

**Informedia-II:
Auto-Summarization and Visualization
Over Multiple Video Documents and Libraries**

**NSF Cooperative Agreement No. IIS-9817496
Annual Progress Report
February 2000**

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

Principal Investigator: **Howard Wactlar**
School of Computer Science

Co-PI's: **Michael Christel**
Computer Science Dept.

Takeo Kanade
Robotics Institute

Christos Faloutsos
Computer Science Dept.

John Lafferty
Computer Science Dept.

Alexander Hauptmann
Computer Science Dept.

Yiming Yang
Language Technologies Inst.

1 Overview

The Informedia-II Project will change the paradigm for accessing digital video libraries through meaningful, manipulable overviews of video document sets, multimodal queries, and adaptive summarizations of very large amounts of video from heterogeneous distributed sources.

Video information collages are the key technology in Informedia-II and will be built by advancing information visualization research to effectively deal with multiple video documents. A video information collage is a presentation of text, images, audio, and video derived from multiple video sources in order to summarize, provide context, and communicate aspects of the content for the originating set of sources. The collages to be investigated include chrono-collages emphasizing time, geo-collages emphasizing spatial relationships, and auto-documentaries which preserve video's temporal nature. Users will be able to interact with the video collages to generate multimodal queries across time, space, and sources. Video collages are made adaptive by giving preference to the concepts and query terms in the user's interaction history. The synthesis and summarization functions underlying these collages will be made possible through extensions of text clustering and expectation maximization algorithms to video and audio features.

2 Research and Testbed Summary

2.1 Video Information Collages: Adaptive Visualization and Summarization

2.1.1 Adaptive query-based and adjustable filmstrips

In order to deal with the problem of user frustration in having to vertically scroll through too many thumbnails representing a particular filmstrip, a toolbar was added to support this zooming operation as well as toggling the information frame. The user can reduce the amount of space used by the filmstrip window by "zooming out" to display each thumbnail at one-half its resolution in both the horizontal and vertical dimensions. The user can also resize the filmstrip window and the filmstrip will redraw itself accordingly, fitting as many images as possible horizontally in the window before starting another row.

Even with the changes noted above, the filmstrips still could contain over a hundred images, limiting their effectiveness as a quick abstraction for a video segment. By utilizing information derived from speech recognition and alignment along with the user's query context, we reduced the number of images included in the filmstrip window to a much smaller set.

We applied the same sort of runtime adjustment to the filmstrip abstraction, so that filmstrips could be presented differently based on the query. For a user's query, the information retrieval engine identifies the matching words within a video segment's descriptive text. This descriptive text could be a transcript of its narrative, timed tightly to the video through automatic speech alignment. The text could also be generated through "video OCR" processing, which detects and translates into ASCII format the text overlaid on video frames. Users annotating sections of video with their own notes could produce searchable text as well. Regardless of the text origination, the matching words have associated video times, enabling the matches to be plotted

against the shot and hence enabling matching words and their respective locations to be shown on the filmstrip itself.

By displaying only the shots where matches occurred, along with the opening and closing shot for the segment to provide some temporal context for that video, the number of shots can be greatly reduced.

For more detailed information, please refer to [Christel1999].

2.1.2 Dynamic adjustable video skims

The goal for video skims in the Informedia interface goes beyond motivating a viewer to watch a full video segment, instead seeking to communicate the essential content of a video in an order of magnitude less time. Based on the query context, video skims are adjusted to emphasize the audio and video surrounding match locations. Rather than being precomputed like the early work of video skim in Informedia-I, these new style video skims are generated dynamically so that context can be used to assemble better skims, e.g., following a query the skim will be assembled to emphasize the locations in the video where match words or images are found. In addition, users are able to determine the compression ratio, i.e., rather than precomputing 20:1 skims or 10:1 skims. During skim playback users can dynamically adjust their decision, e.g., if the start of a video skim proves interesting a user might decrease the compression ratio from 20:1 to 5:1 as the skim plays to get more details in the remainder of the skim.

For more details, please refer to [Christel1999].

2.1.3 Video digests

Video digests summarize sets of stories from the library, providing users with a visual mechanism for interactive browsing and query refinement. These digests are generated dynamically under the direction of the user based on automatically derived metadata from the video library. In addition to previously developed Informedia's VIBE digests (emphasizing word relationships) and timelines digests (showing trends against time), we have added maps digests (showing geographic correlations) and topic digests (showing topic relationships). We present these two new digests in the following two subsections. The availability of different kinds of video digests gives users choice of orientation in digesting retrieval video stories. Multiple digests can be combined into a single view or animated into a temporal presentation.

For more details, please refer to [Christel1999a] and [Hauptmann1999].

2.1.4 Interactive maps

A new extension to the Informedia video processing is the extraction of geographic references from these descriptors. We developed an 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 arrow icon in the toolbar for the map window is used to drag a rectangular region on the map that serves to identify the user's region of interest.

While a query region shaped as a rectangle is currently operational, we recognize the value in supporting more powerful spatial query mechanisms. For example, the user might click on a country, shift-click a number of countries, or use a bounding polygon to select a number of countries within a region. All matches to cities, political entities, countries, and other geographic references within that area would then be included in the returned set of video segments. This extension from rectangular searches to searches using more map information such as country boundaries, is provided in the features for many geographic information systems (GIS), including the ESRI MapObjects library.

For more details, please refer to [Christel1999b].

2.1.5 Topic digests

Building upon the analyzed result of automatic topic generation to all video stories, we developed a Topic Digest viewing tool to leverage topics (up to five in each story) when browsing through the library. For example, consider a user wishing to overview the past month's news. A thousand stories, henceforth referred to as documents, are part of this set, which would be too time-consuming to progress through via title-browsing or other linear scanning mechanism. The recent news can be organized to highlight the most prevalent topics, however, allowing the user to get a quick summary of recent events as well as navigation aids for focusing subsequent examinations.

The user can interact with the topic digest to tailor the presentation into a more manageable, interpretable state. This interface allows the user to see the scatterplot of documents to topics change as the slider meaning is changed. The immediate feedback of the plot area guides the user in the investigation; the direct manipulation control enables the user to make informed decisions and to dynamically produce a topic summary. A number of sliders are available by default in the Informedia-II digests, including query engine relevance score, broadcast date, and story duration. These sliders can be used to color-code and size-code attributes in the document plot.

The utility of topic digests increases when topics are viewed in conjunction with other aspects of the video library data, such as geographic references. The topic digest could be used to narrow down a set of documents, with a map digest then showing an interesting visualization of video library coverage.

For more details, please refer to [Christel2000].

2.1.6 User's interaction history

In order to develop *adaptive* video information collages that provide more focused information of relevance to the user given his or her current context, we started to create a user history component to the Informedia Digital Video Library System (DVLS). The component is able to

create user session and record all queries from all sessions. The component is also able to restore a particular query from the history. The current functionality enables a user to prioritize queries using features other than the time features according to the history.

2.1.7 Panoramic image for video summarization – the static representation

We developed a technique to create on-demand panoramic image over user-selected video segment. The technique selects an anchor point, which acts as a reference image, and aligns images in the selected video segment to construct the panoramic image. The construction process considers two topologies (stitching methods): painting action when panning the camera and springing action when zooming the camera. After stitching all images together, the technique performs a global alignment to counter all misalignment common happened close to the edges of images.

The on-demand panoramic image creation is essential to summarization that removes redundant information in images in a video segment. We performed initial evaluation on some less “bumpy” sequences of video images. The results were satisfactory to users.

2.1.8 Using statistical machine translation to distill a given document into a query

We developed a new probabilistic approach to information retrieval based upon the ideas and methods of statistical machine translation. The central ingredient in this approach is a statistical model of how a user might distill or “translate” a given document into a query. To assess the relevance of a document to a user’s query, we estimate the probability that the query would have been generated as a translation of the document, and factor in the user’s general preferences in the form of a prior distribution over documents. We also developed a simple, well motivated model of the document-to-query translation process, and an algorithm for learning the parameters of this model in an unsupervised manner from a collection of documents. As we have shown, one can view this approach as a generalization and justification of the “language modeling” strategy recently proposed by Ponte and Croft. In a series of experiments on TREC data, a simple translation-based retrieval system performs well in comparison to conventional retrieval techniques.

For more details, please refer to [Berger1999] and [Berger1999a].

2.2 Information Analysis and Semantic Summarization

2.2.1 Named faces

We combined a variety of approaches – face recognition in images, OCR over the text on screen, and Web spiders – to associate people’s names with corresponding face images in video. The technique detects and recognizes superimposed text in the video. It then verifies or repairs the text with a dynamic programming match to a large list of automatically generated names found in news stories. Faces found in the video where superimposed names were recognized are

tracked, extracted, and associated with the superimposed text. The system gathers news stories and images from the Web to improve its performance.

For more details, please refer to [Houghton1999].

2.2.2 Automatic title generation using statistical machine-translation approach

Our prototype automatic title generation system inspired by statistical machine-translation approaches treats the document title like a translation of the document. Titles can be generated without extracting words from the document. A large corpus of documents with human-assigned titles is required for training title “translation” models. Our approach shows a higher precision and recall than another approach based on Bayesian probability estimates.

Extractive summarization is the most common approach to generate titles or short summaries of text data. The interesting phrases are usually determined through a variant of a TFIDF (Term Frequency by Inverse Document Frequency) word score for each document sentence. Highly interesting phrases are included in the headline summary. Our approach is non-extractive; a summary does not have to consist of phrase snippets taken from the document. Instead of a statistical approach to summarization using naïve Bayesian estimates, we use an Estimation/Maximization algorithm (EM).

In our EM approach, the system first estimates an alignment of document words to title words (i.e. which document word would likely translate into which title word) and then determines how well (i.e. with what probability) this alignment produces titles for the training documents. In successive iterations the EM algorithm improves by maximizing the probabilities of good title words through better alignment.

To evaluate our title generation approach, we trained a word-pair model $P(dw|tw)$ for 3 iterations using the approach outlined above on a corpus of 40000 transcripts of broadcast-news stories with human-assigned titles and also built a standard trigram language model, as $P(\text{title})$, from just the titles in the corpus.

Using a held-out test set of 100 news stories, we selected the top 50 title words from each document that maximized $\sum_{dw \in \text{doc}} P(dw|tw)$ where dw denotes a document word and tw likewise denotes a title word. Recall and precision were computed as the percentage of words in the original (manual) reference title compared to the automatically generated list of the top 50 candidate title words.

Average precision (40%) and recall (31.5%) for the EM at 3 iterations compares favorably with the naïve Bayesian approach (precision 28%, recall 20.5%).

To create a linearized “English-like” title, a lattice was formed consisting of a regular set of 6 columns, each column being a copy of the top 50 list of title word candidates with corresponding probabilities. The lattice-rescorer is run with this lattice and the trigram language model for titles as input. The output of the lattice rescorer is taken as the finished title, subject to a procedure to eliminate repeated words therefrom.

For more details, please refer to [Kennedy2000].

2.2.3 Automatic topic generation

Short topics of one to seven words are automatically assigned to each of the stories in the Informedia-II library, where the stories average just less than two minutes in playback time. A K-nearest-neighbor strategy is used to assign up to five topics to each story, making use of a training phase and a classification phase. The training phase uses a commercial broadcast news CD-ROM for a given year as input, which is indexed by a TFIDF scheme using the SMART search engine. The classification phase uses the SMART search engine to index each story and find ten stories in the training stories that have the minimum distances computed by the cosine similarity measure. Topic scores in the chosen ten training stories are summed up. Up to five topics above a minimum threshold are then assigned to the video story.

For more details, please refer to [Christel2000] and [Hauptmann1999a].

2.2.4 Density biased sampling for summarization

We have developed a single-pass algorithm that does sampling using *reverse* bias; that is, it obtains fewer samples from dense areas and thus leaves more room for sparser areas. Uniform random sampling is frequently used in practice and also frequently criticized because it will miss small clusters. Many natural phenomena are known to follow Zipf's distribution and the inability of uniform sampling to find small clusters is of practical concern. Density Biased Sampling is proposed to probabilistically under-sample dense regions and over-sample light regions. A weighted sample is used to preserve the densities of the original data. Density biased sampling naturally includes uniform sampling using only a single scan of the data. We empirically evaluated density biased sampling using synthetic data sets that exhibit varying cluster size distributions. Our proposed method scales linearly and out performs uniform samples when clustering realistic data sets.

For more details, please refer to [Plamer2000].

2.2.5 Fast creation of spatial associations between objects

We discovered a surprising law governing the spatial join selectivity across two sets of points. An example of such a spatial join is "find the libraries that are within 10 miles of schools". Our law dictates that the number of such qualifying pairs follows a power law, whose exponent we call "pair-count exponent" (PC). We showed that this law also holds for self-spatial-joins ("find schools within 5 miles of other schools") in addition to the general case that the two point-sets are distinct. Our law holds for many real datasets, including diverse environments (geographic datasets, feature vectors from biology data, galaxy data from astronomy).

In addition, we introduced the concept of the Box-Occupancy-Product-Sum (BOPS) plot, and we showed that it can compute the pair-count exponent in a timely manner, reducing the run time by orders of magnitude, from quadratic to linear. Due to the pair-count exponent and our analysis, we can achieve accurate selectivity estimates in constant time ($O(1)$) without the need for

sampling or other expensive operations. The relative error in selectivity is about 30% with our fast BOPS method, and even better (about 10%), if we use the slower, quadratic method.

For more details, please refer to [Faloutsos2000].

2.2.6 Automatic body segmentation from video

In attempting to analyze and automatically extract semantic content from video, we started with identifying human bodies, which may be presented in different orientations, colors (from clothing and lighting), sizes, and motions. Our approach is different from others like background subtraction with the assumption of still background comparing against moving body or predefined body parts with manual initialization. Our approach is to find pivot points, where a mass of pixel motions move around them.

To segment body, the approach first defines a model of a body pivot point with a Body Tree Joint, which is a cyclic graph to connect a set of possible triplets (three nodes in the Body Tree Joint). An energy term is then calculated depending on the geometrical relation, relative motion, and image information for all possible configurations of triplets. The goal is to find body joints, which have minimum energy terms. In order to reduce the complexity of calculating energy terms for a Body Tree Joint, we used a Clique Tree, which is a non-cyclic graph, to represent the Body Tree Joint. Clique Tree is a reduction mechanism without changing the calculation of the energy terms. We then applied dynamic programming on the Clique Tree to find the optimal configuration of joints for body.

2.2.7 Context Analysis: concept association

We have used a statistical co-occurrence technique to perform context analysis with concepts (term phrases) extracted from news transcripts. The co-occurrence technique has been previously applied in scientific, medical, and engineering domains and has shown semantic-resembled association comparable to a man-made thesaurus. The initial empirical observation of the result with 6000 video stories has shown that this co-occurrence technique can be applied to the general news domain. However, further testing and tuning are needed to reduce the “noisy” association.

2.3 Multimodal Query: Beyond Query/Browse by Text to Video Exploration

2.3.1 Multimodal queries

We have started to create a simple multimodal query interface by combining text and image media. In addition, we are working on a zoomable map with level of details ranging from city, country, to region.

2.3.2 Automatic image classification

We investigated the use of perceptual color to cluster video images. The initial result [Gong98] did not show a promising result relative to standard approaches. We are currently investigating other methods.

2.4 Information Extraction and Metadata Creation

2.4.1 Learning to recognize speech by examples

Our technique gathers large amounts of speech from open broadcast sources and combines it with automatically obtained text or closed captioning to identify suitable speech-training material. We use broadcast news data from TV sources to improve acoustic models by combining them with imperfect closed captions. Our initial efforts provided only limited success with small amounts of data.

The information gathered from accurately transcribed speech data serves as training data for the acoustic-model component of most high-accuracy speaker-independent speech-recognition systems. The error-ridden closed-captioned text aligns with the similarly error-ridden speech-recognition output. We assume matching segments of sufficient length are reliable transcriptions of the corresponding speech. We then use these segments as the training data for an improved speech recognizer. Our CCtrain acoustic model achieved improvements over our baseline system: an overall decrease in WER (word error ÷ total number of words in the test set) from 32.82% to 31.19%.

For more details, please refer to [Jang1999].

2.4.2 Geocoding news stories

For geocoding purposes, one additional source of information is used. Often, the location where a news story was filmed is not actually mentioned in the story. Instead, a line of text superimposed on the video may indicate the name of the correspondent, as well as the location from which he/she is reporting, such as "Ted Koppel, Jerusalem". In order to capture this potentially useful information video optical character recognition (VOCR) is used to scan frames within the video for possible text, and then to extract the text and include it in the meta data. As with the transcript, text derived via VOCR is synchronized with the video.

The address coverage used for geocoding in Informedia is a subset of ESRI's world gazetteer. This subset currently consists of all countries and administrative areas worldwide, as well as approximately 81,000 cities, towns and villages. Each record in the address coverage includes other information on a place. Columns used for geoprocessing in Informedia are the country name, the type of place, the administrative area and the continent. For all countries we have added geographic term expansions to account for the different ways in which a country might be mentioned in a news broadcasts. For example, the term "Germany" is expanded to also include "German" and "Germans".

Once words, or sequences of words (such as "South Carolina") are tagged as locations they are extracted from the closed captioned text and become candidates for address matching. For each candidate, an attempt is made to find at least one match in the address coverage. If no match is found, the candidate is discarded (in most instances this is due to a false positive during entity extraction; in a few cases it is due to a place not being in the address coverage, even though it exists). If one match is found we assume that it is the correct place. If more than one match is found the extracted place is ambiguous.

In order to resolve which of the places matching the name of an ambiguous location is the correct one, we look for clues in the transcript of the segment that might help us resolve the issue and assign scores to each place based on how many clues were found. First, we scan the transcript for any mention of the administrative region. For example, if there are two instances of Salem in the address coverage, one in Ohio and the other in Massachusetts, we scan the transcript for any mention of Ohio or Massachusetts. Each time one of the two states are mentioned, the corresponding Salem is given a point. If, after scanning the text one of the two places has more points than the other, then we assume that the one with more points is the correct location.

In order to determine the accuracy of the geoprocessing algorithm we randomly selected 200 CNN news segments representing about 5 hours of video from the Informedia library and geocoded them. None of the transcripts contained in the 200 segments were used during training of the entity extractor. A total of 357 places were mentioned in the 200 segments. Of these, the geocoding algorithm correctly identified and matched 269, or 75%. This is approximately on par with street address matching systems. Of the 88 locations (25%) that were incorrect the following error sources were identified:

- Place not in address coverage: 16 (18%)
- Disambiguation error (correctly identified as a place, but wrong coordinates): 30 (34%)
- Place misspelled in transcript: 4 (5%)
- False positive: 31 (35%)
- Missed: 15 (17%)

For more details, please refer to [Olligschlaeger1999].

2.4.3 Integration of continuous speech recognition and information retrieval for mutually optimal performance

Traditionally, indexing and searching of speech content in multimedia databases have been achieved through a combination of separately constructed speech recognition and information retrieval engines. Although each technology has a legacy of research, only recently have efforts been made to study the potential suboptimality of this strategy, and none of these efforts specifically addresses the presence of uncertainty in automatically generated transcriptions.

The research in integrating continuous speech recognition and information retrieval for mutually optimal performance develops a refinement of the most common information retrieval relevance formula, TFIDF, to incorporate uncertainty as a retrieval feature, along with a set of techniques

to acquire this uncertainty from multiple hypotheses produced by existing speech recognition data structures. In the process a greater amount of evidence is extracted than is available in the most likely transcription hypothesis, and overall retrieval precision and recall are improved.

The term weighting scheme known as the inverse document frequency is shown to be a special case of the mutual information between the document set and the term, the former requiring a Boolean characterization of term occurrence information and the latter permitting fractional probabilities. The relevance between a query and document from speech recognition is then modeled as a random variable arising from the statistical nature of the speech recognition system. The statistics of this model are then derived from the word lattices and the N-Best lists from the output of the recognizer.

In analyzing the word lattices, the path probabilities for each node are summed. The relative rankings of competing terms of these summed probabilities are shown to be indicative of the probability of term occurrence. A model of this relationship is used to predict term presence and term count, reducing the degradation in retrieval quality due to speech recognition by 24%. In a separate model, the Top-N distinct text-processed hypotheses from the word lattices are used to estimate the term probability and term count. This strategy reduces the degradation in retrieval quality due to speech recognition by 63%. Experiments were performed on a standardized test of broadcast news stories that had been transcribed manually and judged against a set of natural language queries.

For more details, please refer to [Siegler99].

2.4.4 Fast access to multimedia objects

The design of a new *metric index tree*, Slim-tree, improves the accessing performance of traditional M-trees. Such trees are vital for multimedia indexing in Informedia. The Slim-tree is a dynamic tree for organizing metric datasets in pages of fixed size. The Slim-tree uses the “fat-factor” which provides a simple way to quantify the degree of overlap between the nodes in a metric tree. It is well-known that the degree of overlap directly affects the query performance of index structures. There are many suggestions to reduce overlap in multidimensional index structures, but the Slim-tree is the first metric structure explicitly designed to reduce the degree of overlap.

Moreover, we have developed new algorithms for inserting objects and splitting nodes. The new insertion algorithm leads to a tree with high storage utilization and improved query performance, whereas the new split algorithm runs considerably faster than previous ones, generally without sacrificing search performance. Results obtained from experiments with real-world data sets show that the new algorithms of the Slim-tree consistently lead to performance improvements. After performing the Slim-down algorithm, we observed improvements up to a factor of 35% for range queries.

For more details, please refer to [Traina2000] and [Traina2000a].

2.4.5 Statistical models for text segmentation

A new statistical approach was introduced for automatically partitioning text into coherent segments. The approach is based on a technique that incrementally builds an exponential model to extract features that are correlated with the presence of boundaries in labeled training text. The models use two classes of features: *topicality* features that use adaptive language models in a novel way to detect broad changes of topic, and *cue-word* features that detect occurrences of specific words, which may be domain-specific, that tend to be used near segment boundaries. Assessment of our approach on quantitative and qualitative grounds demonstrates its effectiveness in two very different domains, *Wall Street Journal* news articles and television broadcast news story transcripts. Quantitative results on these domains are presented using a new probabilistically motivated error metric, which combines precision and recall in a natural and flexible way. This metric is used to make a quantitative assessment of the relative contributions of the different feature types, as well as a comparison with decision trees and previously proposed text segmentation algorithms.

The models were trained on two million words of CNN transcripts furnished with segment boundaries, and tested on one million words of previously unseen text. A 100 feature exponential model combining cue-word and topicality features had an error rate of 0.132 comparing to that of 0.346 from the TextTiling algorithm, evaluated with a window size of 498 words, equal to half of the average segment size.

For more details, please refer to [Beeferman1999].

2.4.6 Concept extraction in text

In addition to generating topics to each video story using external training data, we started to investigate the use of a simple automatic phrase formation technique to extract term phrases to represent concepts in a story. The phrase formation technique is based on an indexing technique in information retrieval using stop words and a stemming technique. We are currently testing and tuning the phrase formation technique to yield “cleaner” term phrases.

2.5 Interoperable Heterogeneous Distributed Libraries

2.5.1 Web-based Informedia

The Informedia Project is migrating to an XML/XSLT delivery infrastructure, where users via web browsers will be able to query the library and retrieve data in XML format corresponding to the metadata associated with video in the library. Depending on user locale and the rights associated with the video clips, the video itself may also be accessible. We are working with other Carnegie Mellon researchers to improve MPEG-1 streaming, and plan to take advantage of commercially available lower bandwidth video streaming solutions for a subset of the video library. The architecture of the Informedia library system is being re-engineered, migrating from a two-tier client-database server system to a multi-tier client (user interface)/presentation transformations (e.g., XSL to convert XML into HTML presentation for web-based client)/video

library logic (e.g., ideal use of shot images within slide shows and synchronized presentations containing text fragments)/database.

2.5.2 Metadata publishing and sharing

We have published and shared the Informedia video metadata with another DLI2 project: “Simplifying Interactive Layout and Video Editing and Reuse” (<http://www.cs.cmu.edu/~7Esilver/#About%20SILVER>) at Carnegie Mellon University. We have also shared the metadata with Gary Marchionini and his students at UNC-Chapel Hill who are working on the “Video Repository Framework” (<http://ils.unc.edu/idl/projects.html>).

2.6 External Testbeds, Corpora, and Usability Evaluations

2.6.1 Open video testbed establishment

Michael Christel presented the Informedia Project at the Video Retrieval Evaluation and Testbed Symposium at UNC-Chapel Hill on October 21, 1999. Other multimedia content-based indexers, digital video evaluators and potential digital video library patrons discussed the value of sharing descriptive data (metadata) associated with digital video repositories, as well as making those repositories and metadata available over the internet. After conducting a property rights survey of the Informedia library, we identified 25 hours of Informedia-processed video clips which we could immediately make available on the Internet without restrictions. This data was shared with Gary Marchionini and his students at UNC-Chapel Hill who are working on the “Video Repository Framework” (<http://ils.unc.edu/idl/projects.html>). The database schema was also further optimized for web-based delivery of low-bandwidth metadata (also referred to as "surrogates" elsewhere), and this schema was shared with Marchionini and his group. The goal will be the establishment of an Open Video Project where researchers and users alike can examine, evaluate and make use of the digital video and associated metadata.

For more details, please refer to http://ils.unc.edu/idl/details/Symposium_Overview.html.

2.6.2 Corporeal coverage extension

We have collected 250+ hours of “NewsHour with Jim Lehrer” from PBS since March 1999. This collection gives in-depth news analysis to complement the continuously growing Informedia corpus of 1000+ hours of CNN news video from “CNN World View” and “CNN World Today” since 1996.

2.6.3 Skim evaluation

Currently, we are conducting a user study on evaluating video skims created by a new technique, which uses a variation of TFIDF model. The evaluation is to compare the summarization performance of the new technique, judged by users, with that of the current automatic skim generation method and manually crafted video skim. We will use the user evaluation result to determine whether we should adopt the new technique to produce video skims as video summary for each video story.

3 Notable Outreach and Inclusion Activities

East-West Informedia

The Informedia project has established a cooperative research and technology transfer relationship with three Chinese research and educational institutions in order to further extend the system's use, increase its capabilities, and enrich its distributed corpus with multilingual Chinese video content. The four institutions involved are Carnegie Mellon University, The Chinese University of Hong Kong (PI's: Jerome Yen, Joseph Hui, Michael Lyu), the South China University of Technology in Guang Zhou (PI: Ling Zhang), and the Academia Sinica in Taiwan (PI: Der-Tsai. Lee). We have transferred an independently operating demonstration library with seed content to Hong Kong and established specific research relationships amongst individual researchers at the cooperating institutions to replicate and extend the underlying video analysis and extraction techniques in both Mandarin and Cantonese variants. Wactlar and Hauptmann visited each of the Chinese institutions to deliver a series of talks and establish a cross-institutional research plan.

ECHO

The Informedia project is the only American partner for the European Commission Information Societies Technology (IST) program sponsored by European Chronicles On-line (ECHO). The project is coordinated by Consiglio Nazionale delle Ricerche -- Istituto di Elaborazione della Informazione (CNR-IEI) and includes the national film archives of Italy, the Netherlands, France and Switzerland. Technical partners include Centre National de la Recherche Scientifique, University of Mannheim, and Universiteit Twente. Industrial partners include Tecmath GmbH & CoI, MediaSite Ltd., and Eurospider Information Technology.

The main objectives of the project are to develop a long term reusable software infrastructure to support digital film archives, to provide web-based access to collections of historical documentary films of great international value, and to increase the productivity and cost effectiveness of producing digital film archives. The project will develop and demonstrate an open architecture approach to distributed digital film archive services. The open architecture will support service extensibility and interoperability. The distinct features of the ECHO system will be: semi-automatic metadata extraction and acquisition from digital film information, non-English speech recognizers (Italian, French, Dutch, German) for the purpose of indexing, searching and retrieval, cross-language retrieval capabilities, intelligent access to digital films, automatic film summary creation, collection mechanisms, and privacy and billing mechanisms.

TIC

Another independently operating system has been installed in a classified environment as part of DARPA's Genoa Project. Under this project work is proceeding to implement a version of the system that will operate entirely in a WindowsNT environment.

Fall 1999 Multimedia Course (20-859). CMU Institute for E-Commerce, Masters of Science in E-Commerce Program.

20-859 Multimedia Course – Course Description:

Until recent years, most computing tasks dealt with numerical, text, and symbolic data, and Computer Science has emphasized these data types. Digital representations of audio, video, and images are now becoming quite common. With the advent of relatively cheap, large online storage capacities, network transmission speeds and advances in digital compression, comprehensive sources of multiple media (Text, Image, Video and Audio) can be easily stored and made available. Collecting and intelligently integrating these multiple media opens up opportunities for novel business applications. Consequently, an understanding of multimedia is essential for many e-commerce businesses.

This course teaches students to work with multiple media on computers. Students learn the issues involved to capture, process, compress, search, index, store, and retrieve various kinds of continuous media. Projects require work with audio, scanned images, digital video, and other media, all in digital form. Readings and lectures provide a conceptual and technical framework for multimedia work.

After completing this course, students will be able to:

- appreciate the role of multimedia in Ecommerce
- understand the concepts underlying multimedia creation, representation and transmission
- create media for the Web using various software tools to manipulate audio, images and video

Fall 1999 Topics in Information Retrieval Seminar Course (11-743), Language Technologies Institute Masters and Ph.D. program. Guest lecture on Digital Libraries, "Informedia Digital Video Library: Search and Summarization in the Video Medium."

11-743: Topics in Information Retrieval – Course Description:

Topics in Information Retrieval is a seminar that focuses on current research in Information Retrieval. The seminar will cover recent research on subjects such as retrieval models, text classification, information gathering, fact extraction, information visualization, summarization, text datamining, information filtering, collaborative filtering, question answering systems, and portable information systems. Other topics will be drawn from recent SIGIR, Digital Libraries, TREC, Machine Learning, and AAAI conferences.

One goal of the seminar is to identify topics for M.S. and Ph.D. students interested in doing research in Information Retrieval, so the syllabus is likely to be adjusted during the semester, based on the interests of participants. The seminar will be a mix of lectures by the instructors, lectures by guest lecturers, and presentations by participants. Significant classroom discussion will be encouraged.

4 Project Director's Narrative

The Informedia-I project built a solid foundation on indexing, searching, and viewing multimedia video documents. The progression of Informedia-II project which emphasizes adaptive summarization and visualization on video documents and libraries requires the effort to go beyond information analysis at the syntactic level – extracting and integrating patterns for retrieval and presentational purposes. To support the idea of creating adaptive and context-dependent video information collages, we have the following research and experimentation challenges:

- Adaptive and adjustable presentation of library content to meet user information need
- Context analysis on library content as well as user preferences and interaction history
- Multimodal query
- Content extraction from video documents

In order to help the user gain insight into a large set of retrieved video stories, we developed and tested different visualization methods for viewing a set of retrieved result as a whole as well as each individual story. In its first year, we developed various independent summarization and visualization methods – VIBE, timeline, map and topic digests – to examine and correlate the members of the search result set. We also converted several static or batch-mode video processing techniques like the video skim and filmstrip to real-time functionalities, which enable the user to control the level of abstraction when viewing video stories.

To support the integration of video information collages presented in different features, information analysis is required at the semantic level. The complexity of content and context analysis goes along with the levels of abstraction that we can conceptually recognize. A video document contains a set of concepts – basic semantic elements like people, objects, places and time, as well as abstracted semantic elements like topics, events, reasons, actions, effects, relationships, etc. In addition, occurrences, usages and relationships of those concepts in different documents create various levels of abstraction across time and space. We investigated various methods to extract concepts – human bodies, motions, faces, topics, etc. – in video documents. We also examined various association methods to correlate various concepts within the same and across different media.

5 Journal and Conference Proceeding Publications

[Beeferman1999] “Statistical Models for Text Segmentation,” Beeferman, D., Berger, A., and Lafferty, J. Machine Learning, Special Issue on Natural Language Learning, C. Cardie and R. Mooney, eds., Vol. 34, Nos. 1-3, pages 177-210, 1999.

- [Berger1999] "Information Retrieval as Statistical Translation," Berger, A. and Lafferty, J. 1999 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), pages 222-229, August 1999.
- [Berger1999a] "The Weaver System for Document Retrieval," Berger, A. and Lafferty, J. Proceedings of TREC-8, Gaithersburg, MD, 1999.
- [Christel1999] "Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library," Christel, M.G., Hauptmann, A.G., Warmack, A.S., and Crosby, S.A. Proceedings of the IEEE Advances in Digital Libraries Conference (Baltimore, MD), pages 98-104, May 1999.
- [Christel1999a] "Visual Digests for News Video Libraries," Christel, M.G. Proceedings of the ACM Multimedia '99 Conference (Orlando, FL), pages 303-311, November 1999.
- [Christel1999b] "Interactive Maps for a Digital Video Library," Christel, M.G. and Olligschlaeger, A.M. Proceedings of the IEEE International Conf. On Multimedia Computing and Systems, D. Martin, ed. (Florence, Italy), pages 381-387, June 1999.
- [Christel2000] "Accessing News Video Libraries via Topics," Christel, M.G., Begun, A.P., and Hauptmann, A.G. Technical Report, January 2000.
- [Faloutsos2000] "Spatial Join Selectivity Using Power Laws," Faloutsos, D., Seeger, B., Traina, A.J., and Traina Jr., Caetano. To be published in SIGMOD 2000.
- [Hauptmann1999] "Integrating and Using Large Databases of Text, Image, Video and Audio," Hauptmann, A.G. IEEE Intelligent Systems, Vol. 14, No. 5, pages 34-35, 1999.
- [Hauptmann1999a] "Topic Labeling of Multilingual Broadcast News in the Informedia Digital Video Library," Hauptmann, A.G., Lee, D., and Kennedy, P.E. ACM DL/SIGIR MIDAS Workshop, Berkeley, CS, June, 1999.
- [Houghton1999] "Named Faces: Putting Names to Faces," Houghton, R. IEEE Intelligent Systems, Vol. 14, No. 5, pages 45-50, September/October 1999.
- [Jang1999] "Learning to Recognize Speech by Watching Television," Jang, P.J. and Hauptmann, A.G. IEEE Intelligent Systems, Vol. 14, No. 5, pages 51-58, 1999.
- [Kennedy2000] "Automatic Title Generation using EM," Kennedy, P.E. and Hauptmann, A.G. Submitted for publication in ACM Digital Libraries '00, San Antonio, Texas, June 2000.
- [Olligschlaeger1999] "Multimodal Information Systems and GIS: The Informedia Digital Video Library," Olligschlaeger, A.M. and Hauptmann, A.G. 1999 User Conference, San Diego, CA, July 27-29, 1999.

- [Palmer2000] “Density Biased Sampling: An Improved Method for Data Mining and Clustering,” Palmer, C.R. and Faloutsos, C. To be published in SIGMOD 2000.
- [Siegler1999] “Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance,” Siegler, M.A. PhD’s thesis, Carnegie Mellon University, Electrical and Computer Engineering, December 1999.
- [Traina2000] “Slim-trees: High Performance Metric Trees Minimizing Overlap Between Nodes,” Traina Jr., C., Traina, A., Seeger, B., and Faloutsos, C. To be published in Proceedings of the International Conference on Extending Database Technology EDBT 2000, Konstanz, Germany, 27-31 March, 2000.
- [Traina2000a] “Distance Exponent: A New Concept for Selectivity Estimation in Metric Trees,” Traina Jr., C., Traina, A., and Faloutsos, C., In Proceedings of the IEEE 16th Intl. Conference on Data Engineering in San Diego, CA, February 29 – March 3, 2000.
- [Wactlar1999] “Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library,” Wactlar, H., Christel, M., Gong, Y. and Hauptmann, A. IEEE Computer, Vol. 32, No. 2, pages 66-73, 1999.
- [Wactlar1999a] “New Directions in Video Information Extraction and Summarization,” Wactlar, H. 10th DELOS Workshop, Santorini, Greece, June 24-25, 1999.
- [Wactlar2000] “Complementary Video and Audio Analysis for Broadcast News Archives,” Wactlar, H., Hauptmann, A., Christel, M., Houghton, R., and Olligslaeger, A. Communications of the ACM, Vol. 43, No. 2, pages 42-47, February 2000.
- [Wactlar2000a] “Informedia – Search and Summarization in the Video Medium,” Wactlar, H. Proceedings of Imagina 2000 Conference, Monaco, January 31 – February 2, 2000.

6 Presentations, Demonstrations, and Industry Visitors

- May 1999 IEEE Advances in Digital Libraries Conference 99, Baltimore, Maryland. Presentation and paper - “Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library”
- May 1999 Helen Meng, Chinese University of Hong Kong. Research discussion/demo.
- June 1999 East-West Digital Library Initiative project meeting and presentation at Academia Sinica in Taipei, Taiwan - “New Directions in Video Information Extraction and Summarization”
- June 1999 East-West Digital Library Initiative project meeting and presentation at South China University of Technology in GuangZhou, P.R. China - “New Directions in Video Information Extraction and Summarization”

- June 1999 East-West Digital Library Initiative project meeting and presentation at the Chinese University of Hong Kong - "New Directions in Video Information Extraction and Summarization"
- June 1999 Tenth DELOS Workshop on Audio-Visual Digital Libraries, Santorini, Greece. Presentation/paper - "New Directions in Video Information Extraction and Summarization"
- June 1999 IEEE Int'l Conference on Multimedia Computing and Systems, Florence, Italy. Chaired session and paper - "Interactive Maps for Digital Video Library"
- July 1999 Carnegie Mellon Research Institute staff meeting. Informedia presentation and demo.
- Sept 1999 Boeing Connestoga Project, developing technology roadways for information technologies to support decision systems. Invited talk.
- Sept 1999 InfoTEST Sector Board Meeting, Carnegie Mellon Research Institute, Pittsburgh, Pennsylvania. Invited talk - "An Overview of the Informedia Digital Video Library Project".
- Sept 1999 Monterrey Institute of Technology of Mexico, Digital Library Project. Research discussion/demo.
- Oct 1999 First International Workshop on Multimedia Intelligent Storage and Retrieval Management, Orlando, Florida. Invited talk - "Informedia Digital Video Library Accomplishments and Future Directions".
- Oct 1999 Presentation to MIT Lincoln Lab, "Automatic Title Generation for Documents".
- Oct 1999 Video Retrieval Evaluation and Testbed Symposium, University of North Carolina. Invited talk - "Reflections on the Informedia Digital Video Library Interface"
- Nov 1999 2nd Asian DL Conference, Taipei. Invited talk - "The Informedia System Architecture for Continuous Video Information Capture, Extraction and Cataloging".
- Nov 1999 ABC/ESPN. Research discussion/demo.
- Nov 1999 TREC-8 1999 Conference, Washington, DC.
- Nov 1999 Presentation at ASIS (American Society for Information Science) 1999 Conference. Session panel and presentation on "Information Retrieval from Speech"
- Nov 1999 Presentation to MITRE, "Automatic Title Generation for Documents".
- Dec 1999 Intel China Research Center. Research discussion/demo.

- Jan 2000 Imagina 2000, Monaco Film and Broadcast Festival, Monte Carlo, Monaco. Invited talk - "Informedia – Search and Summarization in the Video Medium"
- Feb 2000 School of Computing, National University of Singapore. Research discussion/demo.
- Feb 2000 NSF ECHO (European Chronicles) Project Kickoff Meeting, Pisa, Italy.

7 Statement

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).

Principal Investigator Signature: _____