# Complementary Video and Audio Analysis for Broadcast News Archives

Howard D. Wactlar, Alexander G. Hauptmann, Michael G. Christel, Ricky A. Houghton and Andreas M. Olligschlaeger

## Abstract

The Informedia Digital Video Library system extracts information from digitized video sources and allows full content search and retrieval over all extracted data. This extracted 'metadata' enables users to rapidly find interesting news stories and to quickly identify whether a retrieved TV news story is indeed relevant to their query. This article highlights two unique features: *named faces* and *location analysis*. *Named faces* automatically associate a name with a face, while *location analysis* allows the user to visually follow the action in the news story on a map and also allows queries for news stories by graphically selecting a region on the map.

## 1 The Informedia Digital Video Library Project

The Informedia Digital Video Library project [1], initiated in 1994, uniquely utilizes integrated speech, image and natural language understanding to process broadcast video. The project's goal is to allow search and retrieval in the video medium, similar to what is available today for text only. To enable this access to video, fast, high-accuracy automatic transcriptions of broadcast news stories are generated through Carnegie Mellon's Sphinx speech recognition system and closed captions are incorporated where available. Image processing determines scene boundaries, recognizes faces and allows for image similarity comparisons. Text visible on the screen is recognized through video OCR and can be searched. Everything is indexed into a searchable digital video library [2], where users can ask queries and retrieve relevant news stories as results. The *News-on-Demand* collection in the Informedia Digital Library serves as a testbed for automatic library creation techniques of continuously captured television and radio news content from multiple countries in a variety of languages. As of October 1999, the Informedia project had about

1.5 terabytes of news video indexed and accessible online, with over 1600 news broadcasts containing about 40,000 news stories dating back to 1996.

The Informedia system allows information retrieval in both spoken language and video or image domains. Queries for relevant news stories may be made with words, images or maps. Faces are detected in the video and can be searched. Information summaries can be displayed at varying detail, both visually and textually. Text summaries are displayed for each news story through topics and titles. Visual summaries are given through thumbnail images, filmstrips and dynamic video skims. Every location referenced in the news stories is labeled for geographic display on a map and the corresponding news item can be retrieved through a map area selection. The system also provides for extraction and reuse of video documents encoded in MPEG-1 format for web-based access and presentation.

A multi-lingual component, currently implemented for Spanish and Serb/Croatian corpora, translates English language queries for text search into the target language. English language topics are also assigned to news stories. A user can add spoken or typed annotations to any news story, which are immediately searchable. News clips can be cut and pasted into HTML or PowerPoint presentations.

### 1.1 Speech Recognition

Speech recognition in the Informedia Digital Video Library is done in two different passes. To get an initial transcript into the library as quickly as possible, a 20,000-word vocabulary version of the Sphinx-II recognizer is applied [5]. In a second pass, we use the slower, but more accurate Sphinx-III speech recognition system. In the 1997 DARPA broadcast news evaluations, the CMU Sphinx-III system achieved overall word error rate of 24% when multiple passes are applied [5]. However, this result was obtained at processing speeds of several hundred times real time. To make the speech recognition reasonably fast, we restrict the beam of the recognizer, resulting in a word error rate of about 34% on the broadcast news evaluation data. To obtain better performance, we use

additional training data extracted from closed-captioned news transcripts for which the speech recognizer has a high confidence that the data is correct [7]. In addition, we also build a new language model every day, which interpolates a standard broadcast news transcription corpus [5] with current online web news reports and actual transcripts available on the CNN website (cnn.com). With help from both the improved acoustic models and by using the daily language model, we obtain a word error below 20%. Since we can parallelize the processing of a news story, the actual transcript data appears in our library within 2.5 hours from the broadcast time.

## 2 Information Extraction and Use

### 2.1 Named Faces

Within the Informedia system, there is also a module for automatically associating names with faces occurring in news video. Face detection and tracking are used to find faces, optical character recognition of text overlaid on the video is used to label faces, and face similarity matching is used for the retrieval. Dynamic programming techniques are used together with the output of a named-entity tagger [10] run over the speech-recognized (or closed-captioned) transcript to improve the detection of names in video captions. Together these pieces create a database of named faces that is used for labeling faces found in video or for retrieving faces of people via a text query. The named face processing proceeds in two parts: **mining** of names and faces from news video to create a database and **retrieval** of names or faces from the database. The system supports querying across modes: a face image can be used as a image search key to find video where the name of the face is mentioned, and a name can be used as a text search key to retrieve corresponding faces.

During the **mining** of faces, the system identifies the location of faces in video frames using [9]. We have also successfully used a commercial face recognizer [3] for this task. To obtain corresponding names for the found faces, we recognize the text that is overlaid on the screen [4], sometimes referred to as Chyron text, which frequently includes names and job titles. It is necessary to utilize Chyron data in addition to audio transcripts, since only about 50% of the words in the

Chyron text are also spoken in the audio track. Since the Chyron text only appears in the video as part of the image and not part of the audio stream, extracting this data requires optical character recognition procedures. In contrast to closed caption (CC) data, which is encoded into the video stream as a non-visible channel and can be easily converted to ASCII with an off-the-shelf CC decoder, overlaid Chyron text must be extracted from the image. A technique called Video OCR [4], video optical character recognition, is being used to extract this information. The Informedia video optical character recognition system identifies and recognizes captioned text that appears on the video. For efficiency reasons, a rough and quick text region detection algorithm searches for horizontal rectangular structure of clustered sharp edges using horizontal differential filtering techniques. Potential text regions are then sequentially filtered across all detected frames, effectively increasing the resolution of the each caption. Filtering is performed to reduce background noise. The potential text region is then extracted as a tiff image and submitted to a commercial optical character recognition package for the final stage of recognizing the text.

Names are extracted from the speech transcript using named entity extraction algorithms described in [10]. While this extraction yields names, locations, organizations and percentage amounts and dates, we are currently only using extracted names and locations within the system. Names can also be harvested from the online www pages of CNN and other news agencies, as described in [11].

Given that we have likely names extracted from web news stories and named-entity extraction of the speech transcripts, we now use this information to select and correct the video OCR results of the Chyron text for a named faces database.

With a word error rate of 65% in the video OCR, the probability of correctly recognizing a first-name/last-name pair drops to about 12%. To improve this error rate, we use dynamic programming (also known as dynamic time warping) [8] with a set of possible names to improve the video OCR output. It is very useful for computing similarity or a “fuzzy” match of two strings. After the lists of names have been generated, Each line of video OCR output is then compared against each name in

the name lists generated from the named entity extraction and the census data. A character by character dynamic programming match checks if a name from the list provides a sufficiently good match to the video OCR line.

For two CNN news shows, 177 lines of video OCR results were derived from video OCR. Of these 32 represented the name of the person on the screen. The word error rate for these 32 names was computed to be about 65%. After the dynamic programming match repair (see Table 1), the video OCR output word error rate dropped to about 37%. Further improvements to this process are under investigation.

| *VOCR Output* | *Repaired VOCR Output with DP match* |
|---|---|
| A Time Warnef Compan~ | |
| Jole Chen | JOIE CHEN |
| Behind Closed Door~ | |
| Paul McNulty Counsel to Judiciary Cmte | PAUL MCNULTY, COUNSEL TO JUDICIARY CMTE. |
| Rep Ste phep Bu~r H Indiana | REP. STEPHEN BUYER (R), INDIANA |
| FT Hyde and Seek | |
| Rep Tom DeL~y R Majority Wh_ip | REP. TOM DELAY (R), MAJORITY WHIP |
| Joe Lockhar W~H Deputy Press Secretary | JOE LOCKHART, W.H. DEPUTY PRESS SECRETARY |
| Roqcr Clinton President Clinton S Brothei | |
| Rep Este ban Tortes D California | REP. ESTEBAN TORRES (D), CALIFORNIA |
| Candy Cro vl ~ flI fj~FF~~ V | |
| Clintons Shadow | |
| Sen_ Paul We us tone D Minnesota | SEN. PAUL WELLSTONE (D), MINNESOTA |
| Prof_ Ste pheP Sa~Itzburq George W~5 hingt6n Univ Law Schoo | PROF. STEPHEN SALTZBURG, GEORGE WASHINGTON UNIVERSITY LAW SCHOOL |
| Melanie Kibler Congressional Statte | MELANIE KIBLER, CONGRESSIONAL STAFFER |

*Table 1 Sample VOCR Output and results after repairing with DynamicProgramming and list of named entities.*

There are two modes for **retrieval** of named face information (see Figure 1 below), querying for the name associated with a face and finding a face corresponding to a name. If the name of a new face is desired, a commercial face recognizer [3] is used to match the novel image to each image in the database. An N-best list of names with a score representing the quality of the match to that name is returned. A more accurate approach extends this model to consider the temporal nature of video. Instead of submitting one face image from a video scene, multiple face images from the desired scene are submitted making the recognition more reliable. For example if a turned face happens to

match (incorrectly) with a random person in the database, the match with the same incorrect person is unlikely to score well when the face is oriented in a different pose. The process of submitting multiple images of the same face thus reduces spurious false alarms. The second retrieval mode is to return the face of a given name. This merely amounts to a text search in the named-face database, retrieving the image associated with a particular name.

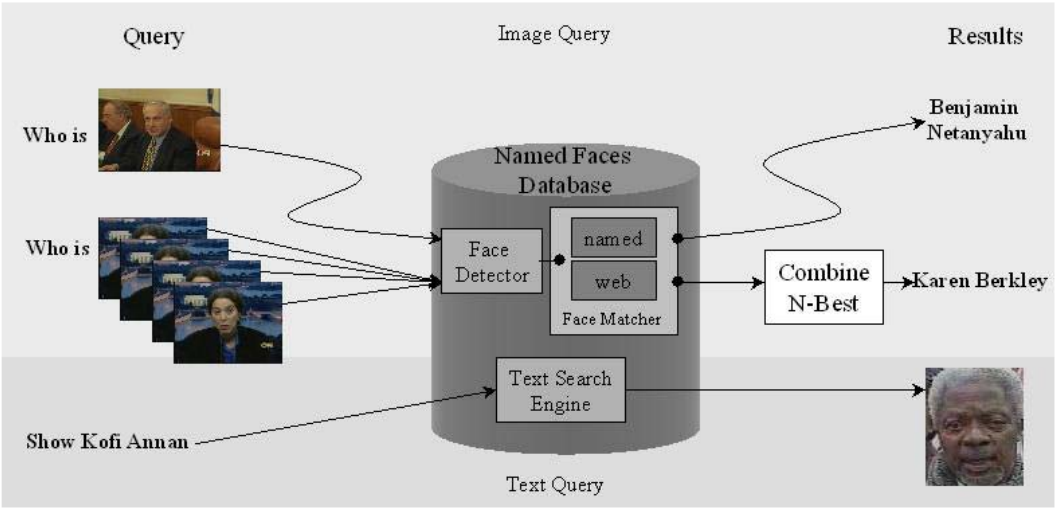


*Figure 1. Using Named Faces to identify an unknown face and to find a face for a name.*

## 2.2 Locations

In addition to the naming of faces, the named-entity extraction process is also used to provide location data derived from the speech transcripts for geocoding and map displays and searches.

### 2.2.1 Location Identification

Using a named entity tagger implemented as described in Schwartz et al. [10], we also extract possible location phrases from the audio transcript. These possible locations are then cross-referenced against a gazetteer of about 80,000 places and their locations [12], which include countries, cities, villages and states or provinces from all over the world. Currently excluded from this list are water areas, mountains and other non-political geographical data.

Since we will ignore all locations that cannot be found in the gazetteer with a latitude and longitude, the accuracy of the named entity extraction process per se is not as critical. What is more relevant is the question whether we have identified the correct coordinates for a location. If the location is not in the gazetteer, then we have no chance of providing the coordinates of this entity. If the words from the phrase describing the location were not in the speech recognition vocabulary, we again have no chance of providing the proper coordinates, even if we have properly identified the words as denoting a location. However, often locations are ambiguous in their coordinates. We then need to disambiguate different references to locations, e.g. "Washington" may refer to a number of cities in the United States, or it can refer to a state. A simple hierarchical disambiguation scheme has been implemented to distinguish among candidate coordinates: If a location is determined to be ambiguous, then we first check if other location references within the current news story disambiguate among the alternate locations. If the location is still ambiguous, then we see if other location references within a state or province favor a particular state. This way we can, for example, distinguish between different, initially ambiguous, references to Memphis in different states within the United States. If there is no disambiguating evidence at the state/province level, we check to see if the location can be distinguished on the basis of country reference elsewhere in the news story transcripts. The mention of France in conjunction with Paris would, for example, distinguish Paris, Texas from Paris, France, under the assumption that either Texas or France would be mentioned elsewhere in the news story. If the location is still ambiguous, we check for reference to locations within continents that might disambiguate the locations.

Having the coordinates identified unambiguously, allows us to use the location information in two ways:

1. Locations can be dynamically displayed on a map, allowing the user to "follow along" with the geographic focus of a news story.

2. We can use a query that graphically specifies a rectangle on the map to find all news stories that refer to any of the locations within this map area.

Figure 2 below shows an automatically generated map for a news detailing President Clinton's trip to Africa in March of 1998. In the news story, the transcript talks about the various stops in Ghana, Senegal, Uganda, Rwanda, Botswana and South Africa. This screen snapshot was taken as the narrator detailed the events at Clinton's next stop in Uganda. The system highlighted the country and any cities currently in focus, while showing less boldly the other countries and places


*Figure 2. A map automatically generated for a news story detailing President Clinton's trip to Africa. Uganda is currently mentioned in the audio. Selecting an area on the map allows a search for news stories that refer to the places in the selected area.*

mentioned in this news story. A user might then select an area on this map with the mouse, and initiate a spatial query to retrieve any news stories related to the specified locations.

Thus the location information allows both active maps, where the maps change according to the mention of the current location in the audio, and interactive maps, where the user makes a request by selecting an area on the map.

## 3 Conclusions

Extracting information from the audio track of news stories makes the data much more useful for the Informedia system. The Informedia Digital Video Library News-on-Demand system goes beyond the simple extraction of named entities from the audio stream and presents a demonstration that it is the ultimate use of this extracted information that matters, not the improved percentage point in extraction accuracy of one information extraction approach over another. We emphasize that it is the integration and presentation of the extracted metadata that makes the Informedia video library usable and useful. It is critical to have multiple different types of errorfully extracted information, to allow the user to quickly scan of the information presented in different forms, so that metadata redundancy can overcome the errors of any one type of metadata.

## 4 References


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

2. Hauptmann, A.G. and Witbrock, M.J., *Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval*, In Maybury, M. (ed.), "Intelligent Multimedia Information Retrieval", AAAI Press, 1997.

3. Visionics Corporation. FaceIt, Face Detector and Face Recognizer SDK http://www.faceit.com

4. Sato, T., Kanade, T., Hughes, E., and Smith, M. Video OCR for Digital News Archive. In Proc. Workshop on Content-Based Access of Image and Video Databases. (Los Alamitos, CA, Jan 1998), 52-60.

5. K. Seymore, S. Chen, M. Eskenazi, and R. Rosenfeld. “Language and Pronunciation Modeling in the CMU 1996 Hub 4 Evaluation,” *Proc. Spoken Language Systems Technology Workshop*. Morgan Kaufmann Publishers, 1997.

6. Primary Source Media, Broadcast News CDROM, Woodbridge, CT, 1995, 1996, 1997, 1998.

7. Jang, P.J. and Hauptmann, A.G., Learning to Recognize Speech by Watching Television, IEEE Intelligent Systems, Vol. 14, No. 5, 1999 pp. 51 - 58

8. Sakoe H. and Chiba. Dynamic programming algorithm optimization for spoken word recognition IEEE Trans on ASSP, 26(1): 43-49, 1978.

9. Rowley, H., Baluja, S., and Kanade, T. “Human face detection in visual scenes”, CMU, 1995. Technical Report CMU-CS-95-158.

10. Kubala, F., Schwartz, R., Stone, R., and Weischedel, R. (1998). Named entity extraction from speech, *Proceedings of DARPA Broadcast News Workshop*, Landsdowne, VA.

11. Houghton, R.A., Automatic Accumulation of a Scalable Named Face Database, IEEE Intelligent Systems, Vol. 14, No. 5, 1999 pp. 45 – 50.

12. Environmental Systems Research Institute, Inc., MapObjects and ArcView, http://www.esri.com