

**Informedia
Digital Video Library System**

**NSF Cooperative Agreement IRI-9411299
Annual Progress Report
February 1998**

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

Principal Investigators:

Takeo Kanade
Robotics Institute

Raj Reddy
School of Computer Science

Marvin Sirbu
Information Networking Inst.

Scott Stevens
Entertainment Technology Ctr.

Doug Tygar
Computer Science Department

Howard Wactlar
School of Computer Science

1. Overview

The Informedia project is building a large, online digital video library (DVL) by developing intelligent, automatic mechanisms that can populate the library and support full-content and knowledge-based search and retrieval via desktop computer, wide-area reserved networks, and the Internet.

The distinguishing feature of our technical approach is its 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 industrial partners committed to providing substantial resources and base technology. They are also evaluating commercial opportunities for the underlying technology and for providing information services. Together these companies span the requisite disciplines for commercializing DVL technology.

2. Research and Testbed Summary

Work this contract year has focused on increasing the scale, functionality, and robustness of the IDVLS collection and its automatic processing. Through improved speech recognition and title generation, better text and image segmentation, and video OCR, autoproccessing now generates more complete and accurate metadata, indexed for user search. User queries now return more relevant and useful results, and user-supplied information — in the form of annotations — can now be indexed and searched by others. We have improved image- and face-matching and enhanced information retrieval through larger language models for transcribing speech automatically. Additionally, we have investigated how user preferences might be determined automatically and query results tailored to the individual's needs.

2.1 Creating and Using the Digital Video Library

2.1.1 Recognizing speech

Improving information retrieval

One source of Informedia's power lies in its ability to match user queries to automatically transcribed speech. In our system, as in typical applications, the speech recognizer processes each utterance in much the same way a human transcriber would. The machine, however, uses only the most probable decoding of an acoustic signal, selected from a potentially large number of hypotheses considered during the process. It is a relatively simple matter in most recognizers to obtain a list of these slightly different hypotheses, ranked in order of decreasing likelihood. We have investigated improving speech recognition by considering less-probable hypotheses.

Using the n -best hypotheses seems promising for information retrieval, since it offers the hope of including terms that would otherwise be missed by the speech recognizer in documents, allowing them to match with query terms and increase document recall. On the other hand, words incorrectly identified may cause spurious matches with query terms, decreasing query precision.

Our experiments showed that retrieval effectiveness is only slightly improved when the second and third most likely decodings for each utterance are included in documents. However, additional hypotheses decrease effectiveness. While it is encouraging that an improvement in retrieval can be obtained at all by this method, it is clear that further work will be required if the promise of this idea is to be realized.

Faster speech recognition and language understanding

Early in this contract year we developed a means to parallelize the speech-recognition process, essential for generating transcripts as well as timing data. Using a socket-based program to manage several recognizers, we effectively reduced speech-recognition processing time linearly by a factor equal to the number of speech servers (with modest overhead, up to 30 servers).

With this enhanced computational capability, we were able to implement a larger broadcast-news-specific language model for speech processing. Used with our parallel speech recognizer, the new language model provides better coverage, although it is slower.

2.1.2 Understanding text and language

Generating titles

When Informedia returns search results for a user query, it also creates titles that pop-up over each “poster frame” representing a successful match. These titles help identify content and, in previous system versions, were generated on a word-by-word basis: Starting from the beginning of the match segment, words were chosen whose weight, based on a combination of IDF and a chi-squared-based measure of document discrimination, passed a threshold, until a sufficiently long title had been generated. Headlines of this sort frequently included useful terms, but suffered from two flaws:

- They often lacked syntactic coherence. Words were simply gathered together into a list. In short, the headlines sometimes made little sense
- Word choice was based solely on corpus properties and ignored repeated appearances of a word within a document, even though repetition (term frequency) is an important indicator of topic salience.

The new title generator replaced the lexical approach with a whole sentence approach. For each sentence in the document, the average TF/IDF weight of the words in the sentence is computed, and sentences are ranked. Excessively short sentences are excluded. The highest ranked sentence is subjected to a template-based editing process which attempts to reduce the sentence to more “headline” like form by means such as shortening titles, replacing entity names with acronyms and removing sentence adverbials. Sentences are added to the prospective title in this way until it reaches the maximum allowable length (60 characters in the current interface).

Segmenting text

Segment boundaries can be located roughly using simple image statistics (such as detecting changes in color histograms, or the presence of black frames), but more accurate and useful segmentation requires concurrent use of image, speech, and language information. One approach on which we focused is extracting relevant features incrementally and automatically using a new machine learning paradigm based on exponential models in statistics. We use language models in a novel way to gauge topic changes in the text stream. We applied these ideas to the subproblem of text segmentation and evaluated it on transcripts of CNN news broadcasts, with very promising results. Our approach was directly compared to competing methods based on Hidden Markov Models and information retrieval techniques in the Topic Detection and Tracking (TDT) study, and was shown to outperform them in the broadcast news domain.

We are working to extend our text segmentation experiments to multimedia segmentation, and have completed the initial work of extracting the relevant audio and video features in order to complement the information available through closed captions. We have also hand-labeled the segment boundaries in more than 12 hours of CNN WorldView shows, in order to calibrate the problem and enable us to compare human performance (and consistency) to the machine output. In addition to the algorithms we have developed using exponential models, we are training more traditional classifiers based on decision trees and neural networks..

Recognizing text within imagery

Text contained within video imagery represents another source of DVL text content. To exploit such data, we have developed an optical character reader for video data. Our Video OCR system is based on a conventional OCR method to which we add preprocessing (such as image enhancement using linear interpolation or video data properties) and postprocessing (which corrects the OCR result by comparing it with dictionaries). Our initial implementation correctly recognized 69.5% of words overlaid on video.

We improved our Video OCR preprocessing by developing a character extracting filter, which is realized as a combination of four line extractors. This new technique improved text extraction from the video background, and increased our character recognition rate to 85%. We are also evaluating Caere and Xerox OCR Development Toolkits for use as the pattern recognition engine to recognize all text fonts and sizes that appear in television broadcasts.

We also combined Video OCR with our Name-It system that matches names to faces. This experiment suggest that a multimodal approach to face/name association can not only improve the name-recognition rate, but can also learn names automatically using video captions.

We are now implementing semantic parsing of recognized text and investigating alternatives to the filtering mechanism that extracts text from an image. Our original approach employed a horizontal-differencing filter; replacement possibilities include morphological filtering and matched filtering. Preliminary results are very encouraging, resulting in better text detection with fewer false alarms. Our demonstration of Video OCR at the June DLI Meeting attracted considerable interest.

2.1.3 Understanding video images

Segmenting video

Our new "Spot-It" segmentation technique divides video data into small image-text associated data segments and makes a new representation of a news story as a flow of small elements. Currently, we can divide a news program into stories and segment a story into small video elements. We find that this data structure for news video is valuable for precise video skimming or integration of news stories across time.

Analyzing video structure

Determining the physical and semantic structure of an extended video sequence is essential for providing appropriate processing, indexing, and retrieval capabilities for video databases. To this end we are exploring a novel technique, developed at the University of Maryland, that reduces a sequence of MPEG-encoded video frames to a trail of points in a low dimensional space. In this space one can cluster frames, analyze transitions between clusters, and compute properties of the resulting trail. By classifying portions of the trail as either stationary or transitional, gradual edits between shots can be detected. Furthermore, tracking the interaction of clusters over time lays the groundwork for the complete analysis and representation of the video's physical and semantic structure.

Image matching based on perceived color

We also investigated a new image matching method based on human color perception. This new approach will produce a new content-based image retrieval methodology and a corresponding new video segmentation method with more accurate scene change detection.

Any color image with 24-bit color resolution consists of up to 16 million colors. However, human eyes are not sensitive enough to discriminate colors with trivial differences.

Research has shown that colors with Godlove distance below 3 in the Munsell color space are perceived as the same color by humans. Based on this observation, our new method performs color clustering for each input image in the Munsell color space, where colors with Godlove distance below 3 are all grouped into the same cluster. The number of clusters from each image is variable, depending on the color distribution of the image. The cluster set is then stored in the database, and used as the collective index for the image.

With image retrieval the user forms a database query by specifying a sample image. The system clusters colors in the sample image, displays the obtained cluster set, and asks the user to select which cluster or combination of clusters to use for image matching. After the user selects the desired cluster/cluster set, the system conducts the nearest neighbor search for each of the user-specified clusters. Suppose an image set $U(i)$ is retrieved from the database based on cluster i , where $i=1,2,\dots,n$. Then, the final retrieved image set is obtained by taking the intersection of the sets among $U(i)$.

Initial experimental evaluations of the new method have shown impressive image retrieval results. Our next task is to introduce more user controls on the image retrieval process, such as setting more weight for hue, reducing weight for brightness, etc., in performing nearest-neighbor search for the user-specified clusters.

Region matching for content-based image retrieval

During this year, we developed a compound region match scheme for content-based image retrieval. This match scheme enables users to retrieve images that contain a particular shape S . Shape S may have an internal structure, or contain several component regions; there are unlimited ways of forming shape S . For example, in image I , S might be composed of the region set $A=\{a_1,\dots,a_m\}$, while in image J it might be composed of the region set $B=\{b_1,\dots,b_n\}$. To match $S=A$ with $S=B$, we developed an effective method to determine quickly each combination of the component regions that collectively form a shape similar to S . Experimental evaluations have shown that this scheme has considerably enhanced the image retrieval capability of our content-based image retrieval system.

Detecting video shot boundaries

Researchers have proposed numerous video-segmentation methods to detect shot boundaries in video images. Some methods are too sensitive to camera/object motions and minor illumination changes, causing false alarms, while others cannot detect shot boundaries when the frames on both sides of the boundaries have a similar color distribution but different spatial structures. We have proposed a new video-segmentation method that is highly sensitive to shot boundaries, while less susceptible to illumination changes and camera motions. Our approach effectively suppresses both intraframe and interframe color variations caused by noise or minor illumination changes, as well as camera and object motions. We use a novel method of extracting color distribution from video frames and two metrics devised to monitor both spatial and color changes between subsequent frames. To combine the two metrics for shot boundary detection systematically, our method includes a multivariate discriminant analysis to find the optimal coefficient for them. Our comparison study shows promising performance for this method.

2.2 Data Organization, Networking Architecture, and Interoperability

Most of the infrastructure issues in this year addressed scaling, as our project has grown in several important dimensions: incrementally accumulating online content, extending the user base, and serving remote users over longer distances. These three requirements have forced the Informedia architecture to become highly distributed and thus have driven the following activities during this period.

To support remote playback over a low-bandwidth network — until the Internet's inherent bandwidth problems are resolved, we have investigated several interim strategies:

- We demonstrated a prototype, Web-based slideshow (continuous audio with synchronized video images) using Netscape's new Media Server and latest browser (Netscape 4.0).
- We evaluated Oracle's "Video Server" product to stream MPEG-1 over TCP sockets.
- We evaluated new Internet playback packages that use proprietary video encoding, optimized for the Internet: Vivo, Vxtreme, ActiveMovie (Microsoft), RealVideo (Progressive Networks).

None of these, unfortunately, provides the functionality that Informedia's video navigation requires.

We redesigned our Web interface to remove any Java applet requirements because of their slow performance. The new interface uses dynamic HTML and Javascript as a faster alternative. Some interface issues remain to be solved outside of Javascript, e.g., synchronizing a scrolling transcript with the audio playback.

To implement "slideshows," we chose the RealAudio player for playback. The "slides" are timestamped "scenebreak" images, provided in Version 1.41 of the IDVLS system, and are synchronized to the audio. This method performs well even over low-bandwidth modem connections.

We videotaped the DLI conference at Carnegie Mellon with the intention of demonstrating our Web client using the the conference talks as the corpus. Unfortunately, the capture quality was too poor to provide a useful range of experimental data.

2.3 IDVLS Client and Database

2.3.1 Client Version 1.41 — Annotations, Shotbreaks, Smooth Skimming, and Database Conversion

Version 1.41 of the Informedia Digital Video Library System, released during the first calendar quarter, featured improvements in face matching, image matching, and filmstrip presentations (representing a video as a sequence of thumbnail images). Early prototyping was done to allow user notes to be added dynamically as the library is used, and for these notes to be immediately available through the query interface for searching and browsing. Other changes included better support for audio-only data and an improved color-histogram matcher that operated on filmstrip images.

A utility was created to hand-generate skim videos so that human-generated skims could be compared to automatically-produced ones.

We introduced a new metadata type to represent “shotbreaks.” These are similar to what we previously stored as “filmstrip” data; however, we additionally stored the begin/end time stamps which delimit each “shot” using rules of cinematography. These timestamped images allowed us to further deconstruct the video more finely (that is, to “shots” within a segment).

Representative poster frames for video segments were dynamically chosen based on closest shotbreak frame to query matches, instead of using a predetermined static image.

We scaled the image (histogram) and (eigen)face databases dramatically in size. We now include all shotbreak images (average approximately 600 static images per hour of video), whereas we previously only included poster frames for segments (average approximately 25 per hour of video).

Previously we implemented skims as playlists that the client would use to instruct the media player which parts of the movie to play. This strategy had two drawbacks; first, the video data and audio track were necessarily aligned from the original MPEG. This is not always desirable in a video skim, where in some places, it is useful to include a montage of short video segments aligned to one longer audio phrase. Secondly, the play-lists caused many media players to playback inaccurately, since they were quickly being asked to “seek-play-stop” in rapid succession. To try to solve these problems we designed a strategy to create a “static” skim, i.e., create a new MPEG file by dissecting the original audio and video tracks, finding the most relevant portions for the skim (independantly of each other), and then merging the resultant segments back together.

2.3.2 Version 2.06d — Conversions, Parallelization, and Staged Processing

We demonstrated the use of Video OCR and topic annotations in IDVLS v.2.06d. Also, we successfully demonstrated the dynamic addition of annotations (user notes) to the database, with those notes being immediately searchable and viewable. We have prototyped the use of commercial database filters to reduce a search result set to video segments meeting certain criteria, like date ranges and video play duration ranges. We removed many dependencies on 3rd party .ocx software (e.g., we replaced Dolphin Systems Inc. dssock32.ocx for TCP/IP control with Microsoft Winsock control 5.0, and replaced MediaArchitects' MediaKnife .ocx and .dll with picture box manipulations within Visual Basic).

We have expended considerable effort making the system more robust and more time efficient, with some restructuring effort for future parallelization work. We have implemented multilayer backups that allow for failsafe recovery in the presence of corrupt data and failed backups.

System components that can be parallelized have been identified and structured more efficiently. The goal is to target three types of parallel environments:

- Multi-CPU machines, where tasks are simply forked
- LSF-PVM environments, where these packaged systems deal with resource allocation
- A controlled environment, where parallel tasks are assigned to other machines and operated through shell scripts.

We have also begun investigating “hurry-up” algorithms, where the system does processing analysis as quickly as possible to generate usable data and then when resources permit, a more costly analysis is performed.

2.3.3 Library Database

We chose to move the Informedia code to a relational database for three fundamental reasons: extensibility, scalability (over, for example, data, data types, and users), and reliability, albeit at a considerable cost in overall performance.

We produced a prototype of the Informedia DVL system structured on a commercial RDBMS, to which we transitioned all analysis and metadata. We are currently using Informix and Interbase, but IDVLS can be implemented on many standard database systems (we have also tested on Oracle and Sybase).

To minimize client changes necessitated by the transition, we also constructed a new API that mirrors our current data API. We are also implementing a distributed architecture that enables a separation of database, query mechanism, and video. This architecture will provide replication of function and data for performance and robustness.

2.4 Testbeds, Specialized Corpora, and User Studies

In September we conducted an empirical study into the effectiveness of the “skim” multimedia abstraction. Twenty-five Carnegie Mellon students participated. Results indicated that “selective” skims — in which skim components were based on the most highly weighted phrases — receive significantly higher satisfaction scores than do subsampled skims composed of video extracted at fixed intervals. Significant benefits were found for skims built from audio sequences meeting certain criteria, with video synchronization playing a role.

We are also collaborating proactively with several faculty and staff at our testbed site, a local K-12 school, and are supporting our partners there as they develop “educational artifacts” that apply Informedia technology in practical teaching contexts.

2.5 NetBill Authentication and Billing System

During the past year NetBill completed implementing and testing a capability for handling compound goods and deployed a merchant software version that allows a merchant to fetch goods from an unrelated web server. Netbill used this new fetch capability to implement a new merchant that provides site-licensed access to *Pittsburgh Post Gazette* news stories for users holding a Carnegie Mellon credential.

The NetBill Alpha trial system and the PC MoneyTool are now generally available to users on the Internet, and we improved system automation to minimize the ongoing operations support required.

3. Significant Event

The Informedia system's new ability to "read" video captions and annotations offers users significant help in locating relevant topics in digital news and documentary video archives. To provide this functionality, we implemented and incorporated a technique called Video OCR, which automatically recognizes the presence of embedded text in digital video data then extracts, recognizes, and interprets its content in context.

Video OCR can provide unique information regarding the content of video news data, but poses some challenging technical problems. We addressed the preprocessing problems of low-resolution data and complex backgrounds by combining subpixel interpolation on individual frames and multiframe integration across time. We also evaluated new techniques for character abstraction and segmentation using specialized filtering and recognition-based character segmentation. Postprocessing techniques further improved accuracy, and overall word recognition rates reach 70% on news-caption data.

Information gained through Video OCR is often unobtainable from other video-understanding techniques. The approach can not only enhance conventional video libraries that rely on text-based search but can also enable new modes of understanding video content, methods not currently possible, such as: matching faces to names, associating different types of information, identifying advertisements, tracking financial or other statistics across time, and capturing the content of a video sequence described only by music and captions.

4. Publications

[Christel, Winkler, and Taylor 97a]

Christel, M.G., D.B. Winkler, and C.R. Taylor.

Improving Access to a Digital Video Library.

In *Human-Computer Interaction: INTERACT97, the 6th IFIP Conference on Human-Computer Interaction*, pages 524-531. IFIP, July, 1997.

Sydney, Australia.

[Christel, Winkler, and Taylor 97b]

Christel, M.G., D.B. Winkler, and C.R. Taylor.

Multimedia Abstractions for a Digital Video Library.

In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 21-29. ACM, July, 1997.

Philadelphia, PA.

- [Christel, Winkler, and Taylor 98]
Christel, M.G., D.B. Winkler, and C.R. Taylor.
Evolving Video Skims into Useful Multimedia Abstractions.
In *CHI-98: Human Factors in Computing Systems*, pages 171-178.
ACM SIGCHI, April, 1998.
Los Angeles, CA.
- [Hauptmann and Witbrock 97]
Hauptmann, A. and M. Witbrock.
Informedia News-On-Demand: Using Speech Recognition to Create
a Digital Video Library.
In *Working Notes for AAAI-97 Spring Symposium on Intelligent Integration
and Use of Text, Image, Video and Audio Corpora*, pages 120-126.
AAAI, March, 1997.
Stanford, CA.
- [Hauptmann, Witbrock, and Christel 97]
Hauptmann, A., M. Witbrock, and M.G. Christel.
Artificial Intelligence Techniques in the Interface to a Digital Video Library.
In *CHI-97: Human Factors in Computing Systems*.
ACM SIGCHI, March, 1997.
Atlanta, GA
- [Nakamura and Kanade 97]
Nakamura, Y. and T. Kanade.
Spotting by Association in News Video.
In *Proceedings of the AAAI 1997 Spring Symposium: Intelligent Integration and Use
of Text, Image, Video and Audio Corpora*.
- [Satoh and Kanade 97]
Satoh, S and T. Kanade.
Name-It: Association of Face and Name in Video.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,
1997.
- [Smith and Kanade 97]
Smith, M. and T. Kanade.
*Video Skimming and Characterization through the Combination of Image and
Language Understanding Techniques*.
Technical Report CMU-CS-97-111, School of Computer Science, Carnegie Mellon
University, February, 1997.
Also to appear in CVPR97.
- [Witbrock and Hauptmann 97]
Witbrock, M. and A. Hauptmann.
Speech Recognition in a Digital Video Library.
JASIS: Journal of the American Society for Information Science, 1997, in press.
- [Witbrock and Hauptmann 97]
Witbrock, M. and A. Hauptmann. "Using Words and Phonetic Strings for Efficient
Information Retrieval from Imperfectly Transcribed Spoken Documents," to appear in
DL97, The Second ACM International Conference on Digital Libraries, July 23 - 26,
1997, Philadelphia.

[Witbrock and Hauptmann 97]

Witbrock, M. and A. Hauptmann. "Improving Acoustic Models by Watching Television," in Working Notes for AAAI-97 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora. March 24-26 1997, Stanford, pp 61-63

[Witbrock and Hauptmann 97]

Witbrock, M. and A. Hauptmann. "Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents," in Proceedings of the 1997 DARPA Speech Recognition Workshop, Feb 1997.

Witbrock, M.J. and Hauptmann, A.G. Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents, in DL97, The Second ACM International Conference on Digital Libraries, July 23 - 26, 1997, Philadelphia.

5. Presentations, Demonstrations, and Industry Visitors

2/13/97 Telecom Italia, Stefano Berccia, Saracco, Cortesi

3/5/97 MAYA Design Group. Steven Little, Director, Human Science Group;

Philip Ashe, Director, Strategic Planning, Storage Products, Digital and Applied Imaging.

3/5/97 Eastman Kodak Company. Stanley Caplan, Human Factors Engineer, Systems Concept Center.

3/5/97 Open University. Jeremy Chapple, Vice President of Operations.

3/6/97 NACSIS. Prof. Atsuhiko Takasu, Dr. Norio Katayama.

3/11/97 Curtis Industries. Maurice P. Andrien, Jr., President/CEO; A.

Keith Frewett, President, A & I Division; Idelle K. Wolf, VP of Finance and CFO; William Beaver, Director of Marketing; Anthony Morabito, Director of Information Systems; Corrine Reali, Director of Purchasing; James P. Waters, Director of Finance.

3/26/97 Andersen Consulting, Joseph McCarthy, Senior Research Associate.

3/31/97 NHK. Akihiko Chigono, Engineering Administration Department

HAYAMA; Minoru, Deputy Director, Media Development Department, Library &

Data Services Division; IZUMI Hiroshi, Senior Officer, Information Systems Bureau; TSUCHIYA Kaoru, Senior Manager, System Development Department; CHIGONO Akihiko, Senior Engineer, Engineering Administration Department, Engineering Development Center.

3/31/97 NASA Ames, Mark Leone, Gene Mia.

4/05/97 Information Management and Fusion Group, Dr. Lakshmi Narasimhan, Head, Information Technology Division, Salisbury, South Australia, lakshmi.narasimhan@dsto.defence.gov.au

4/7/97 MITRE, Lynette Hirshman.

4/13/97 DEC SRC. Michael Schroeder.

4/14/97 Mitsubishi, John Howard.

4/14/97 Intel, Laura Good.

4/16/97 ATR, Dr. Nakatsu, President.

4/22/97 Weirton Steel; Craig Costello, Earl Davis, Pat Pathipati, Len Jenkins, Pat Stewart, Debo Aichbhaumik, Dick Riederer.

4/29/97 AT&T, Joel Winthrop.

5/1/97 Scuola Superiore Sant'Anna, Italy. Student presentation.

Presentations:

3/2-5/97 Stevens, S. and Marinelli, D. "Synthetic Interviews," ACM97, San Jose, CA.

4/7/97 National Association of Broadcasters. Integrating Speech, Natural

Language, Image Processing and Information Retrieval: The Informedia Digital Video Library Project, Machines that Learn Conference, Snowbird, Utah.

4/17-20/97 Academia Europaea: The Impact of Electronic Publishing on the Academic Community, April 17-20, 1997. Stockholm, Sweden.

4/21-24/97 Wactlar, H., and Hauptmann, A. "Indexing and Search of Multimodal Information," ICASSP '97 International Conference on Acoustics, Speech, and Signal Processing. Munich, Germany.

6/27/97 Eiji Sugawara, Assistant Div. Manager, Space Communications Corporation;

Hiroyuki Izumi, Sakura Bank;

Hitoshi Hatanaka, Sales Manager, NEC Corporation;

Izumi Tamuki, Assistant Manager, NEC Corporation;

Katsumi Kanai, Matsushita Electric Industrial Co.,Ltd.;

Keiji Okumura, Manager, Mitsubishi Electric Corporation;

Kozo Tsuchiya, Benesse Corporation;

Masakazu Imai, Associate Professor, Nara Institute of Science and Technology;

Masaki Abe, Assistant to Dept. General Manager, Mitsubishi Electric Corp.;

Masanobu Fukawa, Assistant Manager, Mitsubishi Corporation;

Masayo Oka, General Manager, Ikegami;

Setsuji Arika, Assistant General Manager, Space Communications Corp;

Shigeru Takemoto, Japan Radio Co.bLtd.;

Shin-ichi Tanaka, Japan Satellite Systems Inc.;

Toru Takanashi, Benesse Corporation;

Yuuko Nakamura, Nomura Research Institute, Ltd.;

Mr. Yasuhito Iwasaki, Deputy General Manager, Maruzen Co.

6/30/97 Dr. Masatsugu Kidode, Senior Vice President and General Manager,
Advanced Information Technology Center, TOSHIBA America, Inc.

8/01/97 Phil Chomsky, CMU benefactor, and ~15 students and faculty from
Pittsburgh's Electronic Institute (EE/CS vocational technical school)

Presentations:

Name-It Demo in DLI Meeting '97 at CMU

Spot-It presentation and demonstration in DLI Meeting '97 to illustrate importance of semantic analysis such as "Spot-it" to find smaller video segments. As an end of the first phase, we submitted a paper to ACM Multimedia '97 and it was accepted.

Informedia DVL at DLI Meeting '97

10/97 Dr. Janusz Marszalec, VTT Electronics, Optoelectronics

10/97 Dr. Masafumi Hagiwara, Keio University, Japan

9/10/97 Eastman Kodak Company: Dr. Craig S. Willand, Systems Concept Center,
Imaging Research & Advanced Development. Kodak Imagination Works: Jim
Stoneham, Managing Director; Donald Olson, Director, Software.

9/19/97 Sony: Norisha Suzuki, President, Chief Technology Officer. Along
with C. Yankowski, T. Hasebe.

9/19/97

Mdm. Norkhayati Hashim, National Library of Malaysia

Mdm. Norpishah Mohd. Noor, Dir. of Library Development Division, National
Library of Malaysia

Mdm. Norkhayati Hashim, Director of Library and Information Technology
Division

Mr. Johnny Kueh, Sarawak State Librarian

Mr. Ong Chai Lin, Penang State Librarian

Mrs. Ku Joo Bee, Senior Librarian from Sabah State Library

Mdm. Norsham Abu Hassan, Officer from Subang Jaya Municipal Council

Mdm. Asha from STAR Library

10/21/97 Karen Fullerton, Pen-Dor Project, University of Pittsburgh.

10/22/97 Andrew Johnson, Principal Engineer. Interactive Multimedia Services.

Telstra.

Presentations:

8/21/97 Pittsburgh High Technology Council. Demo.

9/1/97 Digital Libraries conference, Pisa, Italy. "Search and Discovery in the Video Medium". Presentation and demo.

9/24/97 Voice on the Net Conference, Boston, MA. "Search and Discovery in the Video Medium". Presentation and demo.

PROJECT SUMMARY

DATE PREPARED:

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.

RECENT ACCOMPLISHMENTS:

Dynamic annotations

new corpus materials: CNN, documentaries, humanities

Infrastructure evolution and growth

perceptual color clustering

parallelization --> scaling

PLANS:

Video OCR experiments will continue, concentrating on semantic parsing of recognized text. We will investigate possibilities for enhancing text recognition itself by trying additional filtering mechanisms, ideally resulting in better text detection with fewer false alarms.

Video tracing remains an exciting possibility for determining the physical and semantic structure of an extended video sequence. We will further explore this technique to detect gradual edits between shots.

We will work to develop a distributed architecture for Informedia.

TECHNOLOGY TRANSITION, SHARING, PARTNERING, ETC.:

We now have the IDVLS installed over the ATDnet on OC3 datarate links. Multiple hosts in Washington D.C. metropolitan area, including locations at DARPA headquarters and Fort Mead Maryland, providing remote delivery over high bandwidth connections. At these locations, users have full client access, with all data optionally served from CMU, including the speech recognition server which provides recognition of spoken queries.

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: _____