Concept-Based Video Retrieval

Concept-Based Video Retrieval

Cees G. M. Snoek

University of Amsterdam
Science Park 107, 1098 XG Amsterdam
The Netherlands
cgmsnoek@uva.nl

Marcel Worring

University of Amsterdam
Science Park 107, 1098 XG Amsterdam
The Netherlands
worring@uva.nl

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is C. G. M. Snoek and M. Worring, Concept-Based Video Retrieval, Foundation and Trends® in Information Retrieval, vol 2, no 4, pp 215–322, 2008

ISBN: 978-1-60198-234-6

Foundations and Trends® in Information Retrieval

Volume 2 Issue 4, 2008

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2008, Volume 2, 4 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 2, No. 4 (2008) 215–322

DOI: 10.1561/1500000014

Concept-Based Video Retrieval

Cees G. M. Snoek[1] and Marcel Worring[2]

[1] *University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands, cgmsnoek@uva.nl*
[2] *University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands, worring@uva.nl*

Abstract

In this paper, we review 300 references on video retrieval, indicating when text-only solutions are unsatisfactory and showing the promising alternatives which are in majority concept-based. Therefore, central to our discussion is the notion of a semantic concept: an objective linguistic description of an observable entity. Specifically, we present our view on how its automated detection, selection under uncertainty, and interactive usage might solve the major scientific problem for video retrieval: the semantic gap. To bridge the gap, we lay down the anatomy of a concept-based video search engine. We present a component-wise decomposition of such an interdisciplinary multimedia system, covering influences from information retrieval, computer vision, machine learning, and human–computer interaction. For each of the components we review state-of-the-art solutions in the literature, each having different characteristics and merits. Because of these differences, we cannot understand the progress in video retrieval without serious evaluation efforts such as carried out in the NIST TRECVID benchmark. We discuss its data, tasks, results, and the many derived community initiatives in creating annotations and baselines for repeatable experiments. We conclude with our perspective on future challenges and opportunities.

Contents

1

Introduction

1.1 How to Retrieve Video Content?

This question is highly relevant in a world that is adapting swiftly to visual communication. Online services like YouTube and Tudou show that video is no longer the domain of broadcast television only. Video has become the medium of choice for many people communicating via Internet and their mobile phones. Digital video is leading to an abundance of narrowcast repositories, with content as diverse as Al Jazeera news, concerts of the Royal Philharmonic Orchestra, and the baby panda at your local zoo, to name just three examples. A nation's broadcast archive can be expected to contain petabytes of video data, requiring careful treatment for future preservation and disclosure. Personal video archives are likely to be much smaller, but due to the amount of effort involved, the willingness to archive for the purpose of future retrieval will be lower. At all stages and for all target groups, effective and efficient video retrieval facilities will be necessary, not only for the public broadcasters, but also for any private broadcaster and narrowcaster-to-be.

User needs determine both the effectiveness and efficiency of video search engines. To understand what are the user needs for video retrieval, we draw inspiration from the video production process. According to Jain and Hampapur [106], the purpose for which a video is created is either entertainment, information, communication, or data analysis. For all these purposes, the user needs and demands vary substantially. A consumer who wants to be entertained, for example, will be satisfied if a complete movie is accessible from an archive through a mobile phone. In contrast, a cultural anthropologist studying fashion trends of the eighties, a lawyer evaluating copyright infringement, or an athlete assessing her performance during training sessions might be more interested in retrieving specific video segments, without going through an entire video collection. For accessing complete video documents, reasonable effective commercial applications exist, YouTube and Netflix being good examples. Video search applications for consumers and professionals targeting at retrieval of specific segments, however, are still in a nascent stage [112]. Users requiring access to video segments are hardly served by present-day video retrieval applications.

In this paper, we review video search solutions that target at retrieval of specific segments. Since humans perceive video as a complex interplay of cognitive concepts, the all-important step forward in such video retrieval approaches will be to provide access at the semantic level. This is achievable by labeling all combinations of people, objects, settings, and events appearing in the audiovisual content. Labeling things has been the topic of scientific endeavor since Aristotle revealed his "Categories." Following in this tradition are Linnaeus (biology), Werner (geology), Mendeleev (chemistry), and the Human Genome Project (genetics) [263]. In our information age, Google labels the world's textual information. Labeling video content is a grand challenge of our time as humans use approximately half of their cognitive capacity to achieve such tasks [177]. Two types of semantic labeling solutions have emerged: (i) the first approach relies on human labor, where labels are assigned manually after audiovisual inspection; (ii) the second approach is machine-driven with automatic assignment of labels to video segments.

1.2 Human-Driven Labeling

Manual labeling of (broadcast) video has traditionally been the realm of professionals. In cultural heritage institutions, for example, library experts label archival videos for future disclosure using controlled vocabularies [56, 148]. Because expert labeling [50, 155] is tedious and costly, it typically results in a brief description of a complete video only. In contrast to expert labor, Web 2.0 [172] has launched social tagging, a recent trend to let amateur consumers label, mostly personal, visual content on web sites like YouTube, Flickr, and Facebook. Alternatively, the manual concept-based labeling process can be transformed into a computer game [253] or a tool facilitating volunteer-based labeling [198]. Since the labels were never meant to meet professional standards, amateur labels are known to be ambiguous, overly personalized, and limited [69, 73, 149]. Moreover, unlabeled video segments remain notoriously difficult to find. Manual labeling, whether by experts or amateurs, is geared toward one specific type of use and, therefore, inadequate to cater for alternative video retrieval needs, especially those user needs targeting at retrieval of video segments [204].

1.3 Machine-Driven Labeling

Machine-driven labeling aims to derive meaningful descriptors from video data. These descriptors are the basis for searching large video collections. Many academic prototypes, such as Medusa [22], Informedia classic [255], and Olive [51], and most commercial video search engines such as Baidu, Blinkx, and Truveo, provide access to video based on text, as this is still the easiest way for a user to describe an information need. The labels of these search engines are based on the filename, surrounding text, social tags, closed captions, or a speech transcript. Text-based video search using speech transcripts has proven itself especially effective for segment-level retrieval from (English) broadcast news, interviews, political speeches, and video blogs featuring talking heads. However, a video search method based on just speech transcripts results in disappointing retrieval performance, when the audiovisual content is neither mentioned, nor properly reflected in the associated text. In addition, when the videos originate from non-English speaking countries, such as China, and the Netherlands, querying the content becomes much harder as robust automatic speech recognition results and their accurate machine translations are difficult to achieve.

It might seem that video retrieval is the trivial extension of text retrieval, but it is in fact often more complex. Most of the data is of sensory origin (image, sound, video) and hence techniques from digital signal processing and computer vision are required to extract relevant descriptions. In addition to the important and valuable text data derived from audio analysis, much information is captured in the visual stream. Hence, a vast body of research in machine-driven video labeling has investigated the role of visual content, with or without text. Analyzing the content of visual data using computers has a long history [195], dating back to the 1960s. Some initial successes prompted researchers in the 1970s to predict that the problem of understanding visual material would soon be solved completely. However, the research in the 1980s showed that these predictions were far too optimistic. Even now, understanding visual data is a major challenge. In the 1990s a new field emerged, namely content-based image retrieval, where the aim is to develop methods for searching in large image archives.

Research in content-based retrieval has resulted in a wide variety of image and video search systems [17, 32, 34, 61, 67, 72, 114, 143, 180, 197, 207, 214, 258, 266]. A common denominator in these prototypes is their dependence on low-level visual labels such as color, texture, shape, and spatiotemporal features. Most of those early systems are based on query-by-example, where users query an archive based on images rather than the visual feature values. They do so by sketches, or by providing example images using a browser interface. Query-by-example can be fruitful when users search for the same object under slightly varying circumstances and when the target images are available indeed. If proper example images are unavailable, content-based image retrieval techniques are not effective at all. Moreover, users often do not understand similarity expressed in low-level visual features. They expect semantic similarity. This expected semantic similarity, is exactly the major problem video retrieval is facing.

The source of the problem lies in the *semantic gap*. We slightly adapt the definition by Smeulders et al. [213] and define it as: “The lack of correspondence between the low-level features that machines extract from video and the high-level conceptual interpretations a human gives to the data in a given situation." The existence of the gap has various causes. One reason is that different users interpret the same video data in a different way. This is especially true when the user is making subjective interpretations of the video data related to feelings or emotions, for example, by describing a scene as *romantic* or *hilarious* [76]. In this paper, those subjective interpretations are not considered. However, also for objective interpretations, like whether a *windmill* is present in a video, developing automatic methods is still difficult. The main difficulties are due to the large variations in appearance of visual data corresponding to one semantic concept. Windmills, for example, come in different models, shapes, and colors. These causes are inherent to the problem. Hence, the aim of video retrieval must be to bridge the semantic gap.

1.4 Aims, Scope, and Organization

In this paper, we review state-of-the-art video retrieval methods that challenge the semantic gap. In addition, we also address the important issue of evaluation. In particular, we emphasize *concept-based video retrieval*. A recent breakthrough in the field, which facilitates searching in video at a segment-level by means of large sets of automatically detected (visual) concepts, like a *telephone*, a *flamingo*, a *kitchen*, or one of the concepts in Figure 1.1. Evidence is accumulating that when large sets of concept detectors are available at retrieval time, such an approach to video search is effective [36, 219, 227]. In fact, by using a simulation study, Hauptmann et al. [85] show that even when the individual detectors have modest performance, several thousand detectors are likely to be sufficient for video search in the broadcast news domain to approach standard WWW search quality. Hence, when using concept detectors for video retrieval, we might be able to reduce the semantic gap for the user.

In contrast to other reviews on video retrieval, which emphasize either content-based analysis [221], machine learning [158], text and image retrieval [283], search strategies [118], interactive browsing

Fig. 1.1 Visual impression of 101 typical semantic concepts [230] for which automatic detection, retrieval, and evaluation results on video data are described in this paper.

models [86], or challenges for the future [129], the aim of this paper is to cover the semantic gap completely: from low-level features that can be extracted from the video content to high-level interpretation of video segments by an interacting user. Although these concepts have a visual nature, concept-based video retrieval is different from still-image concept detection as the concepts are often detected and retrieved using an interdisciplinary approach combining text, audio, and visual information derived from a temporal video sequence. Our review on concept-based video retrieval, therefore, covers influences from information retrieval, computer vision, machine learning, and human–computer interaction. Because of this interdisciplinary nature, it is impossible for us to provide a complete list of references. In particular, we have not attempted to provide an accurate historical attribution of ideas. Instead, we give preference to peer-reviewed journal papers, over earlier published conference papers, and workshop papers, where possible. Throughout the review we assume a basic familiarity with computer science and information retrieval, but not necessarily the specific topic of (visual) video retrieval. For in depth, technical details on the fundamentals underlying many concept-based video retrieval methods, the interested reader is referred to recent books [16, 18, 19, 74, 128, 147, 199], review papers [48, 103, 140, 213, 241, 261], special issues [75, 270], and online proceedings [171], that provide further entry points into the literature on specific topics.

We organize the paper by laying down the anatomy of a concept-based video search engine. We present a component-wise decomposition of such an interdisciplinary multimedia system in our aim to bridge the semantic gap. The components exploit a common architecture, with a standardized input–output model, to allow for semantic integration.

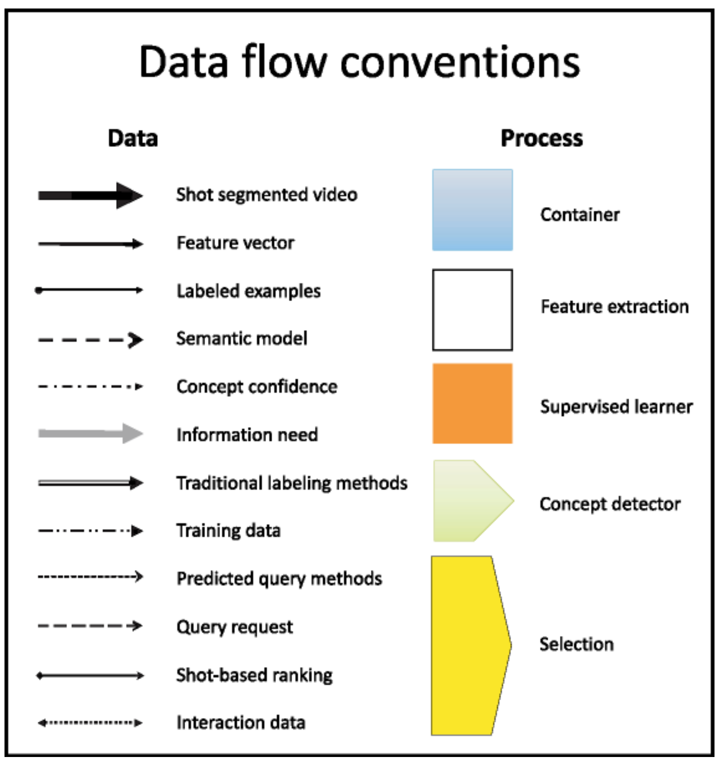

Fig. 1.2 Data flow conventions as used in this paper. Different arrows indicate difference in data flows.

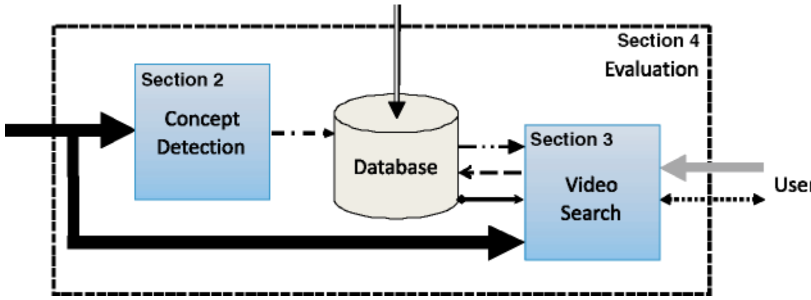

Fig. 1.3 We organize our review by laying down the anatomy of a concept-based video search engine, building upon the conventions introduced in Figure 1.2. First, we detail generic concept detection in Section 2. Then, we highlight how concept detectors can be leveraged for video search in combination with traditional labeling methods and an interacting user in Section 3. We present an in depth discussion on evaluating concept-based video retrieval systems and components in Section 4.

The graphical conventions to describe the system architecture are indicated in Figure 1.2. We will use the graphical conventions throughout this paper. Based on these conventions, we follow the video data as they flow through the computational process, as sketched in Figure 1.3. We start in Section 2, where we present a general scheme for generic concept detection. We cover the common concept detection solutions from the literature and discuss how they interconnect for large-scale detection of semantic concepts. The availability of a large set of concept detectors opens up novel opportunities for video retrieval. In Section 3, we detail how uncertain concept detectors can be leveraged for video retrieval at query time and how concept detectors can be combined with more traditional labeling methods. Moreover, we discuss novel visualizations for video retrieval and we highlight how to improve concept-based video retrieval results further by relying on interacting users. In Section 4, we turn our attention to evaluation of concept-based video search engines and their most important components. We introduce the de facto benchmark standard, its most important tasks, and evaluation protocols. In addition, we highlight the many community efforts in providing manual annotations and concept-based video retrieval baselines against which scientific progress in the field is measured. We conclude the review with our perspective on the challenges and opportunities for concept-based video search engines of the future.

References

[1] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y. Lin, A. P. Natsev, M. R. Naphade, C. Neti, H. J. Nock, H. H. Permuter, R. Singh, J. R. Smith, S. Srinivasan, B. L. Tseng, T. V. Ashwin, and D. Zhang, "IBM research TREC-2002 video retrieval system," in *Proceedings of the 11th Text Retrieval Conference*, Gaithersburg, USA, 2002.

[2] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 170–185, 2003.

[3] J. Adcock, M. Cooper, and F. Chen, "FXPAL MediaMagic video search system," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 644–644, Amsterdam, The Netherlands, 2007.

[4] J. Adcock, M. Cooper, and J. Pickens, "Experiments in interactive video search by addition and subtraction," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 465–474, Niagara Falls, Canada, 2008.

[5] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, pp. 832–843, 1983.

[6] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. R. Naphade, A. P. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang, "IBM research TRECVID-2003 video retrieval system," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2003.

[7] S. Ayache and G. Quénot, "Evaluation of active learning strategies for video indexing," *Image Communication*, vol. 22, nos. 7–8, pp. 692–704, 2007.

[8] S. Ayache and G. Quénot, "Video corpus annotation using active learning," in *European Conference on Information Retrieval*, pp. 187–198, Glasgow, UK, 2008.
[9] S. Ayache, G. Quénot, and J. Gensel, "Classifier fusion for SVM-based multimedia semantic indexing," in *European Conference on Information Retrieval*, pp. 494–504, Rome, Italy, 2007.
[10] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Transactions on Multimedia*, vol. 4, pp. 68–75, 2002.
[11] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted American football video by highlight selection," *IEEE Transactions on Multimedia*, vol. 6, pp. 575–586, 2004.
[12] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in *International Joint Conference on Artificial Intelligence*, pp. 805–810, Acapulco, Mexico, 2003.
[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008.
[14] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A multimedia information network for knowledge representation," in *Proceedings of SPIE Conference on Internet Multimedia Management Systems*, Boston, USA, 2000.
[15] M. Bertini, A. D. Bimbo, and C. Torniai, "Automatic video annotation using ontologies extended with visual information," in *Proceedings of the ACM International Conference on Multimedia*, pp. 395–398, Singapore, 2005.
[16] A. D. Bimbo, *Visual Information Retrieval*. Morgan Kaufmann, 1999.
[17] A. D. Bimbo and P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 121–132, 1997.
[18] C. M. Bishop, *Pattern Recognition and Machine Learning. Information Science and Statistics*. Springer, 2006.
[19] H. M. Blanken, A. P. de Vries, H. E. Blok, and L. Feng, eds., *Multimedia Retrieval*. Springer, 2007.
[20] T. Bompada, C.-C. Chang, J. Chen, R. Kumar, and R. Shenoy, "On the robustness of relevance measures with incomplete judgments," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 359–366, Amsterdam, The Netherlands, 2007.
[21] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 55–73, 1990.
[22] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck-Jones, and S. J. Young, "Automatic content-based retrieval of broadcast news," in *Proceedings of the ACM International Conference on Multimedia*, San Francisco, USA, 1995.
[23] R. Brunelli, O. Mich, and C. M. Modena, "A survey on the automatic indexing of video data," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 78–112, 1999.

[24] E. Bruno, N. Moenne-Loccoz, and S. Marchand-Maillet, "Design of multimodal dissimilarity spaces for retrieval of multimedia documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1520–1533, 2008.

[25] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 25–32, Sheffield, UK, 2004.

[26] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, pp. 13–47, 2006.

[27] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[28] G. J. Burghouts and J.-M. Geusebroek, "Performance evaluation of local color ivariants," *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.

[29] D. Byrne, A. R. Doherty, C. G. M. Snoek, G. J. F. Jones, and A. F. Smeaton, "Validating the detection of everyday concepts in visual lifelogs," in *Proceedings International Conference on Semantics and Digital Media Technologies*, pp. 15–30, Berlin, Germany: Springer-Verlag, 2008.

[30] M. Campbell, A. Haubold, S. Ebadollahi, D. Joshi, M. R. Naphade, A. P. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tešić, and L. Xie, "IBM research TRECVID-2006 video retrieval system," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[31] J. Cao, Y. Lan, J. Li, Q. Li, X. Li, F. Lin, X. Liu, L. Luo, W. Peng, D. Wang, H. Wang, Z. Wang, Z. Xiang, J. Yuan, W. Zheng, B. Zhang, J. Zhang, L. Zhang, and X. Zhang, "Intelligent multimedia group of Tsinghua University at TRECVID 2006," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[32] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1026–1038, 2002.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm/, 2001.

[34] S.-F. Chang, W. Chen, H. J. Men, H. Sundaram, and D. Zhong, "A fully automated content-based video search engine supporting spatio-temporal queries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 602–615, 1998.

[35] S.-F. Chang, J. He, Y.-G. Jiang, E. E. Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky, "Columbia University/VIREO-CityU/IRIT TRECVID-2008 high-level feature extraction and interactive video search," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[36] S.-F. Chang, W. Hsu, W. Jiang, L. S. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky, "Columbia University TRECVID-2006 video search and high-level feature extraction," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[37] O. Chapelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*. Cambridge, USA: The MIT Press, 2006.

[38] M.-Y. Chen, M. G. Christel, A. G. Hauptmann, and H. Wactlar, "Putting active learning into multimedia applications: Dynamic definition and refinement of concept classifiers," in *Proceedings of the ACM International Conference on Multimedia*, pp. 902–911, Singapore, 2005.

[39] M.-Y. Chen and A. G. Hauptmann, "Multi-modal classification in digital news libraries," in *Proceedings of the Joint Conference on Digital Libraries*, pp. 212–213, Tucson, USA, 2004.

[40] M. G. Christel and R. M. Conescu, "Mining novice user activity with TRECVID interactive retrieval tasks," in *CIVR*, pp. 21–30, Springer-Verlag, 2006.

[41] M. G. Christel and A. G. Hauptmann, "The use and utility of high-level semantic features," in *CIVR*, pp. 134–144, Springer-Verlag, 2005.

[42] M. G. Christel, A. G. Hauptmann, H. D. Wactlar, and T. D. Ng, "Collages as dynamic summaries for news video," in *Proceedings of the ACM International Conference on Multimedia*, pp. 561–569, Juan-les-Pins, France, 2002.

[43] M. G. Christel, C. Huang, N. Moraveji, and N. Papernick, "Exploiting multiple modalities for interactive video retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1032–1035, Montreal, Canada, 2004.

[44] T.-S. Chua, "Towards the next plateau — Innovative multimedia research beyond TRECVID," in *Proceedings of the ACM International Conference on Multimedia*, Augsburg, Germany, 2007.

[45] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives — Techniques, experience and trends," in *Proceedings of the ACM International Conference on Multimedia*, pp. 656–659, New York, USA, 2004.

[46] T.-S. Chua et al., "TRECVID-2004 search and feature extraction task by NUS PRIS," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2004.

[47] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, and M. G. Strintzis, "Knowledge-assisted semantic video object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1210–1224, 2005.

[48] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences and trends of the new age," *ACM Computing Surveys*, vol. 40, pp. 1–60, 2008.

[49] G. Davenport, T. G. A. Smith, and N. Pincever, "Cinematic principles for multimedia," *IEEE Computer Graphics & Applications*, vol. 11, pp. 67–74, 1991.

[50] M. Davis, "Editing out video editing," *IEEE MultiMedia*, vol. 10, pp. 54–64, 2003.

[51] F. M. G. de Jong, J. L. Gauvain, J. den Hartog, and K. Netter, "OLIVE: Speech-based video retrieval," in *European Workshop on Content-Based Multimedia Indexing*, Toulouse, France, 1999.

[52] O. de Rooij, C. G. M. Snoek, and M. Worring, "Query on demand video browsing," in *Proceedings of the ACM International Conference on Multimedia*, pp. 811–814, Augsburg, Germany, 2007.

[53] O. de Rooij, C. G. M. Snoek, and M. Worring, "Balancing thread based navigation for targeted video search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 485–494, Niagara Falls, Canada, 2008.

[54] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 800–810, 2001.

[55] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 881–884, Toronto, Canada, 2006.

[56] P. Enser, "Visual image retrieval: Seeking the alliance of concept-based and content-based paradigms," *Journal of Information Science*, vol. 26, pp. 199–210, 2000.

[57] J. Fan, A. K. Elmagarmid, X. Zhu, W. G. Aref, and L. Wu, "ClassView: Hierarchical video shot classification, indexing and accessing," *IEEE Transactions on Multimedia*, vol. 6, pp. 70–86, 2004.

[58] J. Fan, H. Luo, Y. Gao, and R. Jain, "Incorporating concept ontology for hierarchical video classification, annotation and visualization," *IEEE Transactions on Multimedia*, vol. 9, pp. 939–957, 2007.

[59] C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*. Cambridge, USA: The MIT Press, 1998.

[60] J. M. Ferryman, ed., *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. Rio de Janeiro, Brazil: IEEE Press, 2007.

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

[62] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, nos. 1–2, pp. 89–108, 2002.

[63] J.-M. Geusebroek, R. Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1338–1350, 2001.

[64] J.-M. Geusebroek and A. W. M. Smeulders, "A six-stimulus theory for stochastic texture," *International Journal of Computer Vision*, vol. 62, nos. 1–2, pp. 7–16, 2005.

[65] T. Gevers, "Adaptive image segmentation by combining photometric invariant region and edge information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 848–852, 2002.

[66] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, pp. 453–464, 1999.

[67] T. Gevers and A. W. M. Smeulders, "PicToSeek: Combining color and shape invariant features for image retrieval," *IEEE Transactions on Image Processing*, vol. 9, pp. 102–119, 2000.

[68] K.-S. Goh, E. Y. Chang, and W.-C. Lai, "Multimodal concept-dependent active learning for image retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 564–571, New York, USA, 2004.

[69] S. A. Golder and B. A. Huberman, "The structure of collaborative tagging systems," *Journal of Information Science*, vol. 32, pp. 198–208, 2006.

[70] R. Goulden, P. Nation, and J. Read, "How large can a receptive vocabulary be?," *Applied Linguistics*, vol. 11, pp. 341–363, 1990.

[71] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu, "Multi-layer multi-instance learning for video concept detection," *IEEE Transactions on Multimedia*, vol. 10, pp. 1605–1616, 2008.

[72] A. Gupta and R. Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, pp. 70–79, 1997.

[73] M. Guy and E. Tonkin, "Folksonomies: Tidying up tags?," *D-Lib Magazine*, vol. 12, Available at: http://www.dlib.org/dlib/january06/ guy/01guy.html, 2006.

[74] A. Hanjalic, *Content-Based Analysis of Digital Video*. Boston, USA: Kluwer Academic Publishers, 2004.

[75] A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith, "The holy grail of multimedia information retrieval: So close or yet so far away?," *Proceedings of the IEEE*, vol. 96, pp. 541–547, 2008.

[76] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, pp. 143–154, 2005.

[77] A. Haubold and J. R. Kender, "VAST MM: Multimedia browser for presentation video," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 41–48, Amsterdam, The Netherlands, 2007.

[78] A. Haubold and A. P. Natsev, "Web-based information content and its application to concept-based video retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 437–446, Niagara Falls, Canada, 2008.

[79] A. G. Hauptmann, R. V. Baron, M.-Y. Chen, M. G. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C. G. M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H. D. Wactlar, "Informedia at TRECVID-2003: Analyzing and searching broadcast news video," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2003.

[80] A. G. Hauptmann and S.-F. Chang, "LIBSCOM: Large analytics library and scalable concept ontology for multimedia research," http://www.libscom.org/, 2009.

[81] A. G. Hauptmann and M. G. Christel, "Successful approaches in the TREC video retrieval evaluations," in *Proceedings of the ACM International Conference on Multimedia*, New York, USA, 2004.

[82] A. G. Hauptmann, M. G. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proceedings of the IEEE*, vol. 96, pp. 602–622, 2008.

[83] A. G. Hauptmann and W.-H. Lin, "Assessing effectiveness in video retrieval," in *CIVR*, pp. 215–225, Springer-Verlag, 2005.

[84] A. G. Hauptmann, W.-H. Lin, R. Yan, J. Yang, and M.-Y. Chen, "Extreme video retrieval: Joint maximization of human and computer performance," in

Proceedings of the ACM International Conference on Multimedia, pp. 385–394, Santa Barbara, USA, 2006.

[85] A. G. Hauptmann, R. Yan, W.-H. Lin, M. G. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news," *IEEE Transactions on Multimedia*, vol. 9, pp. 958–966, 2007.

[86] D. Heesch, "A survey of browsing models for content based image retrieval," *Multimedia Tools and Applications*, vol. 40, pp. 261–284, 2008.

[87] D. Heesch and S. Rüger, "Image browsing: A semantic analysis of NN^k networks," in *CIVR*, pp. 609–618, Springer-Verlag, 2005.

[88] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66–75, 1994.

[89] M. A. Hoang, J.-M. Geusebroek, and A. W. M. Smeulders, "Color texture measurement and segmentation," *Signal Processing*, vol. 85, pp. 265–275, 2005.

[90] L. Hollink, M. Worring, and G. Schreiber, "Building a visual ontology for video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 479–482, Singapore, 2005.

[91] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer, "Video content annotation using visual analysis and a large semantic knowledgebase," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 327–334, Madison, USA, 2003.

[92] R. Houghton, "Named faces: Putting names to faces," *IEEE Intelligent Systems*, vol. 14, pp. 45–50, 1999.

[93] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Reranking methods for visual search," *IEEE MultiMedia*, vol. 14, pp. 14–22, 2007.

[94] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, pp. 179–187, 1962.

[95] J. Huang, S. R. Kumar, W.-J. Z. M. Mitra, and R. Zabih, "Color-spatial indexing and applications," *International Journal of Computer Vision*, vol. 35, pp. 245–268, 1999.

[96] T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, "Active learning for interactive multimedia retrieval," *Proceedings of the IEEE*, vol. 96, pp. 648–667, 2008.

[97] M. Huijbregts, R. Ordelman, and F. M. G. de Jong, "Annotation of heterogeneous multimedia content using automatic speech recognition," in *Proceedings International Conference on Semantics and Digital Media Technologies*, pp. 78–90, Berlin: Springer-Verlag, 2007.

[98] W. Hürst, "Video browsing on handheld devices — Interface designs for the next generation of mobile video players," *IEEE MultiMedia*, vol. 15, pp. 76–83, 2008.

[99] B. Huurnink and M. de Rijke, "Exploiting redundancy in cross-channel video retrieval," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 177–186, Augsburg, Germany, 2007.

[100] B. Huurnink, K. Hofmann, and M. de Rijke, "Assessing concept selection for video retrieval," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 459–466, Vancouver, Canada, 2008.

[101] E. Hyvönen, S. Saarela, A. Styrman, and K. Viljanen, "Ontology-based image retrieval," in *Proceedings of the International World Wide Web Conference*, Budapest, Hungary, 2003.

[102] G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. P. Khudanpur, D. Klakow, M. R. Krause, R. Manmatha, H. J. Nock, D. Petkova, B. Pytlik, and P. Virga, "Joint visual-text modeling for automatic retrieval of multimedia documents," in *Proceedings of the ACM International Conference on Multimedia*, pp. 21–30, Singapore, 2005.

[103] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–37, 2000.

[104] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern Recognition*, vol. 24, pp. 1167–1186, 1991.

[105] A. K. Jain and A. Vailaya, "Shape-based retrieval: A case study with trademark image databases," *Pattern Recognition*, vol. 31, pp. 1369–1390, 1998.

[106] R. Jain and A. Hampapur, "Metadata in video databases," *ACM SIGMOD Record*, vol. 23, pp. 27–33, 1994.

[107] W. Jiang, S.-F. Chang, and A. C. Loui, "Context-based concept fusion with boosted conditional random fields," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 949–952, Honolulu, USA, 2007.

[108] W. Jiang, E. Zavesky, S.-F. Chang, and A. C. Loui, "Cross-domain learning methods for high-level visual concept classification," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 161–164, San Diego, USA, 2008.

[109] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 494–501, Amsterdam, The Netherlands, 2007.

[110] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo, "CU-VIREO374: Fusing Columbia374 and VIREO374 for large scale semantic concept detection," Technical Report 223-2008-1, Columbia University ADVENT Technical Report, 2008.

[111] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods: Support Vector Learning*, (B. Schölkopf, C. Burges, and A. J. Smola, eds.), pp. 169–184, Cambridge, USA: The MIT Press, 1999.

[112] M. Kankanhalli and Y. Rui, "Application potential of multimedia information retrieval," *Proceedings of the IEEE*, vol. 96, pp. 712–720, 2008.

[113] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text and vehicle detection and tracking in video: Data, metrics and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 319–336, 2009.

[114] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database — Query by visual example," in *Proceedings of the International Conference on Pattern Recognition*, pp. 530–533, The Hague, The Netherlands, 1992.

[115] J. R. Kender and M. R. Naphade, "Visual concepts for news story tracking: analyzing and exploiting the NIST TRECVID video annotation experiment," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1174–1181, Washington, DC, USA, 2005.

[116] L. S. Kennedy, "Revision of LSCOM event/activity annotations," Technical Report 221-2006-7, Columbia University ADVENT Technical Report, 2006.

[117] L. S. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 333–340, Amsterdam, The Netherlands, 2007.

[118] L. S. Kennedy, S.-F. Chang, and A. P. Natsev, "Query-adaptive fusion for multimodal search," *Proceedings of the IEEE*, vol. 96, pp. 567–588, 2008.

[119] L. S. Kennedy, A. P. Natsev, and S.-F. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proceedings of the ACM International Conference on Multimedia*, pp. 882–891, Singapore, 2005.

[120] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 489–497, 1990.

[121] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[122] M. Larson, E. Newman, and G. Jones, "Overview of VideoCLEF 2008: Automatic generation of topic-based feeds for dual language audio-visual content," in *Working Notes for the Cross-Language Evaluation Forum Workshop*, Aarhus, Denmark, 2008.

[123] L. J. Latecki, R. Lakaemper, and U. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 424–429, Hilton Head Island, USA, 2000.

[124] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, New York, USA, 2006.

[125] H. Lee and A. F. Smeaton, "Designing the user-interface for the Físchlár digital video library," *Journal of Digital Information*, vol. 2, 2002.

[126] D. Lenat and R. Guha, eds., *Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project*. Reading, USA: Addison-Wesley, 1990.

[127] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the International Conference on Systems Documentation*, pp. 24–26, Toronto, Canada, 1986.

[128] M. S. Lew, ed., *Principles of Visual Information Retrieval*. Springer, 2001.

[129] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on*

Multimedia Computing, Communications and Applications, vol. 2, pp. 1–19, 2006.

[130] J. Li, W. Wu, T. Wang, and Y. Zhang, "One step beyond histograms: Image representation using markov stationary features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[131] X. Li, C. G. M. Snoek, and M. Worring, "Annotating images by harnessing worldwide user-tagged photos," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009.

[132] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 603–610, Amsterdam, The Netherlands, 2007.

[133] C.-Y. Lin, B. L. Tseng, and J. R. Smith, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2003.

[134] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, pp. 267–276, 2007.

[135] W.-H. Lin and A. G. Hauptmann, "News video classification using SVM-based multimodal classifiers and combination strategies," in *Proceedings of the ACM International Conference on Multimedia*, Juan-les-Pins, France, 2002.

[136] H. Liu and P. Singh, "ConceptNet: A practical commonsense reasoning toolkit," *BT Technology Journal*, vol. 22, pp. 211–226, 2004.

[137] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li, "Video search re-ranking via multi-graph propagation," in *Proceedings of the ACM International Conference on Multimedia*, pp. 208–217, Augsburg, Germany, 2007.

[138] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE Transactions on Multimedia*, vol. 10, pp. 240–251, 2008.

[139] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[140] G. Lu, "Indexing and retrieval of audio: A survey," *Multimedia Tools and Applications*, vol. 15, pp. 269–290, 2001.

[141] H.-B. Luan, S.-Y. Neo, H.-K. Goh, Y.-D. Zhang, S.-X. Lin, and T.-S. Chua, "Segregated feedback with performance-based adaptive sampling for interactive news video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 293–296, Augsburg, Germany, 2007.

[142] H. Luo, J. Fan, J. Yang, W. Ribarsky, and S. Satoh, "Analyzing large-scale news video databases to support knowledge visualization and intuitive retrieval," in *IEEE Symposium on Visual Analytics Science and Technology*, pp. 107–114, Sacramento, USA, 2007.

[143] W.-Y. Ma and B. S. Manjunath, "NeTra: A toolbox for navigating large image databases," *Multimedia Systems*, vol. 7, pp. 184–198, 1999.

[144] J. Magalhães and S. Rüger, "Information-theoretic semantic multimedia indexing," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 619–626, Amsterdam, The Netherlands, 2007.

[145] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 836–842, 1996.

[146] B. S. Manjunath, P. Salembier, and T. Sikora, eds., *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[147] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[148] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proceedings ACM International Conference on Hypertext and Hypermedia*, pp. 31–40, Odense, Denmark, 2006.

[149] K. K. Matusiak, "Towards user-centered indexing in digital image collections," *OCLC Systems & Services*, vol. 22, pp. 263–296, 2006.

[150] K. McDonald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *CIVR*, (W.-K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, eds.), pp. 61–70, Heidelberg, Germany: Springer, 2005.

[151] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z. Zha, Y. Liu, Z. Gu, G. Qi, M. Wang, J. Tang, X. Yuan, Z. Lu, and J. Liu, "MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2007.

[152] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *Computer Vision and Image Understanding*, vol. 94, nos. 1–3, pp. 3–27, 2004.

[153] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, pp. 121–143, 2008.

[154] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1632–1646, 2008.

[155] F. Nack and W. Putz, "Saying what it means: Semi-automated (news) media annotation," *Multimedia Tools and Applications*, vol. 22, pp. 263–302, 2004.

[156] M. R. Naphade, "On supervision and statistical learning for semantic multimedia analysis," *Journal of Visual Communication and Image Representation*, vol. 15, pp. 348–369, 2004.

[157] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering and retrieval," *IEEE Transactions on Multimedia*, vol. 3, pp. 141–151, 2001.

[158] M. R. Naphade and T. S. Huang, "Extracting semantics from audiovisual content: The final frontier in multimedia retrieval," *IEEE Transactions on Neural Networks*, vol. 13, pp. 793–810, 2002.

[159] M. R. Naphade, L. S. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. G. Hauptmann, "A light scale concept ontology for multimedia understanding for TRECVID 2005," Technical Report RC23612, IBM T. J. Watson Research Center, 2005.

[160] M. R. Naphade, I. V. Kozintsev, and T. S. Huang, "A factor graph framework for semantic video indexing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 40–52, 2002.

[161] M. R. Naphade, A. P. Natsev, C.-Y. Lin, and J. R. Smith, "Multi-granular detection of regional semantic concepts," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 109–112, Taipei, Taiwan, 2004.

[162] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proceedings of the ACM International Conference on Multimedia*, New York, USA, 2004.

[163] M. R. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. S. Kennedy, A. G. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, pp. 86–91, 2006.

[164] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 991–1000, Augsburg, Germany, 2007.

[165] A. P. Natsev, M. R. Naphade, and J. Tešić, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proceedings of the ACM International Conference on Multimedia*, pp. 598–607, Singapore, 2005.

[166] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *CIVR*, (H. Sundaram et al., eds.), pp. 143–152, Heidelberg, Germany: Springer-Verlag, 2006.

[167] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin, "ETISEO, performance evaluation for video surveillance systems," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 476–481, London, UK, 2007.

[168] G. P. Nguyen and M. Worring, "Interactive access to large image collections using similarity-based visualization," *Journal of Visual Languages and Computing*, vol. 19, pp. 203–224, 2008.

[169] G. P. Nguyen, M. Worring, and A. W. M. Smeulders, "Interactive search by direct manipulation of dissimilarity space," *IEEE Transactions on Multimedia*, vol. 9, pp. 1404–1415, 2007.

[170] H. T. Nguyen, M. Worring, and A. Dev, "Detection of moving objects in video using a robust motion similarity measure," *IEEE Transactions on Image Processing*, vol. 9, pp. 137–141, 2000.

[171] NIST, "TRECVID video retrieval evaluation — Online proceedings," http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html, 2001–2008.

[172] T. O'Reily, "What is web 2.0," http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web%-20.html, 2005.

[173] P. Over, G. Awad, T. Rose, J. Fiscus, W. Kraaij, and A. F. Smeaton, "TRECVID 2008 — Goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[174] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton, "TRECVID 2005 An Overview," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2005.

[175] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton, "TRECVID 2006 an overview," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[176] P. Over, A. F. Smeaton, and P. Kelly, "The TRECVID 2007 BBC rushes summarization evaluation pilot," in *Proceedings of the International Workshop on TRECVID Video Summarization*, pp. 1–15, 2007.

[177] S. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, USA: The MIT Press, 1999.

[178] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proceedings of the ACM International Conference on Multimedia*, pp. 65–74, Boston, USA, 1996.

[179] Z. Pecenovic, M. Do, M. Vetterli, and P. Pu, "Integrated browsing and searching of large image collections," in *Proceedings of the International Conference on Advances in Visual Information Systems*, Lyon, France, 2000.

[180] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, pp. 233–254, 1996.

[181] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2004.

[182] E. G. M. Petrakis, A. Diplaros, and E. Milios, "Matching and retrieval of distorted and occluded shapes using dynamic programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1501–1516, 2002.

[183] K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, Y. Kompatsiaris, and S. Staab, "Knowledge representation and semantic annotation of multimedia content," *IEE Proceedings of Vision, Image and Signal Processing*, vol. 153, pp. 255–262, 2006.

[184] J. C. Platt, "Probabilities for SV machines," in *Advances in Large Margin Classifiers*, (A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, eds.), pp. 61–74, Cambridge, USA: The MIT Press, 2000.

[185] E. Pogalin, A. W. M. Smeulders, and A. H. C. Thean, "Visual Quasi-Periodicity," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[186] G. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang, "Correlative multilabel video annotation with temporal kernels," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 5, 2009.

[187] G. M. Quénot, D. Moraru, L. Besacier, and P. Mulhem, "CLIPS at TREC-11: Experiments in video retrieval," in *Proceedings of the 11th Text Retrieval Conference*, (E. M. Voorhees and L. P. Buckland, eds.), Gaithersburg, USA, 2002.

[188] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[189] T. Randen and J. H. Husøy, "Filtering for texture classification: A comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 291–310, 1999.

[190] N. Rasiwasia, P. L. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, pp. 923–938, 2007.

[191] M. Rautiainen, T. Ojala, and T. Seppanen, "Cluster-temporal browsing of large news video databases," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.

[192] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The AMI and AMIDA projects," in *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, pp. 115–118, Trento, Italy, 2008.

[193] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *International Joint Conference on Artificial Intelligence*, pp. 448–453, Montréal, Canada, 1995.

[194] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4," in *Proceedings of the Text Retrieval Conference*, pp. 73–96, Gaithersburg, USA, 1996.

[195] A. Rosenfeld, "Picture processing by computer," *ACM Computing Surveys*, vol. 1, pp. 147–176, 1969.

[196] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[197] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 644–655, 1998.

[198] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[199] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, USA: McGraw-Hill, 1983.

[200] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," *IEEE MultiMedia*, vol. 6, pp. 22–35, 1999.

[201] A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga, "Ontology-based photo annotation," *IEEE Intelligent Systems*, vol. 16, pp. 66–74, 2001.

[202] N. Sebe and M. S. Lew, "Texture features for content-based retrieval," in *Principles of Visual Information Retrieval*, (M. S. Lew, ed.), pp. 51–86, Springer, 2001.

[203] F. J. Seinstra, J.-M. Geusebroek, D. Koelma, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "High-performance distributed image and video content analysis with parallel-horus," *IEEE MultiMedia*, vol. 14, pp. 64–75, 2007.

[204] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu, "Watch what I watch: Using community activity to understand content," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 275–284, Augsburg, Germany, 2007.

[205] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of the IEEE Computer

Society Conference on Computer Vision and Pattern Recognition, pp. 1–8, Anchorage, USA, 2008.

[206] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *International Journal of Computer Vision*, vol. 67, pp. 189–210, 2006.

[207] J. Sivic and A. Zisserman, "Efficient visual search for objects in videos," *Proceedings of the IEEE*, vol. 96, pp. 548–566, 2008.

[208] A. F. Smeaton, "Large scale evaluations of multimedia information retrieval: The TRECVid experience," in *CIVR*, pp. 19–27, Springer-Verlag, 2005.

[209] A. F. Smeaton, "Techniques used and open challenges to the analysis, indexing and retrieval of digital video," *Information Systems*, vol. 32, pp. 545–559, 2007.

[210] A. F. Smeaton, C. Foley, D. Byrne, and G. J. F. Jones, "iBingo mobile collaborative search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 547–548, Niagara Falls, Canada, 2008.

[211] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 321–330, 2006.

[212] A. F. Smeaton, P. Over, and W. Kraaij, "High level feature detection from video in TRECVid: A 5-year retrospective of achievements," in *Multimedia Content Analysis, Theory and Applications*, (A. Divakaran, ed.), Springer, 2008.

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

[214] J. R. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE MultiMedia*, vol. 4, pp. 12–20, 1997.

[215] J. R. Smith, M. R. Naphade, and A. P. Natsev, "Multimedia semantic indexing using model vectors," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 445–448, Baltimore, USA, 2003.

[216] J. R. Smith, A. P. Natsev, J. Tešić, L. Xie, and R. Yan, "IBM multimedia analysis and retrieval system," http://www.alphaworks.ibm.com/ tech/imars/, 2008.

[217] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Transactions on Multimedia*, vol. 9, pp. 975–986, 2007.

[218] C. G. M. Snoek, J. C. van Gemert, J.-M. Geusebroek, B. Huurnink, D. C. Koelma, G. P. Nguyen, O. de Rooij, F. J. Seinstra, A. W. M. Smeulders, C. J. Veenman, and M. Worring, "The MediaMill TRECVID 2005 semantic video search engine," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2005.

[219] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring, "The MediaMill TRECVID 2006 semantic video search engine," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[220] C. G. M. Snoek and M. Worring, "Multimedia event-based video indexing using time intervals," *IEEE Transactions on Multimedia*, vol. 7, pp. 638–647, 2005.

[221] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, pp. 5–35, 2005.
[222] C. G. M. Snoek, M. Worring, O. de Rooij, K. E. A. van de Sande, R. Yan, and A. G. Hauptmann, "VideOlympics: Real-time evaluation of multimedia retrieval systems," *IEEE MultiMedia*, vol. 15, pp. 86–91, 2008.
[223] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, and F. J. Seinstra, "On the surplus value of semantic video analysis beyond the key frame," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Amsterdam, The Netherlands, 2005.
[224] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1678–1689, 2006.
[225] C. G. M. Snoek, M. Worring, and A. G. Hauptmann, "Detection of TV news monologues by style analysis," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.
[226] C. G. M. Snoek, M. Worring, and A. G. Hauptmann, "Learning rich semantics from news video archives by style analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, pp. 91–108, 2006.
[227] C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Transactions on Multimedia*, vol. 9, pp. 280–292, 2007.
[228] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the ACM International Conference on Multimedia*, pp. 399–402, Singapore, 2005.
[229] C. G. M. Snoek, M. Worring, A. W. M. Smeulders, and B. Freiburg, "The role of visual content and style for concert video indexing," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 252–255, Beijing, China, 2007.
[230] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proceedings of the ACM International Conference on Multimedia*, pp. 421–430, Santa Barbara, USA, 2006.
[231] C. G. M. Snoek et al., "The MediaMill TRECVID 2008 semantic video search engine," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.
[232] M. J. Swain and D. H. Ballard, "Color Indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.
[233] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *IEEE International Workshop on Content-based Access of Image and Video Databases, in Conjunction with ICCV'98*, Bombay, India, 1998.
[234] J. Tague-Sutcliffe, "The pragmatics of information retrieval experimentation, revisited," *Information Processing & Management*, vol. 28, pp. 467–490, 1992.
[235] S. Tang et al., "TRECVID 2008 high-level feature extraction By MCG-ICT-CAS," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.
[236] C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. A. Bouman, and E. J. Delp, "ViBE: A compressed video database structured for active browsing

and search," *IEEE Transactions on Multimedia*, vol. 6, pp. 103–118, 2004.

[237] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki, "Structured video computing," *IEEE MultiMedia*, vol. 1, pp. 34–43, 1994.

[238] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1958–1970, 2008.

[239] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, 2007.

[240] B. L. Tseng, C.-Y. Lin, M. R. Naphade, A. P. Natsev, and J. R. Smith, "Normalized classifier fusion for semantic visual concept detection," in *Proceedings of the IEEE International Conference on Image Processing*, pp. 535–538, Barcelona, Spain, 2003.

[241] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, pp. 177–280, 2008.

[242] J. Urban, X. Hilaire, F. Hopfgartner, R. Villa, J. M. Jose, S. Chantamunee, and Y. Gotoh, "Glasgow University at TRECVID 2006," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2006.

[243] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H.-J. Zhang, "Image classification for content-based indexing," *IEEE Transactions on Image Processing*, vol. 10, pp. 117–130, 2001.

[244] A. Vailaya, A. K. Jain, and H.-J. Zhang, "On image classification: City images vs. landscapes," *Pattern Recognition*, vol. 31, pp. 1921–1936, 1998.

[245] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluation of color descriptors for object and scene recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008.

[246] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *European Conference on Computer Vision*, Marseille, France, 2008.

[247] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, "Robust scene categorization by learning image statistics in context," in *International Workshop on Semantic Learning Applications in Multimedia, in Conjunction with CVPR'06*, New York, USA, 2006.

[248] J. C. van Gemert, C. G. M. Snoek, C. Veenman, and A. W. M. Smeulders, "The influence of cross-validation on video classification performance," in *Proceedings of the ACM International Conference on Multimedia*, pp. 695–698, Santa Barbara, USA, 2006.

[249] V. N. Vapnik, *The Nature of Statistical Learning Theory.* New York, USA: Springer-Verlag, 2nd ed., 2000.

[250] R. C. Veltkamp and M. Hagedoorn, "State-of-the-art in shape matching," in *Principles of Visual Information Retrieval*, (M. S. Lew, ed.), pp. 87–119, Springer, 2001.

[251] T. Volkmer, J. R. Smith, A. P. Natsev, M. Campbell, and M. R. Naphade, "A web-based system for collaborative annotation of large image and video

collections," in *Proceedings of the ACM International Conference on Multimedia*, pp. 892–901, Singapore, 2005.

[252] T. Volkmer, J. A. Thom, and S. M. M. Tahaghoghi, "Modelling human judgement of digital imagery for multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 9, pp. 967–974, 2007.

[253] L. von Ahn, "Games with a purpose," *IEEE Computer*, vol. 39, pp. 92–94, 2006.

[254] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, USA: The MIT Press, 2005.

[255] H. D. Wactlar, M. G. Christel, Y. Gong, and A. G. Hauptmann, "Lessons learned from building a terabyte digital video library," *IEEE Computer*, vol. 32, pp. 66–73, 1999.

[256] D. Wang, X. Li, J. Li, and B. Zhang, "The importance of query-concept-mapping for automatic video retrieval," in *Proceedings of the ACM International Conference on Multimedia*, pp. 285–288, Augsburg, Germany, 2007.

[257] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang, "Video diver: Generic video indexing with diverse features," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 61–70, Augsburg, Germany, 2007.

[258] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 947–963, 2001.

[259] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proceedings of the ACM International Conference on Multimedia*, pp. 862–871, Augsburg, Germany, 2007.

[260] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "AnnoSearch: Image auto-annotation by search," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1483–1490, New York, USA, 2006.

[261] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, pp. 12–36, 2000.

[262] X.-Y. Wei, C.-W. Ngo, and Y.-G. Jiang, "Selection of concept detectors for video search by ontology-enriched semantic spaces," *IEEE Transactions on Multimedia*, vol. 10, pp. 1085–1096, 2008.

[263] D. Weinberger, *Everything is Miscellaneous*. New York, USA: Times Books, 2007.

[264] M.-F. Weng and Y.-Y. Chuang, "Multi-cue fusion for semantic video indexing," in *Proceedings of the ACM International Conference on Multimedia*, pp. 71–80, Vancouver, Canada, 2008.

[265] T. Westerveld, "Using generative probabilistic models for multimedia retrieval," PhD thesis, University of Twente, 2004.

[266] T. Westerveld and A. P. de Vries, "Multimedia retrieval using multiple images," in *CIVR*, pp. 344–352, Springer-Verlag, 2004.

[267] P. Wilkins, T. Adamek, N. E. O'Connor, and A. F. Smeaton, "Inexpensive fusion methods for enhancing feature detection," *Image Communication*, vol. 7–8, pp. 635–650, 2007.

[268] P. Wilkins, A. F. Smeaton, N. E. O'Connor, and D. Byrne, "K-Space interactive search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 555–556, Niagara Falls, Canada, 2008.

[269] P. Wilkins et al., "K-Space at TRECVid 2008," in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2008.

[270] M. Worring and G. Schreiber, "Semantic image and video indexing in broad domains," *IEEE Transactions on Multimedia*, vol. 9, pp. 909–911, 2007.

[271] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Novelty and redundancy detection with multimodalities in cross-lingual broadcast domain," *Computer Vision and Image Understanding*, vol. 110, pp. 418–431, 2008.

[272] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proceedings of the ACM International Conference on Multimedia*, pp. 572–579, New York, USA, 2004.

[273] Y. Wu, B. L. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.

[274] L. Xie and S.-F. Chang, "Pattern mining in visual concept streams," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 297–300, Toronto, Canada, 2006.

[275] L. Xie, H. Sundaram, and M. Campbell, "Event mining in multimedia streams," *Proceedings of the IEEE*, vol. 96, pp. 623–647, 2008.

[276] X. Xie, L. Lu, M. Jia, H. Li, F. Seide, and W.-Y. Ma, "Mobile search with multimodal queries," *Proceedings of the IEEE*, vol. 96, pp. 589–601, 2008.

[277] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Transactions on Multimedia*, vol. 10, pp. 421–436, 2008.

[278] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Transactions on Multimedia*, vol. 10, pp. 1342–1355, 2008.

[279] H. Xu and T.-S. Chua, "Fusion of AV features and external information sources for event detection in team sports video," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, pp. 44–67, 2006.

[280] R. Yan, M.-Y. Chen, and A. G. Hauptmann, "Mining relationship between video concepts using probabilistic graphical models," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, pp. 301–304, Toronto, Canada, 2006.

[281] R. Yan and A. G. Hauptmann, "The combination limit in multimedia retrieval," in *Proceedings of the ACM International Conference on Multimedia*, Berkeley, USA, 2003.

[282] R. Yan and A. G. Hauptmann, "Probabilistic latent query analysis for combining multiple retrieval sources," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 324–331, Seattle, USA, 2006.

[283] R. Yan and A. G. Hauptmann, “A review of text and image retrieval approaches for broadcast news video,” *Information Retrieval*, vol. 10, nos. 4–5, pp. 445–484, 2007.

[284] R. Yan, A. G. Hauptmann, and R. Jin, “Negative pseudo-relevance feedback in content-based video retrieval,” in *Proceedings of the ACM International Conference on Multimedia*, Berkeley, USA, 2003.

[285] R. Yan, J. Yang, and A. G. Hauptmann, “Learning query-class dependent weights for automatic video retrieval,” in *Proceedings of the ACM International Conference on Multimedia*, New York, USA, 2004.

[286] A. Yanagawa, S.-F. Chang, L. S. Kennedy, and W. Hsu, “Columbia University’s baseline detectors for 374 LSCOM semantic visual concepts,” Technical Report 222-2006-8, Columbia University ADVENT Technical Report, 2007.

[287] J. Yang, M.-Y. Chen, and A. G. Hauptmann, “Finding person X: Correlating names with visual appearances,” in *CIVR*, pp. 270–278, Springer-Verlag, 2004.

[288] J. Yang and A. G. Hauptmann, “(Un)Reliability of video concept detection,” in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 85–94, Niagara Falls, Canada, 2008.

[289] J. Yang, R. Yan, and A. G. Hauptmann, “Cross-domain video concept detection using adaptive SVMs,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 188–197, Augsburg, Germany, 2007.

[290] E. Yilmaz and J. A. Aslam, “Estimating average precision when judgments are incomplete,” *Knowledge and Information Systems*, vol. 16, pp. 173–211, 2008.

[291] J. Yuan, H. Wang, L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang, “Tsinghua University at TRECVID 2005,” in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2005.

[292] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, “A formal study of shot boundary detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 168–186, 2007.

[293] J. Yuan et al., “THU and ICRC at TRECVID 2007,” in *Proceedings of the TRECVID Workshop*, Gaithersburg, USA, 2007.

[294] E. Zavesky and S.-F. Chang, “CuZero: Embracing the frontier of interactive visual search for informed users,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 237–244, Vancouver, Canada, 2008.

[295] E. Zavesky, S.-F. Chang, and C.-C. Yang, “Visual islands: Intuitive browsing of visual search results,” in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 617–626, Niagara Falls, Canada, 2008.

[296] Z. J. Zha, T. Mei, Z. Wang, and X.-S. Hua, “Building a comprehensive ontology to refine video concept detection,” in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 227–236, Augsburg, Germany, 2007.

[297] H.-J. Zhang, A. Kankanhalli, and S. W. Smoliar, “Automatic partitioning of full-motion video,” *Multimedia Systems*, vol. 1, pp. 10–28, 1993.

[298] H.-J. Zhang, S. Y. Tan, S. W. Smoliar, and Y. Gong, "Automatic parsing and indexing of news video," *Multimedia Systems*, vol. 2, pp. 256–266, 1995.

[299] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, pp. 213–238, 2007.

[300] R. Zhang, R. Sarukkai, J.-H. Chow, W. Dai, and Z. Zhang, "Joint categorization of queries and clips for web-based video search," in *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 193–202, Santa Barbara, USA, 2006.

[301] W.-L. Zhao and C.-W. Ngo, "LIP-VIREO: Local interest point extraction toolkit," Software available at http://www.cs.cityu.edu.hk/∼wzhao2/lip-vireo.htm, 2008.

[302] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, pp. 536–544, 2003.

[303] X. Zhu, "Semi-supervised learning literature survey," Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.