ENHANCED PERSPECTIVES FOR HISTORICAL AND CULTURAL DOCUMENTARIES USING INFORMEDIA TECHNOLOGIES

Howard D. Wactlar
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
+1 412 268 2571
wactlar@cs.cmu.edu

Ching-chih Chen
Graduate School of Library and Information Science
Simmons College
Boston, MA 02115 USA
+1 617 521 2804
chen@simmons.edu

ABSTRACT
Speech recognition, image processing, and language understanding technologies have successfully been applied to broadcast news corpora to automate the extraction of metadata and make use of it in building effective video news retrieval interfaces. This paper discusses how these technologies can be adapted to cultural documentaries as represented by the award-winning First Emperor of China videodisc and multimedia CD. Through automated means, efficient interfaces into documentary contents can be built dynamically based on user needs. Such interfaces enable the assemblage of large video documentary libraries from component videodisc, CD, and videotape projects, with alternate views into the material complementing the original sequences authored by the materials' producers.

Keywords
Digital video library, summarization, named-entity extraction, historical and cultural materials.

INTRODUCTION
Documentary producers take great effort to communicate a message with a sequence of still images and video of artifacts and locations, audio narration, expert commentary, and music. Likewise, producers of multimedia holdings on CD or videodisc take great care in assembling material to best communicate a desired set of messages. One such collection is the *First Emperor of China*, a multimedia videodisc and CD [1, 2]. Through automated processing, the Informedia Project at Carnegie Mellon University (CMU) can generate video surrogates, which represent the multimedia materials in an abbreviated manner [3, 6]. For example, a still image can be identified for every shot in the video, i.e., every contiguous set of video recorded from a single camera. Displaying these still images in time-ordered sequence as a storyboard conveys the visual flow of the documentary as first assembled by the documentary producer. In addition, though, these same Informedia technologies enable full content search and retrieval of the video materials, provide tighter alignment of narrative text with video imagery, and facilitate the extraction of additional metadata that can be used to build alternate views into the video library. Rather than be limited to the original producer's view, materials can be seen in different perspectives generated by the user. This paper discusses how one particular technique, named entity extraction, can be used to explore Chinese cultural documentaries.

Figure 1 shows an overview of video clips about the Great Wall, with thumbnails overlaid to represent the wall itself, stories from the First Emperor's time for the northern wall with his label "Qin Shi Huang Di", stories from 2002 about the real estate boom for new construction in the mountains near the Great Wall labeled "remarkable changes", and a displayed video of Nixon's 1972 visit there. This is analogous to the video digests views into news that were primarily focused on geography and time, appropriate dimensions for broadcast news [3]. For documentaries, experts' views are of importance, as well as cultural perspectives and the importance and relative timing of historical and political events.

Figure 1. Thumbnail overviews of Great Wall stories, displayed geographically and clustered by time.

NAMED ENTITY EXTRACTION
The ability to extract names of organizations, people, locations, dates and times (i.e. "named entities") is essential for correlating occurrences of important facts, events, and other metadata in a documentary video library, and is central to generating multiple views into the corpus. We

can automatically derive text metadata from speech recognition systems working on the audio narrative, and video OCR transcribing the overlay text appearing in the video image stream [6]. Our focus will be to extract named entities from such text, integrating across the aural and visual modalities to achieve better results. Current approaches have significant shortcomings. Most methods are either rule-based [5], or require significant amounts of manually labeled training data to achieve a reasonable level of performance [4]. The methods may identify a name, company, or location, but this is only a small part of the information that should be extracted; we would like to know further, for example, that a particular person is a politician and that a location is a vacation resort.

One of the significant challenges for named-entity extraction from video is that the output of the speech recognizer lacks case and punctuation, has numbers spelled out, and does not share many of the other common attributes of written text that can be useful for natural language processing. While rule-based systems suffer significant degradations in going from mixed case to this style of text, hidden Markov model (HMM) approaches have proven to be more robust, suffering only a degradation of 4% to 5% in F-measure [4]. We will implement a variation of the HMM approach in which the output distributions are exponential models that weight various features of the words (numbers, titles, etc.), affording a natural and powerful way of smoothing the distributions. The end result will be the automated generation of the following descriptors for video:

- Speakers (by folding in speaker recognition systems working from the audio to cluster speeches by the same person)
- People referenced
- Faces
- Organizations
- Places
- Dates
- Times
- Numbers
- Percentages
- Monetary references

ON-DEMAND INTERACTIVE VISUALIZATIONS

With the data from named entity extraction, numerous views into the documentary videos can be generated. A subset of video can be identified through a text query, or geographic, time-based, or topic query. This material could include a documentary about archeology, but the query focus is on farming methods. The display should reflect the user's focus, highlighting timelines, maps, personality profiles, genealogies, statistical charts, or other views that emphasize the dimensions of interest.

CONCLUSIONS

Informedia indexing and collage visualizations are intended to accurately extract the information content of such productions. These alternative Informedia representations that reveal temporal, spatial, person and institutional relationships, may enable the inference of trends, event relationships and even causality. These aids to helping the viewer interactively gain insight to complex events through a documentary's combined historical, geographic and sociological information may increase understanding and improve learning (yet to be evaluated in user studies and evaluations in an educational setting). On the other hand, these summaries may diminish or lose an embedded or evolving story line. The interactive summaries and visual indices while efficient in time and (screen) space, may not preserve the aesthetic values or pathos of the original production. These perspectives should not be considered replacements for the original in the Internet medium, but rather their complement.

The preliminary exploration from the use of *The First Emperor of China* videodisc and associated resource collection suggests the potential for future collaborative research. With its extensive database of metadata, videos in multilingual formats, and comprehensive descriptive annotations and reference linking, as well as geographical references, future research can concentrate more on task-oriented, user-focused approaches to information extraction, summarization and visualization.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under Cooperative Agreement No. IRI-9817496. More details about Informedia research can be found at http://www.informedia.cs.cmu.edu/. The content development activity of the Emperor collection is a part of NSF/IDLP Project, *Chinese Memory Net*, under Grant No. IIS-9905833.

REFERENCES

1. Chen, Ching-chih. *First Emperor of China*. Voyager CD-ROM, 1994, http://www.voyagerco.com/cdrom/.
2. Chen, Ching-chih. Different Cultures Meet: Lessons Learned in Global Digital Library Development, in *Proceedings of JCDL '01* (Roanoke, VA, June 2001), ACM Press, 90-93.
3. Christel, M.G. Visual Digests for News Video Libraries, in *Proceedings of ACM Multimedia '99* (Orlando, FL, November 1999), ACM Press, 303-311.
4. Makhoul, J., Kubala, F., Liu, D., Nguyen, L., Schwartz, R., and Srivastava, A. Speech and Language Technologies for Audio Indexing and Retrieval. *Proc. IEEE 88*, 8 (August 2000), 1338-1353.
5. Mani, I., House, D., Maybury, M. and Green, M. Towards Content-Based Browsing of Broadcast News Video, in *Intelligent Multimedia Information Retrieval*, M. Maybury, Ed. AAAI Press/MIT Press, Menlo Park, CA, 1997.
6. Wactlar, H., Christel, M., Gong, Y., and Hauptmann, A. Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library. *IEEE Computer 32*, 2 (February 1999), 66-73.