International Conference on Image and Video Retrieval,
National University of Singapore, Singapore, July 20-22, 2005.

Lecture Notes in Computer Science, Volume 3568, August 2005, pp. 1-10.

Lessons for the Future from a Decade of Informedia Video Analysis Research

Alexander G. Hauptmann

School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA 15213
hauptmann@cs.cmu.edu

Abstract. The overarching goal of the Informedia Digital Video Library project has been to achieve machine understanding of video media, including all aspects of search, retrieval, visualization and summarization in both contemporaneous and archival content collections. The base technology developed by the Informedia project combines speech, image and natural language understanding to automatically transcribe, segment and index broadcast video for intelligent search and image retrieval. While speech processing has been the most influential component in the success of the Informedia project, other modalities can be critical in various situations. Evaluations done in the context of the TRECVID benchmarks show that while some progress has been made, there is still a lot of work ahead. The fundamental "semantic gap" still exists, but there are a number of promising approaches to bridging it.

1 A Brief History of the Informedia Digital Library Project

Vast amounts of video have been archived, and more is produced daily, yet remains untapped as an archival information source for on-demand access because of the difficulty and tedium involved in processing, organizing, filtering, and presenting huge quantities of multimedia information. The overarching goal of the Informedia initiatives has been to achieve machine understanding of video, including all aspects of search, retrieval, visualization and summarization in both contemporaneous and archival content collections.

For the last ten years, Informedia has focused on information extraction from broadcast television news and documentary content. Multiple terabytes of video have been collected, with automatically generated metadata and indices for retrieving videos from this library continuously available online to local users. The base technology developed by the Informedia project combines speech, image and natural language understanding to automatically transcribe, segment and index broadcast video for intelligent search and image retrieval.

Initially funded with a small seed grant from the Heinz foundation, the Informedia Digital Video Library was one of six projects funded by the first NSF Digital Library Initiative in 1994. At the time, CMU uniquely boasted state of the art technology in speech, image and language technologies, so applying them all to the video analysis

W.-K. Leow et al. (Eds.): CIVR 2005, LNCS 3568, pp. 1–10, 2005.

of digital video libraries seemed a natural fit. In the 90's, it was clear that multimedia would soon be available on all personal computers. At that time, the promise of combining speech recognition technology, image understanding and language processing seemed to open boundless opportunities in film and video production and archives, education, sports, and home entertainment.

An early demonstration system and video made with manually transcribed, synchronized and indexed data proved to be very convincing. Many aspects of the current Informedia system were already included: text transcripts, visual summaries, titles and a free text search interface. It took several years for reality to catch up with this target demonstration. Eventually, the effectiveness of the concept was demonstrated with the "News-on-Demand" application, which automatically processed broadcast news shows for the Informedia archive. The second phase of the Digital Libraries Initiative provided the project with the opportunity to extend single video abstractions to summarizing multiple documents, in different collections and visualizing very large video data sets. Over the years, follow-on projects extended this to multi-lingual broadcast news. Following a different line of research we established cross-cultural video archive collaborations in China and Europe, as well as specialized cultural history archives in the U.S. An early focus on education, which prompted us to install a version of the

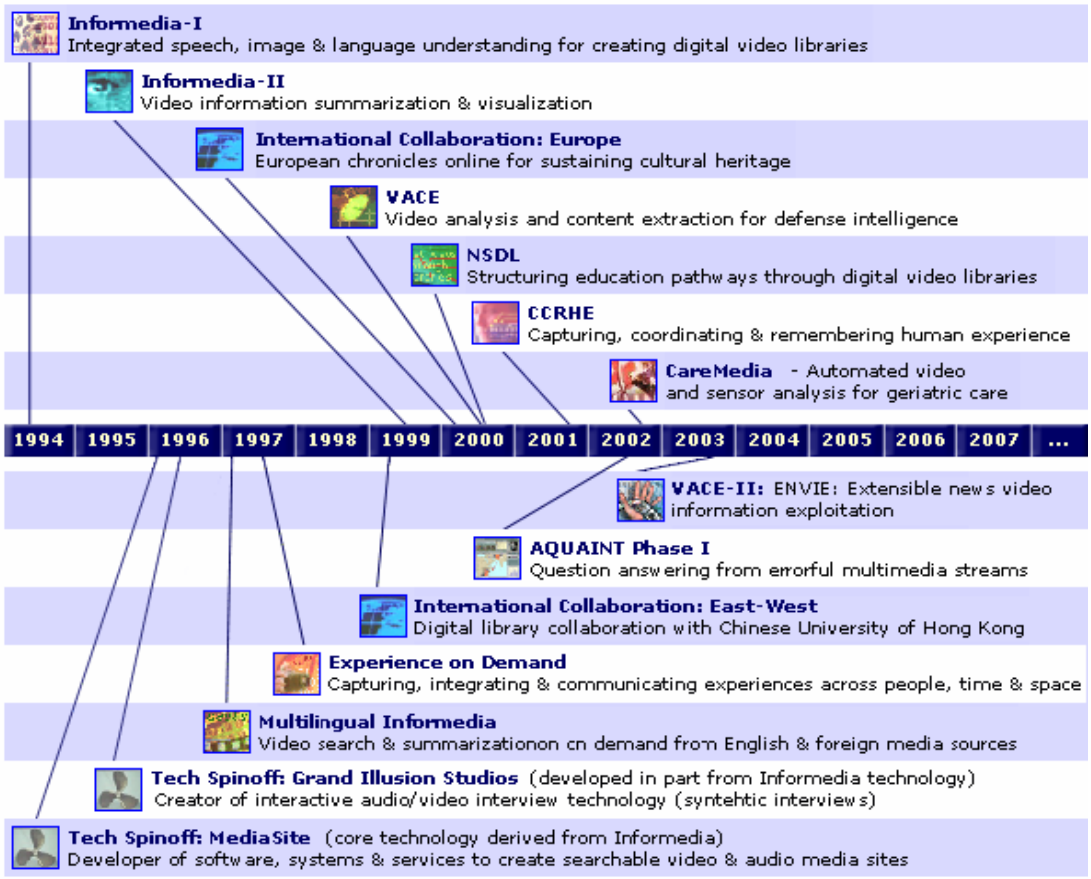

Fig. 1. A graphic timeline of major Informedia project efforts

system in the library of a Pittsburgh K-12 school, has continued to persist throughout the project. The biggest obstacles to adoption of video archive exploration in schools appear to be the smooth integration of the technology into the curriculum schedule and ongoing classroom practices, without increasing the teachers work load.

Several Informedia spin-off research projects explored the analysis of video and sensor data captured with wearable cameras. The promise of capturing your whole life was tantalizing, but the difficulties in getting useful video data out of an unsupervised, continuously moving and recording camera proved formidable enough, that despite progress in a number of areas, the overall vision of complete "digital human memory" in video archive form was not realized.

As was typical during the Internet boom, a couple of technology companies were created, which initially did well, but all floundered during the ensuing bust.

More recently, in the aftermath of the September 11, 2001 terrorist attacks, the ability to analyze open source video broadcasts from foreign countries has sparked great interest by government organizations, especially the intelligence analysis community. The problems faced by government analysts are very diverse, ranging from "footage of anything military related" or "anything about this person" to "scenes in this neighborhood" and "detailed shots of this event". Research efforts are also currently underway to apply Informedia technology in the domain of health care, where we believe video observations and archives can have a large societal, economic and scientific impact.

2 Lessons Learned

It is difficult to sum up the literally hundreds of research papers generated by the Informedia project over the last decade. Instead, I will try and give my impressions of the most significant insights that have enabled success and provided a basis for understanding video and accessing large video archives.

- **Speech recognition and audio analysis.** Speech analysis has perhaps provided the clearest benefits for accessing information in a video archive. From the very beginning, we had a clear connection: automatic speech recognition could transform the audio stream into words, and then it is well known how to index text words for information retrieval. The challenges of speech recognition relate to the recognizer accuracy (or word error rate) in different conditions. Currently, the best recognizers trained for broadcast news have a word error rate of about 15% on studio recorded anchor speech. The error rate is higher for other studio speakers, increases further for reporters in the field and remains fairly high for news subjects interviewed outdoors. Foreign accents or emotional factors such as crying during the interview further degrade the performance. Commercial advertisements have music mixed with singing, dramatic speech and specific product names, which gives them very high error rates. In evaluations of spoken broadcast news retrieval, it was not a coincidence that the best performing systems simply identified the commercials and eliminated them completely, rather than trying to recognize the contents. The big problem with speech recognition is the lack of robustness across

domains. Many speech systems can be trained to work well on specific collections, but this does not transfer to other types of data. In general, as long the recordings were done professionally, and the spoken audio is well separated from music and other environmental noises, good speech recognition with a word error rate less than 40% is possible, sometimes with specialized re-training of acoustic models.

The currently standard speech recognition vocabularies of 64,000 words also appear sufficient to cover more than 99% of the English vocabulary in most broadcast news cases. Special vocabularies and pronunciations can usually be added if domains require it. Subword matching (i.e. trying to find a word that was not in the lexicon based on its sequence of phonemes) is an option that frequently leads to worse performance than full word recognition despite missing words, and should only be used during retrieval in specific circumstances. Similarly, language models, which specify the transition probabilities between words, can be easily adapted to many domains using sample texts. In addition to creating text, speech recognition allows alignments of recognized words to existing transcripts or closed captions [1]. This enables time-accurate search down to the individual word.

Beyond speech recognition, audio analysis can be useful for speaker identification, segmentation, and for computational auditory scene analysis. While we have evidence that these audio analysis techniques can contribute to the effectiveness of the video retrieval system [2], they have remained fairly error-prone in our implementations. As a result, their value to effective retrieval tends to be small, in the form of modules that contribute additional useful data, but not critical to overall success.

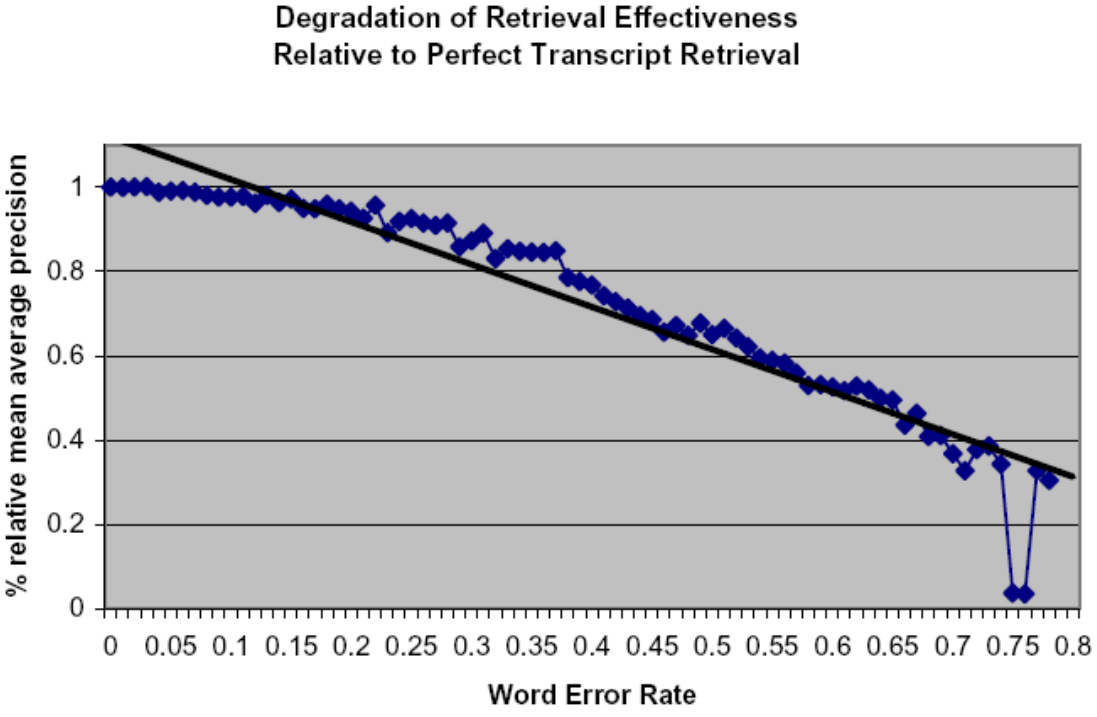

Fig. 2. Degradation of retrieval effectiveness as a function of word error rate in a basic SDR system without query or document expansion. The linear regression trendline shows that degradation is less than expected at lower error rates

- **Information Retrieval.** One of the fundamental lessons we learned from information retrieval experiments was that even relatively high word error rates in the speech recognition nevertheless permit relatively effective information retrieval [3]. The graph in figure 2.shows that at high recognition accuracy, retrieval from speech is as effective as retrieval from perfect text transcripts. Only at lower recognizer accuracy (> 40% word error rate) does retrieval drop off substantially. This was a fundamental premise of the Informedia project, and it was reassuring to see the empirical research results validate the initial conjecture. Using standard information retrieval techniques, the current state of the art in speech recognition was adequate for information retrieval from broadcast news type video. Additional information retrieval techniques, such as relevance feedback and document expansion, specialized for application to speech transcripts, can further improve the retrieval results by a few percent [19].
- **Linguistic Analysis.** The Informedia project regularly uses language analysis for labeling news stories with topics [4]. Our archives initially assigned over a thousand topic labels with high accuracy, but over time we found the training data became historically outdated and accuracy was greatly reduced. The British Princess Diana no longer figures as prominently in today's news as she did in 1997 or 1998, yet the trained topic classifier relies still give stories about her a high (prior) probability. Most importantly, linguistic analysis plays a critical role in identifying named entities (people, places, organizations), which can be used to summarize events, compile people profiles, identify faces, and create map displays of geographic information. From almost the very beginning of the Informedia project, we derived great benefit from automatically creating headlines summarizing the key phrases in a news story [1]. Linguistic cues also provide information about segmentation, especially in broadcast news [5] and query classification [6].
- **Image Processing.** The first successful application of image processing in Informedia was the detection of shot boundaries, which enabled keyframe selection extracted from the shots in a video segment for a rich overview display. This was followed by face detection [7, 8], which is possibly still the most useful image processing result for the project. Video OCR proved helpful in some retrieval applications [9]. We implemented several different types of image similarity retrieval [10, 11], but we were generally disappointed by the results. The diversity of the imagery in the collection was so large, that only virtually identical images could be found, while all other "nearby" images always contained irrelevant material. Our experiments with image segmentation [12] gave few useful results on broadcast news. More detailed image analysis for keyframe selection and skims [13] also did not prove to be of great benefit.
- **Interfaces and Integration.** Probably the biggest reason for the success of the Informedia project can be attributed to the quality of the interface. In the course of the project, much research effort was devoted to the automatic creation of multimedia visualizations and abstractions [14]. Especially when combined with empirical proofs of their effectiveness [15], we were able to improve the interface to allow users to access data in many different ways, tailoring presentations based on context [17]. Depending on the specific user task, either collapsed temporal presentations in the form of "video skims" [13], collages, or storyboards with

semantic class filters [15] may be appropriate as the most efficient way for the user to browse and drill down into the data.

Beyond traditional timelines, geographic visualizations provided one dramatic breakthrough in the presentation of large result sets for a video archive. The map displays can be dynamically generated from extracted metadata [18] based on locations in named entities and a gazetteer that maps these entities into the map (or latitude/longitude) locations. The maps are both active, in that they could highlight the locations mentioned in the transcript or recognized in through VOCR, and interactive, in the sense that a user can select and area or country to filter results from a larger set.

- **Integration.** The Informedia systems draws its lifeblood from integration of all modalities, integrating in different ways: Named faces combine text or VOCR analysis with face identification, manual metadata created externally is merged with automatically extracted information, multimedia abstractions allow users to see text, keyframes and metadata in flexible ways, as well as integration of modalities for improved retrieval, where prompted by the TRECVID semantic feature classification tasks, the utility of a few *reliable* semantic features in broadcast news, mainly anchors, sports and weather news has shown itself to be useful for integrated retrieval. While text often provides strong clues, many semantic classifications rely on color, face and features as the most robust and reliable low-level features for automatic classification [22].

One major benefit of the Informedia project, rarely credited in research publications, was derived from an infrastructure that allows daily processing without any manual intervention. This has forced us to develop a robust toolkit of components. Daily processing also underscores many issues that are easy to ignore when publishing a single research paper claiming success with one evaluation. During routine processing, it quickly becomes clear which components break easily, which are too computationally expensive and which have been overtrained on a particular data set, resulting in unacceptably low accuracy during general use. Advances in computer speed and storage costs have helped make processing affordable, we now no longer have to devise algorithms that "forget" unneeded videos to save room for incoming data.

3 Evaluations and TRECVID

A number of the Informedia projects successes have been motivated or refined by the NIST TREC and later TRECVID evaluations. In the early phases or the project, the TREC Spoken Document Retrieval track demonstrated that utility of combining speech transcription with information retrieval [19]. There can be a wide difference in recognition rates for anchor speech and others, but fortunately, the news anchor usually introduces a news story, using many keywords relevant for retrieval, So if the anchor speech is recognized well, it becomes easy to find the relevant story.

TRECVID [23] encourages research in information retrieval specifically from digital video by providing a large video test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. TRECVID benchmarking covers both interactive and manual searching by end users, as well as the bench-

marking of some supporting technologies including shot boundary detection, extraction of some semantic features, and the automatic segmentation of TV news broadcasts into non-overlapping news stories. Evaluations done in the context of the TRECVID benchmarks show that while some progress has been made, there is much work ahead before video search rivals text search in accuracy. Generally, we have found that speech transcripts provide the single most important clue for successful retrieval. However, automatically finding the individual shots and images is still basically an unsolved challenge. It has been disappointing for us to repeatedly find that none of the multimedia analysis and retrieval techniques provide a significant benefit over retrieval using only textual information such as ASR transcripts or closed captions. This is consistent with findings in the earlier TRECVID evaluations in 2001 and 2002, where the best systems were based exclusively on retrieval using automatic speech recognition. However, we should also point out that it is not the case that "nothing works" here. In interactive systems, we do find significant differences among the top systems, indicating that interfaces can make a huge difference for effective video search. For interactive tasks we have developed efficient interfaces that require few key clicks, but display large numbers of keyframes for visual inspection by the user. The text search finds the right context in general, but to select specific relevant shots we need good interfaces to easily browse the storyboard keyframes.

In general, TRECVID has motivated us to be honest about what we don't know how to do well (sometimes through painful failures), and has focused us to work on substantial improvements to the actual task of video retrieval, as opposed to flashy demos based on technological capabilities.

4 Current Opportunities and Roadblocks

Intellectual property concerns has inhibited and will continue to inhibit the growth of centralized digital video libraries. While the Informedia library contains some public domain video, the majority of the contributions from CNN, the British Open University, WQED Communications and other sources were restricted for access only by local users and could not be published to the web. Often, the content providing organization would have liked to agree to broader access, but was not sure how to retroactively classify and pass along these rights.

The Informedia project has attempted to field general purpose solutions, serving a broad class of users accessing wide-ranging video data. In retrospect, this approach may be more limiting rather than liberating. Many processing techniques, such as video skim creation, work best if heuristics can be applied based on the subclass of the particular video. On the other extreme, many other research groups have shown that special case applications can be made to work well if good researchers with clever solutions approach them. In particular, the last few years of CIVR and ACM Multimedia conferences have seen a plethora of multimedia analysis on sports broadcasts and other specialized domain applications.

Over time, we have also been amazed by the speed at which components decay. Speech recognition vocabularies need to be updated regularly to reflect current lan-

guage use, topics beyond a core set of a few dozen are time dependent, broadcasters will change their formats thus affecting carefully tuned video OCR algorithms, story segmentation and even shot detection. Even the countries in the gazetteer have changed over time, Yugoslavia is no longer the country it was a decade ago.

Image and video imagery processing remains the biggest unsatisfied promise of the original Informedia Digital Video Library vision. We have found that most research from computer vision has not been robust enough to be usable. The general problem of automatically characterizing all objects and their interrelationships in video scenes remains our most challenging research issue [20].

In many ways our research has only just begun. So far, we have harvested a number of the low-hanging fruit. In retrospect, perhaps we have only done the obvious things until this point. Now the challenge is to transform a collection of clever and obvious tricks into a serious body of science applicable to large-scale video analysis and retrieval. The fundamental "semantic gap" still exists, and there are a number of promising approaches to bridging it:

1) It has been argued that we should give up on the idea of automatic video analysis, and instead allow millions of internet users to annotate video and images, perhaps within the framework of the semantic web.

2) The computer vision community is still focused on solving the harder problems of complete understanding of images and scenes at a fairly detailed level of granularity. To the extent this community can make progress and find sufficient solutions that scale to the diversity and volume of video archives, any success here will directly transfer to improved video retrieval.

3) The machine learning community is building increasingly sophisticated models for learning the relationship between low-level feature vectors and the content represented in the video or image. Their approach is that with enough annotated training data, sophisticated learning approaches will converge on the right models needed to understand video or image collections.

4) My currently favorite approach is to give up on general, deep understanding of video – that problem is just too hard for now. Instead we should focus on reliable detection of semantic concepts, perhaps a few thousand of them [21]. These concepts can be combined into a taxonomy, perhaps even an ontology that could be used in video retrieval. These concepts would represent a set of intermediate (textual) descriptors that can be reliably applied to visual scenes. Many researchers have been developing automatic feature classifiers like face, people, sky, grass, plane, outdoors, soccer goals, and buildings [22], showing that these classifiers could, with enough training data, reach the level of maturity needed to be effective filters for video retrieval.

Of course, this splits the semantic gap between low-level features and user information needs into two, hopefully smaller gaps: (a) mapping the low-level features into the intermediate semantic concepts and (b) mapping these concepts into user needs. I believe this divide-and-conquer approach using semantic concepts as an intermediate layer will allow us to develop thousands of concepts that can be reliably identified in many contexts, and with sufficient numbers of these concepts available, covering a broad spectrum of visible things, users will finally be able to bridge the semantic gap.

Acknowledgements

This work was supported by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037. The paper benefited tremendously from discussions at the March 2005 Dagstuhl Seminar on the future of multimedia and I would like to particularly thank Marcel Worring, Marc Davis, Lloyd Rutledge, Mubarak Shah, Tat-Seng Chua and many other participants.

References

1. Hauptmann, A.G., Witbrock, M.J. and Christel, M.G. Artificial Intelligence Techniques in the Interface to a Digital Video Library, Extended Abstracts of the ACM CHI 97 Conference on Human Factors in Computing Systems, (New Orleans LA, March 1997), 2-3.
2. Christel, M., Smith, M., Taylor, C.R., and Winkler, D. Evolving Video Skims into Useful Multimedia Abstractions, Proc. of the ACM CHI 98 Conference on Human Factors in Computing Systems, Los Angeles, CA, April 1998, 171-178.
3. Hauptmann, A.G. and Wactlar, H.D. Indexing and Search of Multimodal Information, International Conference on Acoustics, Speech and Signal Processing (ICASSP-97), Munich, Germany, April 21-24, 1997.
4. Hauptmann, A.G. and Lee, D., Topic Labeling of Broadcast News Stories in the Informedia Digital Video Library, DL-98 Proc. of the ACM Conference on Digital Libraries, Pittsburgh, PA, June 24-27, 1998.
5. Tat-Seng Chua, Shih-Fu Chang, Lekha Chaisorn and Winston Hsu. Story Boundary Detection in Large Broadcast News Video Archives – Techniques, Experience and Trends. ACM Multimedia 2004. Brave New Topic Paper. New York, Oct 2004.
6. Yan, R., Yang, J., Hauptmann, A., Learning Query-Class Dependent Weights in Automatic Video Retrieval, Proceedings of ACM Multimedia 2004, New York, NY, pp. 548-555, October 10-16, 2004
7. Rowley, H., Baluja, S. and Kanade, T. Human Face Detection in Visual Scenes. Carnegie Mellon University, School of Computer Science Technical Report CMU-CS-95-158, Pittsburgh, PA.
8. H. Schneiderman. "A Statistical Approach to 3D Object Detection Applied to Faces and Cars." Ph.D. Thesis. Carnegie Mellon University. CMU-RI-TR-00-06
9. Satoh, S., and Kanade, T. NAME-IT: Association of Face and Name in Video. IEEE Conference on Computer Vision and Pattern Recognition (CVPR97), (San Juan, Puerto Rico, June, 1997).
10. Gong, Y. Intelligent Image Databases: Toward Advanced Image Retrieval. Kluwer Academic Publishers: Hingham, MA, 1998.
11. W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos and G. Taubin, The QBIC Project: Querying Images By Content Using Color, Texture and Shape SPIE 1993 Intl. Symposium on Electronic Imaging: Science and Technology, Storage and Retrieval for Image and Video Databases, Feb. 1993.
12. Jianbo Shi, Jitendra Malik: Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8): 888-905 (2000)
13. Smith, M., Kanade, T., "Video Skimming for Quick Browsing Based on Audio and Image Characterization," Carnegie Mellon University technical report CMU-CS-95-186, July 1995. Also submitted to PAMI Journal (Pattern Analysis and Machine Intelligence), 1995.

14. Christel, M.; Conescu, R.: Addressing the Challenge of Visual Information Access from Digital Image and Video Libraries. ACM/IEEE JCDL 2005
15. Christel, M.; Moraveji, N.: Finding the Right Shots: Assessing Usability and Performance of a Digital Video Library Interface. Proc. ACM Multimedia, ACM Press (2004), 732–739
16. Christel, M., Huang, C., Moraveji, N., and Papernick, N. " Exploiting Multiple Modalities for Interactive Video Retrieval," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, 2004, pp.1032-1035.
17. Wactlar, H.D., Christel, M.G., Gong, Y., and Hauptmann, A.G. "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library," IEEE Computer, 32(2): pp.66-73, 1999.
18. Olligschlaeger, A.M., and Hauptmann, A.G., Multimodal Information Systems and GIS: The Informedia Digital Video Library, 1999 ESRI User Conference, San Diego, CA, July 27-29, 1999
19. J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In Eighth TextRetrieval Conference, pages 107–129, Washington, 2000
20. R.V. Cox, B.G. Haskell, Y. Lecun, B. Shahraray, and L. Rabiner, “Applications of Multimedia Processing to Communications,” Proceedings of the IEEE, May 1998, pp. 754-824.
21. Hauptmann, A.G., Towards a Large Scale Concept Ontology for Broadcast Video. 3rd International Conference on Image and Video Retrieval (CIVR'04), Dublin City University, Ireland, pp. 674-675, July 21-23, 2004
22. M. Naphade, J. R. Smith, “On Detection of Semantic Concepts at TRECVID,” ACM Multimedia (ACM MM-2004), Oct., 2004.
23. Kraaij, W., Smeaton, A.F., Over, P., Arlandis, J.: TRECVID 2004 – An Introduction. TRECVID 2004 Proceedings, http://www-nlpir.nist.gov/projects/trecvid/