

Interactive Maps for a Digital Video Library

Michael G. Christel, Andreas M. Olligschlaeger,
and Chang Huang
Carnegie Mellon University

To improve library access, the Informedia Digital Video Library uses automatic processing to derive descriptors for video. A new extension to the video processing extracts geographic references from these descriptors. The operational library interface shows the geographic entities addressed in a given story, highlighting the regions discussed in the video through a map display synchronized with the video playback. The map can also serve as a query mechanism, allowing users to search the terabyte library for stories taking place in a selected area of interest.

The Informedia Project at Carnegie Mellon University investigates the utility of speech recognition, image processing, and natural language-processing techniques for improving search and discovery in the video medium.¹ Since 1994, the project has digitized, in MPEG-1 format, news video from CNN as well as documentary and educational video from the British Open University, QED Communications, the Discovery Channel, and numerous US government agencies such as NASA, the National Park Service, and the US Geological Survey. The resulting digital video library now contains over 2,000 hours of video and continues to grow at a rate of more than ten hours per week.

The sheer volume of the video data reveals new issues for digital-video-library interfaces of the future. A good query engine is not sufficient for exploration because often the candidate result sets grow in number as the library grows. Interfaces for browsing both the library and defined library subsets—such as the results from a query—become increasingly important.

Users are interested in quickly finding the set of video stories or segments relevant to their needs. When the library was on the order of a hundred hours, a statistical word query engine adequately provided this focus. Users entered text queries and received a small set of segments sorted by the query engine's relevance score. The engine presented alternate representations of the video that users could view in less time. This aided in deciding which segments were worth a full viewing.² Figure 1 shows this style of library interface.

When the library grew to a thousand hours, queries returned hundreds of segments, overwhelming users much like Web search engines that return lists whose length and default ordering no longer meet user needs. We developed an information-visualization interface to let the user browse the whole result space without resorting to the time-consuming and frustrating traversal of a list of results.³ The visualization techniques employed let the user browse and retrieve video from the Informedia library based on date (when) and word occurrences (what). However, we realized that a potentially rich vein of information was overlooked in our corpus. Many documentaries and most news stories deal with location information (where). We could also use this information dimension in presenting overviews of the video content, in summarizing multiple video segments, and as a query mechanism in finding segments dealing with a particular region of interest. This article discusses using interactive maps with the Informedia Digital Video Library.

Extracting video geographic references

The transcript of the narrative is the greatest source of geographic reference information for videos in the Informedia library. We use the Carnegie Mellon University Sphinx speech-recognition engine to transcribe the content of video material. Our word error rate is proportional to the amount of processing time devoted to the task. For example, a processing effort of 30 times real time using evening news broadcasts results in a word error rate of approximately 35 percent (including insertions, deletions, and substitutions), whereas an effort of 300 times real time yields a word error rate of about 24 percent.¹ If closed-captioned text exists for a video, we integrate it with the output of the recognizer. The final text transcript is synchronized at a word level to the video through Sphinx processing. Hence, if the narrative mentioned, "Heavy snows in Switzerland caused...", this process would capture the video time when "Switzerland" was mentioned.

While the transcript provides the primary source of geographic references, it's not the sole source. Often a location name and perhaps a person's name are overlaid on the video, especially for news. The Informedia video optical character recognition (VOCR) process⁴ identifies video frames containing probable text regions, in part through horizontal differential filters with binary thresholding. VOCR then filters the probable text region across the multiple video frames to

improve the image quality used for OCR-processing input. Commercial OCR software converts the final filtered image of alphanumeric symbols into text. The VOOCR-produced text is another potential source of geographic references. For example, a video segment discussing volcanic activity includes shots of lava with the overlaid text stating “Mount Etna, Italy.” While the transcript text was associated to video times through Sphinx speech alignment, the VOOCR text is associated to video times through image processing, which identifies the frames containing the probable text regions.

Other descriptors for the Informedia library contents—that is, metadata—include production notes, automatic topic identification, and user annotations whereby the user can type or speak comments pertaining to a specified portion of video. These additional sources of text may also include location information such as a user comment like “add Los Angeles to my itinerary.”

Figure 2 shows the process for adding geographic references to video segments. Matching addresses, or in this case, named places, to their spatial coordinates is known as geocoding.⁵ We begin geocoding by using the text metadata as the source material to process. We use a set of known places along with their spatial coordinates, known as a gazetteer, to resolve geographic references. The Informedia library gazetteer is derived from a data subset contained in the world gazetteer from Environmental Systems Research Institute (ESRI)⁶ consisting of approximately 300 countries, states, and administrative entities, and 17,000 major cities. We add a postprocessing step that expands the gazetteer to include related terms. For example, “Canada” will identify that country, but so will “Canadian.” Further postprocessing removes gazetteer entries that are also common English words from subsequent matching. For example, the word “of” is removed from consideration, even though there is a city “Of” in Turkey. The text metadata associates text with video times and is then matched to terms in the so-called geographic codebook, which maps geographic text terms to latitude and longitude. The end result is the tagging of video sequences with latitude and longitude.

Geocoding addresses in fixed field format with known address components such as street num-



Figure 1. Informedia interface with images representing six video segments and text headline for the second segment.

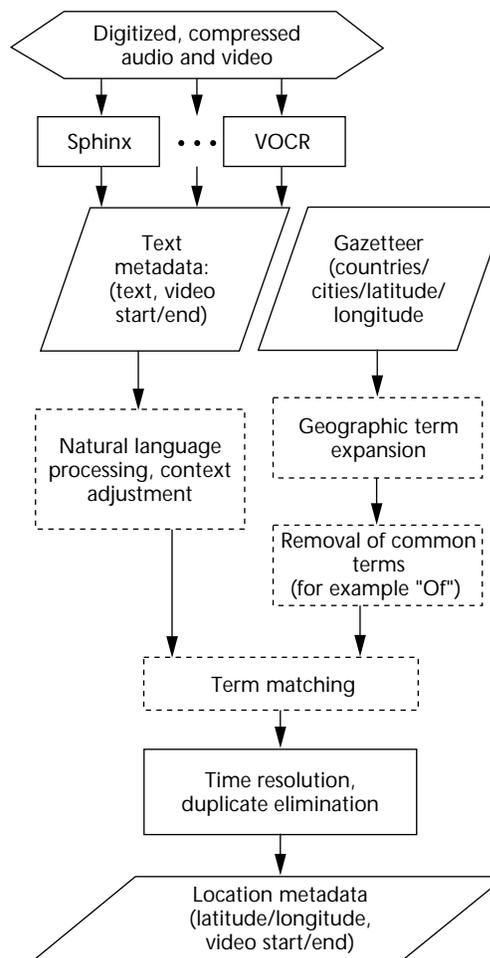


Figure 2. Process for adding location information to video.

Table 1. Example entries (not all rows and columns shown) for results of geocoding.

Text	Type	X	Y	Administrative Entity	Country	Segment ID	Start Time (ms)	End Time (ms)
Kenya	Country	37.915	2.605	Eastern	Kenya	CWT0Z4	204704	366366
India	Country	82.410	20.605	Madhya Pradesh	India	CWT0Z4	362095	362095
Nairobi	City	36.804	-1.270	Nairobi	Kenya	CWT0Z4	195329	375209
China	Country	108.986125	36.628	Shaanxi	China	CWT1F16	1665665	1667667
South Korea	Country	127.772	36.711	Ch'ungCh'ong-Bukto	South Korea	CWT1F16	1738805	1738805
Japan	Country	137.795	35.551	Chubu	Japan	CWT1F16	1635068	1721054
North Korea	Country	126.538	39.105	P'Yongan-Namdo	North Korea	CWT1F16	1629530	1754087
Brussels	City	4.368	50.837	Bruxelles-Brussel	Belgium	CWV1031	3248748	3250250

ber, street name, and city has been well researched and documented.⁵ However, extracting named places from free-form text such as video transcripts is a relatively new idea and far more complicated. Addresses tend to be unique, for example, there is only one place called 123 Main Street in Anytown, and, providing they are in a consistent format, we can simply match them against a known set of addresses with known coordinates. With free-formatted text, on the other hand, it's not known a priori which portion of the transcript contains references to places.

It's therefore necessary to parse the text metadata to extract candidate portions of sentences that may represent places. Currently we achieve this by eliminating words that are most commonly used, such as "and" and "the," and examining the remainder of the text. The items surrounded by dotted boxes in Figure 2 indicate portions of the process that can greatly improve with the addition of new knowledge, including this parsing step.

If we know that "Of" is part of the phrase "Of, Turkey" in the text metadata, then we use that contextual information to resolve that "Of" is a city, where "of" in the phrase "things of the past" wouldn't resolve to any location information. Similarly, context can help with term matching. The proper noun "Washington" could refer to a US western state, the US capital city, or to a person's name. Contextual cues such as "Seattle, Washington," "Washington, DC," and "since the time of George Washington" distinguish these different meanings. Through contextual analysis on the source metadata, the system can better classify proper nouns as persons' names or as places and more accurately assign location information.

We've found that hidden Markov models (HMMs) can achieve this level of analysis. They are effective for automatically tagging entities, including locations, in text output from speech

recognizers, where such text lacks punctuation cues.⁷ We are currently exploring the use of HMMs for the tasks in Figure 2's dotted boxes. We expect that the HMM approach will deliver more accurate geographic referencing than our current quick processing baseline, which uses little context adjustment, removes all common and ambiguous terms from the gazetteer match set, and enforces strict term matching. We need further studies to determine the accuracy and benefits associated with the additional processing.

Enabling geographic reference use

The unit of information retrieval in the Informedia library is the video segment, which (when segmentation strategies work to perfection) contains a single story. On average, each hour of broadcast news consists of 20 segments. For each segment, a list is constructed during the geocoding process consisting of places mentioned in a segment. A place may be named more than once in a segment, but it's represented only once in the segment's list. The number of references is included in each place's entry to enable subsequent interfaces to emphasize locations visually, based on how frequently the places are mentioned.

The geocoding process establishes a relationship between the video and place names. For a given video time interval, it identifies the place names referenced in that interval, and for a given place, it quickly accesses its time interval. For places identified in transcript metadata, the sentence or sentences where each place is mentioned is tracked. If a place is only mentioned once, the beginning and ending times of the sentence it's cited in are used as the time interval. Time is measured in milliseconds from the beginning of the video. If a place is mentioned more than once, the time span from the start of the first sentence to the end of the last sentence citing the place is used. The timing of transcript words to the video

is accessible via Informedia's Sphinx speech-recognition processing. If a place is identified from other text metadata such as VOCR, then the start and end time associated with the text for that metadata is used. For VOCR, this time span approximates the duration when the overlaid text appears in the video. As with transcripts, if a place occurs multiple times, then its time span extends from the start time for its first mention to the end time for its final mention in the video segment.

The resulting list is written to a database (currently in DBASE III .dbf format), and converted to a shape file in ESRI format using the geocoding capabilities of that company's ArcView and gazetteer. Finally, the shape file is indexed geographically to optimize spatial searches. Table 1 shows an example of the output from the geocoding algorithm.

Map representation of a video segment

Figure 3 shows an example of how the geocoded information is incorporated into the Informedia interface. When a user plays back a video segment, an optional window pops up that contains a map displaying all places discussed in that segment. We use functionality provided by the ESRI MapObjects library in creating this interface, written using Microsoft Visual Basic. This overview shows that the given segment covers the countries of Kenya and India and the city of Nairobi.

The user can interact with the map through toolbar icons that enable zooming in and out, panning, selecting search areas, and returning to the full map extent showing the countries of the world.

Aside from the spatial search, these icons and their underlying operations are supplied directly by the MapObjects library. By integrating this functionality into a digital-video-library interface, the user can access details relevant to the video content. For example, the video segment in Figure 3 is paused where the CNN footage shows a map, illustrating that the video's producer recognizes

the importance of geographic detail. The geocoding process of Figure 2, generic map functionality like that provided by MapObjects, and user interaction can automatically extract and display this same detail. In Figure 3 the user chose to enable the tips text display. When the mouse is paused over Ethiopia, that country's name appears in a tips text window. The resulting map display looks very similar to the produced CNN shot.

The real value of geocoding and of the map interface lies in displaying location information for video segments to which the producer didn't add a map. Not every news story has an embedded map that becomes part of the broadcast, but through our geocoding maps can be automatically produced to

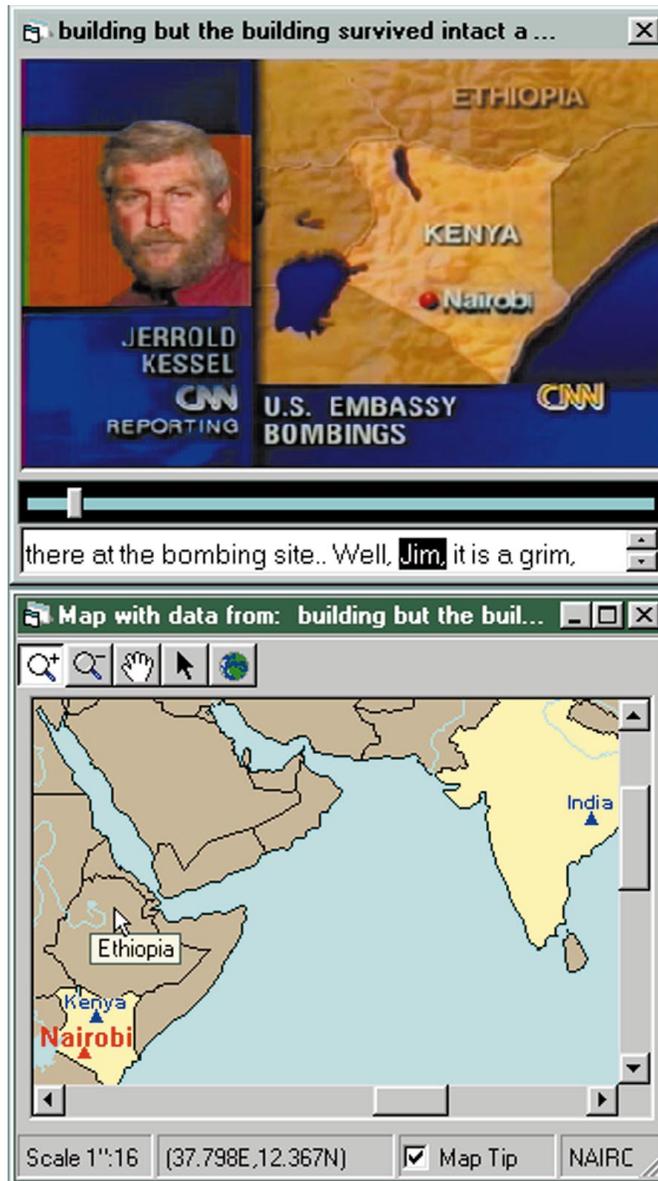


Figure 3. Map overview of location references for a video segment.

Figure 4. Animated map that highlights as video plays (also showing scrolling transcript).



reflect the areas discussed within each story. Another benefit is the user can interact with the interface map using the toolbar icons to get additional detail. No such interaction is possible when an image of a map is encoded as part of the video stream. For example, the user could zoom into the Kenya area of Figure 3 to see a map display with

other Kenyan details such as the port city of Mombassa.

The maps accompanying videos are not static displays. They are animated in synchronization with video playback. Places are highlighted on the map when they are discussed. For countries and administrative areas such as states or provinces, the interface highlights areas contained within their respective bounding polygon. Areas covered at some time during a video segment are colored yellow, and when the video frames play in which an area is discussed, that area is colored orange. For cities and other places, the marker color is changed from blue to red and the name is shown. A user can glance at the map to see the areas of current focus. For example, Figure 4 shows a story where the current focus is on North Korea and Japan, with China, South Korea, and Russia mentioned elsewhere in the story. The highlighting changes over time to show the story flow within the video segment.

A classic research area in cartography studies the accurate and effective display of data on a map; animated maps are just beginning to be addressed. Depending on the number of features, it's important to avoid so-called noisy maps that show a lot of detail but are difficult to read. Thus, we are currently limiting the appearance of text labels for city and administrative areas to those times when they are discussed in the accompanying video. Country labels are always displayed. In Figure 3, the video is paused at a point where Nairobi is being discussed, so that label appears in red. If the segment of video no longer actively references Nairobi, as indicated by the times stored in the database (see

Table 1), then the Nairobi location marker remains visible but in a different color and without its text. This strategy is useful when a video segment has many city references across a broad area, making display of all the references labels difficult without significant overlap.

Since the amount of information displayed for

each video segment varies, we plan to let the user change the map-display default settings. These settings include the types of entities to mark (cities and countries), the symbols and labels to use, and the colors and styles for marking both the overview (as in Figure 3) and the highlights for a given video time (as in Figure 4). In addition, the default settings themselves may depend on the type of video shown. For example, our current default settings may be adequate for broadcast news footage that tends to mention only major cities and countries. However, a corpus that contains only videos for a European soccer league may have only city labels displayed by default.

Accessing video through spatial queries

The maps in the Informedia interface are not merely for presentation but can also specify a location query. By choosing an arrow icon in the toolbar for the map window, the user can drag a rectangular region on the map that serves to identify a region of interest.

In Figure 5 the user has selected the region encompassing the Netherlands, Belgium, and Luxembourg. Functionality provided by the MapObjects library performs the query against the video library's associated geographic references. The Informedia library currently contains around 40,000 segments, with geocoding producing nearly 20,000 location references such as those shown in Table 1. Searching against this corpus provides, within a few seconds, results displayed with headlines and representative thumbnail images, just as results for text queries are displayed (as shown in Figure 1). Feedback is provided on the map to indicate the locations within the specified query that actually produced results. In Figure 5 this feedback is shown as white circles for the three countries, indicating that each country was found somewhere in the resulting set of 46 video segments. The headline for the sixth result (shown in the text box) indicates that it's a story on a Belgian roller coaster. If this video is subsequently played, the map would change to an image like that of Figure 4, with Belgium initially colored to indicate that it's mentioned somewhere in the story and with that color changing during the frames when it's actually discussed.

Figure 4 originated with the spatial query shown in Figure 6. When users search with words, the matching word locations are marked with the time it appears.² These match locations are indicated with vertical lines drawn in the video scrollbar shown beneath the video playback area. The

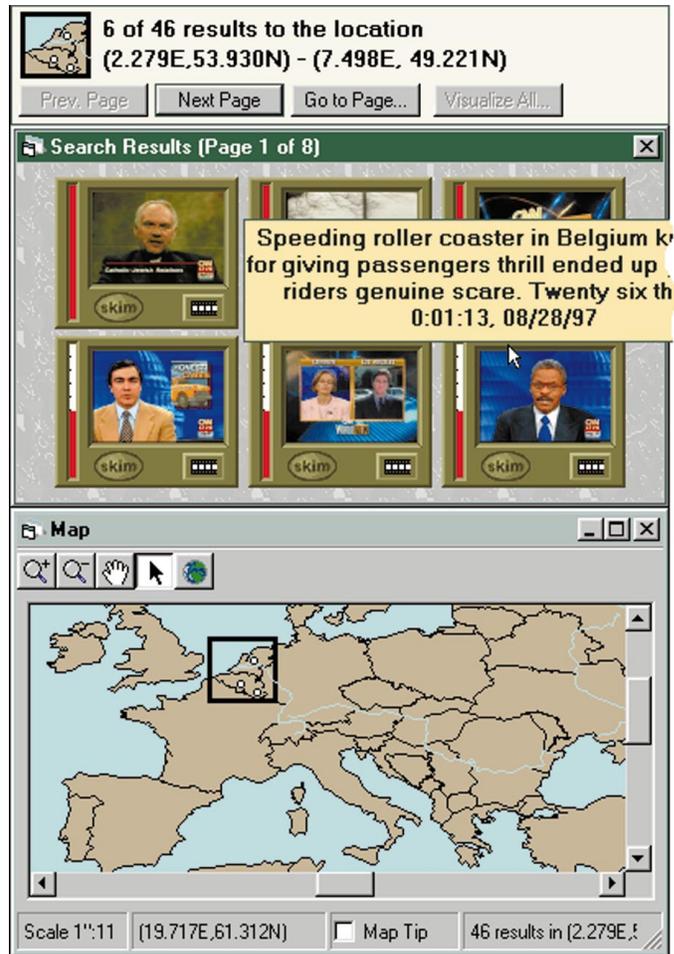


Figure 5. Using the map to request video segments for a specified location.

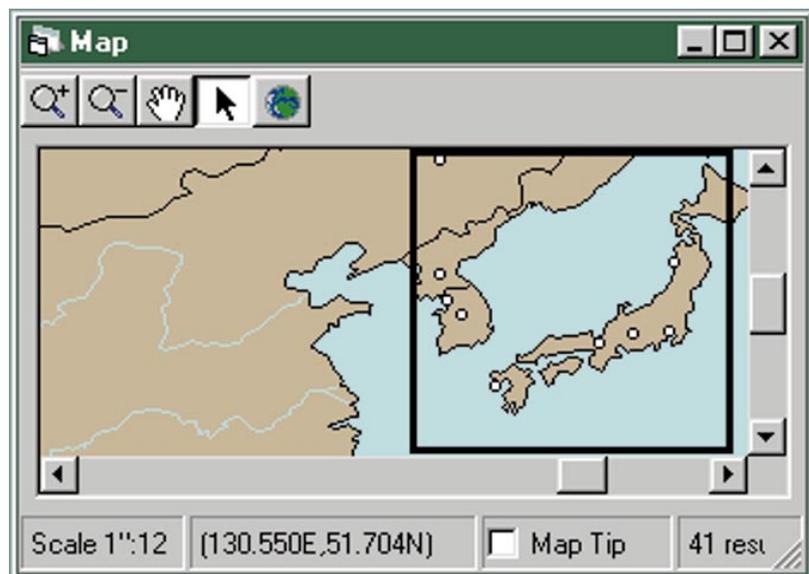


Figure 6. Spatial query that produced the video segment of Figure 4.

We can improve the presentation of information by using not only the information dimensions of date, time, and topic, but also the location dimension.

same indication is used when the video segment is found with a spatial query. Numerous matching places may appear within a given rectangular query area with each place having associated video times. These times are used to draw the match lines. In Figure 4, the three match lines on the video scroll bar correspond to the times when North Korea, Japan, and South Korea are mentioned in the segment, these three places being in the region of interest defined by the query shown in Figure 6. Through the use of additional controls—the left and right arrow buttons—the user can quickly seek and go to the section of video where the locations of interest are discussed.

While this discussion has focused on query regions shaped as rectangles, we recognize the value of supporting more powerful spatial query mechanisms. For example, the user might click on a country to select it, shift-click to designate a number of countries, or use a bounding polygon to choose a number of countries within a region. All matches to cities, political entities, countries, and other geographic references within that area would then be included in the returned set of video segments. Our Informedia interface also allows for spatial searches that leverage from additional map information, such as country boundaries.

There is also a hidden assumption by the user that the drawn area will specify a request for all locations contained within that area. In specifying an area request the user might expect an “overlaps” or “contains” relationship rather than the “contained by” relationship. The work of the Alexandria Digital Library,^{8,9} addressing this ambiguity in formulating spatial queries holds great promise for use with geocoded digital video libraries.

Future work

We can improve the presentation of information by using not only the information dimensions of date, time, and topic, but also the location dimension. Results from word queries could be visualized on a map, revealing interesting patterns for discovery. For example, a search on volcanoes might show the stories are concentrated in a ring around the Pacific Ocean. Empowering the user to manipulate all the metadata for the video library could reveal time-based patterns.

Another interesting area of work for the Informedia interface is enabling mixed modal query. For example, it could find all the video segments mentioning “famine” for a specified area on the map, or finding all the people who match a given face image for a specified country.

Ongoing work by Informedia researchers has focused on automatically providing summaries across collections of video segments rather than relying solely on user ingenuity and diligence. We can produce automatic visual digests, referred to as “video collages,” from richer sources of automatically derived metadata such as people, events, affiliations, and topics as well as location and time.¹⁰ These collages reflect the distillation of information across multiple video segments. For example, a search on El Niño effects could show hot spots of activity in one geographic area for early 1997 but new hot spots in different areas for late 1998. The collage becomes a video magazine that summarizes all the salient information, while also responding to user interaction by emphasizing particular details of interest. With the inclusion of geographic references into such collages, the Informedia interface can better serve users in their quest for search and discovery in the video medium. MM

Acknowledgments

This article is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under SPAWAR contract No. N66001-97-D-8502 and on work supported by the National Science Foundation (NSF) under Cooperative Agreement No. IRI-9817496 (<http://www.dli2.nsf.gov>). Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of DARPA or NSF. The contributions of Informedia partners and team members have been invaluable; visit <http://www.informedia.cs.cmu.edu> for a complete list of participants, along with further information on Informedia-related efforts.

References

1. H.D. Wactlar et al., "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library," *Computer*, Vol. 32, No. 2, 1999, pp. 66-73.
2. M.G. Christel, D.B. Winkler, and C.R. Taylor, "Multimedia Abstractions for a Digital Video Library," *Proc. ACM Conf. on Digital Libraries*, ACM, New York, 1997, pp. 21-29.
3. M.G. Christel and D.J. Martin, "Information Visualization within a Digital Video Library," *J. Intelligent Info. Systems*, Vol. 11, No. 3, 1998, pp. 235-257.
4. T. Sato et al., "Video OCR for Digital News Archive," *Proc. Workshop on Content-Based Access of Image and Video Databases*, IEEE Computer Society, Los Alamitos, CA, 1998, pp. 52-60.
5. A. Olligschlaeger, *Spatial Analysis of Crime Using GIS-Based Data: Weighted Spatial Adaptive Filtering and Chaotic Cellular Forecasting with Applications to Street Level Drug Markets*, doctoral dissertation, H. John Heinz III School of Public Policy and Management, Carnegie Mellon University, Pittsburgh PA, 1997.
6. Environmental Systems Research Institute, <http://www.esri.com/>.
7. F. Kubala et al., "Named Entity Extraction from Speech," *Proc. DARPA Workshop on Broadcast News Understanding Systems*, Distr. by Morgan Kaufmann, copyright 1998 by DARPA, pp 287-292.
8. K. Beard, T. Smith, and L. Hill, "Meta-information Models for Georeferenced Digital Library Collections," *Proc. Second IEEE Metadata Conf.*, IEEE Computer Society, Los Alamitos, CA, 1997, <http://computer.org/conferen/proceed/meta97/papers/kbeard/kbeard.html>.
9. T.R. Smith et al., "A Digital Library for Geographically Referenced Materials," *Computer*, Vol 29, No. 5, 1996, pp. 54-60.
10. M.G. Christel, "Visual Digests for News Video Libraries," *Proc. ACM Multimedia Conf.*, ACM, New York, 1999, pp. 303-311.



Michael G. Christel is a senior systems scientist at Carnegie Mellon University's School of Computer Science. His research focus includes multimedia interfaces and information visualization. He has developed and evaluated Informedia digital video library interfaces since 1994. He remains interested in the educational use of technology and has explored the use of video for synthetic interviews.



Andreas M. Olligschlaeger is a systems scientist at Carnegie Mellon University's School of Computer Science and has a joint appointment with the Heinz School of Public Policy and Management, also at Carnegie Mellon. His primary areas of research are information systems for law enforcement, geographic information systems, and space and time forecasting models.



Chang Huang is a research programmer with the Informedia Project in the Computer Science Department at Carnegie Mellon University. She has designed and developed map interfaces for broadcast video libraries, as well as for unstructured collections of field-captured video coupled with GPS data. Her research interests include geoprocessing technology and scalability of information-access interfaces.

Readers can contact Christel at Carnegie Mellon University, Computer Science Dept. and HCI Institute, Pittsburgh, PA 15213, e-mail christel@cs.cmu.edu.