

**FINAL TECHNICAL REPORT**  
*High Performance Distributed Services Technology*

CDRL A001  
CONTRACT N66001-97-D-8502, Delivery Order 0002

01 August 2000

**Multilingual Informedia Research and Demonstration Support**

SUBMITTED TO

Receiving Officer  
e-mail address: "hpdstnrad@spawar.navy.mil"

Rich Laverty  
619-553-2918  
laverty@nosc.mil

Frank Schindler  
619-553-2845  
fschindl@nosc.mil

Bob Medearis  
619-553-6377  
medearis@nosc.mil

SUBMITTED BY

**Carnegie Mellon University**  
**Pittsburgh, PA 15213**

Principal Investigator  
Howard Wactlar  
(412) 268-2571  
fax: (412) 268-5576  
wactlar@cmu.edu

Administrative Contact  
Colleen Everett  
(412) 268-7674  
everett@cs.cmu.edu

Contract/Financial Contact  
Karen M. Faber  
(412) 268-5838  
(412) 268-5841  
faber+@andrew.cmu.edu

Distribution authorized to DOD components only.  
Do not distribute to DTIC or other data depositories.

## Section 1

### 1.1 Report Title

Multilingual Informedia Research and Demonstration Support

### 1.2 Notices

Carnegie Mellon University will retain all rights to all intellectual property, hardware and software developed under this proposed contract, but will license them to the United States government at no charge for its own use.

### 1.3 Abstract

The purpose of this report is to summarize Informedia's efforts to develop a real-world usable broadcast news indexing system, focusing on:

- Providing full access to Informedia as a network service over an appropriately high bandwidth link.
- Enhancing the Informedia methodology to handle alternative languages enable cross-lingual retrieval and provide tools for rudimentary analysis of foreign-language material.

### 1.4 Table of Contents; List of Tables, Figures and Illustrations.

Section 2	
2.1 Summary .....	3
2.2 Introduction .....	3
2.3 Methods, Assumptions and Procedures .....	4
2.4 Results and Conclusions .....	7
2.5 Recommendations .....	7
Section 3	
3.1 Bibliography .....	9
Exhibit A: Research Papers .....	11

### 1.5 Preface and Acknowledgements.

We acknowledge the contribution of our colleagues at the Interactive Systems Labs at the University of Karlsruhe, Germany, and the Interactive Systems Lab at Carnegie Mellon.

### 1.5 List of Symbols, Abbreviations, and Acronyms

MMR            Maximal Marginal Relevance

## Section 2

### 2.1 Summary

The purpose of the Multilingual Informedia project was to develop automated systems and tools enabling multilingual and multimedia information capture, search, retrieval, summarization and reuse. The system, built on the underlying Informedia Digital Video Library system concepts, technology and infrastructure, is designed to access textual, audio (radio) and video (TV) information, to index, categorize, retrieve, summarize and analyze it, in one or multiple languages.

We implemented and demonstrated a prototype multilingual browser of text, video and radio material that accepts English queries and returns the most relevant Serbo-Croatian, German and English language reports or segments in their original language, in full or summary form. For example, this would enable an analyst to compare divergent American and foreign reporting of the same event or topic.

We built and delivered functional broadcast news-focused systems to multiple, network-connected, offsite locations including DARPA and NSA. Network delivery issues were being addressed and system architecture was being redesigned to improve performance when anticipated project funding was curtailed.

### 2.2 Introduction

From its inception in 1995, the Informedia project's goal has been to allow search and retrieval in the video medium, similar to what is available today for text only. To enable this access to video content, speech recognition is used to provide a text transcript for the audio track; and image processing determines scene boundaries, recognizes faces and allows for image similarity comparison. Everything is indexed into a searchable digital video library, where users can ask queries and receive relevant news stories as results.

The Multilingual Informedia Project pursues a seamless extension of the Informedia approach to search and discovery across video documents in multiple languages. Previously, we successfully demonstrated that current speech recognizers allow accurate information retrieval for automatically processed English news TV broadcasts. The multilingual system performs speech recognition on foreign language news broadcasts, segments it into stories and indexes the foreign data in parallel with existing English news data. This first multilingual prototype should be easily extensible to other languages.

Under this contract, DO#0002, we developed network delivery strategies and techniques to enable remote access to large, centrally-located libraries of data. Additionally, we explored indexing and accessing material in alternate source languages (e.g. Spanish, Mandarin, Chinese).

We built upon an existing technological base at Carnegie Mellon that includes:

- **Speech recognition.** (JANUS and SPHINX). Both are state-of-the-art large vocabulary continuous speech recognition systems and are both available as a client server architecture
- **Single word spotting.** Performance 1st and 2nd in ARPA benchmark tests, 1992, 1993
- **Speech translation.** The JANUS system, now handles two channel spontaneous human-human dialogs without push-to-talk button
- **Information retrieval.** CMU developed the pursuit engine used in LYCOS.

- **Statistical language modeling** for speech recognition and translation; proven methods in Pangloss, Sphinx, etc.
- **Informedia digital video library** built at CMU includes automated indexing of text, audio and video, full-retrieval, initial summarization (gisting) of video segments
- **Linguistic resources:** dictionaries, phrase books in German, Spanish, French, Japanese; Serbo-Croatian accessible, including access to Serbo-Croatian, German, and other broadcasts and resources.
- **Machine translation:** KANT (large-scale accurate interlingual translation within domain), Pangloss and EBMT, example-based machine translation (general-purpose translation assistants).
- **Summarization:** Initial metrics and methods for automated on-demand summary generation for text and gisting for video.

## 2.3 Methods, Assumptions, and Procedures

### Foreign Language Topic Detection

Initial experiments with foreign-language demonstrated that returning a foreign language result to the user was not sufficient. The users were unable to tell if a particular news clip was actually relevant to their query, or if it was returned due to poor query translation or inadequate information retrieval techniques. To allow the user at least *some* judgment about the returned stories, we attempted to label each foreign news story with an English-language topic.

Topic identification was done using the Informedia translation facility to translate the whole story into English words. This translation became the *topic query*. Separately, we had a collection of about 35,000 English language news stories from Prime Source Media, which had topics manually assigned to them. Using the SMART information retrieval system, we now used the translated *topic query* to retrieve the ten most relevant labeled English stories. Each of the topics for the labeled stories that were retrieved was weighted by the relevance of its respective *topic query* result and the weights for each topic were summed. The most favored topics above a threshold were then used to provide a topic label for the Croatian news story. This topic label allows the user to identify the general topic area of an otherwise incomprehensible foreign language text and determine if it is relevant at least in its topic area coverage.

We found the approach to be promising and the initial results were encouraging. In particular, a high-quality machine translation provided topic assignment results comparable to perfect translations. The quick-and-dirty phrase translation system, however, showed noticeable degradation in the topic assignment.

### Multidocument Summarizer

As a tool for analysts using Multilingual Informedia, we built a synthetic multidocument summarizer based on query/profile relevance. The primary result of this effort has been the creation of the Maximal Marginal Relevance (MMR) metric and its successful implementation as a means of improving document retrieval, single-text summary generation, and most recently multi-text summary fusion. In essence, MMR combines relevance with novelty in both document selection for IR, and passage selection for constructing synthetic multi-text summaries.

Relevance is established by the IR system that retrieved that document, e.g. cosine similarity between query and word vectors or any other computable metric. In essence, we search for the document that is both most different from the one already scanned but that still scores high on query relevance. This

method promotes a maximal-diversity search, still providing report(s) of the most relevant event first, but then switching to reports of other relevant events. In essence, the new method ranks the documents dynamically by the marginal query-relevant information gain per additional report. In this manner, both relevance-to-query and diversity from already scanned information are considered, and a tunable function of the two is optimized when selecting the next document in the ranking. The objective is simply to provide the analyst with a maximal-diversity sampling of information pertinent to the query. The maximal marginal relevance (MMR) metric determines the differential relevance of a document with respect to documents already seen.

Beyond single-document summarization, a synthesized summary of a set of documents -- such as those output by the retrieval engine with respect to an analyst's query -- often proves more desirable. We first apply the same method for localization of relevant document segments as single-document summarization to each passage in each document in the relevant retrieved set. Then, we filter these for redundancy (noting, if desired, the various sources) using the MMR method developed for improved ranking. Finally, we assemble by topic-cohesion the non-redundant and query-relevant document segments into a synthetic multi-document summary at the desired level of granularity.

We extended the summarizer to full language-independence, so that it can summarize an English document in English, Spanish in Spanish, Korean in Korean, etc. MMR's statistical basis has made this enhancement relatively straightforward.

### **Dynamic Language Modeling**

We developed an automated "daily language" modeling capability that emphasized recently used vocabulary. This technique first examines Web text corresponding to CNN, AP, and Reuters news stories of the day. By interpolating that text with general English data, we build a language model specialized to current usage. Speech-recognition errors dropped to approximately 19% on news stories, validating our earlier experiments with simulated daily data.

### **Segmentation and Commercial Detection in Broadcast News**

Segmentation is an integral process in the Informedia Digital Video Library. The success of Informedia hinges on two critical assumptions; that we can extract sufficiently accurate speech recognition transcripts from the broadcast audio, and that we can segment the broadcast into video paragraphs or stories that are useful for information retrieval.

The story segmentation step for the Informedia Digital Video Library splits full-length news broadcasts into individual news stories. During this phase the system also labels commercials as separate "stories". Informedia takes advantage of closed captioning which is frequently broadcast with the news, extracts timing information by aligning captions with the result of speech recognition, and integrates closed-caption cues with the results of image and audio processing.

### **Spotting by Association**

Spotting by association is a novel method to detect video segments with typical semantics. Video data contains various kinds of information through continuous images, natural language and sound. For videos to be stored in retrieved in a digital library, it is essential to segment the video data into meaningful pieces.

### **VOCR**

Video OCR is a technique that can greatly help to locate topics of interest in a large digital news video archive via the automatic extraction and reading of captions and annotations. News captions generally provide vital search information about the video being presented -- the names of people and places or description of objects. Our research addressed two difficult problems of character recognition for video: low resolution characters and extremely complex backgrounds. We apply an interpolation filter, multi-frame integration and a combination of four filters to solve these problems. Segmenting characters is done by a recognition-based segmentation method and intermediate character recognition results are used to improve the segmentation. The overall recognition results are good enough for use in news indexing. Performing video OCR on news video and combining its results with other video understanding techniques can improve the overall understanding of the news video content.

### **Network/Bandwidth Issues**

A multimedia database like the core of the Informedia system contains a large number of short segments - video clips, sound bites from TV news, movie story boards that provide access to individual scenes and shots, commercials, etc. Today the size of the data collection is about 1 TeraByte. To allow *remote* access to such a data collection, researchers at Carnegie Mellon investigated transmission of video over existing best-effort networks that make up the Internet. The key problem is ensuring that the player receives a continuous feed in the presence of variable congestion. The goal becomes maximizing the transfer of comprehensible information given the available, varying bandwidth.

The MPEG format is a widely used vehicle for the distribution of video and audio material over the Internet. However, the hierarchical structure of MPEG systems complicates the task of delivering continuous, synchronized streams of video and audio in a best-effort environment (today's Internet). If the network throws away packets on encountering congestion, the video and audio stream may lose synchronization for a number of frames. Therefore, adapting the resource demands of an MPEG system must be done by an entity that is knowledgeable of the MPEG system structure: an MPEG system filter. Experience to date indicates that mid-range PCs can host such a filter, and that the filter succeeds in adapting the resource requirements of an MPEG system in response to changes in the network load.

The final remote NoD installation experienced fairly severe system performance degradation. We isolated the problem to insufficient bandwidth for handling the query traffic from a remote search engine and client to a local library. The client was being reengineered as funding was curtailed.

### **Semantic Analysis of Video Content**

Video data contains various kinds of information: continuous images, natural language, sound, etc. For videos to be stored and retrieved in a digital library, it is essential to segment the video data into meaningful pieces. We developed a novel method to detect video segments with semantically-similar contents. Our method identifies a correspondence between *image clues* detected by image analysis and *language clues* detected by NL analysis. That video segment is tagged automatically with one of several labels that reflect the semantic contents indicated by the correspondence. Similar content then becomes another search constraint.

### **Relevant Informedia System Improvements**

### Map Interface

Through automatic processing, descriptors are derived for the video to improve library access. A new extension to the video processing is the extraction of geographic references from these descriptors. The operational library interface shows the geographic entities addressed in a given story, highlighting the regions discussed at any point in the video through a map display synchronized with the video playback. The map can also be used as a query mechanism, allowing users to search the terabyte library for stories taking place in a selected area of interest.

The real value to the geocoding and map interface is in displaying location information for video segments for which the producer has not previously added a map within the video data. Every news story does not have an embedded map that becomes part of the broadcast. Through our geocoding, however, maps can be produced *automatically* to reflect the areas discussed within each story. Another benefit is that the user can interact with the interface map using toolbar icons to get additional detail; no such interaction is possible with an image of a map encoded as part of the video stream.

The maps accompanying videos are not static displays. They are animated in synchronization with video playback. Places discussed are highlighted on the map. For countries and administrative areas such as states or provinces, the areas contained within their respective polygon boundaries are highlighted by changing the color with which they are shaded. Areas covered at some time during a video segment are colored yellow; when the video frames during which an area is discussed are played, that area is then colored orange. For cities and other places, the marker color is changed from blue to red and the label is shown. A glance at the map can then show the areas of current focus.

### Multimodal Search and Relevance Feedback

We investigated multimodal queries, specifically, searching with text while also searching indicated geographic areas. A weighting scheme was developed to combine results produced from these different modalities, with direct manipulation interfaces provided so that the user can adjust the weights to match specific needs. In addition, features are extracted from the result set for use in relevance feedback, whereby the user informs the system which results are most relevant to his needs. The features common to those results marked as relevant are used to reorder the result set accordingly. Experiments will be conducted in the next phase of Informedia research to measure the effectiveness of multimodal map and text queries and relevance feedback strategies, as compared to video information retrieval interfaces with text-only query, and text query with relevance feedback.

## **2.4 Results and Conclusions**

We demonstrated the viability of enabling for news video all the functionality and capability existing for textual information retrieval, while leveraging its temporal and visual qualities for richer information delivery. The next phase of research needs to introduce new paradigms for video information access and understanding. We will need to aggregate and integrate video content on-demand to enable summarization and visualization that provides responses to queries in a useful broader context, perhaps with historic or geographic perspectives.

## **2.5 Recommendations**

There are some issues in the Multilingual Informedia client that would merit further work. One of these issues is to allow the combination of target language transcript and target language OCR, together with English language user annotations, titles and topics. Other issues include: How should a

search proceed over video library data that has mixed language information? How should the architecture be modified to handle this? What about searches over larger collections of both English and foreign language video material? How should the search results from each language be combined?

In addition, non-language information (such as visual features, overlaid text faces, color, shapes, etc.) has added importance, as it is not subject to a language-specific interpretation. We need to further advance and better integrate visual feature extraction and image-based queries.

## Section 3

### 3.1 Bibliography

Christel M., Hauptmann, A., Witbrock M. *Artificial Intelligence Techniques in the Interface to a Digital Video Library*. Proceedings of the CHI-97 Computer-Human Interface Conference, New Orleans, LA, March 1997.

Christel, M., Martin, D. *Information Visualization Within a Digital Video Library*, Journal of Intelligent Information Systems 11(3); 235-257 (1998)

Christel, M., Smith M., Taylor, R., Winkler, D. *Evolving Video Skims into Useful Multimedia Abstractions* Proceedings of the CHI'98 Conference on Human Factors in Computing System, C. Karat, A. Lund, J. Coutaz, and J. Karat, eds. pp 171-178, Los Angeles, CA, April 1998.

Christel, M., Winkler, D., and Taylor, R. *Multimedia Abstractions for a Digital Video Library*. Proceedings of ACM Digital Libraries '97 Conference, Philadelphia, PA July 1997.

Christel, M., Winkler, D., and Taylor, R. *Improving Access to a Digital Video Library*. Human Computer Interaction, Interact 97. The IFIP Conference, Sydney Australia, July 14 1997.

Gueter, P., Finke, M., Scheytt, P., Waibel, A., Wactlar, H., *Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexical Adaptation*. DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA., Feb 8-11, 1998.

Hauptmann, A., Jones, R., Seymore, K., Slattery, S., Witbrock, M., Siegler, M. *Experiments in Information Retrieval from Spoken Documents*. DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA., Feb 8-11, 1998.

Hauptmann, A., Witbrock, M. *Story Segmentation and Detection of Commercials in Broadcast News Video*. Proceedings of ADL-98 Advances in Digital Libraries Conference, Santa Barbara, CA., April 22-24, 1998

Hauptmann, A. Wactlar, H. *Indexing and Search of Multimodal Information*. International Conference on Acoustics, Speech and Signal Processing (ICASSP-97), Munich, Germany, April 1997.

Hauptmann, A., Witbrock, M. *Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval*, in "Intelligent Multimedia Information Retrieval", Mark T. Maybury, Ed., AAAI Press, pps. 213-239, 1997.

Hauptmann, A. and Witbrock M. *Informedia News-On-Demand: Using Speech Recognition to Create a Digital Video Library*. May 1997.

Hauptmann, A.G., Witbrock, M. and Christel, M., *News-on-Demand-An Application of Informedia Technology*. D-LIB Magazine, September, 1995.

Hauptmann, A., Kennedy, P.E., Lee, D. *Topic Labeling of Multilingual Broadcast News in the Informedia Digital Video Library* ACM DL / SIGIR MIDAS Workshop, Berkeley, CA, June, 1999

Houghton, R. *Named Faces: Putting Names to Faces*

IEEE Intelligent Systems Magazine, Vol 14, No. 5, pp 45-50, September/October, 1999

Jang, P., Hauptmann, A. *Learning to Recognize Speech by Watching Television* IEEE Intelligent Systems Magazine, Vol 14, No. 5, pp 51-58, September/October, 1999

Ishikawa, Y., Subramanya, R., Faloutsos, C., *MindReader: Querying Databases Through Multiple Examples*. Submitted to VLD98 Conference (Very Large Databases), New York, August 24-27, 1998.

Kanade, T., *Immersion into Visual Media: New Applications of Image Understanding*. In IEEE Expert Intelligent Systems and Their Applications, Vol. 11, No. 1, pages 73-80, IEEE Computer Society, 1996. "The SR-tree: An Index Structure for High-dimensional Nearest Neighbor Queries."

Katayama, N., Sato, S., Proceedings of ACM SIGMOD, 1997.

Lafferty, J., Della Pietra, S., Della Pietra, V., *Statistical Learning Algorithms Based on Bregman Distances*. Proceedings of the 1997 Canadian Workshop on Information Theory, Toronto, Canada., 1997.

Nakamura, Y., Kanade, T., *Semantic Analysis for Video Contents Extraction -- Spotting by Association in News Video*. Proceedings of the Fifth ACM International Multimedia Conference, Nov 1997.

Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R., Thayer, E., *The 1996 Hub-4 Sphinx-3 System*.

Submitted to proceedings of DARPA Spoken Systems Technology Workshop, Feb. 1997.

Ravishankar, M., *Some Results on Search Complexity vs Accuracy*. Submitted to Proceedings of DARPA Spoken Systems Technology Workshop, February, 1997.

Rowley, H., Baluja, S., Kanade, T., *Neural Network-Based Face Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, January 1998 and In Proceedings of International Conference on Computer Vision and Pattern Recognition, pages 203-208, San Francisco, CA.

Rowley, H., Baluja, S., Kanade, T., *Rotation Invariant Neural Network-Based Face Detection*. CMU CS Technical Report, CMU-CS-97-20, 1997

Sato, T., Kanade, T., Hughes, E., Smith, M., *Video OCR for Digital News Archives*. IEEE Workshop on Content-Based Access of Image and Video Databases (CAIVD'98), Bombay, India, January, 1998.

Sato, S., Kanade, T., *NAME-IT: Association of Face and Name in Video*. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 17-19 June, 1997.

Sato, S., Nakamura, Y., Kanade, T., *Name-It: Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing*. In Proceedings of IJCAI-97, 1997.

Seymore, K. Chen, S., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M., Stern, R., Thayer, E. *The 1997 CMU Sphinx-3 English Broadcast News Transcription System*. DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA., Feb 8-11, 1998.

Siegler, M.A., *Integration of Continuous Speech Recognition and Information Retrieval for Mutually*

*Optimal Performance* PhD thesis, Carnegie Mellon University, Electrical and Computer Engineering, December 15, 1999.

Smith, M., Kanade, T., *Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques* Computer Vision and Pattern Recognition Conference, San Juan, Puerto Rico. Pp. 775-781, June 1997

Smith, M., Kanade, T. IEEE International Workshop on Content-Based Access of Image and Video Databases, ICCV98, Bombay, India, 1998.

Wactlar, H., Hauptmann, A., Gong, Y., Christel, M., *Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library* IEEE Computer 32(2): 66-73, 1999

Wactlar, H., *New Directions in Video Information Extraction and Summarization* 10th DELOS Workshop, Santorini, Greece, June 24-25, 1999

Wactlar, H., Hauptmann, A., Witbrock, M., *Informedia: News-on-Demand Experiments in Speech Recognition*. Proceedings of ARPA Speech Recognition Workshop. Arden House, Harriman, NY, Feb 18-21, 1996.

Witbrock, M., Hauptmann, A., *Speech Recognition for a Digital Video Library*. Journal of the American Society for Information Science (JASIS) 49(7), 1998

Witbrock, M., Hauptmann, A., *Improving Acoustic Modes by Watching Television*. CMU CS Technical Report, CMU-CS-98-110, March 19, 1998.

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