New Directions in Video Information Extraction and Summarization

Howard D. Wactlar
wactlar@cmu.edu
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

The Informedia Digital Video Library project provided a technological foundation for full content indexing and retrieval of video and audio media. New directions for this research extend to: (1) search and retrieval in multilingual video corpora, (2) analysis and indexing of continuously captured, unstructed and unedited field-collected video, and (3) summarization of video-based content across multiple stories based on the user's perspective.

Informedia Digital Video Library Foundation Work

The Informedia Digital Video Library focused on the development and integration of technologies for information extraction from video and audio content to enable its full content search and retrieval. Over a terabyte (1600 hours, 4,000 segments) of online data was collected, with automatically generated metadata and indices for retrieving video segments from this library. 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 [Wactlar96,99a, 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 [Gong98].
- Video processing: key frame selection, skims [Smith96,97], Video OCR [Sato98], and Video Trails [Kobla97].

Building on this base technology, extensions are being pursued in the following areas:

- Multilingual video libraries and cross-lingual search
- Indexing of continuously captured, unstructured, unedited field video
- Auto-summarization and visualization over multiple video documents and libraries

Multilingual Informedia

The Multilingual Informedia Project demonstrates a seamless extension of the Informedia approach to search and discovery across video documents in multiple languages. The new system performs speech recognition on foreign language (non-English) news broadcasts, segments it into stories and indexes the foreign data together with English news data from English language sources.

The Components of Multilingual Informedia

There are three components in the Multilingual Informedia System [Hauptmann98d] that differ significantly from the original Informedia system:

1. The speech recognizer recognizes a foreign language, specifically Serbo-Croatian [Guetner 97,98, Scheytt97]. This component will not be described here.
2. A phrase-based translation module transforms English queries into Serbo-Croatian, allowing a search for equivalent words in a joint corpus of English and Serbo-Croatian news broadcasts.
3. English topic labels for the foreign language news stories allow a user to identify a relevant story in the target language [Hauptmann98e].

The Informedia Translation Facility

The current version of the Translation Facility attempts to translate large chunks of phrases it finds in the Serbo-Croatian text. The Multilingual Informedia System allows a query to be posed in English and the query will be translated into the target corpus language and used for retrieval there. This system takes advantage of multi-word phrase entries in a machine-readable dictionary [Brown97]. It parses the source-language text for phrases using a simple recursive algorithm. It first scans the dictionary for a translation(s) of the entire text as one phrase; if that fails it searches for phrasal translations of substrings one word smaller, then one more word smaller than that, and so on. The first phrasal translation thus obtained (or set thereof if there are multiple alternative translations) is kept as part of the output string, and the process is recursively invoked on the pieces of the text preceding and following the substring just translated. The recursion continues until a set of chunks and individual words is produced covering the text string, for which translations have been found for all the chunks and may or may not have been found for the individual words. The concatenated results become the output string. In general, this text-translation facility will work with any language pair so long as a bilingual machine-readable dictionary is available in the format the program understands.

The DIPLOMAT "example-based" machine translation system [Carbonell97, Frederking97] developed at Carnegie Mellon University was also put to use for "high-quality" story translation from Serbo-Croatian into English. However, machine translation of errorful (20-50% word error rate) speech recognition generated transcripts of naturally spoken language produces results significantly degraded from that of fully accurate written text.

Foreign Language Topic Detection

After initial experiments with the Serbo-Croatian news, it became clear that returning a foreign language result to the user was not sufficient. The users were unable to tell if a particular news clip was actually relevant to their query, or if it was returned due to poor query translation or inadequate information retrieval techniques. To allow the user at least some judgment about the returned stories, we attempted to label each Serbo-Croatian news story with an English-language topic.

The topic identification was done using the query translation facility to translate the whole story into English words. This translation became the *topic query*. Separately, we had indexed about 35000 English language news stories, which had manually assigned topics assigned to them. Using the SMART information retrieval system, we now used the translated *topic query* to retrieve the most relevant 10 labeled English stories. Each of the topics for the labeled stories that were retrieved was weighted by its relevance to the *topic query* and the weights for each topic were summed. The most favored topics, above a threshold, were then used to provide a topic label for the Serbo-Croatian news story. This topic label allows the user to identify the general topic area of an otherwise incomprehensible foreign language text and determine if it is relevant at least in the topic area.

Informedia Experience-on-Demand

The Informedia Experience-on-Demand Project (EOD) [Wactlar99b] develops tools, techniques and systems allowing people to capture a record of their activities unobtrusively, and share them in collaborative settings spanning both time and space. Users may range from rescue workers carrying personalized information systems in operational situations to remote crisis managers in coordinating roles. Personal EOD units record audio, video, Global Positioning System (GPS) spatial information, and other sensory data, which can be annotated by human participants. The EOD environment synthesizes data from many EOD units into a "collective experience" – a global perspective of ongoing and archived personal experiences. Distributed collaborators can be brought together over time and space to share meaning and perspectives.

Each constituent EOD unit captures and manages information from its unique point of view. This information is transferred to a central site where the integration of multiple points of view provides greater detail for decision-making and event reporting. A longer term goal, dependent on advances in communication technology, is for each portable EOD unit to be not only a data collector but also a data access device, interoperating with the other EOD units and allowing audio and video search and retrieval.

We have built a prototype EOD system that builds on the core Informedia technologies by addressing continuously captured, unstructured, unedited video in which location data is added as another information dimension.

By tailoring speech recognition for mobile, active talkers, we hope to improve the quality of the resulting text transcript that is used to index material in the multimedia experience database. We are exploring enhancements to existing speech recognition algorithms to filter out the background noise typical in audio collected in outdoor environments. By optimizing language models for information retrieval accuracy rather than transcript word error rate [Witbrock97], we hope to further improve the utility of the speech recognition output for indexing the experience database.

Similarly, we are modifying Informedia image processing modules to better work with field-captured motion video. Our current object detectors for recognizing and matching faces and overlaid text work well on broadcast news given certain assumptions, such as a well-lit face looking directly at the camera. These assumptions are less likely to be met with field video, and so we are investigating more robust techniques for object detection within video having varying shades of lighting and where the object of interest may appear at varying resolutions.

The data is by nature voluminous in size yet sparse in information content, with tremendous redundancy along the temporal and spatial dimensions and across points of view. We deal with this by developing filtering techniques that scan for change, yet retain all salient time, location and image information in the metadata.

Figure 1 shows how continuously recorded GPS data can be used to tightly synchronize a playing video to a map. As the video plays, the location for the

Figure 1. Video with map showing trajectory of motion (as dotted line) for full video segment and location (upper left of dotted line) associated with the shown video frame

displayed video image is highlighted on the map, with the area covered within the video segment shown on the map as well. The map is both *active*, changing as the video plays, and *interactive*, allowing the user to modify the map display and use it to issue spatial queries to locate experiences in specified areas. Experience-on-Demand addresses collaboration and summarization of multiple simultaneous information generators integrated across people, time, and space.

Auto-Summarization and Visualization of the Result Set

The Informedia processing provided state of the art access to video by *content*. This new research direction will communicate information trends across time, space, and sources by emphasizing analysis and understanding of *context* as well as content.

Future multi-terabyte digital video libraries present new challenges requiring different approaches. The Informedia interface was optimized to expose content for a single document from a query's result set, as illustrated in Figure 2 which shows 12 documents returned from a text query on "El Niño" with a headline, filmstrip and video opened for one of those documents. This interface proved insufficient as the library grew beyond 1000 hours of video. The new work will utilize video information "collages" to expose content from sets of videos. For example, using the query and results shown in Figure 2, it would allow users to see the countries represented in all 215 results, the key people involved, and minimize the overlap in coverage.

Figure 3 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.

Figure 2: Informedia interface following "El Niño" text query and display of one text title, one filmstrip and one video

Collages as Video Information Synthesis Across Time, Space, and Sources

Text extraction and summarization is a rich area of research [Cowie96, Larkey96, Klavans96, Soderland97, MUC98]. This work will be 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 the geo-collage shown in Figure 4a following a query on "El Niño effects." The dark-colored areas indicate the spatial distribution of the results filtered to political boundaries.

Representative images for geographical areas are shown on demand, allowing the user to see that the Indonesian effects have something to do with fires. The user can drill down into that area, shown as a black rectangular border, producing the more focused collage of Figure 4b. The user has the option to show additional map information and more representative images for the highlighted regions, which show El Niño effects concentrated on two islands.

Collages enable the user to emphasize different aspects or facets of the digital video library. Suppose the user of Figure 4b 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 4c 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

Figure 3: Informedia-II conceptual system overview.

could be added as well via Name-It processing [Satoh97]. 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
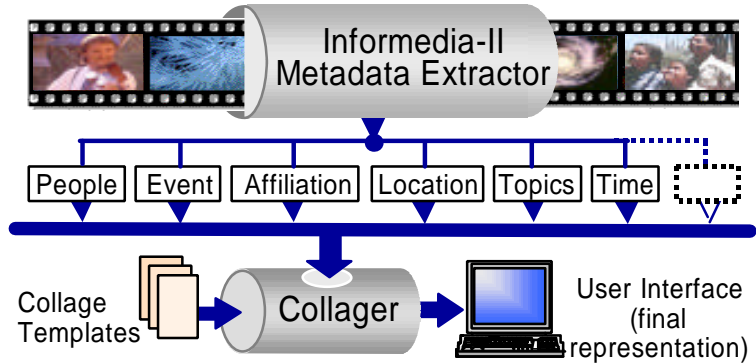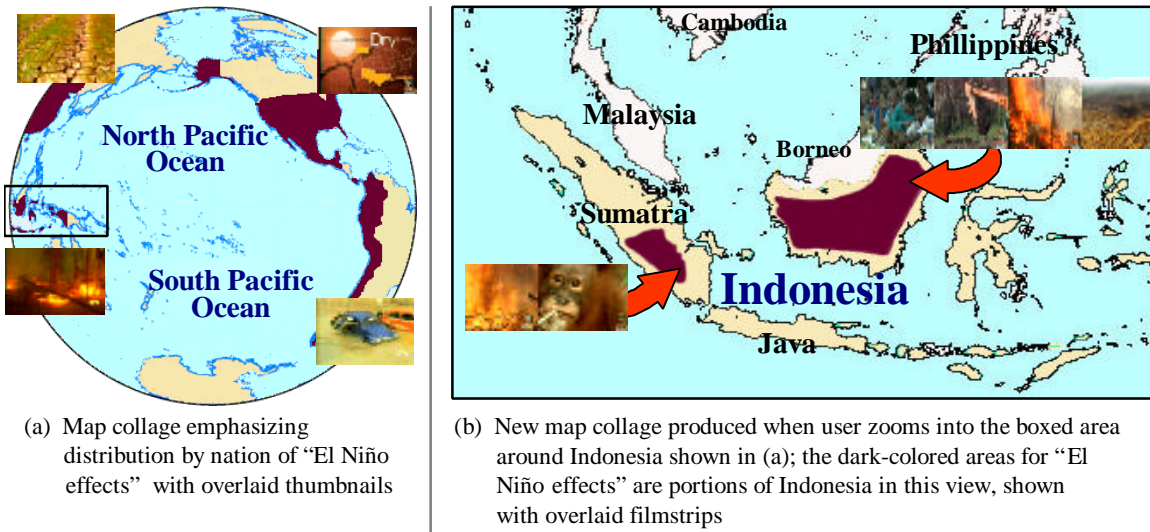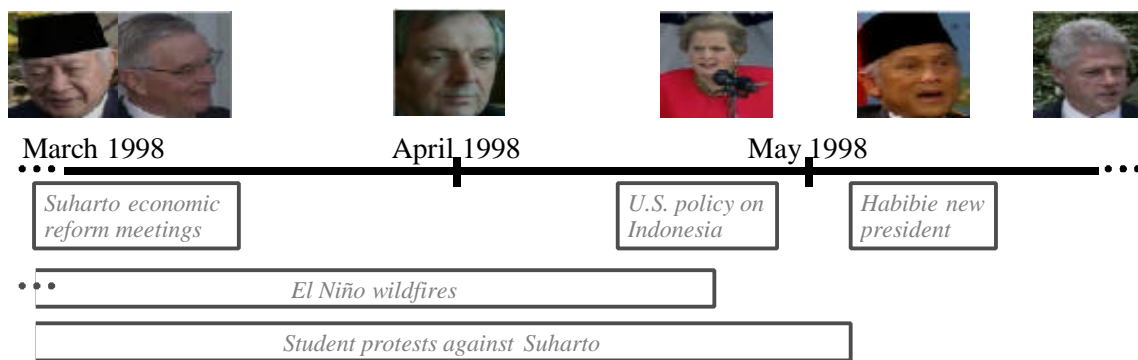

(a) Map collage emphasizing distribution by nation of "El Niño effects" with overlaid thumbnails

(b) New map collage produced when user zooms into the boxed area around Indonesia shown in (a); the dark-colored areas for "El Niño effects" are portions of Indonesia in this view, shown with overlaid filmstrips

(c) Timeline collage emphasizing "key player faces" and short event descriptors, representing the same data shown in the Indonesia map collage in (b)

Figure 4: Multiple video information collages and their interactions

user.

Information layout is obviously important in building the video collages. Information visualization techniques include Cone Trees [Robertson93], Tree Maps [Johnson91], Starfields [Ahlberg94a], dynamic query sliders [Ahlberg94b], and VIBE [Olsen93]. Visualizations such as LifeLines [Freeman95], Media Streams [Davis94], and Jabber [Kominek97], have represented temporal information along a timeline. DiVA [Mackay98] has shown multiple timelines simultaneously for a single video document. In contrast, Informedia-II will make use of templates beyond just timelines, such as the geo-collages emphasizing geographical perspectives information. Our collage templates will significantly advance the field by enabling information visualization across multiple video documents.

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.

Underlying Information Extraction and Metadata Creation

The ability to extract names of organizations, people, locations, dates and times (i.e. "*named entities*") is essential for correlating occurrences of important facts, events, and other metadata in the video library, and is central to production of information collages. Our techniques extract named entities from the output of speech recognition systems and OCR applied to the video stream, integrating across modalities to achieve better results. Current approaches have significant shortcomings. Most methods are either rule-based [Maybury96, Mani97], or require significant amounts of manually labeled training data to achieve a reasonable level of performance [BBN98]. The methods may identify a name, company, or location, but this is only a small part of the information that should be extracted; we would like to know that a particular person is a politician and that a location is a vacation resort.

Geographical references (georeferences) will be 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 will be 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. Geo-collages built from synthesizing georeferences for a set of videos will enable users to spot patterns or trends in the events with respect to location, e.g., to see that El Niño contributed significantly to increased forest fires in Indonesia. 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" will need to be lexically analyzed, as others have done with pictorial captions [Srihari95]. Challenges include varying granularity and relative versus absolute

position information, the synchronization of the location information with the video stream; and the likelihood of inaccurate and errorful identification of named locations.

Similarly, explicit and indirect time and date references need to be detected, resolved and encoded in a consistent manner. Such time references might range from "next month" in a current news story to "before the war" in a documentary retrospective.

Conclusion

The overarching long-term goal of the Informedia initiatives has been to bring to spoken language and visual documentation the same functionality and capability that we have with written communication, including all aspects of search, retrieval, categorization and summarization. New research directions will enable us to take special advantage of the richness of holistic visual and temporal presentation by providing the analysis tools and techniques to extract requisite content, assemble context for responding to user interactions, minimize redundancy, and summarize over multiple dimensions and granularity. For example, this work enables a user to generate a visual perspective of the conflict in Kosovo from multiple reports by the various foreign press corps and contrast it with video vignettes of Balkan culture and history since WWI produced in the native countries.

Perhaps even more importantly, these methodologies may have a societal impact beyond the scientific community as they provide a set of new capabilities and aid in understanding how events evolve and are correlated over time and geographically. Any citizen will potentially be empowered to ask even analytic questions of the global video record our society is creating of itself. The evolution of events can be tracked and perspectives from around the globe can be brought to bear on their understanding, and presented in a medium that is visually rich and engaging.

Acknowledgements

References

[Ahlberg94a] Ahlberg, C. and Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays, *Proc. ACM CHI '94 Conference on Human Factors in Computing Systems*, Boston, 313-322.

[Ahlberg94b] Ahlberg, C. and Shneiderman, B. The Alphaslider: A Compact and Rapid Selector, *Proc. ACM CHI '94 Conference on Human Factors in Computing Systems*, Boston, 365-371.

[BBN98] BBN Corporate Web Site, Speech and Language IdentifFinder, URL http://www.bbn.com/products/speech/identifi.htm.

[Brown97] Brown, R.D., Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. *Proc. of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, Santa Fe, July 23-25, 1997.

[Carbonell97] Carbonell, J., Yang, Y., Frederking, R., Brown, R.D., Geng, Y., and Lee, D. Translingual Information Retrieval: A Comparative Evaluation. *Proc. of Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97),* August 23-29, 1997.

[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.

[Davis94] Davis, M. Knowledge Representation for Video, *Proceedings of AAAI '94*, 120-127.

[Frederking97] Frederking, R., Rudnicky, A., and Hogan, C. Interactive Speech Translation in the DIPLOMAT Project. Spoken Language Translation Workshop of the Association for Computational Linguistics, ACL-97. Madrid, Spain. July 7-12, 1997.

[Freeman95] Freeman, E., and Fertig, S. Lifestreams: Organizing your Electronic Life, *AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval*, November, Cambridge, MA. URL http://www.halcyon.com/topper/jv6n1.htm.

[Gong98] Gong, Y. *Intelligent Image Databases: Toward Advanced Image Retrieval*. Kluwer Academic Publishers: Hingham, MA, 1998.

[Guetner98] Geutner, P., Finke, M., Scheytt, P., Waibel, A., and Wactlar, H., Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexicon Adaptation. To appear in *BNTUW-98 Proc. of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne VA, February 8-11, 1998.

[Guetner97] Geutner, P., Finke, M., Scheytt, P., Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News, Technical Report, Carnegie Mellon University, Pittsburgh, PA, CMU-LTI-97-155, December 1997. [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.

[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.

[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., Scheytt, P., Wactlar H. Multi-Lingual Informedia: A Demonstration of Speech recognition and Information Retrieval Across Multiple Languages, *BNTUW-98 Proc. Of DARPA Workshop on Broadcast News Understanding Systems*, Lansdowne, VA, February 1998.

[Hauptmann98e] Hauptmann, A., Lee. D., Kennedy, P. Semantic Topic Labeling of Multilingual Broadcast News in the Informedia Digital Video Library. *Proc. Of IEEE IFIP'98 Network Operations and Management Symposium*, New Orleans, LA, February 15-10, 1998.

[Johnson91] Johnson, B., and Shneiderman, B. Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. *Proc. IEEE Visualization '91*, (San Diego, October), 284-291.

[Kominek97] Kominek, J., and Kazman, R. Accessing Multimedia through Concept Clustering, *Proceedings of ACM CHI '97 Conference on Human Factors in Computing Systems*, (Atlanta, GA, March, 1997), 19-26.

[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.

[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, Jne 11-13, 1998.

[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.

[Mackay98] Mackay, W.E., and Beaudouin-Lafon, M. DIVA: Exploratory Data Analysis with Multimedia Streams, *Proceedings of the ACM CHI'98 Conference on Human Factors in Computing Systems*, (Los Angeles, CA, April 1998), 416-423.

[Mani97] Mani, I., House, D., Maybury, M. and Green, M. Towards Content-Based Browsing of Broadcast News Video, in *Intelligent Multimedia Information Retrieval*, M. Maybury, Ed. AAAI Press/MIT Press: Menlo Park, CA.

[Maybury96] Maybury, M., Merlino, A., and Rayson, J. Segmentation, Content Extraction and Visualization of Broadcast News Video using Multistream Analysis, *ACM Multimedia Conf.*, Boston, MA.

[MUC98] *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, (Fairfax, VA, April 1998), Morgan Kaufmann Publishers.

[Olsen93] Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., and Williams, J. G. Visualization of a Document Collection: The VIBE System. *Information Processing & Management*, 29(1), 69-81.

[Robertson93] Robertson, G., Card, S., and Mackinlay, J. Information Visualization Using 3D Interactive Animation, *Communications of the ACM,* 36(4), 56-71.

[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.

[Sato98] Sato, T., Kanade, T., Hughes, E., Smith, M. Video OCR for Digital News Archive, *Proc. of the 1998 IEEE International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98),* Bombay, India, Jan. 3, 1998, 52-60.

[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).

[Scheytt97] Scheytt, P., Finke, M., Geutner, P., Speech Recognition on Serbo-Croatian Dictation and Broadcast News Data, Technical Report, Carnegie Mellon University, Pittsburgh, PA, CMU-LTI-97-154, December 1997.

[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.

[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. [Soderland97] Soderland, S., Fisher, D., and Lehnert, W. Automatically Learned vs. Hand-crafted Text Analysis Rules, *CIIR Technical Report TE-44*, URL .

[Soderland97] Soderland, S., Fisher, D., and Lehnert, W. Automatically Learned vs. Hand-crafted Text Analysis Rules, *CIIR Technical Report TE-44*.

[Srihari95] Srihari, R.K. Automatic Indexing and Content-Based Retrieval of Captioned Images, *IEEE Computer*, 28(9), 49-56.

[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.

[Wactlar99a] Wactlar, H., Christe. M., Gong, Y., Hauptmann A. Lessons Learned from Building a Terabyte Digital Video Library. *IEEE Computer*, Special Issue on Digital Libraries, February 1999, 32(2), pp. 66-63.

[Wactlar99b] Wactlar, H., Christel, M., Hauptmann, A., Gong, Y. Informedia Experience-on-Demand: Capturing, Integrating and Communicating Experiences Across People, Time and Space. *ACM Computing Surveys Special Issue on Collaboration Technology*, Vol. 31, March 1999.

[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.

[Witbrock97] Witbrock, M., Hauptmann, A. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken *Documents Proceedings of ACM Digital Libraries '97*, ACM, 30-25. July 23-16, 1997.

[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.