Personalizing Video Recorders using Multimedia Processing and Integration

Nevenka Dimitrova, Radu Jasinschi, Lalitha Agnihotri, John Zimmerman, Thomas McGee, Dongge Li
Philips Research
345 Scarborough Rd
Briarcliff Manor NY 10598
1.914.945.6000
{nevenka.dimitrova, radu.jasinschi, lalitha.agnihotri, johnzimmerman, thomas.mcgee, dongge.li} @ philips.com

ABSTRACT

Current personal Video recorders make it very easy for consumers to record whole TV programs. Our research however, focuses on personalizing TV at a sub-program level. We use a traditional Content-Baaed Information Retrieval system architecture consisting of archiving and retrieval modules. The archiving module employs a three-layered, multimodal integration framework to segment, analyze, characterize, and classify segments. The retrieval module relies on users' personal preferences to deliver both full programs and video segments of interest. We tested retrieval concepts with real users and discovered that they see more value in segmenting non-narrative programs (e.g. news) than narrative programs (e.g. movies). We benchmarked individual algorithms and segment classification for celebrity and financial segments as instances of non-narrative content. For celebrity segments we obtained a total precision of 94.1% and recall of 85.7%, and for financial segments a total precision of 8 1.1% and a recall of 86.9%.

Categories and Subject Descriptors

H.5.1 [Multimedia Information Systems]: Video; 1.2.10 [Vision and Scene Understanding]: Video analysis;

General Terms

Algorithms, Design, Experimentation, Human Factors.

Keywords

Content-based video retrieval, multimodal integration, Bayesian Networks.

1. INTRODUCTION

For many years there has been a television of the future vision where users can "watch what they want, when they want." Hard disk recorders deliver this vision by allowing users to time-shift. These devices make it easy for users to record whole programs, but offer no tools for users to access content within a program.

MM'01, Sept. 30-Oct. 5.2001, Ottawa, Canada.

We designed Video Scout with this problem in mind. Users make content request in their user profiles and then Scout begins recording TV programs. In addition, Scout actually watches the TV programs it records and personalizes program segments. Scout analyses the visual, audio, and transcript data in order to segment and index the programs. When viewing full programs, users see a high-level overview as well as topic specific starting points. For example: users can quickly find and playback Dolly Patton's musical performance within an episode of *Late Night with David Letterman.* In addition, users can access video segments organized by topic. For example: users can quickly find all of the segments on Philips Electronics that Scout has recorded from various financial news programs.

The following are complementary approaches to multimodal processing of visual, audio, and transcript information. Rui et al. discuss an approach that uses low-level audio features to detect excited speech and hits in a baseball game. They employ a probabilistic framework for automatic "highlight" extraction [1]. Syeda-Mahmood et al. present event detection in multimedia presentations from teaching and training videos [2]. Another approach uses semantic concepts in videos interact and appear in context was proposed by Naphade [3].

Section 2 describes the system architecture. We present program analyses based on categorization and the extraction of meaningful segments from which the system automatically derives high-level representations in section 3. Section 4 describes retrieval module where users can explore and directly access segments of interest. Section 5 presents results from multimedia integration. And section 6 offers conclusions and future directions for this research.

2. SYSTEM OVERVIEW

Video Scout consists of an archiving module and a retrieval module (see figure 1). Scout uses encoded video, electronic program guide (EPG) metadata, and the user profile in order to index high-level information. This process involves the following:

1. Archiving Module
   - Pre-selects programs to record based on matches between the EPG metadata and the user profile.
   - **Unimodal Analysis Engine:** Analyzes individual visual, audio, and transcript streams.

. Multimodal **Bayesian** Engine: Integrates visual, audio, and transcript information to classify segments.

2. Retrieval Module
   - **Personalizer:** Matches requests in user profile with indexed content.
   - User interface: Allows users to access video segments organized by TV program or segment topic.
3. User profile: Collects data from the user interface and passes it to the archiving module.

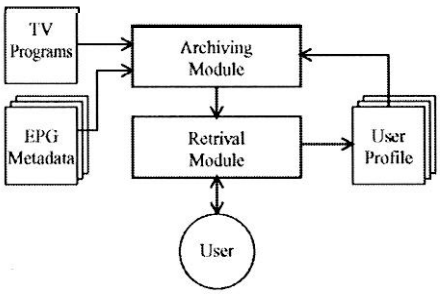

Figure 1. System Diagram

2.1 Archiving Module

An overview of the archiving module is given in Figure 2. We should note that the archiving can be performed during program recording at the set-top box end or at a service provider end. In the second case, content descriptions can be encoded in XML in a proprietary format or using standardized description (e.g. MPEG-7) for later use in retrieval. As shown in Figure 2., the program for analysis is selected based on EPG data and the user interest. Selected fields of the EPG data are matched to a user profile. When a match occurs, the relevant program information is written to a program text tile. The MPEG-2 Video related to the one of selected programs is then retrieved and decoded in software. During the decoding process the individual visual, audio and transcript data are separated out.

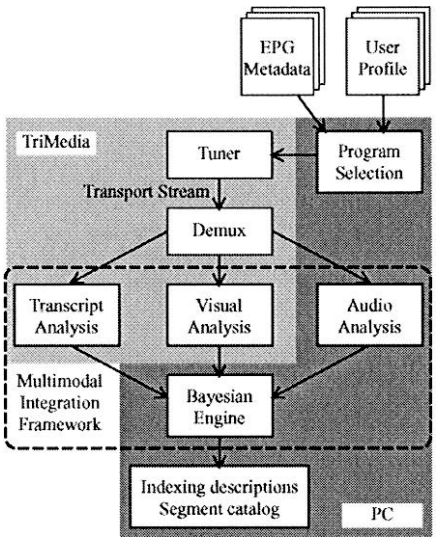

Figure 2. Block Diagram of the system architecture displaying the PC and TriMedia sides. Items within the dashed line comprise the Multimodal Integration Framework.

The Unimodal Analysis Engine (UAE) reads the separated the visual and audio streams from the MPEG-2 video. In addition. it creates the transcript stream by extracting the closed caption data. For visual stream, the UAE generates probabilities for cuts videotext, and faces. For the audio stream, the UAE segments the audio signal and generates probabilities for seven audio categories: silence, speech, music, noise, and speech with background music, speech and noise. For the transcript stream, the UAE looks for the double and triple arrows in text and generates probabilities indicating the category of a segment (e.g "economy", "movie"). The UAE produces a descriptive summary for each program. These probabilities and the summary are then written to text files. The Multimodal Bayesian Engine combines the input fmm the unimodal analysis engine and delivers high level inferences for the retrieval module.

2.2 Retrieval Module

An overview of the retrieval module is shown in Figure 3. The retrieval module contains the Personalizer performing the personalized segment selection and the User Interface. The Personalizer looks for matches between the indexed segments and the user profile. When a match is found, the segment is stored. The user interface allows users to customize their profiles. These profiles use a magnet metaphor, attracting the content users request. In addition, the user interface allows users to access and playback by whole programs and video segments organized by topic.

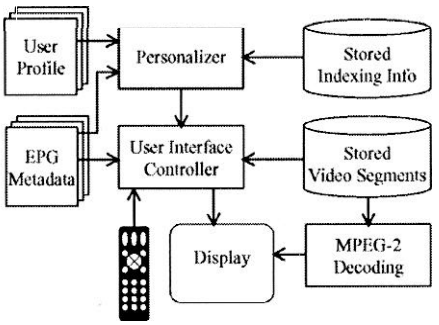

Figure 3. Overview of Video scout retrieval application.

2.3 User Profde

The user profile contains implicit and explicit user requests. Users can request both whole TV programs and topic-based video segments. Implicit requests are inferred from viewing histories. Profiles contain program titles, genres, actors, descriptions, etc. In addition, the profile contains specific topics users are interested in. For example: profiles might contain request for financial news about specific companies. In this case, Scout would collect individual news stories and not whole programs. Profiles function like database queries, performing two functions. First, the Analysis Engine uses the profile to determine which programs to record. Second, the Personalizer uses the profile to select specific segments from the total indexed, recorded content.

3. CONTENT PROCESSING

The Multimodal Integration Framework (MIF) is the most important archiving module. The MIF consists of the Unimodal Engine and the Multimodal Bayesian Engine from the Archiving Module.

3.1 Unimodal Analysis Engine

In the visual domain, the Unimodal Analysis Engine (UAE) searches for boundaries between sequential I-frames using the DCT information. It outputs the keyframes and a list with a probabilities and frame numbers. The keyframe's probability reflects the relative confidence that it represents the beginning of a new shot. The UAE examines these uncompressed keyframes for videotext using an edge-based method and examines keyframes for faces [4]. It then updates the keyframe list indicating which keyframes appear to have faces and/or videotext.

In the audio domain, the UAE extracts low-level acoustic features such as short-time energy, mel-frequency cepstral coefficients and delta spectrum magnitude coefficients. These features are extracted from the audio stream (PCM .wav tiles sampled at 44.1 kHz) frame by frame along the time axis using a sliding window of 20ms with no overlap. A multidimensional Gaussian maximum ***a posteriori*** estimator classifies each segment into one of seven audio categories: speech, noise, music, silence, speech plus noise, speech plus music, and speech plus speech.

The transcript stream is extracted from the MPEG-2 closed captions (CC) data. The UAE generates a timestamp for each line from the frame information. The UAE looks for double and triple arrows to identify events such as a change in topic or speaker. Figure 4 displays an overview of the transcript analysis. The UAE begins by extracting a high-level summary using known cues for the program's genre. The knowledge database and the temporal database embody the domain knowledge and include the temporal sequence of cues. For example: when analyzing a ski race, the system creates segments for each racer. The phrases "now at the gate" or "contestant number" indicate when a contestant starts. We employ a database of categories with associated keywordsikey-phrases for different topics. This database helps find main topics. Finally, the non-stop words (generally nouns and verbs) in each segment are indexed.

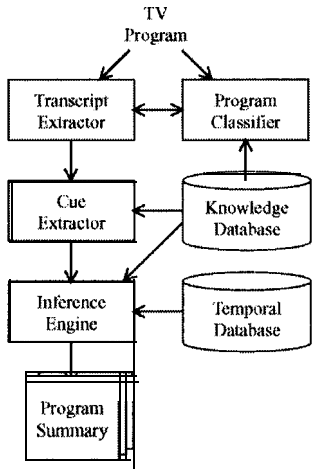

Figure 4. Story segmentation and summarization.

3.2 Multimodal Analysis

The Bayesian Engine (BE), a probabilistic framework, performs multimodal integration [5]. We chose to use a probabilistic framework for the following reasons: (i) allows for the precise handling of certainty/uncertainty, (ii) describes a general method for the integration of information across modalities, and (iii) has the power to perform recursive updating of information. The probabilistic framework we use is a combination of Bayesian networks with hierarchical priors.

The MIF is divided into three layers: low, mid, and high-level. The low-level layer encompasses the following low-level feature: visual (color, edges, shape); audio (20 different signal processing parameters); transcript (CC text). In the process, probabilities are generated for each feature. The mid-level layer encompasses the following features: visual (faces, keyframes, videotext); audio (7 audio categories); transcript (20 CC categories). These features are causally related to features in the low-level layer. Finally, the high-level layer represents the outcome of high-level inferences about a segment's genre.

In implementing the MIF, we used the mid and high-level features to perform high-level inferences. We compute and use probabilities only for these layers. The features in the mid-level layer represent video content information, while the features in the high-level layer represent the results of inferences.

High-level inferences are modeled by a combination of hierarchical priors and Bayesian networks. In order to determine a segment's high-level indexing, the MIF computes two joint probabilities, one for financial news and one for talk shows.

$$P_{\text{fin-topic}} = P_{\text{videotext}} * P_{\text{keywords}} * P_{\text{face}} * P_{\text{audio-fin}} * P_{\text{CC-fin}} * P_{\text{face-text-fin}} \quad (1)$$

$$P_{\text{talk-topic}} = P_{\text{videotext}} * P_{\text{keywords}} * P_{\text{face}} * P_{\text{audio-talk}} * P_{\text{CC-talk}} * P_{\text{face-text-talk}} \quad (2)$$

For example, a TV talk show contains a lot of speech, noise (clap, laughter, etc.) and background music (with or without speech), faces, and a very structured story line.

4. RETRIEVAL AND ACCESS

The retrieval module consists of the Personalizer and the user interface. The Personalizer matches indexed content from MIF with specific requests in the user profile. The user interface allows users to visualize and playback the video segments.

4.1 Personalizer

The Personalizer uses the user profile as a query against the stored, indexed content. The user interface represents these requests using a "magnet" metaphor, allowing users to "attract" different categories within the themes financial and celebrity. For TV financial news requests, the Personalizer searches the transcript and reconciles segment boundaries using the description given by the Bayesian Engine. When a segment has a financial category that appears more than n times (n>= 2), it is indexed under that category name. For example: if a financial news segment mentions the word "Microsoft" more than two times, that segment is indexed as "Microsoft" under the corresponding magnet. For celebrities, the Pcrsonalizer looks for matches between the summarization information produced by the UAE and users' celebrity magnets.

4.2 User Interface

We created a user interface for three reasons: (i) environment for testing different segmentation and indexing techniques, (ii) method for sharing our research with product managers, and (iii) method for testing consumer need for content-based

segmentation and indexing of television in the home. The interface is divided into two sections called Program Guide and TV Magnets. The Program Guide (see Figure 5) allows users to interact with whole TV programs that they can segment in different ways. TV Magnets offers users access to their profiles and access to video clips organized by topic. Users navigate the interface on a TV screen using a remote control.

Figure 5. Program Guide screen showing a stored talk show's segmented by TV commercials.

5. RESULTS

The individual algorithms for unimodal analysis have been benchmarked and the results have previously been published [6, 7]. We benchmarked the Bayesian Engine to investigate how the system would perform on financial and celebrity segments. We automatically segmented and indexed seven TV programs: Marketwatch, Wall Street Week, and Wall Street Journal Report and talk shows hosted by Jay Leno and David Letterman. Initially, each segment was classified as either a program segment or a commercial segment.

Program segments were subsequently divided into smaller, topic-based segments relying mainly on the transcript. The Bayesian Engine performed a bi-partite inference between financial news and talk shows on these sub-program units. Visual and audio information from the mid-level layer was also used in addition to the transcript. Next, a post-processing of the resulting inferences for each segment was performed by merging small segments into larger segments.

We had 84 program segments from financial news and 168 program segments from talk shows, giving us a total of 252 segments to classify. When using all multimedia cues (audio, visual, and transcript), we get the following results: (i) total precision of 94.1% and recall of 85.7% for celebrity segments, and (ii) total precision of 81.1% and a recall of 86.9% for financial segments.

6. CONCLUSIONS

We described Video Scout a content-based retrieval system for personalizing TV at a subprogram level. The archiving module employs a three-layered, multimodal integration framework to segment, analyze, characterize, and classify segments. The retrieval module relics on users' personal preferences to deliver both full programs and video segments.

Video Scout is currently implemented on a 600 MHz PC and a Philips TriCodec board (100 MHz TM1000 TriMedia). On this low-end platform, the visual, audio and transcript analyses take less than the length of the TV programs to analyze and extract the features while the Bayesian engine takes less than one minute per one hour of TV program to give the segmentation and classification information. The rchieval application is not a compute heavy part of the system and therefore can migrate onto the TriMedia. We demonstrated that content-based retrieval technology can be embodied in a consumer application.

In future we plan to explore (i) the use of different features (ii) multimodal integration for narrative content, and (iii) delivery of a full fledged system capable of responding to users' personalization needs. Also, we plan to integrate face learning and person identification methods into the current system.

ACKNOWLEDGEMENTS

We would like to thank the following people for their contribution to this research: Gang Wei of Wayne State University, Serhan Dagtas of Philips Research, George Marmaropoulos, Clive van Heerden and John Milanski of Philips Design.

REFERENCES

[1] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," presented at ACM Multimedia, Marina Del Rey, 2000.

[2] T. Syeda-Mahmood and S. Srinivasan, "Detecting Topical Events in Digital Video," presented at ACM Multimedia, Marina Del Rey, 2000.

[3] M. Naphade and T. Huang, "A Probabilistic Framework for Semantic Indexing and Retrieval in Video," presented at IEEE International Conference on Multimedia and Expo, New York, 2000.

[4] G. Wei and 1. Scthi, "Omni-Face Detection for Video/Image Content Description," presented at Intl. Workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia, 2000.

[5] R. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, and J. Zimmerman, "Video Scouting: an Architecture and System for the Integration of Multimedia Information in Personal TV Applications," presented at ICASSP, Salt lake City, 2001.

[6] N. Dimitrova, L. Agnihohi, C. Dorai, and R. Bolle, "MPEG-7 VideoText Description Scheme for Superimposed Text in Images and Video," ***Signal Processing: Image Comm. Journal,*** pp. 137-155, 2000.

[7] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of General Audio Data for Content-Based Retrieval," ***Pattern Recognition Letters, vol.*** 22, pp. 533-544, 2001.