The role of evaluation in the development of content-based retrieval techniques*

Arjen P. de Vries
Centre for Telematics and Information Technology
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands
arjen@acm.org

1 Introduction

There are no satisfactory methods for measuring the effectiveness of multimedia search techniques. Precision and recall types of metrics have been used in some of the literature but are impractical due to the tedious process of measuring relevances. The process is even more complicated because of human subjectivity in tasks involving multimedia. Also, multimedia collections quickly grow very large, making evaluation expensive with respect to the required hardware. There seem to be no standard corpora or benchmark procedures.

In this document, we review the evaluation approaches that have been taken in the literature about multimedia retrieval systems. Our review covers far from all related publications, but we believe the sample of papers studied is sufficiently large to draw our (sad) conclusions.

2 Quantitative evaluation

Quantitative evaluation uses previously collected relevance judgments about queries against corpora as the ground-truth, against which retrieval systems are evaluated.

2.1 IR evaluation methodology

Information retrieval research has developed a strong scientific evaluation methodology. Using huge data collections, and ground-truth data for a set of queries, the quality of the retrieval systems is measured and then used to compare different approaches. Many different collections are available to evaluate retrieval performance. The most popular are the TREC collections, consisting of several Gigabytes of documents with relevance judgments.

Evaluation in traditional IR is based on the following assumptions. First, we assume that the user reads all the results. Second, we assume that reading one relevant document does not influence the judgment of other documents in the result set. Also, relevance is considered a binary property. Under these assumptions, the quality of the result for a query

*Published as CTIT technical report TR-CTIT-00-19, November 2000, http://www.ctit.utwente.nl/.

can be expressed using recall (how many of the relevant documents in the collection did we retrieve?) and precision (how many of the retrieved documents are relevant?).

For systems that produce ranked output, it is not likely that the user checks *all* the query results. We then choose several cut-off points, to represent the fact that not every user will evaluate as many documents of the result set, and compute average recall and precision values for these sets. A common method is to compute the **11-point average precision**. This measure is computed by averaging the precision over the standard recall points (0%, 10%, 20%, etc.). To get the precision for these standard recall points, we calculate precision and recall for each relevant document in the result set and interpolate.

IR experiments are stochastic experiments affected by random errors. Therefore, to compare the performance of different approaches, we must decide whether the observed difference in performance is statistically significant. It is not sufficient to compare different retrieval approaches by its mean performances, because these can be heavily affected by outliers. Instead, we should compare the distributions of the observations. The statistical significance of the performance difference is best checked using non-parametric tests, because these tests make least assumptions on the experimental data. The sign-test is most widely used in IR. A common alternative – assuming a Gaussian distribution though – is the paired t-test.

Mira, Esprit working group 20039, studies new paradigms for the evaluation of IR [Dra97]. They focus on the role of the user in the evaluation, because the interaction with users plays an important role in current IR techniques and cannot be evaluated very well in the traditional research methodology. Also, the evaluation of the role of new media in information retrieval requires further research. Stephen Robertson announced to develop a multimedia test collection and launch a collaborative project around this.

2.2 Quantitative evaluation in audio retrieval

Speech

Cambridge University together with Olivetti research have developed a video mail retrieval system. For evaluation purposes, they collected relevance assessments for 50 requests on a (fairly small) collection of 300 messages, the VMR1 message set [JFJY96, BFJ+96].

Schäuble and Wechsler did innovative work in speech retrieval with an open vocabulary [SW95, Sch97]. They developed a test collection of 1289 documents with relevance judgments for 26 queries [WS95].

In TREC-6, the Spoken Document Retrieval track was aimed at the evaluation of speech retrieval systems. The collection consists of a human transcribed text and the output of a speech recognizer of the same data. The collection consists of 50 hours of broadcast news, and 49 queries with one relevant document per query [VH97].

General audio

Musclefish is a small company in Silicon Valley that specializes in audio retrieval techniques. They developed feature extraction for audio retrieval, and evaluated the performance against a manually classified collection [WBKW96]. This collection is rather small; it consists of about 400 samples varying from 1 to 15 seconds, that has been manually classified in groups like ‘bells’, ‘crowds’, and ‘laughter’.

Foote adapted machine learning techniques from speaker identification research to achieve a more general form of audio retrieval [Foo97]. For his evaluation, he too used the collection gathered by Musclefish, but only provides details about classifying the data into a small amount of groups. In a demonstration site, he applied the same classifier to search music based on similarity, but this has not been evaluated.

2.3 Quantitative evaluation in image retrieval

Image retrieval

For the texture features of images, the Brodatz texture collection is used in many papers. The complete Brodatz database consists of 112 classes of each 9 images [PMS96]. It has been used a.o. in the papers on image retrieval by texture by [SC94], [LP96], [MM96], [RSZ96], [ZCA97], [PP97], [MP97], [CSZSM97] and [SJ]. Most papers mention only the existence of 112 images, and only a small subset of 13 homogeneous textures is widely used. In some of these, the Brodatz set is used to construct larger data sets, using images composed of several pieces of different textures.

VisTex, provided by the MIT Media Lab, has 167 textures from natural scenes, categorized into 19 mutually exclusive groups. This collection is used in [Min96] and [ZCA97]. The MIT papers also use a collection of 96 vacation photo's, cf. [MP97] and [PM95], and another data set called 'BT collection' [Min96].

Recently, Smith and Burns of University of Queensland, Australia, made available the MeasTex framework[1]. The framework contains software and test suites necessary to measure performance of an algorithm in the MeasTex framework, and implementations of and results for some well known texture classification algorithms. The MeasTex framework rates an algorithm based on its average performance on a test-suite. The data set consists of the Brodatz collection, the VisTex collection, supplemented with the MeasTex images (artificial and natural textures) and the 'Ohanian and Dubes' images.

De Bonet (of the MIT AI Lab) used a collection based on 29 Corel CD, each containing 100 images of some class [Bon97, BV97]. Note that he argues that the Brodatz set is 'too easy' for proper evaluation of texture models [BV98]. In a recent paper he evaluates his texture features on a SAR data set for vehicle classification in radar images [BVI98].

Gevers and Smeulders evaluated the effectiveness of several color models for color invariant retrieval [GS96]. They use a collection of 500 images of 2-D and 3-D domestic objects, recorded with a CCD color camera. They randomly selected a set of 70 objects, and recorded those objects again (so these have different lighting conditions) to create a query set. De Bonet did some similar experiments with the Corel data, by varying visual characteristics of some of the images (such as brightness, contrast, and varying degrees of additional noise) and measuring the effect on retrieval [BV97, Bon97].

New initiatives aim to create huge image collections using the WWW as a source. Sclaroff et al. developed a fully operational system ImageRover [STC97]. Their fleet of 32 robots is estimated to collect approximately 1 million images monthly. To ensure diversity of images, they start the robot at several Yahoo categories. The Dutch national SION-AMIS project[2] has started to construct the large Acoi benchmark [NK98]. The benchmark is geared at provision of 1 million still images, hundreds of video sequences, and thousands of audio tracks. Unfortunately, the current setting of the benchmark only measures the execution performance, not the quality of the retrieved images.

[1] URL `http://www.cssip.elec.uq.edu.au/~guy/meastex/meastex.html`

[2] `http://www.cwi.nl/~acoi/Amis/`

Video segmentation

Zhang et al. [ZKS93] evaluated the quality of the detected scene changes using the video data, but only for three different videos.

Gargi and Kasturi [GK96] evaluated different color spaces and frame difference measures used in video segmentation algorithms. They constructed a test set of 21,000 frames from 9 movies, with 200 ground truth cuts. The human subjects first previewed the video data at full-speed, and then marked the cuts during half-speed viewing. They found that this procedure resulted in the most consistent results.

The MITRE corporation used an adapted version of the Text-tiling algorithm ([Hea94]) on the captions of broadcast news to find program boundaries [MHMG97]. The adapted algorithm did not only find topic boundaries, but also provide topic classifications. They collected captions for 17 1/2 hours, but did not hand-segment topic boundaries, only program boundaries. They did notice that the program changes that they missed had visual cues, so a combined algorithm might work better.

3 Minimal evaluation

A great deal of published multimedia retrieval research barely has an evaluation phase. The techniques are explained, and the results of a small set of example queries are given to convince us that the techniques work.

Some examples are:

- The QBIC image retrieval project by IBM [NBE+93]. In [FBF+94] the performance of the color-based image retrieval has been measured for a relatively small database (1000 images for 10 queries). Retrieval by shape has been evaluated on an even smaller scale (259 test images and 7 queries).

- The Chabot project at Berkeley [OS95] used the Postgres DBMS to implement image retrieval techniques. They evaluated the system using a single query ('yellow flower'), on a database containing 11,643 images (with 22 yellow flowers).

- Vellaikal and Kuo (UCLA) use a set of 3,400 color images with four queries in the ground-truth [VK95]. In [VK96] 12,000 images are used in the evaluation with three queries.

- The MARS project at University of Illinois ([MRC+97]) used a collection from the Getty museum to experiment with shape based retrieval. This collection has 300 images of ancient African artifacts . In [ORC+97], the quality of retrieval has been evaluated for thirteen conceptual queries, like 'stone masks' or 'golden pots'. In other work, they did not evaluate the quality of the retrieval, but they did study the relevance information provided by users from the system [RHMO97a, RHMO97b, RHM98]. Their main conclusion is that different users have very different measures of what images are 'similar', and hence need relevance feedback methods.

- The work on multimedia retrieval from the National University of Singapore. They developed search engines for retrieval of faces (FACEit) and trademarks (STAR) [NL95]. STAR has been evaluated against a database of 500 trademarks, in which two 'ideal' ranked retrieval sets have been constructed by ten people (using voting to get agreement). The results of the system are compared to this ideal result [WLM+97]. The same group developed the more general content-based retrieval engine CORE [WNM+95]. For the evaluation of content-based retrieval of segmented images, Chua et al. used a small collection of about 100 images divided in 10 different categories [CLP94].

- Columbia University, NY, has projects on both image and video retrieval. Smith and Chang evaluate color retrieval in their VisualSEEK system using a collection 3100 images using a single request to select the 83 images of lions in the collection [SC96, SC97]. The video retrieval system VideoQ, allows some dynamic aspects in the query language [CCM+97]. They use a collection of 200 shots, categorized into sports, science, nature, and history. For evaluation they used only 4 queries.

- In the Informedia project at Carnegie Mellon University, a reasonable amount of tools for video retrieval have been implemented [WKSS96]. For evaluation of their News-on-demand application, only the quality of the output of the speech recognizer has been evaluated [HW97].

- Ardizzone and La Cascia describe the JACOB video retrieval system developed at University of Palermo [AC97]. Their evaluation uses a collection of 1500 keyframes extracted from about 500 shots. They use a test set of five example queries, for topics 'water polo', 'interview', 'TV-show', 'TV-movie', and 'cycling race'.

- Sheikholeslami et al. of State University of New York at Buffalo study a.o. clustering techniques to improve retrieval. In [SCZ98a] they evaluate on a 29,400 texture and color feature vectors (without mentioning the number of images used). The database is classified in five classes (cloud, floral, leaves, mountain, water). 'Recall' and 'precision' ratios are used to express the quality of the resulting clusters. Retrieval is evaluated using 19 query images. In [SCZ98b] they evaluate a neural network to learn weights (off-line) for combining feature spaces during retrieval. In this paper, they use judgments for 400 pairs of images, classified as similar or non-similar. They do not mention how many different images are used to construct the set of pairs, or who did classification.

The lack of evaluation makes it very hard to say something useful about the performance of these approaches. Multimedia retrieval research is still in its infancy. Apparently, the introduction of new techniques does not yet have to be supported by thorough experimental evaluation. The community is still in the 'proof-of-concept' phase.

4 Other approaches to evaluation

Lienhart et al. describe the MoCA video abstracting project in [LPE97]. They compare the quality of the generated abstracts to commercial video abstracts used on German television. The evaluation has a setup similar to the Turing-test. The human subjects could not tell the difference between the automatically generated abstracts and the commercial abstracts. Note that they did not evaluate whether it is possible to do worse than this, by using randomly selected fragments of the video data as a baseline performance.

Vosconcelos and Lippman, MIT Media Lab, present a Bayesian video modeling framework for segmentation [VL97]. They perform some experiments to select the best segmentation algorithm, using 24 trailers, of each two minutes. This collections contains about 100 shots (which takes 26 Gb of disk space). They evaluate the characterization of the detected segments in an interesting way. They use the manual classification from the 'Internet Movie Database' to evaluate the quality of the automatic classification.

Thomas Minka, MIT Media Lab, evaluated the algorithms in his Foureyes learning agent on learning time using simulation experiments [Min96, PMS96]. He measures the number of corrections the user would have to supply until the database system 'knows' the correct classification for all objects.

Lakshmi Ratan and Grimson evaluated the performance of their method for the classification of scenes in images without a manually annotated image collection. They compared the retrieval performance of their method to the performance of QBIC on the same data set, counting the number of false positives in the result sets of both systems [RG97]. The collection used is the Corel photo library, consisting of 800 annotated images of natural scenes.

La Cascia et al. also avoid manual annotation [CSS98]. They describe a small user study to measure the performance of their retrieval system collecting images from the web ([STC97]). In an evaluation, they randomly select 100 images from a 10,000 image collection. The subjects then have the task to try and find those images using the retrieval system. A search is considered successful if the subjects could get the image displayed in the top 100 within four iterations of relevance feedback. They vary the following conditions: use visual, textual, or both for content representation, and use visual, textual, or both for the relevance feedback. The experiments have been repeated for larger database sizes, up to 100,000 images. Only two subjects have been studied however, and these were the same in the different conditions.

Sutcliffe et al. did a more traditional user study to investigate how users of a multimedia data set perform a question-answer task [SHDR97]. However, the data collection consisted of just seven documents in the MS Windows environment, without a special retrieval system.

5 Conclusions

In the papers studied in this review, the data sets are very small. Many collections are tailored to evaluate only a very specific low-level task, cf. the evaluation of texture algorithms based on the Brodatz collection. However, the results on such a collection are easily generalized for very different, high-level search tasks. Although some authors seem to realize the relative weaknesses in their evaluations, others are perfectly happy with the 'proof' derived from their experiments.

Most papers claim to present a novel, *better* approach to multimedia query processing. As proof that the novel approach really is 'better', results provided in an 'evaluation' section are vaguely based on concepts borrowed from the scientific evaluation methodology used in IR. The evaluation sections in the papers mentioned in section 3 proof however mainly the authors' lack of understanding of that methodology:

- only a small number of queries is used (often one or two);
- only precision-recall measures for one cut-off point in a ranking are presented;
- a significance test is not applied;

- the data is divided in a small number of classes, and these are considered both the relevance judgments and the complete description of the user's information need;
- relevance judgments are considered completely objective, but usually made by the paper's authors and not by real users.

The inherent subjectivity of multimedia search is usually ignored completely. Almost all papers report experiments with multimedia search in which only the success on a task of object identification is tested: e.g. does the image contain a lion or not. The emotional and aesthetic values that play a role in the evaluation process of the user are overlooked. Or, even worse, the underlying techniques are 'improved' in a such a way that they are less sensitive to exactly those aspects that *are* important for such values.

The lack of evaluation methodology is clearly a limiting factor in the development of the field of multimedia retrieval. The MeasTex initiative of University of Queensland now provides a common framework for evaluation of texture classification algorithms. Although the task of texture classification is only partly related to building multimedia retrieval systems, the availability of a common evaluation framework for this subtask is definitely a step in the right direction.

We are (unfortunately) not convinced that an evaluation methodology for multimedia retrieval exists that can draw valid conclusions based on experiments without real users. The underlying problem is that there exists no objective ground truth in retrieval experiments involving multimedia data. The difficulties (with respect to resources) to construct a thorough method to evaluate multimedia techniques may be alleviated using a combination of ideas described in section 4. Historical data of experiments with real users in common test sets may be crucial to allow comparison between different approaches.

Until we find a better approach to measure the performance of multimedia retrieval systems, it is very important that we realize the limitations of our experimental 'proof'. We should also realize that to the end users, multimedia retrieval often constitutes much more than 'just' the identification of objects of some particular class.

References

[AC97] E. Ardizzone and M. La Cascia. Automatic video database indexing and retrieval. *Multimedia tools and applications*, 1(4):29–55, 1997. 6

[BFJ+96] M.G. Brown, J.T. Foote, G.J.F. Jones, K. Spärck Jones, and S. J. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings ACM Multimedia*, Boston, November 1996. 3

[Bon97] J.S. De Bonet. Novel statistical multiresolution techniques for image synthesis, discrimination, and recognition. Master's thesis, MIT, 1997. 4, 4

[BV97] J.S. De Bonet and P. Viola. Structure driven image database retrieval. In *Advances in Neural Information Processing*, number 10, 1997. 4, 4

[BV98] J.S. De Bonet and P. Viola. Texture recognition using a non-parametric multi-scale statistical model. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 1998. 4

[BVI98] J.S. De Bonet, P. Viola, and J.W. Fisher III. Flexible histograms: A multiresolution target discrimination model. In E.G. Zelnio, editor, *Proceedings of SPIE*, volume 3370, 1998. 4

[CCM+97] S.-F. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong. VideoQ: an automated content based video search system using visual cues. In *Proceedings of ACM Multimedia 1997*, Seattle, November 1997. 6

[CLP94] T.-S. Chua, S.-K. Lim, and H.-K. Pung. Content-based retrieval of segmented images. In *ACM Multimedia 94*, pages 211–218, San Francisco, 1994. 6

[CSS98] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the World Wide Web. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, Santa Barbara, CA, June 1998. 8

[CSZSM97] W. Chang, G. Sheikholeslami, A. Zhang, and T. Syeda-Mahmood. Efficient resource selection in distributed visual information systems. In *ACM Multimedia'97*, Seattle, WA, November 1997. 3

[Dra97] S. Draper. The aims for MIRA workshops. URL http://www.psy.gla.ac.uk/~steve/workshopAims.html (visited Feb 11th, 1998), 1997. 2

[FBF+94] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3:231–262, 1994. 5

[Foo97] J. Foote. A similarity measure for automatic audio classification. In *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, March 1997. 3

[GK96] U. Gargi and R. Kasturi. An evaluation of color histogram based methods in video indexing. In *Proceedings of the first international workshop on Image databases and multi-media search (IDB-MMS '96)*, pages 75–82, Amsterdam, The Netherlands, August 1996. 5

[GS96] T. Gevers and A.W.M. Smeulders. A comparative study of several color models for color image invariant retrieval. In *Proceedings of the first international workshop on Image databases and multi-media search (IDB-MMS '96)*, pages 17–26, Amsterdam, The Netherlands, August 1996. 4

[Hea94] M.A. Hearst. Multi-paragraph segmentation of expository text. In *ACL '94*, Las Cruces, 1994. 5

[HW97] A.G. Hauptmann and M.J. Witbrock. *Intelligent multimedia information retrieval*, chapter Informedia: news-on-demand multimedia information acquisition and retrieval, pages 215–239. AAAI Press/MIT Press, 1997. 6

[JFJY96] G.J.F. Jones, J.T. Foote, K. Spärck Jones, and S.J. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th*

International Conference on Research and Development in Information Retrieval (SIGIR '96), Zürich, Switzerland, August 1996. 3

[LP96] F. Liu and R.W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):722–733, July 1996. 3

[LPE97] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Communications of the ACM*, 40(12), December 1997. 7

[MHMG97] I. Mani, D. House, M. Maybury, and M. Green. *Intelligent multimedia information retrieval*, chapter Towards content-based browsing of broadcast news video, pages 241–258. AAAI Press/MIT Press, 1997. 5

[Min96] T.P. Minka. An image database browser that learns from user interaction. Master's thesis, MIT, 1996. Also appeared as MIT Media Laboratory technical report 365. 4, 4, 7

[MM96] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, August 1996. 3

[MP97] T.P. Minka and R.W. Picard. Interactive learning using a "society of models". *Pattern Recognition*, 30(4):565–581, April 1997. 3, 4

[MRC+97] S. Mehrotra, Y. Rui, K. Chakrabarti, M. Ortega, and Th.S. Huang. Multimedia analysis and retrieval system. In *Proceedings of the 3rd International Workshop on Information Retrieval Systems*, Como, Italy, September 1997. 6

[NBE+93] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutsos. The QBIC project: querying images by content using color, texture and shape. Technical Report RJ 9203, IBM Research Division, 1993. 5

[NK98] N. Nes and M. Kersten. The Acoi algebra: A query algebra for image retrieval systems. In *Advances in Databases. 16th British National Conference on Databases, BNCOD 16*, pages 77–88, Cardiff, Wales, UK, July 1998. 4

[NL95] A. Desai Narasimhalu and M.-K. Leong. Experiences with content based retrieval of multimedia information. In *Final Workshop on Multimedia Information Retrieval (MIRO'95)*, October 1995. 6

[ORC+97] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and Th.S. Huang. Supporting similarity queries in MARS. In *Proceedings of ACM Multimedia 1997*, Seattle, Washington, November 1997. 6

[OS95] V.E. Ogle and M. Stonebraker. Chabot: retrieval from a relational database of images. *IEEE Computer*, 28(9):40–48, September 1995. 5

[PM95] R.W. Picard and T.P. Minka. Vision texture for annotation. *Journal of multimedia systems*, 3:3–14, 1995. 4

[PMS96] R.W. Picard, T.P. Minka, and M. Szummer. Modeling user subjectivity in image libraries. In *IEEE International Conference on Image Processing*, Lausanne, September 1996. 3, 7

[PP97] K. Popat and R.W. Picard. Cluster-based probability model and its application to image and texture processing. *IEEE Transactions on Image Processing*, 6(2):268–284, February 1997. 3

[RG97] A. Lakshmi Ratan and W.E.L. Grimson. Training templates for scene classification using a few examples. In *Proceedings IEEE Workshop on Content-based Access of Image and Video Libraries*, San Juan, Puerto Rico, June 1997. 8

[RHM98] Y. Rui, Th.S. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *Proceedings of IS&T and SPIE Storage and Retrieval of Image and Video Databases VI*, San Jose, CA, January 1998. 6

[RHMO97a] Y. Rui, Th.S. Huang, S. Mehrotra, and M. Ortega. Automatic matching tool selection via relevance feedback in MARS. In *Proc. of The 2nd Int. Conf. on Visual Information Systems*, San Diego, California, December 1997. 6

[RHMO97b] Y. Rui, Th.S. Huang, S. Mehrotra, and M. Ortega. A relevance feedback architecture for content-based multimedia information retrieval systems. In *Proceedings IEEE Workshop on Content-based Access of Image and Video Libraries*, San Juan, Puerto Rico, June 1997. 6

[RSZ96] E. Remias, G. Sheikholeslami, and A. Zhang. Block-oriented image decomposition and retrieval in image database systems. In *The 1996 International Workshop on Multi-media Database Management Systems*, Blue Mountain Lake, New York, August 1996. 3

[SC94] J.R. Smith and S.-F. Chang. Quad-tree segmentation for texture-based image query. In *ACM Multimedia 94*, pages 279–286, San Francisco, 1994. 3

[SC96] J.R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *ACM Multimedia 96*, Boston, MA, 1996. 6

[SC97] J.R. Smith and S.-F. Chang. *Intelligent multimedia information retrieval*, chapter Querying by color regions using the VisualSEEK content-based visual query system, pages 23–41. AAAI Press/MIT Press, 1997. 6

[Sch97] P. Schäuble. *Multimedia information retrieval. Content-based information retrieval from large text and audio databases*. Kluwer Academic Publishers, 1997. 3

[SCZ98a] G. Sheikholeslami, W. Chang, and A. Zhang. Semquery: Semantic clustering and querying on heterogeneous features for visual data. In *ACM Multimedia 98*, Bristol, UK, September 1998. 7

[SCZ98b] G. Sheikholeslami, S. Chatterjee, and A. Zhang. Neuromerge: An approach for merging heterogeneous features in content-based image retrieval systems.

In *4th International Workshop on Multi-Media Database Management Systems (IW-MMDBMS'98)*, Dayton, Ohio, August 5-7 1998. 7

[SHDR97] A. Sutcliffe, M. Hare, A. Doubleday, and M. Ryan. *Intelligent multimedia information retrieval*, chapter Emperical studies in multimedia information retrieval, pages 449–472. AAAI Press/MIT Press, 1997. 8

[SJ] S. Santini and R. Jain. Similarity matching. Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence. 3

[STC97] S. Sclaroff, L. Taycher, and M. La Cascia. Imagerover: A content-based image browser for the world wide web. In *Proceedings IEEE Workshop on Content-based Access of Image and Video Libraries*, San Juan, Puerto Rico, June 1997. 4, 8

[SW95] P. Schäuble and M. Wechsler. First experiences with a system for content based retrieval of information from speech recordings. In *IJCAI Workshop: Intelligent multimedia information retrieval*, August 1995. 3

[VH97] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, number 500-240 in NIST Special publications, November 1997. 3

[VK95] A. Vellaikal and C.-C. J. Kuo. Content-based image retrieval using multiresolution histogram representation. In *SPIE Digital Image Storage and Archiving Systems*, pages 312–323, Philadelphia, October 1995. 5

[VK96] A. Vellaikal and C.-C. J. Kuo. Joint spatial-spectral indexing for image retrieval. In *IEEE International Conference on Image Processing*, pages 867–870, Lausanne, September 1996. 5

[VL97] N. Vasconcelos and A. Lippman. A Bayesian video modeling framework for shot segmentation and content characterization. In *Proceedings of the IEEE workshop on content-based access of image and video libraries*, San Juan, Puerto Rico, June 1997. 7

[WBKW96] E. Wold, Th. Blum, D. Keisler, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3), 1996. 3

[WKSS96] H. Wactlar, T. Kanade, M. Smith, and S. Stevens. Intelligent access to digital video: The Informedia project. *IEEE Computer*, 29(5), May 1996. 6

[WLM+97] J.K. Wu, C.P. Lam, B.M. Mehtre, Y.J. Gao, and A. Desai Narasimhalu. Content-based retrieval for trademark registration. *Multimedia tools and applications*, 1(3), 1997. 6

[WNM+95] J.K. Wu, A. Desai Narasimhalu, B.M. Mehtre, C.P. Lam, and Y.J. Gao. CORE: a content-based retrieval engine for multimedia information systems. *Multimedia Systems*, 3:25–41, 1995. 6

[WS95] M. Wechsler and P. Schäuble. Speech retrieval based on automatic indexing. In *Final Workshop on Multimedia Information Retrieval (MIRO'95)*, October 1995. 3

[ZCA97] A. Zhang, B. Cheng, and R. Acharya. A fractal-based clustering approach in large visual database systems. *Multimedia tools and applications*, 1(3):225–244, 1997. 3, 4

[ZKS93] H.J. Zhang, A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993. 5