

**Informedia  
Digital Video Library System**

**NSF Cooperative Agreement No. IRI-9411299  
Annual Progress Report  
February 1997**

**Carnegie Mellon University  
Computer Science Department  
Pittsburgh, PA 15213-3890**

**Principal Investigators:**

**Takeo Kanade**  
Robotics Institute

**Michael Mauldin**  
Center for Machine Translation

**Raj Reddy**  
School of Computer Science

**Marvin Sirbu**  
Information Networking Inst.

**Scott Stevens**  
Software Engineering Inst.

**Doug Tygar**  
Computer Science Department

**Howard Wactlar**  
School of Computer Science

## **1. Overview**

The Informedia library project will establish a large, on-line digital video library by developing intelligent, automatic mechanisms to populate the library and allow for full-content and knowledge-based search and retrieval via desktop computer and metropolitan area networks.

The distinguishing feature of our technical approach is the integrated application of speech, language and image understanding technologies for efficient creation and exploration of the library. Using a high-quality speech recognizer, the sound track of each videotape is converted to a textual transcript. A language understanding system then analyzes and organizes the transcript and stores it in a full-text information retrieval system. Likewise, image understanding techniques are used for segmenting video sequences by automatically locating boundaries of shots, scenes, and conversations. Exploration of the library is based on these same techniques. Additionally, the user interface will be instrumented to investigate user protocols and human factor issues peculiar to manipulating video segments. We will implement a network billing server to study the economics of charging strategies and also incorporate mechanisms to ensure privacy and security.

The Informedia Project has industry partners who are committed to provide substantial resources and base technology. They will evaluate commercial opportunities for the underlying technology and for the provision of information services. Together these companies span the requisite disciplines for digital video library commercialization.

## **2. Research and Testbed Summary**

### **2.1. Speech, Language and Image Understanding for Library Creation and Exploration**

#### **2.1.1 Speech recognition quality effect on information retrieval of spoken language documents**

Known item retrieval effectiveness on 100 queries was measured for identical source collections — speech recognized, closed captioned, and fully correct text documents — each embedded in a corpus of text documents. Corpus sizes also ranged from 100 to 12000 documents. Retrieval effectiveness was measured using both average inverse rank, and, for a corpus of 600 documents, human relevance judgments. Average inverse rank was a comparable measure to precision in all cases, supporting its use for measurements of retrieval effectiveness on larger corpora.

Speech recognizers with error rates ranging from 0% to 100% were simulated by selectively correcting errors in the SphinxII recognizer output, or by increasing the number of deletion errors. The documents generated in this way were used in IR experiments, allowing for an estimation of retrieval effectiveness at various retrieval accuracies. The results of this model suggest that at speech recognition error rates of 25% or less, information retrieval effectiveness should be substantially equivalent to that for correct text documents. With a word error rate of 50%, retrieval of speech documents was 85-90% as successful as retrieval of text documents for a corpus of 600 documents, but effectiveness fell off more rapidly for speech than for text documents. In the last year, the CMU Sphinx recognition accuracy has improved from about 50% error for broadcast news to about 35% error (SphinxIII) at some cost in speed.

Our pursuit search engine derivative uses keyword spotting, stop words, and synonyms, and grants a bonus to search terms with primacy. Testing various permutations of the search engine features revealed that the best system uses a combination of stemming, stopwords, *tf/idf*, document length normalization, and document weighting.

We further tested our search engine on progressively larger corpora. Each data set used the same 105 prompts for which corresponding stories were either created manually or through a speech recognizer. In this case, though, three corpus sizes were generated by adding fully corrected transcripts. Average rank figures were computed using the best retrieval system. What follows is the average correct rank for the set of 105 stories retrieved from the different size databases.

Average rank based on:	602	2,600	12,000 stories
Manually Prepared Transcript	2.32	5.65	9.34
Speech Generated Transcript	7.89	31.16	60.19

Initial investigations showed that about half the difference in retrieval effectiveness for spoken documents vs. text documents was due to words misrecognized because the correct word was not in the recognition vocabulary. A possible way of correcting this is to use initial speech recognition results to retrieve similar documents from a large corpus, and use those documents to add to the vocabulary before re-recognition. An enlargement of this method is to use the retrieved documents not just as a source of new vocabulary, but as a source of new language model (unigram, bigram and trigram) probabilities that can be interpolated with a general language model to improve recognition. Initial experiments suggested that an improvement of one or two percent in recognition accuracy could be achieved using this method.

### **2.1.2 Natural language processing for information retrieval**

In cooperation with Milind Pandit from Intel, latent semantic indexing (LSI) was used to find semantically similar words within the news corpus. The most similar of these words are currently being used as synonyms in the retrieval system. Further experiments to evaluate the quality of the synonyms generated are planned.

### **2.1.3 Video summarization and abstraction — the video skim**

We continue to improve skim through multiple approaches: improved keyword/keyphrase detection through an enhanced document corpus used for the *tf/idf* relevance weighting; detection of proper names to identify people; detection of object motion to discriminate between actions and camera motion; identification of multiple appearances of individuals and scenes to eliminate redundant images in the skim; and identification of new speakers and proper nouns to detect the introduction of an individual.

We have incorporated our systems for text and face detection to select representative poster frames. When a face appears with captions, this usually indicates a person or affiliation relevant to the video segment. Captions seldom appear in documentaries, so this technology will be primarily used for selecting poster frames in broadcast news. In this domain, captions are used to show a person's name and affiliation, and locations and descriptions of events.

### **2.1.4 Video Spotting and Parsing**

We developed a method for automatic extraction of certain typical, important content from news video, such as the nature of the material (a speech, conversation, commentary, etc.), and type and name of the specific event or incident. For the automatic detection of such semantically rich information, both image analysis and natural language analysis are essential.

Typical phrases in the transcript suggesting the above information were analyzed by an existing natural language parser and manually for comparison. We investigated keyword spotting and detection of people and topic. We were able to detect more than half of the appropriate information in the content we analyzed.

We also analyzed images suggesting the same information, and are investigating utilization of our existing face detection technique for identifying the nature of the news material (i.e., speech, conversation, etc.).

We are developing a method for “linking” images and text. By connecting images and text through dynamic programming, the Infromedia system will be able to generate structured and self-contained data for the type “speech” (for example, data with a speaker's face and the contents of speech that can be referenced by a face, topic, name, and date).

### **2.1.5 Face and name association**

We developed a face and name association method called “Name-It.” The system processes news video footage (including image sequences and transcriptions), extracting faces from image sequences using image understanding techniques while simultaneously extracting names using natural language processing techniques. It then evaluates the “co-occurrence” of faces and names — as well as face similarity — using eigen-image based methods, finally to achieve face and name association. After the system has obtained face and name association from appropriate video footage, it can determine the name of a given unknown face, or produce candidate faces from a given name.

“Name-It” was then enhanced with a face tracking method and an improved name extraction method. Our face tracking method adaptively gets the statistical face color model from a detected face, tracks it within MPEG video, and provides occurrence and duration information for that face. In addition to this, it evaluates the angle of faces to output the best (most nearly frontal) shot of each face. The improved name extraction technique uses dictionaries, a thesaurus, and a parser to analyze transcripts to extract name candidates much more intelligently. Face extraction has been extended to obtain face sequences as well as face occurrence duration.

### **2.1.6 Color image similarity matching and indexing**

We implemented color image similarity matching using hue-based color histograms. We used a database loader program which reads color images and generates color histograms, and a client library which provides image similarity by matching referred color histograms. The library was ported to PC platforms and was linked into Infromedia clients. The technique works well, requiring only a few seconds retrieval time with several thousand images.

However, efficient exploration of a digital video library requires image similarity matching from among *millions* of images. High-dimensional, efficient indexing is necessary since each image is converted into a high-dimensional feature vector in a typical image similarity matching method.

We reviewed existing indexing methods including R\*-tree and SS-tree, which are the most successful and most used in other image matching systems. We discovered problems with these methods, especially when applied to extremely high-dimensional data. To overcome those problems we developed SR-tree, which, according to our evaluations, outperforms other indexing methods in high-dimensional data indexing. SR-tree was tested with an experimental image similarity matching system having over 100,000 images. It acquired 14 - 469 times speed-up compared with a linear search method [Katayama and Sato 96].

### **2.1.7 Color image retrieval system**

We have developed an Advanced Region Based Image Retrieval System (ARBIRS). Having analyzed various existing image retrieval systems, we concluded that prominent regions in the image, along with their associate features, provide the best capability to accomplish a higher-level, content-based, image retrieval system. A major challenge of this approach is that the image retrieval quality depends heavily on the robustness and accuracy of the image segmentation method which detects prominent regions from the image.

We are near to the completion of the second version of ARBIRS. In recent years, image retrieval by content has been actively studied by various research groups. However, most of the research attaches great emphasis to the feature capturing aspects of the problem, while it overlooks the value of integrated efforts in feature capturing, data indexing, database query, and database structure to achieve advanced image retrieval. To achieve advanced content-based image retrieval, ARBIRS v.2 combines efforts from the following two aspects:

1. The feature capturing aspect: To improve the accuracy and robustness of the detection of prominent regions from input images, we have developed a method to automatically detect and separate texture regions from input images, as well as a new image segmentation method that is especially effective in segmenting color images containing non-uniform illumination conditions
2. The database query aspect: Because there is no guarantee that the acquired image features are complete and error-free, we have developed a method to achieve better image retrieval based on imperfect sets of image features.

Initial experimental evaluations have demonstrated how these closely combined efforts work effectively to accomplish advanced image retrieval.

## **2.2. Data Organization, Networking Architecture and Interoperability**

Our efforts in this area for 1996 were largely directed at enhancing Informedia's interoperability with other library systems, and making the Informedia system itself more robust and functional. In support of the latter, we ported the database API and other utilities used in library creation to 32-bit OS (Windows95). This allows more flexibility in the client which often ran into memory barriers in 16-bit Windows. We also changed the format of our catalog and database to binary versions to increase speed and efficiency for use in-client.

### **2.2.1 Web-Based Client**

We began implementing the current library API behind a Web server. Since changes in the API and the library itself are all hidden behind a central location (the Web server), this should allow greater interoperability. In this model, the Web server communicates with the library directly, and exports the search engine, the library browser, and various media types and playback options via HTML. The one missing piece in this model is that typical Web browsers are not equipped with enough multi-media gadgets to support many of the datatypes and playback mechanisms currently in our client.

We completed work on the server side of a Web "gateway" to the Informedia library, and a first cut at a Web client. Initially we chose to maximize portability of the this client by

writing most of the GUI in Java; however, we've begun replacing the Java client interface with ActiveX controls for better performance. While ActiveX itself is still an emerging technology, what we've seen thus far seems to address our need for better run-time performance as well as network scalability.

In order to address the problem of actually playing back content over the limited bandwidth of the Internet, we have begun work to implement a "slideshow." A slideshow would stream continuous audio, and display still images from within the video synchronized to the audio stream. We have been investigating the plethora of commercial web gadgetry available to do this effectively, including ActiveX, RealAudio, InterVu, VDOnet, Xtreme, etc. Ultimately, we hope to "ratchet" the number of frames/second displayed in the slideshow up or down by responding dynamically to the current network throughput.

### **2.2.2 Video Servers**

In an effort to leverage commercial products to deliver the media within our library, we investigated DEC's latest "MediaPlex" product. It held promise both in terms of scalability to a WAN, and additionally provided streaming support over high bandwidth Ethernets. We installed the beta (and later the first production) version of the DEC Video Server, with a test library of approximately seven hours of mpeg video. The product dovetailed nicely with our new Web interface efforts. A prototype of our web client was able to stream mpeg video from the Video Server via embedded HTML links. Furthermore, the server had the capability to stream specified segments within a larger mpeg, which is crucial to our application.

However, the expected externally-developed client side browser plug-in was not completed. This type of control is imperative for Infromedia, since much of what our client displays is two or more synchronized streams, i.e., transcript and filmstrips synchronized to video playback. Without the client side control, our prospects for fully exploiting this technology were minimized and this alternative was deferred. We are continuing the search for commercial streaming products (Oracle, Netscape (LiveMedia), Progressive Technologies (RealMedia), Xing Technologies (StreamWorks)). None seem to be mature enough products to provide our required performance or functionality at this writing.

### **2.2.3 Interoperability**

We conducted an experiment in interoperability with the DLI project at Stanford. We issued queries from our web client to a socket connection to a Stanford "Infobus" server, parsed the returns and displayed the results in the standard Infromedia client. Little (if any) of Stanford's content is multi-media. Likewise, Stanford issued http requests to our library web server to search, browse, and retrieve objects from the Infromedia library for use within their own client interface. This builds on their current work of building "proxies" to scale the amount of data accessible via the Infobus. The impact was twofold:

1. Any existing Infobus client could include Infromedia in their search, along with the existing Infobus repositories such as AltaVista, Dialog, WebCrawler, Lycos, Ultraseek, etc.
2. The Infromedia Web Client was modified to extend its queries to the Infobus, as well as the normal Infromedia libraries. This was demonstrated by including queries to AltaVista and WebCrawler.

## **2.3. Testbeds, Specialized Corpora, and User Studies**

### **2.3.1 User Studies**

We delivered our first testbed system to Winchester Thurston with 8 client stations and a 50 hour video library, and provided training for Winchester Thurston faculty on how to use the system. The second testbed system with additional client stations, enhanced functionality, and a 100 hour corpus was delivered for the fall 1996 semester. We collected feedback, and presented early interface and transaction log results at the NSF site visit and at the DLI meeting in Michigan. Data collected via interviews and transaction logs were used in the refinement of the digital library interface [Christel96].

We conducted a formal empirical study with 30 high school and college students in summer 1996. This work studied multimedia abstractions used by the digital video library as well as how the data is segmented within that library. We learned that quick access to short, relevant segments of video enables more efficient use of a video library. Three interfaces were designed for such access, allowing the user to browse through a set of video segments in support of a fact-finding task. We designed an experiment in which subjects' performance and attitudes were measured to determine the relative effectiveness of these three interfaces. Results show that visual imagery benefits both performance and subjective satisfaction compared to text list presentation. Tailoring the representative images for video segments in a set based on the query which produced that set (query-based poster frames) provides significant improvements.

We are revising work for an empirical study into the effectiveness of a collapsed video, or video "skim," created in four different ways: fixed time intervals, "best" subsets chosen by only considering the video data, "best" subsets chosen by only considering the audio data, and "best" subsets chosen by considering both audio and video content. A pilot test indicated that each of these skim types are useful for locating a section of interest in a video, i.e., all four skims have some utility for navigation. More work will be done to investigate any differences in these skims' ability to summarize the content of a larger video clip, i.e., to investigate the decision-making support of the four skim types.

### **2.3.2 Specialized Corpora**

We installed a client at the Testbed Integration Environment at DARPA. The first client was left at DARPA in late May/early June. We processed a small amount of DARPA-provided video content (approximately 5-7 hours), and installed this library along with the latest client at DARPA in July.

We created a separate corpus from the sessions of the Democratic and Republican national conventions and subsequent debates between candidates. We used the corpus primarily for development work on fielded data and annotations.

The Cable News Network (CNN) has recently become a consortium affiliate and will make available most of its broadcast content for research experimentation and testbed application.

## **2.4. User Interface and Client Implementation**

### **2.4.1 Annotations**

We added the ability for users to "mark-up" the library at their site. These annotations can then be used to refine searches within that library. For example, we have annotated a corpus of the Republican and Democratic conventions as well as the two presidential

debates. The annotations identify the speaker and the location of the speech. Having done that, a user can then issue “fielded” queries such as: “welfare policies” AND Speaker “Dole” AND Location “debate”.

Annotations are also useful in many contexts, enabling users to customize the libraries towards the local usage patterns; e.g., teachers embedding “class notes” in the library’s content.

We also completed an annotation editor which allows the user interactively to mark stop/start times to define the boundaries of the annotation, and to define new annotation types (“speaker\_id”, “class\_notes”, etc.) as well as fields within those types (e.g., for “speaker\_id”, define a “speaker” and “location”). The database API was modified so that library clients can then build search dialogs dynamically, by querying the database for the “fielded data” types.

#### **2.4.2 Client Software Release Version 1.14**

The Informedia client has been extended to include face matching, fielded search, and query-based poster frame and filmstrip creation.

*Face detection/Similarity matching.* We integrated face detection and face similarity matching calls into the database API and library creation scripts, to make this data available in the client. The client is then able to tell if a specific image (poster frame or filmstrip representations) contains a face, and if so is able to find the coordinates of the bounding rectangle. Furthermore, one can search for other similar faces in the database, in much the same way as one can do generic image (histogram) matching. A drag-and-drop interface was developed in the client to exhibit this functionality. This is the precursor to what will become the ability to issue multi-modal searches (i.e., image/face/textual).

*Video Streaming.* The commercial world is slowly making technologies available to stream VHS quality video. Since the available bandwidth for doing so must be in excess of 1.5 mbs per client, we were only able to demonstrate this on our LAN. This was implemented by creating a cgi server script which generated a page with an embedded media player object (Microsoft’s ActiveMovie), that was automatically initialized to the correct mpeg movie and start/end frames specified. VCR controls on the page allowed the user to seek within the segment. It should be noted that this commercial technology is in its infancy, and was not always consistent or reliable.

*Query-based Poster Frames.* We added data to our library which defines an image for each shot-break, much like the filmstrip data. This added data allows us dynamically to choose a representative poster frame from within the same shot containing the most search hits from the user’s query. From the user’s perspective, this should result in more graphically-relevant poster frames than the current static versions we currently use.

#### **2.4.3 Spoken query interface**

Our spoken query interface now has a modified 20,000 word language model. It uses a network server for speech recognition and has been ported to Windows95. We are exploring continuous listening and moving the recognition engine to a PC platform. A related challenge is to enable speech recognition for a web-based interface.

## **2.5. NetBill Authentication and Billing System**

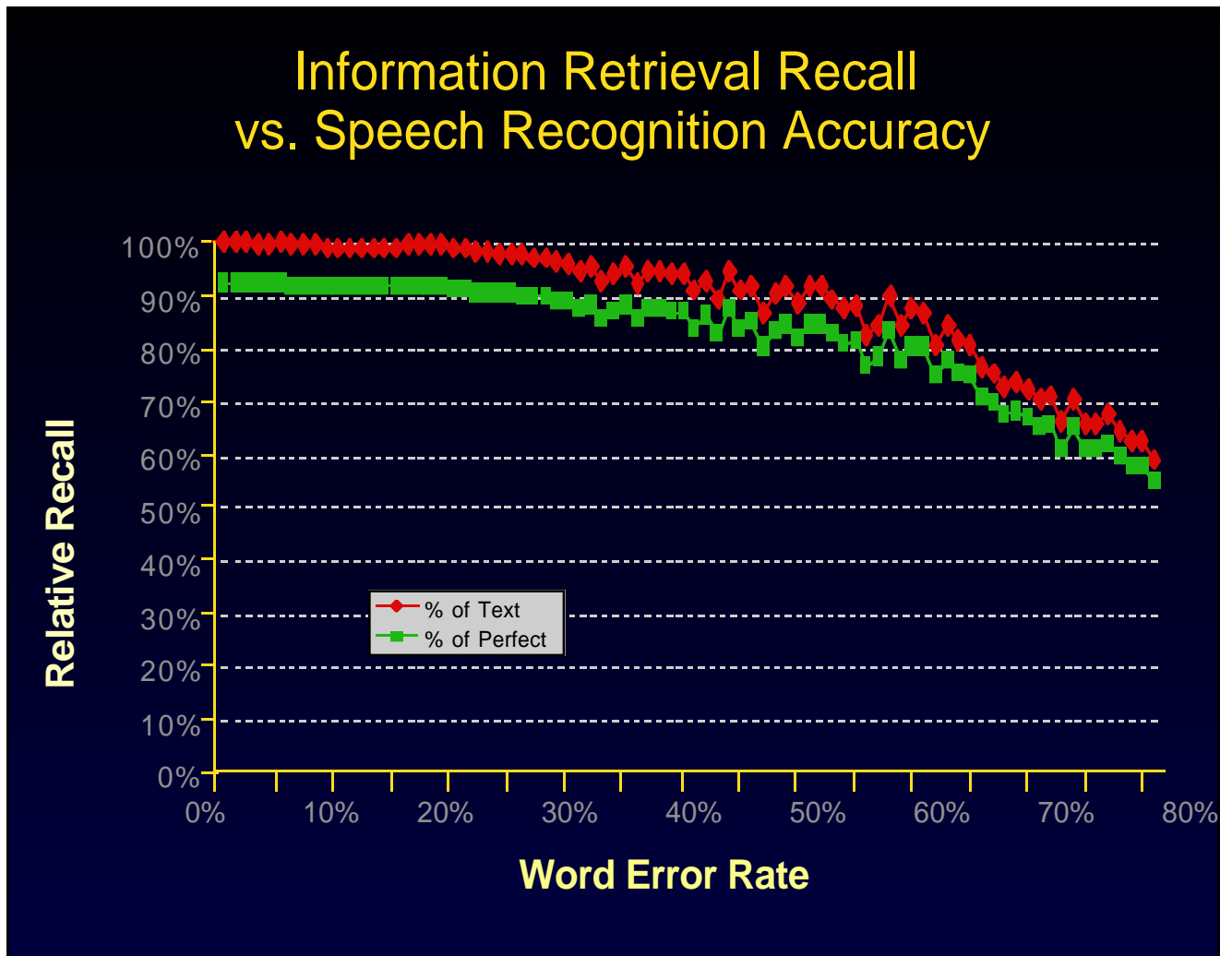
NetBill is a distributed system with software running on the user's (consumer's) system, on the information provider's (merchant's) system, and on a central accounting server. The central server includes a key management subsystem, a transaction processing system, an accounting database, a user and accounts administration system, and a payments system that interfaces to banks and credit card authorization systems.

A complete Alpha version of the NetBill system, including client, information provider, and clearinghouse systems, has been implemented. The latter includes a certificate management system, transaction server, administration server, and database. The Alpha NetBill system has been running successfully on CMU's campus for the last quarter with high availability, although it has not yet been linked to the Infromedia system.

During this period we focused on reimplementing various components for improved performance, flexibility and portability, and made several improvements in our development infrastructure and process.

### 3. Significant Event

The integration of speech recognition, image understanding, natural language processing and information retrieval to overcome limits in each technology is a cornerstone of the Infromedia Digital Video Library system. A fundamental thesis of the Infromedia DVL is that speech recognition generated transcripts will provide a sufficient index into the video content for information search and retrieval. We conducted a study to determine how accurate speech recognition needs to be, compared to fully correct text transcription, in order to be useful and usable for indexing and retrieving spoken language documents. In our experiments, word error rates up to 25% did not significantly impact information retrieval. More surprisingly, word error rates of 50% still provided 85-90% of the recall and precision relative to fully accurate transcripts in the same retrieval system. Our conclusion is that very high speech recognition accuracy, although a desirable goal in itself, is not required in order to achieve information retrieval effectiveness that is similar to retrieval from perfect text transcripts.



## **4. Notable Outreach and Inclusion Activities**

In April, Howard Wactlar presented “New and Emerging Technologies: What the Future will Bring” at the Ernest L. Boyer Technology Summit for Educators, sponsored by the Corporation for Public Broadcasting.

In October, Howard Wactlar presented “Automated Video Segmentation for On-Demand Retrieval from Very Large Video Libraries” to the 138th Society of Motion Picture and Television Engineers (SMPTE) Technical Conference and World Expo, Los Angeles, CA.

## **5. Project Director’s Narrative**

The year's activity reflects a maturation in the infrastructure which enabled us to deepen our research and experimentation in each of the following challenge areas:

- Retrieval performance in the presence of inaccuracy and ambiguity in the underlying cognitive processing
- Approximate match in meaning and visualization
- Presentation and reuse of library content
- Interoperability

We have introduced metrics of performance in the underlying technology and relate them to those in the top level information retrieval tasks. Such metrics are essential to our ability to measure progress and success in the project overall.

Our measurements of retrieval precision and recall, and verification of “average rank” as a useful computable measure, are significant results for spoken language document retrieval. The use of speech understanding generated transcripts were a fundamental premise of this project, for which we have now established some quantitative validation. We achieved a twelve-times speedup in face matching enabling us to incorporate this function in a meaningful size video collection. Work on name/face association through similarity matrices and co-occurrence techniques enabled us to measure the rate of successful correlations for modest video data sets. The effect of applying latent semantic indexing to the corpus will also be gauged through the average rank metric. We have extended the user interface to enable searchable annotations of video content, allowing user “structuring” of the corpus or of a guided path through portions of it (e.g., for programmed learning). We have also made considerable progress towards interoperability through Web interfaces to the data and thus enabled demonstration of two-way access through Stanford Infobus compatibility.

In its second year, our testbed school enabled user studies from student usage profiles. However, it did not provide any basis for measurable educational impact, except through qualitative teacher assessments and anecdotal student evaluations. There have been requests for expanding the testbed corpus beyond the science domain, but that awaits additional resources to incorporate other sources of content with domain related issues. We have provided remote home access to the non-video content of the library to enable teachers to browse the content in order to find relevant material to incorporate in teaching or to which they can direct their students. The testbed site was introduced to the World Wide Web this year as well, and this significantly broadened the students “search and discovery”

activities beyond our video resources. As a result, we will not have isolated data from controlled experiments enabling us to assess educational impact broadly or to quantify the effects of video reinforced learning.

We have managed to implement a portable laptop version of the Infromedia client and server with a limited speech query system, enabling us to deliver the story of our project through hands-on interaction.

The project continues to attract considerable public attention. We have many corporate and academic visitors, give numerous presentations and talks outside of CMU, and receive considerable public press coverage. This continues to be a considerable distraction for the project and its human resources, though a long-term benefit to research relevance and transfer.

## 6. Publications

- [Chen96] “Integrating Digital Signatures into Internet Commerce: A NetBill Example,” Chen, H-Y. Master’s thesis, Carnegie Mellon University, Information Networking Institute, December 1996.
- [Christel96] “Informedia Goes to School: Early Findings from the Digital Video Library Project,” Christel, M. and K. Pendyala. D-Lib Magazine, September 1996. Available from:  
<http://www.dlib.org/dlib/september96/informedia/09christel.html>
- [Christel96a] “Techniques for the Creation and Exploration of Digital Video Libraries,” Christel, M., S. Stevens, T. Kanade, M. Mauldin, R. Reddy, and H. Wactlar. Chapter in Multimedia Tools and Applications, Vol. 2, Borko Furth ed., Boston, MA, Kluwer Academic Publishers, 1996.
- [Christel97] “Improving Access to Digital Video Library,” Christel, M., D. Winkler, and R. Taylor. Submitted to INTERACT97, the 6th IFIP Conference on Human-Computer Interaction, Sidney, Australia, July 1997.
- [Christel97a] “Multimedia Abstractions for a Digital Video Library,” Christel, M., D. Winkler, and R. Taylor. Submitted for publication in ACM Digital Libraries '97, Philadelphia, PA, July 1997.
- [Chuang97] “The Bundling and Unbundling of Information Goods: Economic Incentives for the Network Delivery of Academic Journal Articles,” Chuang, J and M. Sirbu. Submitted to the Conference on Economics of Digital Information and Intellectual Property, Cambridge, MA, January 23-25, 1997.
- [Cox96] “NetBill Security and Transaction Protocol,” Cox, B., J. Tygar, and M. Sirbu. Submitted to the First USENIX Workshop on Electronic Commerce, 1996.
- [Eckert96] “A Pseudonym Server for NetBill,” Eckert, C. Master’s thesis, Carnegie Mellon University, Information Networking Institute, December 1996.
- [Hauptmann97] “Indexing and Search of Multimodal Information,” Hauptmann, A. and H. Wactlar. Submitted to International Conference on Acoustics, Speech and Signal Processing (ICASSP-97), Munich, Germany, April 1997.
- [Heintze96] “Model checking electronic commerce protocols,” Heintze, N., J. Tygar, J. Wing, and H. Wong. Proceedings of the Second USENIX Workshop on Electronic Commerce, pages 147-164. The USENIX Association, Oakland, CA, November 1996.
- [Katayama97] “The SR-tree: An Index Structure for High-dimensional Nearest Neighbor Queries,” Katayama, N. and S. Sato. To appear in Proceedings of ACM SIGMOD, 1997.
- [Kanade96] “Immersion into Visual Media: New Applications of Image Understanding,” Kanade, T. In IEEE Expert Intelligent Systems and Their Applications, Vol. 11, No. 1, pages 73-80, IEEE Computer Society, 1996.

- [Placeway97] "The 1996 Hub-4 Sphinx-3 System," Placeway, P., S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer. Submitted to proceedings of DARPA Spoken Systems Technology Workshop, February 1997.
- [Ravishankar97] "Some Results on Search Complexity vs Accuracy," Ravishankar, M. Submitted to Proceedings of DARPA Spoken Systems Technology Workshop, February 1997.
- [Rowley] "Neural Network-based Face Detection," Rowley, H., S. Baluja, and T. Kanade. In Proceedings of International Conference on Computer Vision and Pattern Recognition, pages 203-208, San Francisco, CA
- [Sato96] "NAME-IT: Association of Face and Name in Video," Sato, S. and T. Kanade. Carnegie Mellon University School of Computer Science Technical Report CMU-CS-96-205, December 1996.
- [Simpson96] "Modeling the Risks and Costs of Digitally Signed Certificates," Simpson, I. In Proceedings of the Second USENIX Workshop on Electronic Commerce, pages 243-250. The USENIX Association, Oakland, CA, November, 1996.
- [Sirbu96] "Public-Key Based Ticket Granting Service in Kerberos," Sirbu, M. and J. Chuang. Internet-Draft, Internet Engineering Task Force, May 1996. Available from: <ftp://ds.internic.net/internet-drafts/draft-sirbu-kerb-ext-00.txt>
- [Sirbu97] "Distributed Authentication Kerberos Using Public Key Cryptography," Sirbu, M. and J. Chuang. Submitted to Proceedings of the Symposium on Networks and Distributed System Security, Internet Society. February 1997.
- [Smith] "Video Skimming for Quick Browsing Based on Audio and Image Characterization," M. Smith and T. Kanade. In Proceedings of The Second Technical Conference on Telecommunications R&D in Massachusetts.
- [Smith96] "Video Skimming and Characterization through Techniques in Language and Image Understanding," M. Smith and T. Kanade. CMU School of Computer Science Technical Report CMU-CS-95-186R (revised 12/96).
- [Su96] "Building blocks for atomicity in electronic commerce," Su, J. and J. Tygar. In Proceedings of the Sixth USENIX Security Symposium, pages 97-104. The USENIX Association, San Jose, CA, July 1996.
- [Tygar96] "Atomicity in Electronic Commerce," Tygar, J.D. In Proceedings of the ACM Symposium on Principles of Distributed Computing. August 1996.
- [Wactlar96] "Intelligent Access to Digital Video: Informedia Project," Wactlar, H., T. Kanade, M. Smith, and S. Stevens. In IEEE Computer, Digital Library Initiative special issue, May, 1996.
- [Wactlar96a] "Automated Video Segmentation for On-Demand Retrieval from Very Large Video Libraries," Wactlar, H., A. Hauptmann, M. Smith, and K. Pendyala. The 138th SMPTE (Society of Motion Picture and Television Engineers) Technical Conference and World Expo, Los Angeles, CA.

[Wactlar96b] "Informedia: News-on-Demand Experiments in Speech Recognition,"  
Wactlar, H., A. Hauptmann, and M. Witbrock. In Proceedings of DARPA  
Speech Recognition Workshop. Arden House, Harriman, NY.

## **7. Presentations, Demonstrations, and Industry Visitors**

- 2/6 Hewlett Packard. Jim Olson, General Manager, Video Communications Division.
- 2/8 ACOM. Col. Jim Wirth, USAF, ACOM Advanced Concept Technology Demo project leader; Maj. Paul Gilles, USMC, assistant project leader.
- 3/4 Corporation for Public Broadcasting. Maria Borges.
- 3/6 Hewlett Packard Computer Research Center. Dick Lampman, CRC Director; Denny Georg, Director of the Computer Systems Lab of CRC; Gary Herman, Director of the Broadband Information Systems Lab in CRC; Moise Zloof; Steven Rosenberg.
- 3/7 CNN America Incorporated. Frank Sesno, Washington News Bureau.
- 3/7 DARPA Headquarters, DC. Demo for members of Intelligence community.
- 3/13 Heinz Technology & Learning Forum Advisory Board, Pittsburgh PA.
- 3/15 Honeywell. Garry Nordenstam, Manager, Training Technologies.
- 3/21 Lawrence Livermore National Laboratory, Advanced Video Research Group.
- 3/96 Digital Equipment Corporation, Video Server Technology Group.
- 4/4 Office of Naval Research. Comdr. Timothy D. Warren, Director, Automated Information Systems; William E. Smith, Sr. Research Physicist, Neural Network Development Lab.
- 4/4 "Informedia Digital Video Library," H. Wactlar. Presentation and demonstration at University of Pittsburgh School of Library and Information Science.
- 4/8 Pixar. Ralph Guggenheim, Vice President, Feature Animation.
- 4/11 NBC News, News Archives. Dr. Richard S. Alben, Business Interface Planning Manager, GE Corporate R&D. NBC Interactive Media - Edmond P. Sanctis, Sr. VP & Executive Producer; David Britton, Director of Production; Julie Buchholz, Director/Sr. Producer Interactive Programming; Mark Kortekaas, Director Technical Operations; Eric Pohl, Principal Engineer, Recording Systems.
- 4/26 Global Field Consortium: Daimler Benz - Dieter Hege, Vice President, IT-Infrastructures; Xerox - Shirley Edwards, Malcolm Goslee.
- 4/29 "New and Emerging Technologies: What the Future will Bring," H. Wactlar. Presented at Ernest L. Boyer Technology Summit for Educators, sponsored by the Corporation for Public Broadcasting.
- 4/96 Princeton Video Library project. Wayne Wolf, PI.

- 5/3 Informedia presentation at the NWIG Conference, Oakridge, TN (Nuclear Weapons Information Group).
- 5/6 Tokyo University of Information Sciences. Toshiaki Itoh, Director, Center for Education and Research Information.
- 5/8 DARPA. Allen Sears, DARPA/ITO Program Manager for HCI, HLS; Ron Larson, DARPA/ITO Program Manager for digital libraries, multilingual technology, University of Maryland; Kevin Mills, DARPA/ITO Program Manager for collaboration technology, NIST; Gary Jones, DARPA Tactical Technology Office Program Manager for simulation-based design (maritime applications); John Silva, M.D., DARPA Defense Science Office Program Manager for health care information systems.
- 5/10 Motorola. Sue Thompson, Manager of Strategic University Relations.
- 5/15 NRaD. Steve Nunn.
- 5/17 "Informedia: News-on-Demand," University of Karlsruhe, Germany.
- 5/21 US Navy Warfare Center. Brad Cope, Larry Keeler.
- 5/22 "Incorporating a Digital Video Library into High School Science Instruction," First Carnegie Mellon University Symposium on Technology Enhanced Learning.
- 5/26 Tokyo University of Art and Design. Prof. Eishi Katsura.
- 5/30 MITI of Japan. Hiroshi Mukaiyama, IT Policy Promotion Department, Japan Information Processing Development Center (JIPDEC); Hidetoshi Hyuga, Department of Technology Application, Information-technology Promotion Agency (IPA); Kanji Kato, Information Systems R&D Division, Hitachi, Ltd.
- 5/96 Participated in a brainstorming session at NASA's CASI (Center for Aerospace Information) about digital libraries in Baltimore at the invitation of RMS Associates.
- 6/6 HBO & Co. Michael Kappel, Senior Vice President, Strategic Product Planning & Marketing; Christine Rumsey, Senior Vice President of Human Resources; Douglas Lucas, Director, Employee Relations and College Recruiting
- 6/10 "Video and Computers: The Future of Digital Libraries," seminar to Pennsylvania Governor's School for the Sciences (90 high school scholarship students).
- 6/11 Boeing Information and Support Services. David P. Himmel and Orlie Brewer, Advanced Computing Technologists.
- 6/13 Informedia presentation, DARE Data Review Group (DDRG) meeting (Data Archival & Retrieval Enhancement) - Defense Nuclear Agency, Washington, DC.
- 6/17 Medical College of Pennsylvania and Hahnemann University. Dr. Leonard Ross, MCPHU Provost; Dr. Glenda Donoghue, MCPHU Vice Provost; Dr. Carol Montgomery, Associate Provost and Director of Academic Informatics.

- 6/20 “Content based Retrieval: Research and Direction,” International Conference on Computer Vision and Pattern Recognition, San Francisco, CA.
- 6/24 Informedia system training, Winchester Thurston school. Trained 18 children (age group 10-14) to use the system as part of their computer Space Camp with an emphasis on reuse of digital library content.
- 6/25 CNBC. Michael Reitman, Vice President, Engineering & Technology. GE Corporate Research and Development, Richard S. Alben, Business Interface Planning Manager.
- 7/8 NTT-Data Communications Systems Corporation. Masaki Yamaoka and Osamu Iwaki, Laboratory for Information Technology, Research and Development Headquarters.
- 7/9 DARPA. Dick Wisner, ISO Assistant Director for Battlefield Awareness & Information Integration; David Gunning, Bob Douglas and John Leon, Program Managers, Information Integration.
- 7/10 “The Informedia Project: An Advanced Digital Video Library,” Information Technology Colloquium, Los Alamos National Laboratory, Los Alamos, NM.
- 7/11 “The Informedia Project: An Advanced Digital Video Library,” Information Technology Colloquium, Lawrence Livermore National Lab., Livermore, CA.
- 7/11 Network Solutions, an SAIC group. Raymond Corson, Vice President; Ivan Yopp, Sr. Staffing Representative.
- 7/16 SAIC. Admiral Bill Owens, Vice Chair; Clint Kelly.
- 7/17 WBN, Worldwide Broadcasting Network. Ali S. Kazaronian, President.
- 7/23 NSF. Sue Kemnitzer, Deputy Director, Engineering Education and Centers Division.
- 7/24 Kodak. James Warnick, Senior Research Scientist, Imaging Science Division.
- 7/96 Lawrence Livermore National Laboratories. Bruce Lownsbery, Project Leader, Computer Applications; David Dirks, Project Manager, Advanced Video Research and Videoconference Systems.
- 8/6 Discovery Channel, Discovery Communications, Inc. Hunter Williams, Senior Manager, Educational Relations; Douglass R. Sawyer, General Manager, Discovery Channel Education; Prof. Gary Marchionini, College of Library & Information Services, HCI Lab, University of Maryland.
- 8/7 “Informedia Digital Video Library,” Federal Webmasters Workshop 1996, Goddard Space Flight Center, Greenbelt, MD.
- 8/9 France Telecom Inc. Aymerik Renard, Manager, Business Development (San Francisco); Dominique Primot, Senior Advisor - Technology Strategy (Paris).
- 8/14 Microsoft. Tom Firman, Director of Technology, Microsoft Network; Bob Dijon, Executive Producer, Microsoft Network.

- 8/16 University of Karlsruhe, Rechenzentrum. Wolfgang Peters, Ministerialrat; Prof. Dr. Adolf Schreiner, Director, Lehrstuhl für Organization von Datensystemen.
- 8/22 TNO Institute of Applied Physics, Netherlands Organization for Applied Scientific Research. Joop van Gent, Multimedia Group, Document Information Technology.
- 8/24 SunGuard Data Systems, Inc. Philip Dowd, CEO.
- 8/28 WRS Motion Picture and Video Laboratory. F. Jack Napor, President.
- 8/28 Wired Magazine. Richard Bierck, Writer/Reporter.
- 9/17 DARPA Project Genoa. Brian Sharkey, DARPA Program Manager; Robert Neches, DARPA Program Manager, ITO; John Poindexter, Project Consultant.
- 9/18 Wright Patterson AFB Labs. Chahira Hopper, Program Manager, WL/AACA; Lt. Mance Harmon, WL/AACF.
- 9/24 The International Monitor Institute. Pippa Scott, Office of the Chair (Netherlands); J. Anthony Young, Office of the Chair (Los Angeles); Anne Herringer, Project Director, Balkan Archive.
- 10/2 Siemens Corporate Research. Thomas Grandke, President.
- 10/7 Speech Understanding open house. Yung-Hwan Oh, Hajin Yu, Hwan-Jin Choi, Ho-Seung Shin, Sang-Mun Chin, Yeon-Jun Kim, Sangho Lee, KAIST CS Dept., Korea.
- 10/9 Kyoto University. Prof. Michihiko Minoh, Prof. Yasuo Okabe, Jiro Kiyama, Sharp Electronics Corporation.
- 10/11 "Informedia Digital Video Library," Tektronix, Inc., Video & Networking Division, R. Bland McCartha, Director, Business Development.
- 10/11 Kwant Woon University. Kim Loon Hyol, Prof., Vice-President of Acoustical Society of Korea; Jin-Woo Hong, Principal Member of Engineering Staff, Human Interfaces, Technology Department, Digital Signal Processing Laboratory, Department of Computer Engineering, Seoul, Korea
- 10/21 Defense Language Institute. Dr. Ray Clifford Deniz Belgin.
- 10/24 CBS News, NY. Josie J. Thomas, Vice President for Business Affairs.
- 10/30 "Informedia Digital Video Library," PREPnet 1996 Conference. On the Horizon: New Facilities, Applications & Opportunities.
- 11/5 Sulzer Technology Corporation. Karl Bochsler.
- 11/14 GE. John McKinley, Chief Technical and Information Officer at GE Capital Services, and Tom Crowley, GE Capital Ventures.
- 11/15 Mellon Bank. John Grego, Director of IS at Mellon Bank, and four colleagues.

- 11/18 Olivetti and Oracle Research Labs, Cambridge, England. Dr. Andy Harter, Principal Research Engineer, Dr. Kenneth Wood, Research Scientist, and Dr. Martin Brown, Research Engineer.
- 11/18 ACM Multimedia Conference in Boston (Nov 18 - 22), entitled “Interoperability for Digital Video Libraries”. Chaired and helped to organize. Notes from the workshop can be found at [http://www.informedia.cs.cmu.edu/acm\\_interop/workshop\\_notes.html](http://www.informedia.cs.cmu.edu/acm_interop/workshop_notes.html).
- 12/6 National Library of Medicine. Dr. John Wilbur, Senior Research Scientist.
- 12/4 Association of Moving Image Archivists Annual Conference. Presented at Digital Video Libraries session. Atlanta, GA.

## PROJECT SUMMARY

DATE PREPARED: 1 February 1997

ORGANIZATION: Carnegie Mellon University

### PRINCIPAL INVESTIGATORS:

Howard D. Wactlar, wactlar@cmu.edu, 412/268-2571, fax:412/268-5576  
Takeo Kanade, takeo.kanade@cmu.edu, 412/268-3016, fax:412/268-5570  
Raj Reddy, raj.reddy@cs.cmu.edu, 412/268-2597, fax:412/683-5348  
Michael Mauldin, mauldin@cs.cmu.edu, 412/268-5293, fax:412/268-6298 (on-leave)  
Scott Stevens, scott.stevens@sei.cmu.edu, 412/268-7796, fax:412/268-5758  
Marvin Sirbu, marvin.sirbu@cmu.edu, 412/268-3436, fax:412/268-7196  
Doug Tygar, doug.tygar@cs.cmu.edu, 412/268-6340, fax:412/268-8320

TITLE OF EFFORT: Informedia Digital Video Library System

ACCESS INFORMATION: <http://www.informedia.cs.cmu.edu>

### OBJECTIVE:

The Informedia digital video library project establishes a large, on-line digital video library by developing intelligent, automatic mechanisms to populate the library and allow for full-content and knowledge-based search and retrieval via desktop computer over local, metropolitan, and wide-area networks. Initially, the library will be populated with 1000 hours of raw and edited video drawn from video documentary, current news and educational video sources. The library will be deployed at a Pittsburgh area K-12 school to study its use, usability, and potential impact on curriculum. Another corpus of broadcast news will provide a "news-on-demand" capability. The library will interoperate with other network-based video and text information systems and repositories through extant and evolving communication, media and naming standards.

### APPROACH:

The approach utilizes several techniques for content-based searching and video sequence retrieval. Content is conveyed in both the narrative (speech and language) and the image. Only by the collaborative interaction of image, speech and natural language understanding technology can we successfully populate, segment, index, and search diverse video collections with satisfactory recall and precision.

This approach uniquely compensates for problems of interpretation and search in error-full and ambiguous data sets. It starts with a highly accurate, speaker-independent, connected speech recognizer which automatically transcribes video soundtracks. A language understanding system then analyzes and organizes the transcript and stores it in a full-text information retrieval system. This text database allows for rapid retrieval of individual video segments which satisfy an arbitrary query based on the words in the soundtrack. Image and language understanding enables one to locate and delineate the corresponding "video paragraph" context by using combined source information about camera cuts, object tracking, speaker changes, timing of audio and/or background music, and change in content of spoken words. Controls allow the user to interactively request corresponding video paragraphs to full volumes, to browse the collection, to intelligently "skim" the returned content, and to annotate the stored video objects for future reuse.

## PROGRESS:

We have introduced performance metrics that are essential to our ability to measure overall progress and success in the project. We have now established quantitative validation for our fundamental premise that video transcripts could be sufficiently automatically generated using speech understanding techniques to enable successful information retrieval. Informedia has incorporated face matching in a meaningful-sized video collection, and we've demonstrated the viability of face/name association. Users can now augment library data with their own searchable annotations, providing personalized accessibility to the video information. Considerable progress in interoperability has been demonstrated with the Stanford Infobus, and through development of Web interfaces.

## RECENT ACCOMPLISHMENTS:

Informedia experiments revealed that word error rates in speech recognizers of up to 25% do not significantly impact information retrieval. Word error rates of up to 50% still provide 85-90% recall and precision relative to a correct text transcript, using the same retrieval system.

A completed design for a platform independent, web-based, Java-based client provides similar function to that of the current MS Visual Basic client, and implementation is proceeding. Browse and search of the Informedia database via the Internet (with only local access to date because of rights issues) are currently enabled.

Users now have the ability to "mark-up" the library at their site. These annotations can then be used to refine searches within that library. Annotations are useful in many contexts as users customize the libraries towards the local usage patterns. For example, teachers in schools might want to include "class notes" in the library.

A successful interoperability experiment with the DLI project at Stanford allowed the Informedia Web client to issue queries to a socket connection to a Stanford Infobus server, parse the returns and to display the results in the same Informedia client. Likewise, Stanford issued http requests to the Informedia library web server, to search, browse, and retrieve objects from the Informedia library for use within their own client interface.

To enable efficient exploration of a digital video library, similarity matching from among millions of images is necessary. High-dimensional, efficient indexing is required since each image is converted into a high-dimensional feature vector in a typical image similarity matching method. To overcome problems with existing indexing methods, SR-tree was developed, outperforming other indexing methods in high-dimensional data indexing.

## PLANS:

Develop the ideas in text tiling into a probabilistic model of segment position, using distances between occurrences of each word as the main indicator of segment break presence. Extend it to LSI identified synonyms, and to distances between possibly different words.

Develop a video content change detection method that is able to detect scene changes based on the context and content of the image sequence, e.g., by object tracking through changes in camera angle and panning.

Extend the existing Web infrastructure supporting an Informedia client and server as a platform to integrate "slideshow" playbacks, NetBill, and hierarchical caching.

## TECHNOLOGY TRANSITION, SHARING, PARTNERING, ETC.:

We delivered a full processing Informedia system to an outside organization, Digital Equipment Corporation, as well as library systems to Boeing Information Systems and the DARPA TIE facilities. Informedia has been in extensive use for the past year at our K-12 testbed, Winchester-Thurston. Additionally, NetBill has been in Alpha test on CMU's campus for the past quarter.



## Search and Discovery in the Video Medium

The realization of full content search and retrieval from digital video, audio and text libraries

The utilization of integrated speech, image and language understanding for their creation and exploration



© Copyright 1997 Carnegie Mellon University

## The Application of Diverse Technologies to a Single Focused Application

---

Speech understanding for automatically derived transcripts and spoken queries

Image understanding for video "paragraphing" (segmentation) and similarity matching

Natural language for processing transcripts and queries

User interfaces for video display, manipulation and reuse

Network authentication and billing for controlled access

Data architectures for network resource interoperability

User studies for validation and assessment

**Builds on existing technology - extends with focused**

## Two Testbed Corpora

	Winchester Testbed	News-on-Demand
<b>Domain</b>	Science Documentaries/Lectures	Broadcast News
<b>Content Creation</b>	Hybrid Process/ Manually Corrected	Fully Automated
<b>Accompanying Transcript</b>	Corrected (not viewable)	Errorful
<b>Corpus Half-Life</b>	Static Archive	Changes Daily
<b>Query</b>	Typed	Spoken
<b>Testbed</b>	K-12 Focus	CMU, Roadshows
<b>Size</b>	45 hours/ 1600 Segments	150 hours/ 20,000 Stories

## Infomedica DVL Major Partners

QED Enterprises (WQED)  
 Digital Equipment  
 Microsoft  
 Bell Atlantic  
 Telecom Italia  
 Intel  
 CNN  
 Boeing  
 Allegheny-Singer Research Institute  
 (Allegheny General Hospital)  
 The Winchester-Thurston School  
 The British Open University

CMU Informedia News On Demand: 08-Dec-95

File Edit Navigate Options Text Video Window Help

Search Request: Anything about new tanks deployed in Bosnia

Search Results: All 25 matches on "Anything about new tanks deployed in Bosnia": (11/30/95) "in depth" -- story, headed soon.

Filmstrip Overview: (11/30/95) "in depth" -- story, headed soon.

training exercises and preparing to move down here on trains and boats and planes from bases in Germany and Italy. The advance units, called enabling teams, could be on the way through a staging point in Hungary with only 96 hours notice. The U.S. Troops will operate in northern Bosnia, headquartered in tuzla.

## Video Characterization

**Raw Audio**

**Text Extraction**

**Keywords** SILENCE they are the jury every toy owner hopes to please MUSIC electric cars are

**Raw Video**

<b>Scene Cuts</b>					
<b>Camera</b>	Static	Zoom	Static	Static	Pan
<b>Objects</b>	Adult Female	Child (2)-Male	Child-Female	Adult-Female	Car-Red
<b>Action</b>	Head Motion	Body Motion	Body Motion	Head Motion	Right Motion
<b>Captions</b>	NONE	Philadelphia	NONE	News	NONE
<b>Scenery</b>	Indoor	Indoor	Indoor	Indoor	Outdoor

**Skim Scene**

**Frame Icon**

I certify that to the best of my knowledge (1) the statements herein (excluding scientific hypotheses and scientific opinions) are true and complete, and (2) the text and graphics in this report as well as any accompanying publications or other documents, unless otherwise indicated, are the original work of the signatories or individuals working under their supervision. I understand that the willful provision of false information or concealing a material fact in this report(s) or any other communication submitted to NSF is a criminal offense (U.S. Code, Title 18, Section 1011).

Project Director Signature: \_\_\_\_\_