Invited speaker at New Information Technology 2001 Conference, Tsinghua University, Beijing, May 29-31, 2001

Multi-Document Summarization and Visualization in the Informedia Digital Video Library

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. The library now contains over 2000 hours of video and is growing daily. A good query engine is not sufficient for information retrieval because often the candidate result sets grow in number as the library grows. Video digests summarize sets of stories from the library, providing users with a visual mechanism for interactive browsing and query refinement. These digests are generated dynamically under the direction of the user based on automatically derived metadata from the video library.

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 two terabytes (2000 hours, 5,000 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].

Auto-Summarization and Visualization of the Result Set

The Informedia processing provided state of the art access to video by *content*. New techniques being pursued 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 1 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 1, it would allow users to see the countries represented in all 215 results, the key people involved, and minimize the overlap in coverage.

Figure 2 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 1: 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 3a 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 3b. 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 3b 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 3c 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

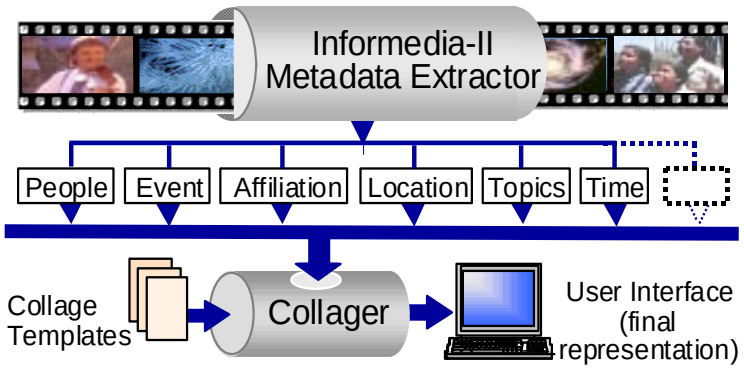
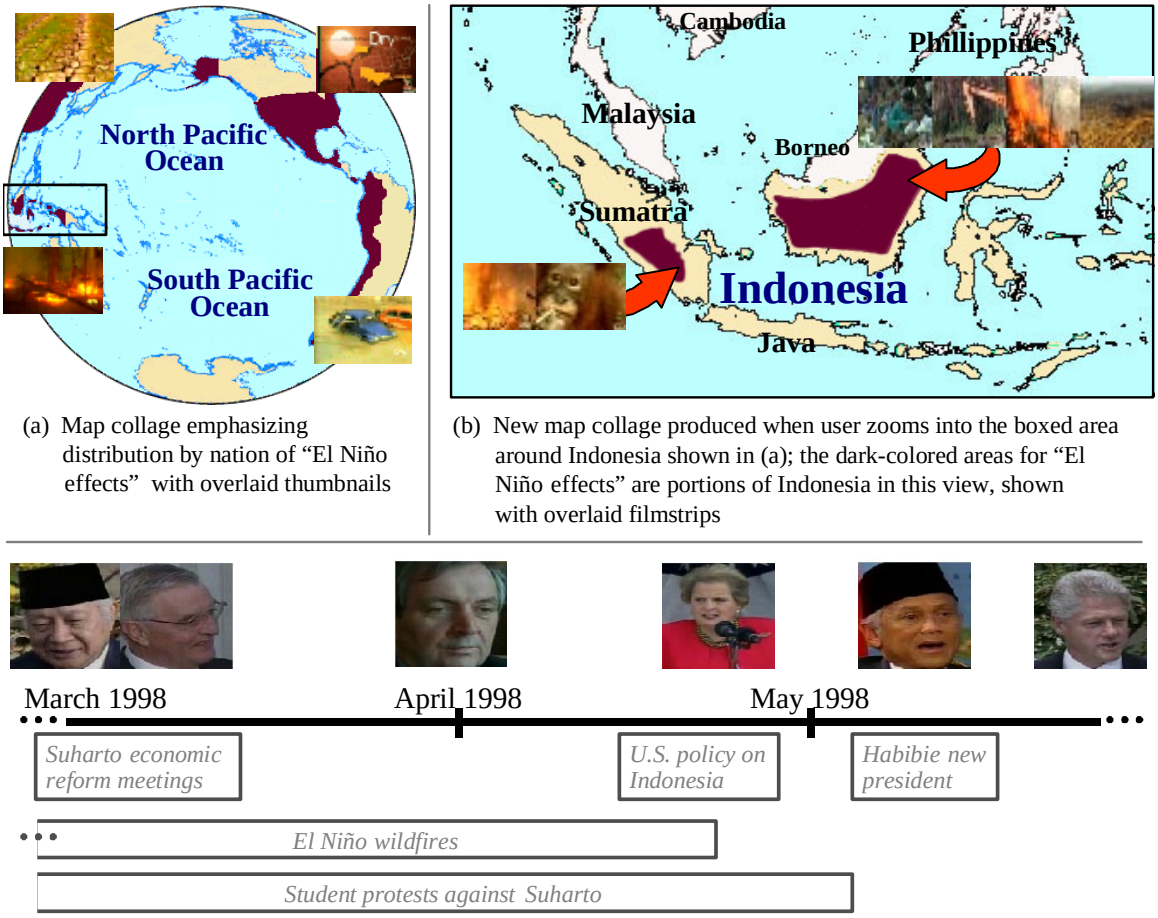


Figure 2: Informedia-II conceptual system overview.



(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 3: Multiple video information collages and their interactions

together discussing economic reform; the text names corresponding to the faces 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 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

This material is based on work supported by the National Science Foundation under Cooperative Agreement No. IRI-9817496.

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.

[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.

[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.

[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.

[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).

[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.

[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.