Proceedings of the 4th International Workshop on Multimedia Information Retrieval in conjunction with ACM Multimedia 2002, Juan-les-Pins, France, December 6, 2002.

# Enriching Perspectives in Exploring Cultural Heritage Documentaries Using Informedia Technologies

Tobun Dorbin Ng, Howard D. Wactlar
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
+1 412 268 4499
{tng, wactlar}@cs.cmu.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 multimedia technologies can be adapted to enrich perspectives in exploring cultural heritage documentaries. Through automated means, efficient interfaces in viewing and summarizing 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.

## Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *video*. H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, standards, user issues*. I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement – *feature representation*.

## General Terms

Management, Design, Human Factors.

## Keywords

Digital video library, summarization, dynamic skim, image feature extraction, face detection, cultural heritage documentaries.

## 1. INTRODUCTION

Cultural heritage materials in different countries have been preserved in documentary films since the beginning of the twentieth century. One such collection is a set of documentary films organized by the European Chronicles On-line (ECHO) project from four European countries – Italy, France, the Netherlands, and Switzerland [12]. Another collection is *The First Emperor of China,* a multimedia videodisc and CD [2, 3]. These collections are of extraordinary value since they have documented different aspects – social, cultural, political, economic, and historical – of life in various countries.

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, media producers of multimedia holdings on CD, DVD, or videodisc take great care in assembling material to best communicate a desired set of messages. In short, a documentary is an aesthetically edited and produced summary presenting in a linear fashion.

Through automated processing, the Informedia Project at Carnegie Mellon University (CMU) can generate video surrogates, which

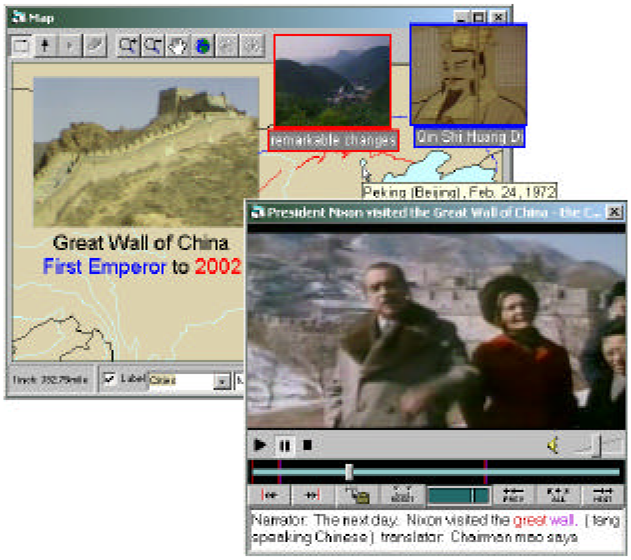
Map
remarkable changes
Qin Shi Huang Di
Peking (Beijing), Feb. 24, 1972
Great Wall of China
First Emperor to 2002
President Nixon visited the Great Wall of China - the C...
Label
Cities
Narrator: The next day, Nixon visited the great wall. [ tong speaking Chinese ] translator: Chairman mao says

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*Conference '00*, Month 1-2, 2000, City, State.

represent the multimedia materials in an abbreviated manner [4, 15]. 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 directed by the user.

This paper discusses how three techniques – dynamic skim, image feature extraction, and face detection – can be used to enrich perspectives in exploring cultural heritage 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 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 [4]. For documentaries, experts' views are of importance, as well as cultural perspectives and the importance and relative timing of historical and political events.

# 2. VIDEO CONTENT ANALYSIS

The Informedia Project has created a scalable infrastructure to continuously incorporate relevant multimedia content analysis techniques to extract various metadata from videos [15]. The metadata extraction process allows high-level abstraction techniques such as video skim to build upon previously generated low-level metadata such as scene breaks, speech recognized text, and audio signal-to-noise ratios. In addition to textual analysis and indexing, the same process extracts features from either whole images or detected objects inside them for image indexing and retrieval.

## 2.1 Dynamic Skim

A video skim is a temporal multimedia abstraction that incorporates both video and audio information from a longer source. A video skim is played rather than viewed statically, e.g., a ten-minute video might be condensed to a one-minute skim [6, 7]. 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 and in the contextual sequence closely adhering to the original editing perspective.

Instead of pre-computing video skims, we enable the dynamic generation of skims based on the availability of contextual information and the desirable condensed playing time [6]. Contextual information comes from users' query, which can be used to emphasize the audio and video surrounding match locations for assembling skims.

The skim is initialized to consist of sequences containing any of the given match locations. After merging closely occurred sequences, the remaining ones in the skim are then expanded: the sequence endpoint with the worst goodness rating is extended out to the next signal-to-noise ratio cutpoint, thus embedding that weak cutpoint into the skim. This process repeats until the target skim playing time is reached. By utilizing the goodness rating, a break between words will likely be sacrificed in favor of a break between full phrases or sentences as sequences are expanded, moving the granularity of the skim components from words to larger syntactic units, albeit through only an analysis of signal-to-noise ratio. With cutpoints occurring on average every 1.5 seconds, the produced skim is typically very close to the desired playing time. In post-processing all sequences less than a minimum length are removed.

During skim playback, users can 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.

## 2.2 Image Feature Extraction

It is a subjective decision on defining similar images because different people interpret similarity differently and even the same person may make different similarity interpretation in various situations. The challenge on searching for similar images is to objectively choose image features to represent them. The possible image features range from low-level color and texture to mid-level shape and segmented blobs to high-level recognizable object and semantics. Once images are converted to their corresponding feature vectors, we can achieve image retrieval by performing similarity search on such feature vectors.

A feature extraction technique can be built upon one or a combination of such image features. The following discusses three image feature extraction techniques utilizing color and texture information. In general, the format of an image feature vector is an array of floating point values. The dimension of feature vectors depends on the corresponding feature extraction technique. According to our own and others' research, there is no single dominant feature extraction technique to cover all kinds of images. For example, color feature is sufficient to support similarity search for some images while texture only feature or the combination of color and texture is sufficient for some others. Nonetheless, the combined use of individual similarity searches based on different image features may yield more relevant results.

### *2.2.1 Munsell Color Histogram*

The essence of this color feature extraction technique is to use Munsell Color Space for generating color histogram to represent an image [10]. The Munsell Color Space uses hue, value, and chroma to represent color [11]. Instead of computing a color histogram for an entire image, our empirical experiment has shown that combining computed color histograms from pre-defined regions to represent an image is more fruitful when performing image retrieval. We first divide an image into 9 (3-by-3) equally sized regions. Then, we use the Munsell Color Space to calculate a 16-bin color histogram for each region. We append all histograms together to form a 144-dimensional vector to represent an image.

### *2.2.2 Texture Histograms*

The second feature extraction technique relies on texture information in images. Similar to the first color feature extraction technique, we first divide an image into 9 (3-by-3) equally sized regions. For each region, we then generate six 16-bin texture histograms. We combine all these histograms to make an 864-dimensional vector to represent an image [1].

### *2.2.3 Cuebik: Color Coding*

The color coding extraction technique, cuebik, is based on one of the behaviors of the IBM's Query By Image Content (QBIC) image search engine [8, 9]. A palette of 255 colors is chosen for a database by marking the strongest colors found in a large sample of images. An image is divided into 256 (16-by-16) equally sized regions. Each region is mapped and coded to one of the palette colors. Consequently, a 256-dimensional feature vector represents the image. In addition to providing the capability of full-size image search, cuebik enables partial image search, which allows a user to choose a set of regions and finds all images that have the same color codes in the selected areas.

## 2.3 Face Detection

The primary challenge in face detection is the amount of variation in visual appearance. The appearance of the face depends on its pose; that is, its position, orientation, and rotation with respect to the camera. The size and resolution of a close-up view of a face will be different from a whole body view of a person from a distance. A side view of a human face will look much different than a frontal view. A tilted head to front, back, left, or right will look different from each other. Moreover, visual appearance depends on the surrounding environment. Light sources will vary in their intensity, color, and location with respect to the face. Nearby objects may cast shadows on the face or reflect additional light on the face. A face detector must accommodate all this variation and still distinguish the face from any other pattern that may occur in the visual world.

To cope with all this variation, Schneiderman used a two-part strategy for face detection [13]. The first part is to use a view-based approach with multiple detectors to deal with variation in pose. Each of these detectors specializes to a specific face orientation. For example, one detector is specialized to the right profile views of faces and one is specialized to the frontal views. Applying these view-based detectors in parallel gives individual results, which are then combined. If there are multiple detections at the same or adjacent locations, the strongest detection will be chosen. The second is to use statistical modeling within each of these detectors to account for the remaining variation. Each of these detectors shares the same underlying statistical form. They differ only in that their models use statistical gathered from different sets of images.

Detected faces are indicated by bounding boxes in their corresponding images. We use Eigenfaces, which are the eigenvectors of face images, as face feature vectors for their indexing and retrieval [14].

# 3. ON-DEMAND INTERACTIVE VISUALIZATIONS

Documentary producers summarize and present cultural heritage materials in story line fashion. The viewing of a documentary film in its entirety may provide the viewer instances of the cultural heritage and historical or thematic context and progression as authored by the documentary producers. The Informedia technologies provide complementary viewing methods for viewers to explore a large set of documentary videos.

## 3.1 Dynamic Linear Abstraction

When browsing the documentary video library, parallel to viewing an original video, users can choose to view the corresponding video abstraction with the desired playing time. The dynamic video skim captures the essential materials and preserves the original linear presentation context. When documentaries are organized and browsed by categories, context-specific dynamic video skims can be generated.

When searching for some specific cultural heritage materials, users can also choose to view dynamic skims of the retrieved results instead of full-length videos. The contextual information in the query will be used to generate context-specific dynamic abstraction with desired playing time.

## 3.2 Interactive Multimodal Query Interface

In addition to traditional text retrieval, Informedia provides multimodal query interface to accept images or faces as search input. Given the difficulties in selecting features to represent image and the differences between what deems as similar by human and computation, image similarity search is far from perfect. Nonetheless, through the iterative use of the Informedia multimodal query interface, users can interactively refine the query and collect relevant results [15]. For example, a user can submit an external image of "Great Wall" as an image query. A set of relevant video segments having similar images will be retrieved. By browsing through the result set, the user may refine the query in the same or different mode to continue the search task.

## 3.3 Interactive Video Collages

The searching mechanism may yield a large set of relevant video clips from the underlying documentary films library. Rather than browsing through the entire result set, users can utilize Informedia's interactive video collages to first obtain a high-level abstraction of all results presented in their choice of viewing perspective – timelines, maps, topics, statistical charts, or other dimensions of interest [5]. Interactive video collages have the summarization and visualization capability to derive context-specific presentations of text and images from multiple video sources, provide an interactive visualization for a set of video documents, and supply a navigation aid for further exploration. The dynamic creation of collages is based on user context, e.g., an originating query, coupled with user-directed viewing windows according to selected time periods or geographic locations as well as automatic processing to refine the candidate imagery and descriptive text. Users can then dynamically manipulate collages to view lower-level abstractions and reveal details.

## 4. CONCLUSIONS

Informedia indexing and collage visualizations are intended to accurately extract the information content of video productions. These alternative Informedia visualizations that reveal temporal, spatial, person, and visual object 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 cultural heritage documentary films, 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.

## 5. 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.

## 6. REFERENCES

[1] Belongie, S., Carson, C., Greenspan, H., and Malik, J. Color- and Texture-Based Image Segmentation Using the Expectation-Maximization Algorithm and Its Application to Content-Based Image Retrieval, in *Proceedings of ICCV '98* (Mumbai, India, 1998), 675-682.

[2] Chen, C. *First Emperor of China*. Voyager CD-ROM, 1994, http://www.voyagerco.com/cdrom/.

[3] Chen, C. Different Cultures Meet: Lessons Learned in Global Digital Library Development, in *Proceedings of JCDL '01* (Roanoke, VA, June 2001), ACM Press, 90-93.

[4] Christel, M.G. Visual Digests for News Video Libraries, in *Proceedings of ACM Multimedia '99* (Orlando, FL, November 1999), ACM Press, 303-311.

[5] Christel, M.G., Hauptmann, A.G., Wactlar, H.D., and Ng, T.D. Collages as Dynamic Summaries for News Video, in *Proceedings of ACM Multimedia '02* (Juan-les-Pins, France, December 2002).

[6] Christel, M.G., Hauptmann, A.G., Warmack, A.S., and Crosby, S.A. Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library, in *Proceedings of IEEE Advances in Digital Libraries Conference '99* (Baltimore, MD, May 1999), 98-104.

[7] Christel, M.G., Smith, M., Taylor, C.R., and Winkler, D. Evolving Video Skims into Useful Multimedia Abstractions, in *Proceedings of CHI '98* (Los Angeles, CA, April 1998), 171-178.

[8] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., and Yanker, P. Query by Image and Video Content: The QBIC System. *IEEE Computer* 28, 9 (September 1995), 23-31.

[9] Hauptmann, A.G. and Papernick, N.D. Video-cuebik: Adapting Image Search to Video Shots, in *Proceedings of JCDL '02* (Portland, OR, July 2002), 156-157.

[10] Miyahara, M., and Yoshida, Y. Mathematical Transform of (r, g, b) Color Data to Munsell (h, v, c) Color Data. *Visual Communication and Image Processing*, 1001 (1988), SPIE, 650-657.

[11] Munsell, A.H. *A Color Notation* (12th Edition). Munsell Color Company, Baltimore, MD, 1971.

[12] Savino, P. and Thanos, C. ECHO – European CHronicles On-line. *Cultivate Interactive* 1, 3 (July 2000), http://www.cultivate-int.org/issue1/echo/.

[13] Schneiderman, H. and Kanade, T. A Statistical Method for 3D Object Detection Applied to Face and Cars, in *Proceedings of IEEE CVPR* (Hilton Head SC, June 2000), 746-751.

[14] Turk, M. and Pentland, A. Face Recognition Using Eigenfaces, in *Proceedings of IEEE CVPR* (Maui HI, June 1991), 586-591.

[15] 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.