 s.	2.62 s.	2.70 s.
$Rank_1$	1.5	1	1	1	1
$R(P(.5))$	.3798	.5520	.6718	.6594	.7049
$Rank$	176.44	152.28	116.13	107.04	104.37
$\widetilde{Rank}$	.1583	.1318	.0921	.0821	.0793
$P(20)$	.5392	.7357	.8642	.8892	.9107
$P(50)$	.4057	.5271	.6085	.6328	.6257
$P(N_R)$	.3883	.5256	.6138	.6640	.6553
$R(100)$	.4839	.6070	.6924	.7279	.7208

Table 1: Overview of the results for the GIFT with the Washington database

The results in Table 1 show that the first two steps of relevance feedback strongly improve the results. Steps three and four only bring minor improvements. The rank of the first relevant image shows that from the initial query there were very good results for every query image. The $P(20)$ shows that after 4 steps of relevance feedback there are an average of more than 18 relevant images in the top 20 which also shows that this is a relatively easy database for similarity queries.

4.2 Image database of the TSR

The image database of the TSR consists of 500 images with only few clusters of really similar images. Five users gave relevance judgments for ten query images, where they had to find all images that they regarded as similar to the query images, with no strict policy given for similarity.

Measure	no RF	RF 1	RF 2	RF 3	RF 4
N_R	10.56	10.56	10.56	10.56	10.56
t	1.83 s	2.44 s	2.65 s	2.77 s	2.92 s
$Rank_1$	14.6	9.96	10.16	9.94	10.1
$R(P(.5))$	.3411	.447	.4575	.4334	.4399
$Rank$	58.14	54.48	53.93	52.64	51.14
$\widetilde{Rank}$	.1065	.0992	.098	.0955	.0925
$P(20)$	.263	.282	.293	.298	.302
$P(50)$	.1572	.1488	.1496	.1528	.1548
$P(N_R)$	.4789	.6068	.6305	.6441	0.65
$R(100)$	.7901	.7988	.8031	.8044	.8123

Table 2: Overview of the results for the GIFT with the TSR database

Table 2 shows a different result from Table 1. The first query step does bring a significant improvement, but afterwards the results do not improve very much. The fact that the first relevant image found does not get to 1 means that for at least one query there was no relevant image in the first $n = 20$ result images. The fact that the precision at 50 images does not improve at all means that the relevance feedback might have improved the ordering of the relevant images within the first 50, but not many new images are shown in there. As there are only an average of 10 relevant images, the maximum precision at this point can be 20 %. It shows that the queries on this database are significantly more difficult than on the database of the University of Washington.

5. CONCLUSION AND FUTURE WORK

This paper presents a working benchmark for content-based image retrieval that can be configured via the WWW (World Wide Web) and also displays the evaluation results on any web browser. The database used in this paper is just an example and we are aware that for a proper CBIRS evaluation larger, free databases are necessary which automatically causes problems for the generation of proper relevance judgments. We hope that the Benchathlon effort will provide this. With such a large database including good ground truth, the web-based benchmark can be a very helpful tool for system developers to test the system performance for new features or new access methods on the fly.

A regular benchmark event like the Benchathlon can of course not be replaced by such a web accessible benchmark, because it is necessary to compare a number of systems. The web-based benchmark is more meant to complement the Benchathlon and give system developers the possibility to try out their system from time to time to be able to identify performance differences. It can also be used to test the MRML interface of a system in an automated way.

For the future we plan to include more performance measures to also be able to compare these measures with respect to their information content. Graphical evaluation methods, like precision/recall graphs, are also being developed for their use on a WWW platform. For a final benchmark there should be a small number of performance measures and, even more important, they should all contain different information. We would be happy to include any database we can get relevance judgments and a URL list for.

6. ACKNOWLEDGMENTS

This work was supported by the Swiss National Foundation for Scientific Research (grant no. 2000-052426.97).

7. REFERENCES

[1] C. W. Cleverdon, L. Mills, and M. Keen. Factors determining the performance of indexing systems. Technical report, Cranfield Project, Cranfield, 1966.

[2] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Target testing and the `PicHunter` Bayesian multimedia retrieval system. In *Advances in Digital Libraries (ADL'96)*, pages 66–75, Library of Congress, Washington, D. C., May 13–15 1996.

[3] A. Dimai. Assessment of effectiveness of content-based image retrieval systems. In Huijsmans and Smeulders [5], pages 525–532.

[4] N. J. Gunther and G. Beretta. A benchmark for image retrieval using distributed systems over the internet: Birds-i. Technical report, HP Labs, Palo Alto, Technical Report HPL-2000-162, San Jose, 2001.

[5] D. P. Huijsmans and A. W. M. Smeulders, editors. *Third International Conference On Visual Information Systems (VISUAL'99)*, number 1614 in Lecture Notes in Computer Science, Amsterdam, The Netherlands, June 2–4 1999. Springer-Verlag.

[6] C. Jörgensen. Classifying images: Criteria for grouping as revealed in a sorting task. In *Proceedings of the 6th ASIS SIG/CR Classification research Workshop*, pages 65–78, Chicago, Il, USA, October 1995.

[7] M. Kokela, J. Laaksonen, S. Laakso, and E. Oja. Evaluating the performance of content-based image retrieval systems. In *International Conference on Visual Information Systems (Visual 2000)*, Lyon, France, November 2–4 2000.

[8] C. Leung and H. Ip. Benchmarking for content-based visual information search. In *International Conference on Visual Information Systems (Visual 2000)*, Lyon, France, November 2–4 2000.

[9] MPEG Requirements Group. MPEG-7: Context and objectives (version 10 Atlantic City). Doc. ISO/IEC JTC1/SC29/WG11, International Organisation for Standardisation, October 1998.

[10] H. Müller, W. Müller, S. Marchand-Maillet, D. M. Squire, and T. Pun. Automatic benchmarking in content-based image retrieval. In *International conference on Multimedia and Exhibition(ICME 2001)*, Tokyo, Japan, October 2001. (to appear).

[11] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: Overview and proposals. *Pattern Recognition Letters*, 2001.

[12] W. Müller, S. Marchand-Maillet, H. Müller, and T. Pun. Towards a fair benchmark for image browsers. In *SPIE Photonics East, Voice, Video, and Data Communications*, Boston, MA, USA, nov 5–8 2000.

[13] W. Müller, H. Müller, S. Marchand-Maillet, T. Pun, D. M. Squire, Z. Pečenović, C. Giess, and A. P. de Vries. MRML: A communication protocol for content-based image retrieval. In *International Conference on Visual Information Systems (Visual 2000)*, Lyon, France, November 2–4 2000.

[14] A. D. Narasimhalu, M. S. Kankanhalli, and J. Wu. Benchmarking multimedia databases. *Multimedia Tools and Applications*, 4:333–356, 1997.

[15] P. S. Salembier and B. S. Manjunath. Audiovisual content description and retrieval: Tools and mpeg-7 standardization techniques. In *IEEE Internation Conference on Image Processing (ICIP 2000)*, Vancouver, BC, Canada, December 2000.

[16] G. Salton. *The SMART Retrieval System, Experiments in Automatic Document Processing.* Prentice Hall, Englewood Cliffs, New Jersey, USA, 1971.

[17] D. M. Squire, W. Müller, and H. Müller. Relevance feedback and term weighting schemes for content-based image retrieval. In Huijsmans and Smeulders [5], pages 549–556.

[18] D. M. Squire and T. Pun. A comparison of human and machine assessments of image similarity for the organization of image databases. In M. Frydrych, J. Parkkinen, and A. Visa, editors, *The 10th Scandinavian Conference on Image Analysis (SCIA'97)*, pages 51–58, Lappeenranta, Finland, June 1997. Pattern Recognition Society of Finland.

[19] E. M. Voorhees and D. Harmann. Overview of the seventh text retrieval conference (TREC-7). In *The Seventh Text Retrieval Conference*, pages 1–23, Gaithersburg, MD, USA, November 1998.