

Informedia – Search and Summarization in the Video Medium

Howard D. Wactlar
Computer Science Department
Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

The Informedia system provides “full-content” search and retrieval of current and past TV and radio news and documentary broadcasts. The system implements a fully automatic intelligent process to enable daily content capture, analysis and storage in on-line archives. The current library consists of approximately a 2,000 hours, 1.5 terabyte library of daily CNN News captured over the last 3 years and documentaries from public television and government agencies. This database allows for rapid retrieval of individual “video paragraphs” which satisfy an arbitrary spoken or typed subject area query based on a combination of the words in the soundtrack, images recognized in the video, plus closed-captioning when available and informational text overlaid on the screen images. There are also capabilities for matching of similar faces and images, generation of related map-based displays. The latest work attempts to produce a visualization and summarization of the content across all the stories in a large retrieval result set. These combined functions enable a time and bandwidth efficient *navigation* of the video space and provide a highly interactive and engaging experience for the user.

Introduction

The overriding goal of the Informedia Digital Video Library project [Wactlar96,99] was to enable meaningful *search* and *discovery* in the video medium. This implied two research goals: (1) fully automated information extraction, and (2) full-content search and retrieval from within and across video productions. The technical approach pursued had three primary thrusts: (1) the application of integrated artificial intelligence techniques, i.e., speech, image and natural language understanding, to the library creation and its navigation, (2) extension and use of textual and content-based information retrieval techniques and metrics, and (3) validation through user testbeds and human-computer interaction studies. The fundamental premise of the research was that the integration of these technologies, all of which are imperfect and incomplete, would overcome the limitations of each, and improve the overall performance in the information retrieval task.

The challenges of the project resulted from the underlying imperfect technology used for analysis and search, and the temporal nature of the video object. Those that we had to overcome in order to provide a viable library included:

- retrieval performance in the presence of inaccuracy and ambiguity in the underlying cognitive processing
- approximate match in meaning and visualization
- presentation and reuse of video content as a new data type with space and time constraints
- interoperability in the presence of restricted use intellectual property and the absence of data and protocol standards

These challenges resulted in the following research and development efforts:

- validate premise of errorful speech generated transcripts for indexing and retrieval
- segment content into meaningfully coherent “video paragraphs” applying integrated speech, image and language understanding
- navigate video with automated marking, abstraction and summarization
- extract image objects and perform similarity matching
- implement data and networking architectures for remote video delivery
- incorporate micro-charging for access control and commerce

Figure 1 illustrates a typical user query and result set.



Figure 1: Typical user query and result set showing automatically generated headlines, topic assignment, video character recognition and text following during video play

Component Technologies

Image Understanding

Although current successful efforts at visual querying of image databases are founded on indirect image statistical methods, they fail to capture and exploit the massive information contained in video. Video is temporal, spatial, and often unstructured; the combined video and audio signal convey an abundance of information not retrievable from either one independently. Image understanding technology was applied to a set of diverse tasks: (1) scene break detection for icon selection, film-strip and skimming summarization, and “paragraph” segmentation, (2) image similarity matching, (3) camera motion determination and object tracking across scenes, (4) video-OCR, detection and recognition of text naturally appearing within or overlaid on the video, (5) face detection and association. Arbitrary object identification and scene characterization remain very long term research goals beyond the scope of this effort.

Video-OCR, the recognition and interpretation of text contained within the video, often contains information not otherwise conveyed in the audio track, such as person’s names, titles and affiliations, or location of the scene and event depicted. Applying multipass differential filtering techniques, the text is detected, filtered through successive frames, OCR’d, and incorporated into the database with a time alignment synchronizing it with where it appeared. This data is then indexed along with all other

metadata representing the video, and becomes part of the searchable index. We have achieved character recognition rates of 80% for overlaid text on news footage with corresponding word error rates approaching 70% when post-processed with a relevant thesaurus [Sato98].

Besides the work in text detection, the system applies neural network arbitration for face detection within scenes. We then use an eigenvector based method (eigenfaces) to compute a distance function between two faces as a way of establishing degree of similarity between them.

Image search and similarity matching remains a complex problem. Many systems incorporate some kind of similarity matching, the dimensions of which are color, texture and shape. The similarity challenge is heightened by its subjective meaning for different people or for the same person in different situations. Most common are images containing similar colors, or those containing similar shapes, but the greatest challenge is in finding images containing similar subject content, independent of color or shape. Informedia has in turn applied color histograms, region based analysis, and now perceptual color clustering based on human sensitivity and perception [Gong98a], which has shown the most striking measured results for image scene change determination.

Speech Understanding

Informedia applies highly accurate, speaker-independent, continuous speech recognizers, Sphinx-II [Hwang94] and Sphinx-III, to automatically generate a transcript of the soundtrack to enable text-based retrieval from spoken language documents. It is also used to provide precise text to audio/video synchronization in the presence of scripts or closed captioning. This data also supplies the necessary information for library segmentation and generating multimedia abstractions of the content.

Speaker recognition accuracy on such a corpus is widely variant, ranging from under 10% word error rate for laboratory environments, to over 80% for commercials with background music and noise, with an average of about 20% for broadcast news. What is important, however, is not the word recognition rate, but rather our overall ability to accurately retrieve segments from the library corresponding to the query. We devised experiments to measure recall (relevant hits returned / all relevant hits) and precision (relevant hits returned / hits returned) from our broadcast news corpus as a function of word error rate [Hauptmann97d]. Figure 2 plots the results for a corpus previously segmented into distinct stories. For speech recognition error rates below 30%, retrieval recall remained within 95% of that obtained on fully accurate transcripts. Even at 50% word error rate, recall was within 85%, but then dropped off more rapidly with increasing speech recognition transcription errors. Some of this results from the high redundancy in spoken language relative to the more carefully constructed written word. In news, documentaries and interviews, similar keywords and names are often repeated even in the brief story segment introductions, body and summation.

Information Retrieval

Our search engine is one distantly related to the original Lycos system. It employs keyword spotting, stop words (those that are ignored), synonyms, bonus for primacy, TF*IDF (text frequency * inverse document frequency) relevance weightings, vector length normalization and word stemming [Salton83]. We have also experimented with semantic query expansion by applying latent

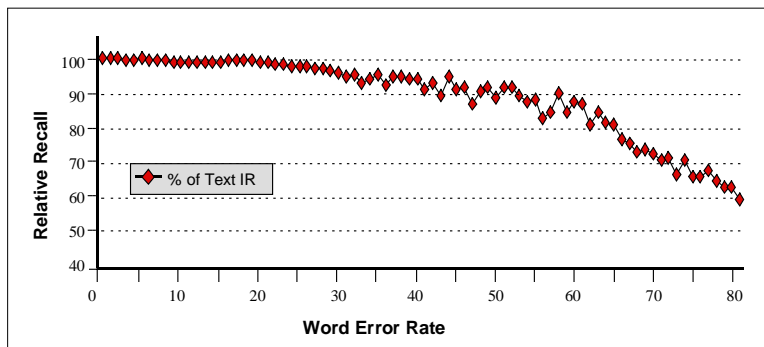


Figure 2: Information retrieval recall as a function of speech recognition accuracy.

semantic indexing to the news corpus (e.g. accident for crash, bacteria for germ, children for kids) but did not observe significant change in retrieval.

Video Navigation

The task of interactively navigating the video medium extends significantly beyond the search for the most closely matched segments. Whereas search often implies a specific information-seeking goal, most users in our studies prefer to browse the related content and seek a greater understanding that results from having a broader context.

Navigation Within the Result Set

Using subjects from both secondary school and college level users, studies were done of the relative value of various abstraction and navigation techniques and their impact on revealing and understanding the content [Christel97b]. It became clear from the outset that an intuitive user interface is as important to the video retrieval task as the search and result set. Efficient selection is especially important with video because it is expensive to transmit and time-consuming to view. The system currently provides multiple levels of abstraction and summarization:

- visual icons - representative video frame with relevance measure in the form of thermometer
- one-line headlines
- static film strip views, one frame per scene change
- active video skims
- transcript following of audio track

Both the film strip view and the time bar of the video player are marked with the occurrences of the query keywords, enabling the user to start play at an arbitrary point within a segment and manually skip to the next or prior keyword occurrence in the video. The active “video skims” are video abstracts that convey the essence of the segment’s content in 15-50% of the full play time by automatically extracting the most relevant phrases, sentences and image sequences, approximately fitting them to the duration specified, and playing the composite video.

Seeking a Broader Context

Topics. To help the user who is seeking a broader context for the segment just selected, each story is assigned a set of related category topics. We have implemented an automatic topic labeling component for the library applied during the library creation process [Hauptmann98d]. Each news story is assigned to one of over 3000 topic categories using a k-nearest neighbor (KNN) classification algorithm based on finding stories with similarly or overlapping content for which topics have already been assigned. In preliminary tests, the system achieved recall of 0.491 with relevance of 0.482 when up to five topics could be assigned to a news story. However, since the topic relationships were derived from a training phase using broadcast news transcripts, application of the same mapping to a broader corpus, such as documentaries, may produce misleading topic associations.

Named faces. Often personalities play a dominant role in the evolution of any story. Discovering where else an individual may have appeared, or whom they may have appeared with, can be important. To this end, we have attempted to detect and identify human faces appearing throughout the video corpus. As an interesting example of collaborative integration of image understanding and natural language processing, we developed a system, called Name-It, that associates faces and names in the videos, given broadcast news content as a knowledge source [Sato97]. The system can infer the name of a given unknown face image, or guess faces which are likely to be those of a selected name. This is achieved by correlating similar faces with names referenced in the audio track near where the face occurs in the video through the use of co-occurrence matrices. Video-OCR may also be used to establish a face name identity with somewhat higher probability if both name and face co-occur in the video image.

Maps. Issues of news and historic interest usually have a geographic setting and relationship to other events and video documents within the library. To this end, a recent extension to the video processing is the automatic extraction of geographic references from the video, audio and image content. The library interface spontaneously generates and displays the geographic entities addressed in a given story, highlighting the regions discussed at any point in the video through a map display synchronized with the video playback. The map can also be used as a query mechanism, allowing users to search the library for stories taking place in a selected area of interest. Figure 3 displays a typical map-based display accompanying the video selection.

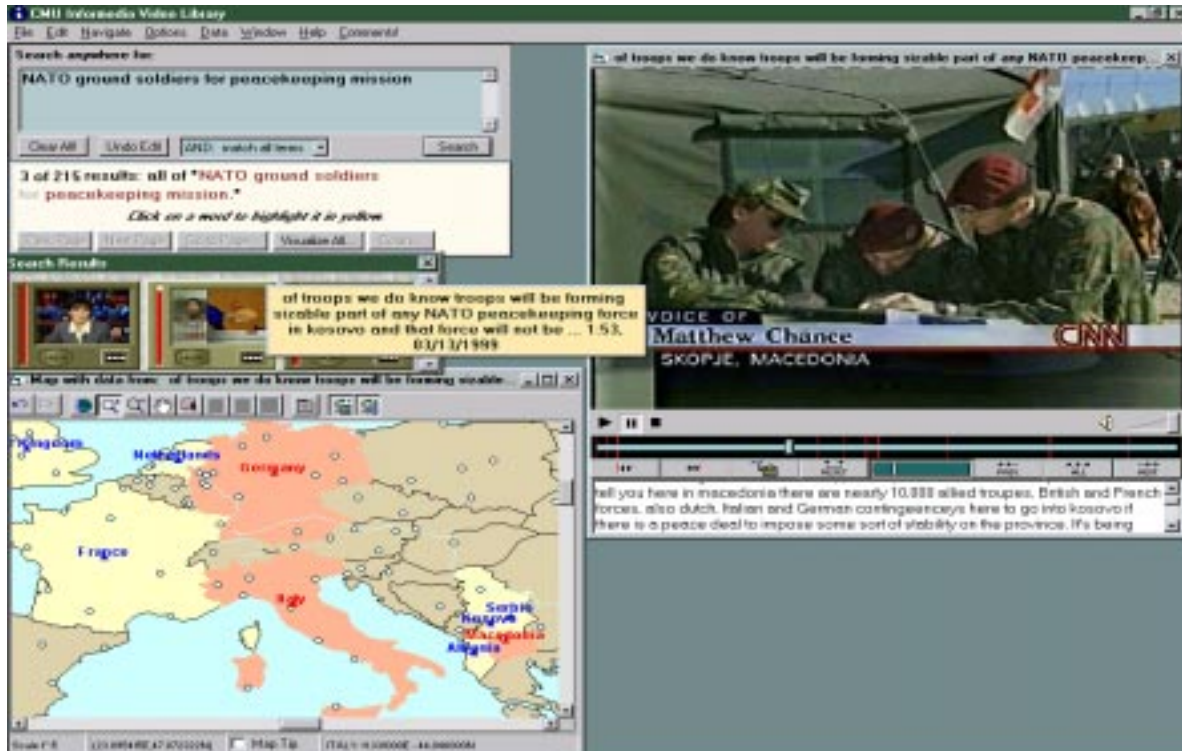


Figure 3: Query and result set displaying geographic context.

Geographical references (georeferences) are associated with each video segment and represented as a single value, a set of distinct values, or range of values corresponding to the locations where the video was situated as well as the locations referred to in the video. The user is thus able to specify a named location or location coordinates in order to query or browse for events at that location or within some “distance” of that location. The distance and location may also be expressed as a region, and refer synonymously, or hierarchically, to political or geographically defined boundaries that determine a region. The named locations, regions and “distances” are resolved, i.e., *geocoded*, to a common notation and metric (latitude and longitude) through integration of robust geographical information systems (GIS). The geocoded data is time-invariant: place and country names can change but their coordinates do not. Geocoded data thus allows for a more accurate display and retrieval of historical data. Prepositional references such as “near”, “above” and “north of” still need to be lexically analyzed and resolved geographically.

Video Summarization Across the Entire Result Set

The Informedia processing provided state of the art access to video by *content*. Current efforts will communicate information trends across time, space, and sources by furthering analysis and understanding of the *context* as well as content.

The Informedia interface was optimized to expose content for a single document from a query's result set, as illustrated in Figure 3 which shows 3 documents returned from a text query regarding NATO ground soldiers with a headline, map and video opened for one of those documents. This interface proved insufficient as the library grew beyond 1000 hours of video. The new work utilizes video information “collages” to expose content from sets of videos. For example, using the query and results shown in

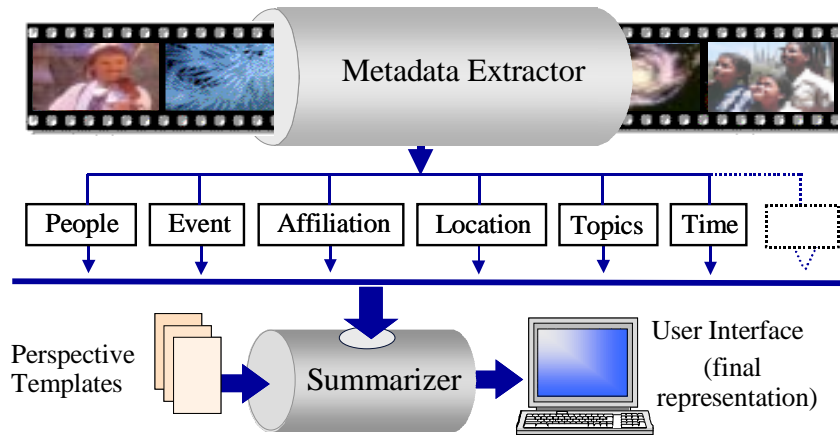


Figure 4: Informedia-II conceptual system overview

Figure 3, it would allow users to see the countries represented in all 215 results, the key people involved, and minimize the overlap in coverage.

Figure 4 presents a schema for the system. Through the extraction of appropriate metadata from diverse video collections, relevant information can be synthesized and presented driven by the user's needs. Currently users may visit numerous video collections in search of an answer that reveals itself only in bits at a time, such as an unfolding

story of a famous criminal trial or a regional political conflict. Video information collages will emphasize dimensions of importance to the user so that the full context can be understood and navigated to narrow the focus to a particular information thread, resulting in only the most useful video pieces then being played.

Collages. Text extraction and summarization is a rich area of research [Cowie96, Larkey96, Klavans96, Soderland97, MUC98]. This work is complemented with information from speech recognition and image processing. Then, *video information collages* can be built from the results of integration of these technologies to achieve information extraction and summarization in the video domain. There will be numerous templates or organizational schemes for collages, including *geo-collages* like maps, *chrono-collages* like timelines, and *auto-documentaries* in which the collage is not viewed all at once but rather is played like a documentary video.

Consider for example, the query “El Niño effects in Indonesia” with a geographic perspective, showing in Figure 5(a) the El Niño effects concentrated on two islands, with representative video icons. Collages enable the user to emphasize different aspects or facets of the digital video library. Suppose the user of Figure 5(a) now wished to see the faces of the key players and short event descriptors for Indonesia during the time period of the El Niño effects. Figure 5(b) shows a stratified chrono-collage emphasizing this information, where the adjacency of the first two faces indicate that those men (Suharto and Mondale) were in a meeting together discussing economic reform; the text names corresponding to the faces could be added as well via Name-It processing [Sato97]. An auto-documentary (not shown) is played rather than viewed all at once. It attempts to sequence together the most relevant and representative audio and visual imagery and present a coherent story that unfolds along the temporal, spatial, or topical dimensions, as controlled by the user.



(a) Map perspective produced with overlaid filmstrips representing constituent stories



(b) Timeline perspective emphasizing “key player faces” and short event descriptors, representing the same data shown in the Indonesia map perspective in (a)

Figure 5: Multiple video information collages and their interactions

Users may wish to further “drill down” to show more detail but perhaps less context, due to limited screen real estate, and “drill up” to show more context but less detail. Video information collages in the Informedia-II system will be designed to be:

Scalable, capable of summarizing a single video, a set of videos, or the whole video library.

• **Semantically zoomed**

- Zooming along the natural dimensions of the collage template. For example, the geo-collage allows zooming from continent to region to country to city. The chrono-collage allows zooming down to days or out to years. This chrono-collage will also support event-descriptor zooming, e.g., zooming into “El Niño wildfires” will reveal that the fires are started by people clearing land but that the drought caused by El Niño results in those fires getting out of control.
- Zooming from the synthesis represented by collages to the specific contributing documents to the Informedia multimedia abstractions for each document.

Foundation References

Informedia successfully pioneered the automatic creation of multimedia abstractions, demonstrated empirical proofs of their relative benefits, and gathered usage data of different summarizations and abstractions. Fundamental research and prototyping was conducted in the following areas, shown with a sampling of references to particular work:

- Integration of speech, language, and image processing: generating multimedia abstractions, segmenting video into stories, and tailoring presentations based on context [Wactlar96,99, Christel97a,97b].
- Text processing: headline generation [Hauptmann97a], text clustering and topic classification [Yang94a,98a, Lafferty98, Hauptmann98b], and information retrieval from spoken documents [Hauptmann97b,97c,98c].
- Audio processing: speech recognition [Witbrock98a,98b], segmentation and alignment of spoken dialogue to existing transcripts [Hauptmann98a], and silence detection for better “skim” abstractions [Christel98].
- Image processing: face detection [Rowley95] and matching based on regions, textures, and colors [Gong98b].
- Video processing: key frame selection, skims [Smith96,97], Video OCR [Sato98], and Video Trails [Kobla97].

The core technology is being commercially developed and productized by MediaSite, Inc. (www.MediaSite.net), Pittsburgh, PA, USA.

Acknowledgements

This paper is based on work supported by the National Science Foundation, DARPA and NASA under NSF Cooperative agreement No. IRI-9411299. Results reported here represent the integrated work of a large number of researchers directly associated with this project and outside of it, both at Carnegie Mellon and other institutions, who have contributed base technology. We are grateful to those who have enabled us to use their content under license in the research and testbeds, especially CNN, QED Enterprises (WQED), and the Open University (U.K.).

References

- [Christel97a] Christel, M., Winkler, D., and Taylor, C. Improving Access to a Digital Video Library, *Human-Computer Interaction: INTERACT97, the 6th IFIP Conf. On Human-Computer Interaction*, Sydney, Australia, July 14-18, 1997, 524-531.
- [Christel97b] Christel, M., Winkler, D., & Taylor, C. Multimedia Abstractions for a Digital Video Library, *Proc. of the 2nd ACM International Conference on Digital Libraries*, (Philadelphia, PA, July, 1997), 21-29.
- [Christel98] 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.
- [Cowie96] Cowie, J. and Lehnert, W. Information Extraction. *CACM*, 39(1), 80-91.
- [Gong98a] Gong, Y.H., Proietti, G., and Faloutsos, C. “Image Indexing and Retrieval Based on Human Perceptual Color Clustering,” *IEEE Int'l Conf on Computer Vision and Pattern Recognition (CVPR98)*, June 1998.
- [Gong98b] Gong, Y. *Intelligent Image Databases: Toward Advanced Image Retrieval*. Kluwer Academic Publishers: Hingham, MA, 1998.
- [Hauptmann97a] 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.
- [Hauptmann97b] 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.

- [Hauptmann97c] Hauptmann, A.G., and Witbrock, M.J. Informedia News-on-Demand: Multimedia Information Acquisition and Retrieval. Chapter 11 in *Intelligent Multimedia Information Retrieval*, M. Maybury, Ed. AAAI Press/MIT Press: Menlo Park, CA, 1997.
- [Hauptmann97d] 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, 1997.
- [Hauptmann98a] Hauptmann, A.G., and Witbrock, M.J., Story Segmentation and Detection of Commercials in Broadcast News Video, *ADL-98 Advances in Digital Libraries*, Santa Barbara, CA, April 22-24, 1998.
- [Hauptmann98b] 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.
- [Hauptmann98c] Hauptmann, A.G., Jones, R.E., Seymore, K., Siegler, M.A., Slattery, S.T., and Witbrock, M.J. Experiments in Information Retrieval from Spoken Documents, *Proc. of the DARPA Workshop on Broadcast News Understanding Systems (BNTUW-98)*, Lansdowne, VA, February 1998.
- [Hauptmann98d] Hauptmann, A.G. and Lee, D. Topic Labeling of Broadcast News Stories in the Informedia Digital Video Library, *DL-98 Proceeding of the ACM Conference on Digital Libraries*, Pittsburgh, PA (in press), 1998.
- [Hwang94] Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., Alleva, F. (1994). Improving Speech Recognition Performance via Phone-Dependent VQ Code-books and Adaptive Language Models in SPHINX-II. *ICASSP-94*, Vol. I, pp. 549-552.
- [Klavans96] Klavans, J.L. and Resnik, P., eds. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press: Cambridge, Massachusetts.
- [Kobla97] Kobla, V., Doermann, D., and Faloutsos, C. Video Trails: Representing and Visualizing Structure in Video Sequences, *ACM Multimedia 97*, Seattle, WA, November, 1997.
- [Larkey96] Larkey, L. and Croft, W. B. Combining Classifiers in Text Categorization, *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*, Zurich, Switzerland, 289-297.
- [Lafferty98] Lafferty, J. and Venable, P. Simultaneous Word and Document Clustering, *Proc. CONALD Workshop on Learning from Text and the Web* (extended abstract), Pittsburgh, PA, June 11-13, 1998.
- [MUC98] *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, (Fairfax, VA, April 1998), Morgan Kaufmann Publishers. [Rowley95] 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.
- [Salton83] Salton, G. and McGill, M.H. "Introduction to Modern Information Retrieval", McGraw-Hill, New York, 1983.
- [Sato98] Sato, T., Kanade, T., Hughes, E., Smith, M. (1998). Video OCR for Digital News Archive, *Proceedings of the 1998 IEEE International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, (Bombay, India, Jan. 3, 1998), 52-60.
- [Sato97] Sato, S., and Kanade, T. "NAME-IT: Association of Face and Name in Video." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR97)*, Puerto Rico, 1997.
- [Satoh97] 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).
- [Smith97] Smith, M. and Kanade, T. Video Skimming and Characterization Through the Combination of Image and Language Understanding Techniques, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR97)*, San Juan, Puerto Rico, June, 1997, 775 – 781.
- [Smith96] Smith, M. and Kanade, T. Video Skimming for Quick Browsing Based on Audio and Image Characterization Carnegie Mellon University, *School of Computer Science Technical Report CMU-CS-95-186R*, Pittsburgh, PA.
- [Soderland97] Soderland, S., Fisher, D., and Lehnert, W. Automatically Learned vs. Hand-crafted Text Analysis Rules, *CIIR Technical Report TE-44*.
- [Wactlar99] Wactlar, H., Christel, M., Gong, Y., and Hauptmann, A. "Lessons Learned from Building a Terabyte Digital Video Library", *IEEE Computer*, 32(2) February 1999, pp. 66-73.
- [Wactlar96] Wactlar, H.D., Kanade, T., Smith, M.A., and Stevens, S.M. Intelligent Access to Digital Video: Informedia Project. *IEEE Computer*, 29(5), 46-52, May 1996.

- [Witbrock98a] Witbrock, M.J., and Hauptmann, A.G. Improving Acoustic Models by Watching Television. Carnegie Mellon University, *School of Computer Science Technical Report CMU-CS-98-110*, Pittsburgh PA, 1998.
- [Witbrock98b] Witbrock, M.J., and Hauptmann, A.G. Speech Recognition in a Digital Video Library, *Journal of the American Society for Information Science (JASIS)*, 47(7), May 15, 1998.
- [Yang94a] Yang, Y. Expert network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval, *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, 13–22, July 3-6, 1994.
- [Yang98a] Yang, Y. Pierce, T., and Carbonell, J. A Study on Retrospective and On line Event Detection, *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, August 24-28, 1998.