Journal of Universal Computer Science, vol. 8, no. 6 (2002), 602-612
submitted: 4/2/02, accepted: 3/5/02, appeared: 28/6/02 © *J.UCS*

Extracting and Visualizing Knowledge from Film and Video Archives

Howard D. Wactlar
(Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
wactlar@cmu.edu)

Abstract: Vast collections of video and audio recordings which have captured events of the last century remain a largely untapped resource of historical and scientific value. The Informedia Digital Video Library has pioneered techniques for automated video and audio indexing, navigation, visualization, search and retrieval and embedded them in a system for use in education and information mining. In recent work we introduce new paradigms for knowledge discovery by aggregating and integrating video content on-demand to enable summarization and visualization in response to queries in a useful broader context, starting with historic and geographic perspectives.

Keywords: Digital video library, visualization, metadata extraction, video summarization, video collage
Categories: H.5.1, H.3.7, I.2.7

1 Introduction

Video information mining is enabled when there exist multiple perspectives of the same event, person, place or object, each adding some content or collateral information. The perspectives may vary by location and time, by resolution and color, by media and format. The result of continuous capture of multiple sources is lots of redundant as well as marginally relevant information. Whether the content is from the 48 million hours/year of unique broadcasts from the 33,000 tv stations operating simultaneously worldwide, or from 4.5 million hours/day of surveillance video from the 14,000 air terminals worldwide, the key to useful access and correlation of this information is the ability to *index*, *search*, and meaningfully *summarize* and *visualize* it as it is captured and queried. For a number of years, Carnegie Mellon's Informedia project has been pursuing the development and integration of core technologies along with validating applications in order to approach these long-term goals.

At the lowest granularity, these combined views and perspectives enable super-resolution of images, composite panoramic synthesis, and 3D reconstruction of people and objects, stable or in motion. At higher levels, they will provide manipulable summarizations and visualizations, enabling traversal by time or geography, with drill down and roll up to any level of detail, eliminating redundancy. The challenges transcend numerous disciplines and call on significant computing and data infrastructure and standards to *capture*, *combine* and *convey* content as it is created.

2 Automated Metadata Extraction and Video Summarization in Informedia

2.1 Multi-modal Information Integration Improves Recognition and Retrieval

The Informedia Project at Carnegie Mellon University pioneered the use of speech recognition, image processing, and natural language understanding to automatically produce metadata for video libraries [Wactlar99a]. The integration of these techniques provided for efficient navigation to points of interest within the video. As a simple example, speech recognition and alignment allows the user to jump to points in the video where a specific term is mentioned, as illustrated in Figure 1.

The benefit of automatic metadata generation is that it can perform post facto analysis on video archives that were previously generated. Such archives will not have the benefit of a rich set of metadata captured from digital cameras and other sources during a digital capture or production process as will be more common in forthcoming capture devices and production environments.

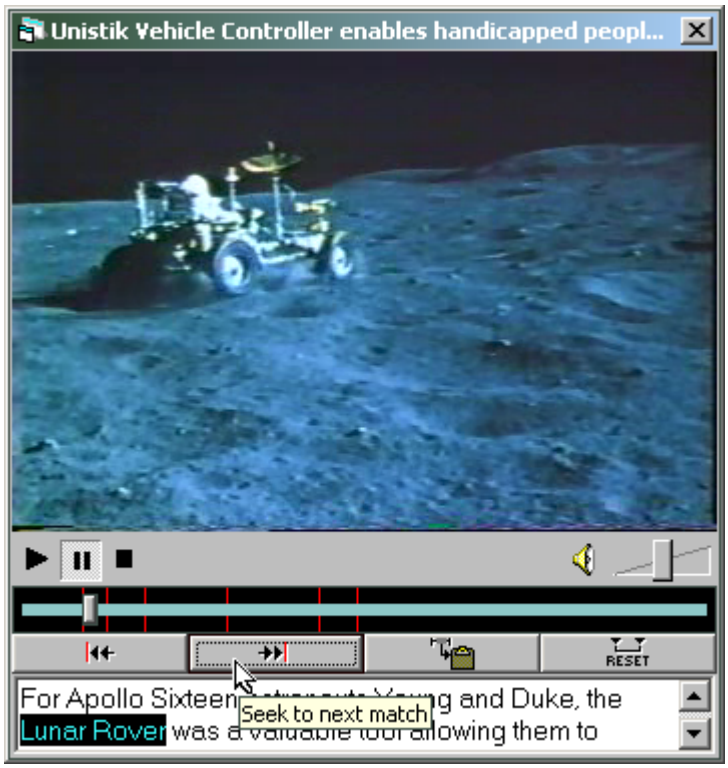


Figure 1: Effects of seeking directly to a match point on "Lunar Rover", courtesy of tight transcript to video alignment provided by automatic speech processing.

The speech, vision, and language processing are imperfect, so the drawback of automatic metadata generation as opposed to hand-edited tagging of data is the introduction of error in the descriptors. However, prior work has shown that errorful metadata can still be very useful for information retrieval, and that integration across modalities can mitigate errors produced during the metadata generation [Wactlar 99a, Witbrock 97].

More complex analysis to extract from video named entities (e.g., places, people, times) which are displayed visually (e.g., street and road signs, placards and billboards, store windows and truck panels) and spoken aurally and use them to produce time and location metadata can lead to exploratory interfaces allowing users to directly manipulate visual filters and explore the archive dynamically, discovering patterns and identifying regions worth closer investigation. For example, using dynamic sliders on date and relevance following an "air crash" query shows that crashes in early 2000 occurred in the African region, with crash stories discussing Egypt occurring later in that year, as shown in Figure 2.

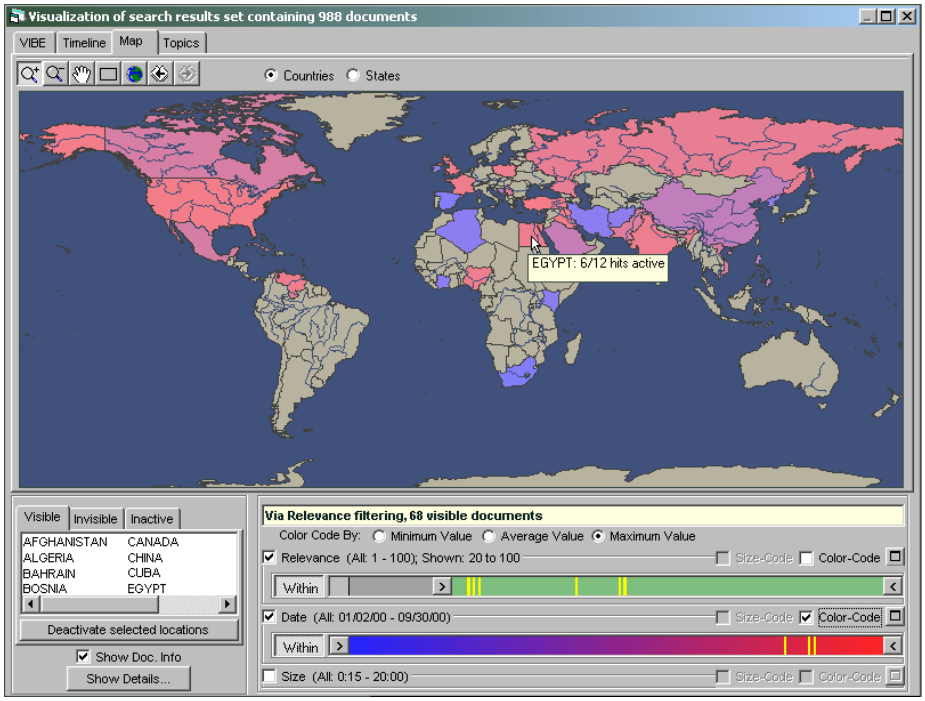


Figure 2: Map visualization for results of "air crash" query, with dynamic query sliders for control and feedback.

2.2 Generating Synthetic Perspectives from Independent Sources Separated by Time and Space

In field trials, capture from mobile video systems collected highly redundant video data. Long sequences of video contained little or no audio, with overlapping visual imagery. Filtering across space for these shots can be accomplished via image processing techniques that exploit location data acquired through GPS. One strategy is to generate a 2-D panoramic view of the environment by combining several independent views based on their time, location, and viewing angle [Gong99]. In Informedia we have used a featureless image mosaicing technique that is able to create an integrated panoramic view for a virtual camera from multiple video sequences which each records a part of a vast scene. The approach results in the following contributions: (1) The panoramic view is synthesized from multiple, independent video sequences, overcoming the limitation of existing image mosaicing techniques. (2) The panoramic view synthesis is seamlessly combined with the virtual environment creation. More specifically, each panoramic view is synthesized according to the virtual camera specified by the user, and can be visualized from an arbitrary viewpoint and orientation by altering the parameters of the virtual camera. (3) To ensure a robust and accurate panoramic view synthesis from long video sequences, a global positioning system (GPS) is attached to the video camera, and its output data is utilized to provide initial estimates for the camera's translational parameters, and to prevent the camera parameter recovery process from falling into spurious local minima. The GPS data acquisition, and the synchronization between the GPS data and the video frames are fully automated.

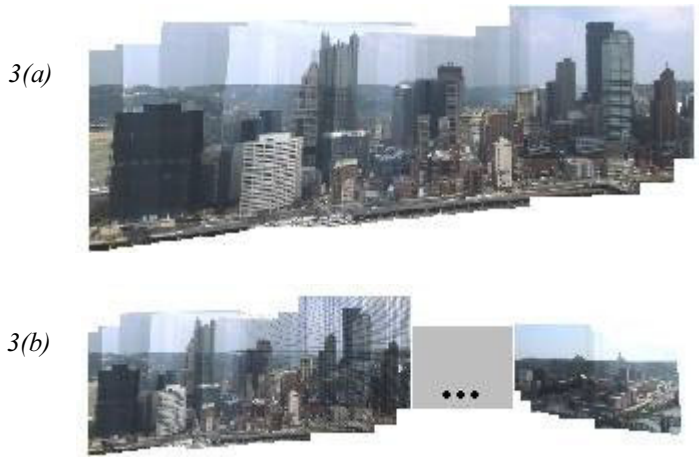


Figure 3: Panorama for: (a) one view, and (b) wider pan generated from multiple views from user-selected viewpoint.

Consider multiple capture systems recording city impressions at various viewing points; one system's output is shown in Figure 3a. GPS for each system is used to

merge the visuals so that a panoramic view as shown in Figure 3b can be constructed. The shade variations have been left in to show that the panorama was generated from individual shots captured at different times with varying amounts of sun and clouds; these shadings could be filtered out to produce a smoother panorama. The box area labeled "..." in Figure 3b indicates a portion of the cityscape for which no viewer has yet contributed information.

This technique is not only suitable for video content summarization, but will also be applicable to the areas of team collaborations in situations such as emergency response and disaster recovery when there is sufficient local computation support for it.

2.3 Enhancing Object Resolution with Separate Images

Another function we incorporate is enhancement of video resolution, exploiting multiple videos taken of the same objects and scene. This area has been generally called super resolution in image processing. Traditional super resolution techniques, however, assume that multiple inputs are subpixel shifted and the techniques utilize the generic image smoothness assumption as the mathematical basis. In contrast, we have been working on class-dependent model-based reconstruction (super resolution) of object images [Baker00a, Baker00b]. Trained with face images, this method has demonstrated conversion of a low-resolution input face image captured in the distance to a higher resolution image with which one can identify the person. We expect the same technique will also be applicable to converting text images from hardly or barely readable to fairly readable.

2.4 Detecting and Classifying Body Motion for Event Recognition and Comparison

A primary role of image understanding in Informedia is to detect and recognize objects, track and interpret changes, and reconstruct and interpret events in the video collected from the environment. By working first in constrained domains we can realize to varying degrees capabilities critical to the task of extracting and identifying individuals and/or their actions. Informedia systems have demonstrated that face detection [Cohn 01, Ratan 98, Rowley 95], based on neural networks, facilitates object-content based video retrieval and video summarization [Christel 98a, Smith 97], as opposed to conventional image-based techniques, such as color histograms. However, unlike the earlier versions of Informedia, where the information source was carefully edited, broadcast-quality video, we must now be able to process video that has been captured continuously from mobile and surveillance-like cameras. For this we will use probabilistic modeling of image properties [Schneiderman00b], image segmentation [Shi 00, Shi 98], and tracking of individuals.

Figure 4 illustrates a system for real-time separation and tracking of individuals moving across realistic backgrounds [Yang 98, Yang 99]. This type of tracking serves as a basis for other more detailed and diverse sensor capture tailored for use in the particular settings.

Figure 4: Real-time segmentation and tracking of individuals.

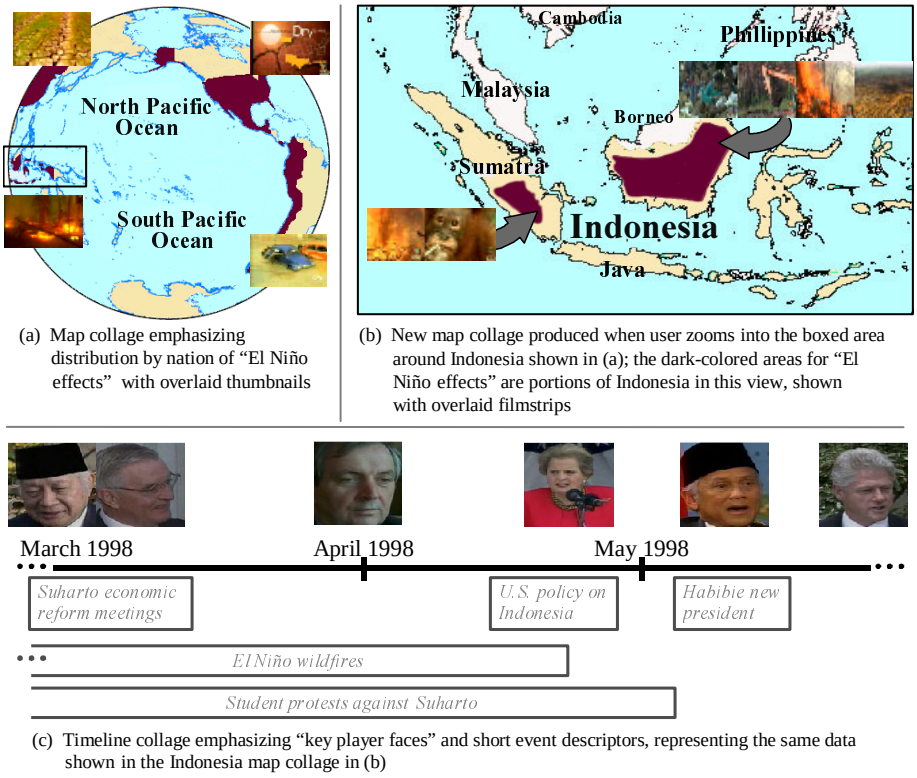
In related NSF Digital Libraries Initiative projects, Ben-Arie has investigated methods of automatically recognizing human movements such as jumping, sitting, standing, and walking [Ben-Arie 01a, Ben-Arie 01b]. Robust recognition is shown for well-framed, posed movements. In Informedia we seek to gain a sense of "activity" from video sequences by deriving and comparing similar patterns of time motion (e.g., direction and frequency) for the extracted objects, compute relative distances traversed by objects in motion, and characterizing interactions between such objects (e.g., moving together, converging, diverging). These dynamic features should enable us to recognize a similarity in the two video clips represented by the frames of Figure 5, where none of the low-level features (e.g., color and texture) are indicative.

Figure 5: Challenge of recognizing semantically similar video content.

2.5 Information Collages Summarizing and Visualizing Many Video Segments

As information capture and access evolve to become contemporaneous, ubiquitous and federated, the response even to a well formed query will generate thousands to millions of results, with a complex sense of relevance ranking. Natural language and image understanding technology may be applied to the comparison of retrieved documents so that duplication of content is eliminated or minimized in the resultant set. Clustering techniques may be applied to grouping them. However, we will need to go beyond management (e.g., ordering, sorting and comparing) of the existing content to the automated generation of new content that summarizes the result set, on-demand, in response to the query. This starts even in current research systems with the automated generation of short summaries or abstracts created using word relevance techniques, both with text and video (see [Mani/Maybury1999] for an overview of state of the art). More expansively, natural language can again be applied to the creation of full synthetic documents that summarize the "story" across multiple

source documents, even across media, and potentially across languages, by detecting differences in content between them. Extracting and resolving named entities and references to them in words and images, is fundamental to realizing such capability. This summarization will need to be of variable granularity, enabling semantic zooming interactively at any point. 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. The synthetic time-series summarization could similarly construct a timeline of events from related content retrievals to show how a story or event unfolded. Geographic and demographic information extracted from the result set may give rise to a sense of progression or causality. Alternative forms of the synthetic summary might include automatically generated, variable length, encyclopedia-like descriptions composed of words, charts and images, or an "auto-documentary" in the video medium created with video, still images and narrative extracted from the result set.



(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 6: Prototype of Informedia-II collage summaries built from video metadata.

The goal of the CMU Informedia-II Project is to automatically produce summaries derived from metadata across a number of relevant videos, i.e., an "auto-

documentary" or "auto-collage", to enable more efficient, effective information access. This goal is illustrated in Figure 6, where visual cues can be provided to allow navigation into "El Niño effects" and quick discovery that forest fires occurred in Indonesia, and that such fires corresponded to a time of political upheaval. Such interfaces make use of metadata at various grain sizes, e.g., descriptions of video stories can produce a story cluster of interest, with descriptions of shots within stories leading to identification of the best shots to represent a story cluster, and descriptions of individual images within shots leading to a selection of the best images to represent the cluster within collages like those shown in Figure 6.

Digital video will remain a relatively expensive media, in terms of broadcast/download time and navigation/seeking time. Surrogates that can pinpoint the region of interest within a video will save the knowledge-seeker time and make the distributed content more accessible and useful. Of even greater interest will be information visualization schemes that collect metadata from numerous video clips and summarize those descriptors in a single, cohesive manner. The consumer can then view the summary, rather than view numerous clips with its high potential for redundant, overlapping content and additional material not relevant to the given information need. Metadata standards are requisite to the implementation of such summaries across documents, allowing the semantics of the video metadata to be understood in support of comparing, contrasting, and organizing different video segments and frames into one presentation.

Information layout is obviously important in building the multimedia summaries. Information visualization techniques include Cone Trees [Robertson 93], Tree Maps [Johnson 91], Starfields [Ahlberg 94a], dynamic query sliders [Ahlberg 94b], and VIBE [Olsen 93]. Visualizations such as LifeLines [Freeman 95], Media Streams [Davis 94], and Jabber [Kominek97], have represented temporal information along a timeline. DiVA [Mackay 98] has shown multiple timelines simultaneously for a single video document. These are predominantly focused on ordering and clustering of terms and concepts. The higher level goals for an automated visual summarizer are captured by Tufte in his renowned works on visual information design [Tufte 87, Tufte 90, Tufte 97]. In his first volume he talks about picturing numbers, but in his second he describes the "art" of visualizing information and in his third he more comprehensively defines visualization as "a narrative that displays the causal relationships between the various working elements". He unfortunately does not describe guidelines to form the basis of an expert system to do the same, but there remains potential to construct such a rule-based process to auto-generate them interactively with the user providing some guidance and relevance feedback to the system-attempted quantization and display.

3 Challenge of a Global Infrastructure for Continuous Capture and Real-Time Analysis

One of the most significant challenges of the "information society" is keeping up with it. The goal must be that of proactive document gathering (in all media) and contemporaneous indexing and incorporation into accessible collections. This implies real-time resources to capture and index newspapers and journals as they are published, radio and television as they are broadcast, and perhaps weather satellite

data as it is transmitted. This requires progress in three domains of electronic information: (1) standards for generating corresponding metadata alongside or embedded within the content, (2) automatic subject identification and tracking for the documents and their individual sub-components, and (3) interoperability of libraries, media types, languages, and databases.

For video analysis, interpretation and indexing, this issue of scale is one of the most perplexing. Consider the challenges of full-content indexing of broadcast television where there are at least some quality control standards. The chart of figure 7, derived from the Berkeley *How Much Information* project [Lyman 00], shows an annual production of approximately 48,000,000 hours, or 24,000 terabytes of storage with lossy MPEG-1 compression. If we extrapolate to the capture from surveillance cameras at the 14,000 air terminals worldwide, that same amount of video (4.8M hours) is generated per day. Even if we were capable of sufficient processing in real-time (1 hour per hour of video) to analyze and index the content, many comparable systems must be running in parallel. If we are to summarize or even just search across these parallel but autonomous systems, common metadata must be extracted with common criteria, in a common lingua. This implies massive distributed computation and storage, with standards for metadata description, criteria for extraction and identification, and protocols for media retrieval and conversion.

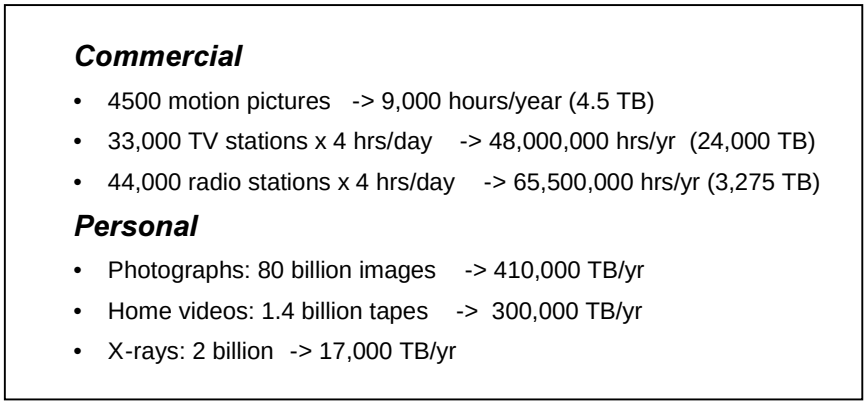


Figure 7: Annual Video and Audio Production

Acknowledgements

This material is based on work supported by the Advanced Research and Development Agency (ARDA) under their Video Analysis and Extraction (VACE) program and by the National Science Foundation (NSF) Digital Libraries Initiatives I and II 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, MA, 1994, pp.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, MA, 1994, pp.365-371.

[Baker00a] Baker, S. and Kanade, T. "Limits on Super-Resoluation and How to Break Them," *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[Baker00b] Baker, S. and Kanade, T. "Hallucinating Faces," *Fourth International Conference on Automatic Face an dGesture Recognition*, March, 2000.

[Ben-Arie01a] Ben-Aire, J., Pandit, P., and Rajaram, S. "Design of a Digital Library for Human Movement," *Joint Conference on Digital Libraries (JCDL'01)*, Roanoke, VA, June 24-28, 2001.

[Ben-Arie01b] Ben-Arie, J., Pandit, P., and Rajaram, S. "Human Activity Recognition Employing Indexing," *The International Conference on Computer Graphics and Imaging (CGIM'02)*, Honolulu, HI, August 13-16, 2001.

[Christel98a] Christel, M., Smith, M., Taylor, C.R., and Winkler, D. "Evolving Video Skims into Useful Multimedia Abstractions," *Proceedings of the ACM CHI'98 Conference on Human Factors in Computing Systems*, Los Angeles, CA, April, 1998, pp.171-178.

[Cohn01] Cohn, R., Shi, J., R. Gross. "Where to go with Face Recognition," *IEEE Conference on Computer Vision and Pattern Recognition 2001 (CVPR'01)*, Third Workshop on Empirical Evaluation Methods in Computer Vision, Hawaii, US, December 9-14, 2001.

[Davis94] Davis, M. "Knowledge Representation for Video," *Proceedings of AAAI '94*, 1994, pp.120-127.

[Freeman95] Freeman, E. and Fertig, S. "Lifestreams: Organizing your Electronic Life," *AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval*, Cambridge, MA, November, 1995. http://www.halcyon.com/topper/jv6n1.htm.

[Gong99] Gong, Y., LaRose, D., Proiett, G. "A Robust Image Mosaicing Technique Capable of Creating Integrated Panoramas," *IEEE 1999 International Conference on Information Visualization*, London, UK, July 14-16, 1999, pp.12-29.

[Johnson91] Johnson, B. and Schneiderman, B. "Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures," *Proc. IEEE Visualization '91*, San Diego, CA, October, 1991, pp.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, pp.19-26.

[Lyman00] Lyman, P., Varian, H. How Much Information, 2000, http://www.sims.Berkeley.edu/how-much-info.

[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, pp.416-423.

[Mani/Maybury1999] Mani, I., Maybury, M., eds. Advances in Automatic Text Summarization, July 1999, MIT Press.

[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*, 1993, **29**(1), 69-81.

[Ratan98] Ratan, A.L., Grimson, W.E.L., and Wells, W.M. "Object detection and localization by dynamic template warping," *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June, 1998.

[Robertson93] Robertson, G., Card, S., and Mackinlay, J. "Information Visualization Using 3D Interactive Animation," *Communications of the ACM*, 1993, **36**(4), 56-71.

[Rowley95] Rowley, H., Baluja, S., and Kanade, T., "Human Face Detection in Visual Scenes," School of Computer Science Technical Report CMU-CS-95-158, Carnegie Mellon University, Pittsburgh, PA, 1995.

[Schneiderman00b] Schneiderman, H. and Kanade, T. "Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Santa Barbara, CA, June, 2000.

[Shi00] Shi, J. and Malik, J. "Normalized Cuts and Image Segmentation," *Accepted and to appear in IEEE PAMI*, 2000.

[Shi98] Shi, J. and Malik, J. "Motion Segmentation and Tracking Using Normalized Cuts," *International Conference on Computer Vision(ICCV)*, Bombay, India, January, 1998.

[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, pp.775 - 781.

[Tufte87] Tufte, E.R. "The Visual Display of Quantitative Information," 1st ed. Graphics Press, 1987.

[Tufte90] Tufte, E.R. "Envisioning Information." Graphics Press, 1990.

[Tufte97] Tufte, E.R. "Visual Explanations: Images and Quantities, Evidence and Narrative." Graphics Press, 1997.

[Wactlar99a] 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*, 1999, **32**(2), 66-73.

[Witbrock97] Witbrock, M.J. and Hauptmann, A.G. "Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents," *Proceedings of the 2nd ACM International Conference on Digital Libraries*, Philadelphia, PA, July, 1997, pp.30 - 35.

[Yang98] Yang, J., Stiefelhagen, R., Meier, U., and Waibel, A. "Visual Tracking for Multimodal Human Computer Interaction," *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, Los Angeles, CA, April 18-23, 1998.

[Yang99] Yang, J., Zhu, X., Gross, R., Kominek, J., Pan, Y., and Waibel, A. "Multimodal People ID for a Multimedia Meeting Browser," *ACM Multimedia '99*, Orlando, FL, October 30-November 5, 1999.