UNIVERSITE DE GENEVE

CENTRE UNIVERSITAIRE
D'INFORMATIQUE
GROUPE VISION

Date: December 1, 1999
N° 99.05

TECHNICAL REPORT

VISION

Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals

Henning Müller Wolfgang Müller* David McG. Squire
Thierry Pun

Computer Vision Group
Computing Science Center, University of Geneva
24 rue du Général Dufour, CH - 1211 Geneva 4, Switzerland

e-mail: Henning.Mueller@cui.unige.ch

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

Abstract

Evaluation of retrieval performance is a crucial problem in content-based image retrieval (CBIR). Many different methods for measuring the performance of a system have been created and used by researchers. This article discusses the advantages and shortcomings of the performance measures currently used. Problems such as a common image database for performance comparisons and a means of getting relevance judgments (or ground truth) for queries are explained.

The relationship between CBIR and information retrieval (IR) is made clear, since IR researchers have decades of experience with the evaluation problem. Many of their solutions can be used for CBIR, despite the differences between the fields. Several methods used in text retrieval are explained. Proposals for performance measures and means of developing a standard test suite for CBIR, similar to that used in IR at the annual Text REtrieval Conference (TREC), are presented.

Keywords: content-based image retrieval, performance evaluation, information retrieval

1 Introduction

Early reports of the performance of CBIR systems were often restricted simply to printing the results of one or more example queries (*e.g.* Flickner et al. (1995)). This is easily tailored to give a positive impression, since developers can chooses queries which give good results. It is neither an objective performance measure, nor a means of comparing different systems. Researchers have subsequently developed a variety of CBIR performance measures, which are discussed in §4. MIR (1996) give a further survey. However, few standard methods exist which are used by large numbers of researchers. Many of the measures used in CBIR (such as *precision*, *recall* and their graphical representation) have long been used in IR. Several other standard IR tools have recently been imported into CBIR, *e.g.* relevance feedback. In order to avoid reinventing pre-existing techniques, it seems logical to make a systematic review of evaluation methods used in IR and their suitability for CBIR.

In the 1950s IR researchers were already discussing performance evaluation, and the first concrete steps were taken with the development of the SMART system in 1961 (Salton (1971b)). Other important steps towards common performance measures were made with the Cranfield test (Cleverdon et al. (1966)). Finally, the TREC series started in 1992, combining many efforts to provide common performance tests. The TREC project (see TRE (1999), Vorhees & Harmann (1998)) provides a focus for these activities and is the worldwide standard in IR. Nevertheless, much research remains to be done on the evaluation of interactive systems and the inclusion of the user into the query process. Such novelties are included in TREC regularly, *e.g.* the interactive track in 1994. Salton (1992) gives an overview of IR system evaluation.

2 Information Retrieval

Although performance evaluation in IR started in the 1950s, here we focus on newer results and especially on TREC and its achievements in the IR community. Not only did TREC provide an evaluation scheme accepted worldwide, but it also brought academic and commercial developers together and thus created a new dynamic for the field.

2.1 Data Collections

The TREC collection is the main collection used in IR. Co-sponsored by the National Institute of Standards and Technology and the Defense Advanced Research Projects Agency, TREC has been held annually since its inception—1999 saw TREC-8. At present TREC participants must index a collection of 2 Gigabytes of textual data at the conference itself. Comparisons of participating systems are given later. A large amount of training data is also provided before the conference. Different collections exist for different topics, and several evaluation methods are used. Special evaluations exist for interactive systems (Over (1998)), spoken language, high-precision and cross-language retrieval. The collections can grow as computing power increases, and as new research areas are added.

2.2 Relevance judgments

The determination of relevant and non-relevant documents for a given query is one of the most important and time-consuming tasks. Using real users, it takes a long time to judge a large number of documents. Since it is unreasonable to expect humans to examine 2 Gb of data, a pooling technique is used for the TREC collection (Sparck Jones & van Rijsbergen (1975)). Only a subset of the collection, which is considered to be complete for a given query, is presented to users for actual relevance judgments.

TREC uses the following working definition of *relevance*: If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant. Only binary judgments ("relevant" or "not relevant") are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document).

2.3 Performance measures

The most common evaluation measures used in IR are *precision* and *recall* (see Eq. 1), usually presented as a *precision vs. recall* graph (PR graph) (*e.g.* Salton (1971a), van Rijsbergen (1979)). Researchers are familiar with PR graphs and can extract information from them without interpretation problems.

$$\begin{aligned} precision &= \frac{\text{No. relevant documents retrieved}}{\text{Total No. documents retrieved}}, \\ recall &= \frac{\text{No. relevant documents retrieved}}{\text{Total No. relevant documents in the collection}}. \end{aligned} \tag{1}$$

Since PR graphs may not contain all desired information (Salton (1992)), several other measures are used at TREC, also based on *precision* and *recall*:

- $P(10), P(30), P(N_R)$ - the *precision* after the first $10, 30, N_R$ documents are retrieved, where N_R is the number of relevant documents for this topic.
- Mean Average Precision - mean (non-interpolated) average *precision*.
- *recall* at .5 *precision* - *recall* at the rank where *precision* drops below .5.
- $R(1000)$ - *recall* after 1000 documents are retrieved.
- Rank first relevant - The rank of the highest-ranked relevant document.

These key numbers offer a set of performance descriptors, so that different systems can be compared meaningfully and objectively.

3 Basic Problems in performance evaluation in CBIR

The current status of performance evaluation in CBIR is far from that in IR. There are many different groups who are working with different sets of specialized images. There is neither a common image collection, nor a common way to get relevance judgments, nor a common evaluation scheme.

3.1 Defining a common image collection

There are several problems which must be addressed in order to create a common image collection. The collection must be available free of charge and without copyright restrictions, so that images can be placed on the web and used in publications. The greatest problem is to create a collection with enough diversity to cater for the diverse, partly specialized domains in CBIR such as medical images, car images, face recognition and consumer photographs.

A common means of constructing an image collection is to use Corel photo CDs, each of which usually contains 100 broadly similar images (*e.g.* Belongie et al. (1998), Ratan et al. (1999), COR (1999)). Unfortunately these images are copyrighted, and are not free. Most research groups use only a subset of the collection, which can result in a collection consisting of several highly dissimilar groups of images, with relatively high within-group similarity. This can lead to great apparent improvements in performance: it is not too hard to distinguish sunsets from underwater images of fish! Another commonly used collection is VisTex, which contains primarily texture images (Vis (1995)). A good candidate for a standard collection could be the images and videos from MPEG-7 (MPEG Requirements Group (1998)). Unfortunately they may not be shown on the web, and the collection is expensive.

An alternative approach is for CBIR researchers to develop their own collection. Such a project is underway at the the University of Washington in Seattle (ANN (1999)). This collection is freely available without any copyright and offers annotated photographs of different regions and topics. It is still small ($\sim$ 500 images), but several groups are contributing to enlarge the data set. Collection size should be sufficiently high that the trade-off between speed and accuracy can be evaluated. In IR it is quite normal to have millions of documents whereas in CBIR most systems work with a few thousand images and some even with fewer than one hundred (*e.g.* Müller & Rigoll (1999)).

3.2 Obtaining relevance judgments

In CBIR there is not yet a common means of obtaining relevance judgments for queries. Even the inclusion of real users in the judgment process (as in IR) is not common, as is shown below.

Use of collections with predefined subsets A very common technique is to use standard image databases with sets of different topics (*e.g.* air-shows, zebras) such as the Corel collection. Relevance "judgments" are given by the collection itself since it contains distinct groups of annotated images. The choice of sets can greatly influence results, since some sets are visually distant from each other and others are visually closely related. Grouping is not always based on global visual similarity, but often on the contained objects. In some studies images which are too visually different are removed from the collection, which definitely improves results (*e.g.* Belongie et al. (1998)).

Image grouping An alternative approach is for the collection creator or a domain expert to group images according to some criteria. The grouping is not necessarily based

only on readily-perceptible visual features. Domain expert knowledge is very often used in medical CBIR (*e.g.* Shyu et al. (1999), Dy et al. (1999)). This can be seen as real groundtruth, because the images have a diagnosis certified by at least one medical doctor. These groups can then be used like the subsets discussed above.

Simulating users Some studies simulate a user, by assuming that users' image similarity judgments are modeled by the metric used by the CBIR system, plus noise (*e.g.* Vendrig et al. (1999)). Such simulations can provide very good results—indeed result quality is controlled by the noise level. Real users are very hard to model: Tversky (1977) has shown that human similarity judgments seem not to obey the requirements of a metric, and they are certainly user- and task-dependent. Such simulations cannot replace real user studies.

User judgments The collection of real user judgments is time-consuming, but only the user knows what he or she expects as a query result. To obtain such judgments, relevance must be defined and the user must examine the entire database or a representative part of it (*cf.* TREC pooling). The user is then given a query image and asked to specify all relevant images in the collection. Experiments show that user judgments for the same image often differ (*e.g.* Squire & Pun (1997), Squire et al. (1999)), which is also observed in IR (Borgman (1989)). This is the only means of obtaining relevance judgments which acknowledges genuine differences between user responses, and does not assume the existence of one "best" query result. These individual differences are especially important if we want to demonstrate the ability of a system to adapt to users' needs by using relevance feedback.

There are fundamental differences between these methods. The ease of obtaining relevance "judgments" is an advantage of using collections with pre-defined groups of similar images. User judgments can still be made for such a collection. Domain expert knowledge should be used when it is available, such as in medicine and other specialized fields. For general CBIR tasks, we believe that the use of real users is essential (see Squire & Pun (1997), Markkula & Sormunen (1998)). For a complete evaluation, the user with his/her expectations is an vital part of the system. The number of images which a user must examine can be reduced by using pooling methods as in IR. It is essential that the user examine a significantly large fraction of the database, and that the relevance judgments are made in advance: users tend to be easily satisfied, even though the result may contain few, or even none, of the images selected as being relevant in advance. The characteristics of the group of users from whom the relevant judgments are obtained are also very important: CBIR system developers of a system have different notions of image similarity from novice users.

4 Performance Evaluation Methods

4.1 User comparisons

User comparison is an interactive method. Users judge the success of a query directly after the query. It is hard to get a large number of such user comparisons as they are time-consuming.

Before-after comparison This is the easiest test method. Users are given two or more different results and allowed to choose the one which is preferred or found to be most relevant to the query. This method needs a base system or another system for comparison.

4.2 Single-valued measures

Rank of the best match Berman & Shapiro (1999) measure whether the "most relevant" image is in either the first 50 or first 500 images retrieved. 50 represents the number of images returned on screen and 500 is an estimate of the maximum number of images a user might look at when browsing.

Average rank of relevant images Gargi & Kasturi (1999) use this measure. It can give a good indication of system performance, although it clearly contains less information than a PR graph. It is vulnerable to outliers, since just one relevant image with a very high rank can adversely affect it. A simpler and more robust measure is the **rank of the first relevant image**, which is used in TREC and it is very useful for CBIR as well.

Precision* and *recall As discussed in §2.3, these are standard measures in IR, which give a good indication of system performance. Either value alone contains insufficient information. We can always make *recall* 1, simply by retrieving all images. Similarly, *precision* can be kept high by retrieving only a few images. Thus *precision* and *recall* should either be used together (*e.g. precision* at .5 *recall*), or the number of images retrieved should be specified, (*e.g. recall* after 1000 images or *precision* after 20 images are retrieved). *Precision* and *recall* are often averaged, but it is important to know the basis on which this is done. Iqbal & Aggarwal (1999) use *precision* and *recall*. Belongie et al. (1998) use *Averaged precision*. Martinez (1999) uses the *recognition rate* which is not defined in the text, but seems to be the *precision* of a query.

Target testing The target testing approach differs significantly from other performance measures. Users are given a target image and the number of images which the user needs to examine before finding the target image is measured. Starting with random images, the user marks images as either relevant or non-relevant. Cox et al. (1996) employ this measure for the `PicHunter` system. Müller et al. (1999) use a more elaborate version of target testing, in which the notion of moving targets is used to evaluate the the ability of the system to track changes in user preferences during a query session.

Error rate Hwang et al. (1999) use this measure, which is common in object or face recognition. It is in fact a single *precision* value, so it is important to know where the value is measured (see above).

$$\text{Error rate} = \frac{\text{No. non-relevant images retrieved}}{\text{Total No. images retrieved}}. \tag{2}$$

Retrieval efficiency Müller & Rigoll (1999) define *Retrieval efficiency* as in Eq. 3. If the number of images retrieved is lower than or equal to the number of relevant images, this value is the *precision*, otherwise it is the *recall* of a query. This definition can be misleading since it mixes two standard measures.

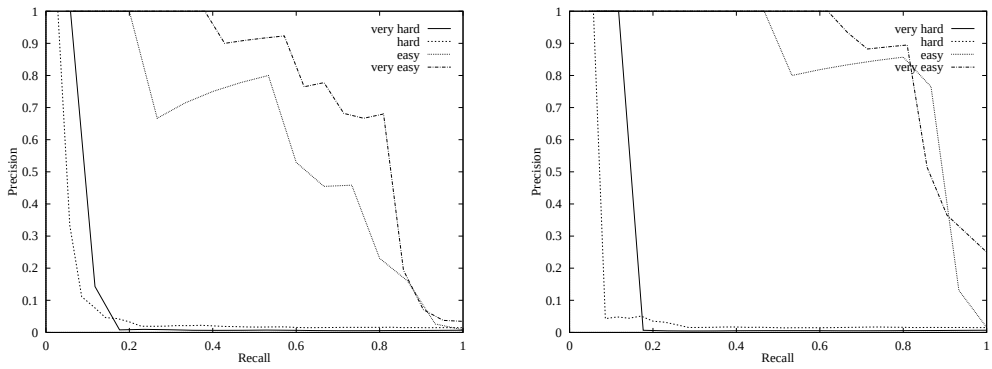
$$\text{Retrieval efficiency} = \begin{cases} \frac{\text{No. relevant images retrieved}}{\text{Total No. images retrieved}} & \text{if No. retrieved} \\ \frac{\text{No. relevant images retrieved}}{\text{Total No. relevant images}} & > \text{No. relevant} \\ & \text{otherwise.} \end{cases}, \tag{3}$$

Correct and incorrect detection Ozer et al. (1999) use these measures in an object recognition context. The numbers of correct and incorrect classifications are counted. When divided by the number of retrieved images, these measures are equivalent to *error rate* and *precision*.

4.3 Graphical representations

***Precision vs. recall* graphs** PR graphs are a standard evaluation method in IR and are increasingly used by the CBIR community (Squire et al. (1999)). PR graphs contain

a lot of information, and their long use means that they can be ready easily by many researchers. He (1997) use the representation with the axes changed (*i.e.* a *recall vs. precision* graph). In the interests of readability, this should be avoided. It is also common to present a partial PR graph (*e.g.* He (1997)). This can be useful in showing a region in more detail, but it can also be misleading since areas of poor performance can be omitted. Interpretation is also harder, since the scaling has to be watched carefully. A partial graph should always be used in conjunction with the complete graph.



(a) without feedback (b) with feedback

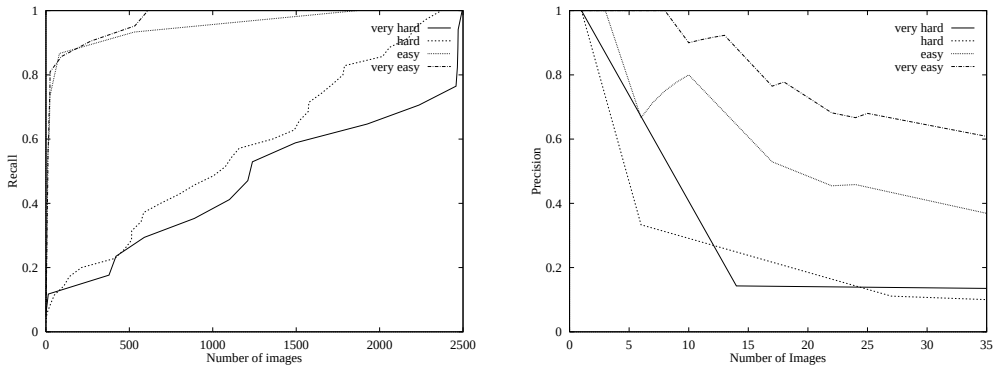
Figure 1: PR graphs for four different queries both without and with feedback.

Figure 1 demonstrates that PR graphs can distinguish well between differing results. The drawback is that the PR graph depends on the number of relevant images for a given query. We can see that the plot for the very hard query starts later than the hard one and looks better, although the decrease of the curve is much faster. Practical information such as *precision* or *recall* after a given number of images have been retrieved can not be obtained.

***Precision vs.* No. images retrieved and *recall vs.* No. of images retrieved graphs** Taken separately, these graphs contain only some of the information in a PR graph. When combined, however, they contain more information and can easily be interpreted. The *recall* graph looks more positive than a PR graph, especially when a few relevant images are retrieved late (Ratan et al. (1999)). The *precision* graph is similar to a PR graph, but it gives a better indication of what might be a good number of images to retrieve. It is more sensitive, however to the number of relevant images for a given query. If only part of the graph is shown it is hard to judge the performance (Aksoy & Haralick (1999)).

We can see in Figure 2 that the *recall graph* can distinguish well between the hard and easy queries, but not too well between the easy and very easy one. A complete *precision graph* does not contain much information in this case, that is the reason for printing a partial one. Here we have the problem with the different numbers of relevant images like in the *PR graph*. The result for the very hard query looks better than the result of the hard query.

Correctly retrieved vs. all retrieved graphs (Vasconcelos & Lippman (1999)) contain the same information as *recall graphs*, but differently scaled. *Fraction correct vs. No. images retrieved* graphs (Belongie et al. (1998)) are equivalent to *precision* graphs. *Average recognition rate vs. No. images retrieved* graphs (Comaniciu et al. (1999)) show the average percentage of relevant images among the first N retrievals. This is equivalent to the *recall* graph.



(a) recall graph (b) precision graph

Figure 2: Recall *vs.* No. of images graph and partial precision *vs.* No. of images graph

Retrieval accuracy vs. Noise graphs Huet & Hancock (1999) use this measure to show the change in retrieval accuracy as noise is added. A noisy image is used as a query and the rank of the original image is observed. This model does not correspond well to many CBIR applications.

5 Proposals

In the preceding sections a large number of different evaluation techniques has been described. It is apparent that many of them are equivalent or contain the same information. Clearly it would be beneficial to the CBIR community if only standardized names were used for performance measures. Since scaling or the use of partial graphs impedes interpretation, these techniques should only be used for emphasis, in conjunction with a complete graph.

We propose a set of performance measures similar to those used in TREC. This set contains mixture of rank-based, single-valued and graphical measures:

- $Rank_1$ and $\widetilde{Rank}$: rank at which first relevant image is retrieved, normalized average rank of relevant images (see below and Eq. 4).
- $P(20)$, $P(50)$ 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 (see §4.2) 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 propose the *normalized average rank*, $\widetilde{Rank}$:

$$\widetilde{Rank} = \frac{1}{NN_R}\left(\sum_{i=1}^{N_R} R_i - \frac{N_R(N_R-1)}{2}\right) \tag{4}$$

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.

Examples of these measures, using the same queries used Figs. 1 and 2, are shown in Table 1. The differences between the measures and the differing information that they contain can be seen.

Table 1: Performance measures for four different queries, in a database of 2500 images.

Query	N_R	$Rank_1$	$\widetilde{Rank}$	$P(20)$	$P(50)$	$P(N_R)$	$R_P(.5)$	$R(100)$
very easy	21	1	0.028	0.73	0.51	0.71	0.82	0.86
easy	15	1	0.067	0.47	0.23	0.60	0.62	0.87
hard	35	5	0.426	0.20	0.08	0.10	0.05	0.14
very hard	17	13	0.558	0.13	0.12	0.14	0.09	0.13

6 Conclusions

This article gives an overview of existing performance evaluation measures in CBIR. The need for standardized evaluation measures is clear, since several measures are slight variations of the same definition. This makes it very hard to compare the performance of systems objectively. To overcome this problem a set of standard performance measures and a standard image database is needed. We have proposed such a set of measures, similar to those used in TREC. A frequently updated shared image database and the regular comparison of system performances would be of great benefit to the CBIR community.

Further work needs to be done to better integrate users in the evaluation process. After all, the ultimate aim is to measure the usefulness of a system for a user. Interactive performance evaluations including several levels of feedback and user interaction need to be developed. We are continuing work in this area, and welcome further discussion and collaboration on this topic.

References

Aksoy, S. & Haralick, R. M. (1999). Graph theoretic clustering for image grouping and retrieval, *in* CVP (1999), pp. 63–68.

ANN (1999). Annotated groundtruth database, Department of Computer Science and Engineering, University of Washington, web page: `http://www.cs.washington.edu/research/imagedatabase/groundtruth/`.

Belongie, S., Carson, C., Greenspan, H. & Malik, J. (1998). Color- and texture-based image segmentation using EM and its application to content-based image retrieval, *Proceedings of the International Conference on Computer Vision (ICCV'98)*, Bombay, India.

Berman, A. P. & Shapiro, L. G. (1999). Efficient content-based retrieval: Experimental results, *in* CBA (1999), pp. 55–61.

Borgman, C. L. (1989). All users of information retrieval systems are not created equal: an exploration into individual differences, *Information Processing and Management* **25**: 225–250.

CBA (1999). *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99)*, Fort Collins, Colorado, USA.

Cleverdon, C. W., Mills, L. & Keen, M. (1966). Factors determining the performance of indexing systems, *Technical report*, Cranfield Project, Cranfield.

Comaniciu, D., Meer, P., Xu, K. & Tyler, D. (1999). Retrieval performance improvement through low rank corrections, *in* CBA (1999), pp. 50–54.

COR (1999). Corel clipart & photos, web page:
`http://www.corel.com/products/clipartandphotos/`.

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

CVP (1999). *Proceedings of the 1999 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'99)*, Fort Collins, Colorado, USA.

Dy, J. G., Brodley, C. E., Kak, A., Shyu, C.-R. & Broderick, L. S. (1999). The customized-queries approach to CBIR using using EM, *in* CVP (1999), pp. 400–406.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. & Yanker, P. (1995). Query by image and video content: The QBIC system, *IEEE Computer* **28**(9): 23–32.

Gargi, U. & Kasturi, R. (1999). Image database querying using a multi-scale localized color representation, *in* CBA (1999), pp. 28–32.

He, Q. (1997). An evaluation on MARS - an image indexing and retrieval system, *Technical report*, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

Huet, B. & Hancock, E. R. (1999). Inexact graph retrieval, *in* CBA (1999), pp. 40–44.

Hwang, W.-S., Weng, J. J., Fang, M. & Qian, J. (1999). A fast image retrieval algorithm with automatically extracted discriminant features, *in* CBA (1999), pp. 8–12.

Iqbal, Q. & Aggarwal, J. K. (1999). Applying perceptual grouping to content-based image retrieval: Building images, *in* CVP (1999), pp. 42–48.

Markkula, M. & Sormunen, E. (1998). Searching for photos - journalists' practices in pictorial IR, *in* J. P. Eakins, D. J. Harper & J. Jose (eds), *The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval*, Electronic Workshops in Computing, The British Computer Society, Newcastle upon Tyne.

Martinez, A. (1999). Face image retrieval using HMMs, *in* CBA (1999), pp. 35–39.

MIR (1996). MIRA: Evaluation frameworks for interactive multimedia retrieval applications. Esprit working group 20039., web page:
`http://www.dcs.gla.ac.uk/mira/`.

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

Müller, S. & Rigoll, G. (1999). Improved stochastic modeling of shapes for content-based image retrieval, *in* CBA (1999), pp. 23–27.

Müller, W., Squire, D. M., Müller, H. & Pun, T. (1999). Hunting moving targets: an extension to Bayesian methods in multimedia databases, *in* S. Panchanathan, S.-F. Chang & C.-C. J. Kuo (eds), *Multimedia Storage and Archiving Systems IV (VV02)*, Vol. 3846 of *SPIE Proceedings*, Boston, Massachusetts, USA. (SPIE Symposium on Voice, Video and Data Communications).

Over, P. (1998). A review of Interactive TREC, *MIRA workshop*, Dublin, Ireland.

Ozer, B., Wolf, W. & Akansu, A. N. (1999). A graph based object description for information retrieval in digital image and video libraries, *in* CBA (1999), pp. 79–83.

Ratan, A. L., Maron, O., Grimson, W. E. L. & Lozano-Perez, T. (1999). A framework for learning query concepts in image classification, *in* CVP (1999), pp. 423–429.

Salton, G. (1971a). Evaluation parameters, in *The SMART Retrieval System, Experiments in Automatic Document Processing* Salton 1971b, pp. 55–112.

Salton, G. (1971b). *The SMART Retrieval System, Experiments in Automatic Document Processing*, Prentice Hall, Englewood Cliffs, New Jersey, USA.

Salton, G. (1992). The state of retrieval system evaluation, *Information Processing and Management* **28**(4): 441–450.

Shyu, C.-R., Kak, A., Brodley, C. & Broderick, L. S. (1999). Testing for human perceptual categories in a physician-in-the-loop CBIR system for medical imagery, *in* CBA (1999), pp. 102–108.

Sparck Jones, K. & van Rijsbergen, C. (1975). Report on the need for and provision of an ideal information retrieval test collection, *British Library Research and Development Report 5266*, Computer Laboratory, University of Cambridge.

Squire, D. M., Müller, W., Müller, H. & Raki, J. (1999). Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback, *The 11th Scandinavian Conference on Image Analysis (SCIA'99)*, Kangerlussuaq, Greenland, pp. 143–149.

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

TRE (1999). Text REtrieval Conference (TREC), web page: `http://trec.nist.gov/`.

Tversky, A. (1977). Features of similarity, *Psychological Review* **84**(4): 327–352.

van Rijsbergen, C. J. (1979). Evaluation, *Information Retrieval*, Prentice Hall, Englewood Cliffs, New Jersey, USA, chapter 7, pp. 112–123.

Vasconcelos, N. & Lippman, A. (1999). Probabilistic retrieval: new insights and experimental results, *in* CBA (1999), pp. 62–66.

Vendrig, J., Worring, M. & Smeulders, A. W. M. (1999). Filter image browsing: Exploiting interaction in image retrieval, *in* D. P. Huijsmans & A. W. M. Smeulders (eds), *Third International Conference On Visual Information Systems (VISUAL'99)*, number 1614 in *Lecture Notes in Computer Science*, Springer-Verlag, Amsterdam, The Netherlands, pp. 147–154.

Vis (1995). VisTex: Vision texture database, Maintained by the Vision and Modeling group at the MIT Media Lab. web page: `http://whitechapel.media.mit.edu/vismod/`.

Vorhees, E. M. & Harmann, D. (1998). Overview of the seventh text retrieval conference (TREC-7), *The Seventh Text Retrieval Conference*, Gaithersburg, MD, USA, pp. 1–23.