Video Scouting Demonstration: Smart Content Selection and Recording

Nevenka Dimitrova
Lalitha Agnihotri

Radu Jasinschi
John Zimmerman
George Marmaropoulos

Thomas McGee
Serhan Dagtas

Philips Research, 345 Scarborough
Briarcliff Manor, NY 10510, USA
Phone +1 914 945-6059
nevenka.dimitrova@philips.com

ABSTRACT

Smart video content selection and recording is the best selling feature of the current personal TV receivers like TiVo. These devices operate at the TV program level in that they use electronic program guides and user's program personal preferences to help consumers record and watch programs that match their interests. In this Video Scouting demonstration, we present a system that allows for the filtering and retrieving of TV sub-programs based on user's content preferences. The filtering process is realized via real-time video, audio, and transcript analysis. The demonstrator personalizes the TV experience in the areas of celebrity and financial information. The technology can translate into differentiating storage and set-top box product features for finding your favorite actors, most interesting personalized financial news of the day, commercial compaction and enhancement, and content augmentation with other sources of information such as Web pages and encyclopedia. The demonstrator also reflects our active involvement in the MPEG-7 standard (Content Description Interface).

Keywords

Video access, content-based retrieval, multimodal integration.

1. INTRODUCTION

Home entertainment is bound to change due to the introduction of personal TV receivers [1][2] in hard disc recorders. Currently, these devices offer personalization at the TV program level mainly using electronic program guides (EPG). The EPG uses the generic form of metadata consisting of program name, cast, direction, etc. However, audio and visual information can be described in terms of its contents, this at different levels of granularity and levels of abstraction. Therefore audio and video content has potentially higher levels of semantic interpretation that can be exploited by more advanced content analysis methods that operates at the TV program segment level, specially when integrated with textual, e.g., close captioning (CC) information. This would allow consumers to access recorded programs or their segments more effectively. For example, multiple interviews in a TV talk show are automatically separated and guests are identified. In the financial area, segments that are related to the investment portfolio of the user are selected (e.g. "IPO," "market capitalization," "Philips," "IBM," "Oracle" etc.). The identified segments matching the viewer's interests can then be presented.

In this demonstration, we present an architecture for content-based video analysis, filtering, and retrieval. Video Scouting personalizes the TV experience at the program segment level, which is a level deeper than program level selection using only EPG. This is broadly shown in Figure 1.

In Figure 1, there are two inputs to the Video Scouting system: the video stream and the EPG/personal profile. The video stream is demuxed. The resulting audio, video (visual) and CC (textual) stream and the user preferences are input to the TV Program Content Selection module. This module is explained next.

2. TV CONTENT SELECTION : MULTI-MODAL CONTENT ANALYSIS

The TV Program Content Selection module consists of the following layers: (i) feature extraction, (ii) tools, (iii) semantic processes, and (iv) user applications. The *feature extraction layer* performs the segmentation of low-level content from the audio, visual, and textual domain. For example, in the audio domain we have the cepstral coefficients, in the video domain we have color, shape, and edges, and in the textual domain we have words. At the *tools layer*, we have extraction of intermediate-level content. For example, in the audio domain we have audio categories, e.g., speech, music, silence, in the video domain we have scene cuts, color histograms, visual text, or (detected) faces, and finally in the textual domain we have keywords. In the next layer – *the semantic processing layer* – information across different domains is integrated. For example, TV commercials are segmented based on audio, scene cut rates or CC information. Also, TV program segments are generated. The top layer – *the user application layer* – generates a final segmentation of specific TV programs by combining the program segments obtained at the previous layer into a single unit that unifies them according the user's profile specifications.

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM Multimedia 2000 Los Angeles CA USA
Copyright ACM 2000 1-58113-198-4/00/10...$5.00

Each of these layers contains a set of elements describing operations on multimedia information. These elements have probabilistic variables associated with them. Taken together, this processing is abstracted by a directed acyclic graph (DAG) for which the nodes represent the individual elements in each layer and the directed arrows represent conditional probabilities. In this demo, the user application layer deals with the representation of two topics: (i) financial, and (ii) celebrities. One of the outputs of this top layer is a probability of occurrence of a specific topic given all dependent information. Initially the multimedia information is processed at the frame level for layers (i) and (ii). At layer (iii) the information is processed at the program segment level. This means that in going from layers (i) and (ii) to layer (iii) all the related multi-modal information is integrated and indexed.

The system runs on a PC-based platform with a Philips Tri-Media multimedia processing board. The input is MPEG-2 video. The extraction of visual features is performed on Tri-Media while audio and textual processing are performed on the PC.

3. RETRIEVAL

For the consumer, the experience is delivered through a simple user interface called the "raindrop model" that removes the difference between live TV and stored content (see Figure 2.) The users will be given the power to design their own "magnets" for desirable or undesirable content. The search power tools include "money seeker" and "celebrity seeker".

4. APPLICATIONS

The applications "inside" the box include personal video content selection and recording, searchable TV, live alerts, commercial detection and enhancement, content augmentation [3], parental control, personalized news retrieval [1], video editing, and content manipulation.

5. SUMMARY

In this demonstration we present an integrated system for audio, visual, and transcript analysis of video content. The function is to provide personalized content selection and recording for TV programs at the subprogram level.

6. REFERENCES

[1] N. Dimitrova, H. Elenbaas, T. McGee, "PNRS - Personal News Retrieval System," *Proc. SPIE Conference on Multimedia Storage and Archiving Systems*, Boston, (September 1999)

[2] N. Dimitrova, T. McGee, L. Agnihotri, S. Dagtas, R. Jasinschi, "On Selective Video Content Analysis and Filtering," *Proc. SPIE on Image and Video Databases*, San Jose, (January 2000).

[3] N. Dimitrova, Y. Chen, L. Nikolovska, "Visual Associations in DejaVideo," *Proc. Asian Conference on Computer Vision*, Taipei, (January 2000).

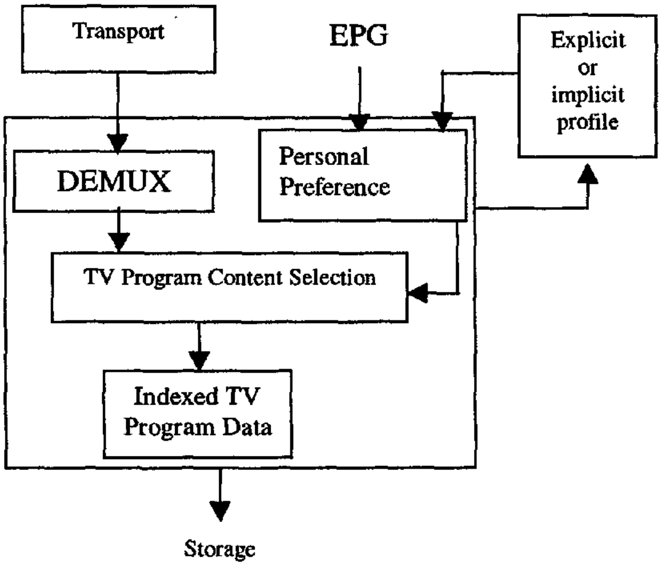

Figure 1. Overall Video Scouting system elements.

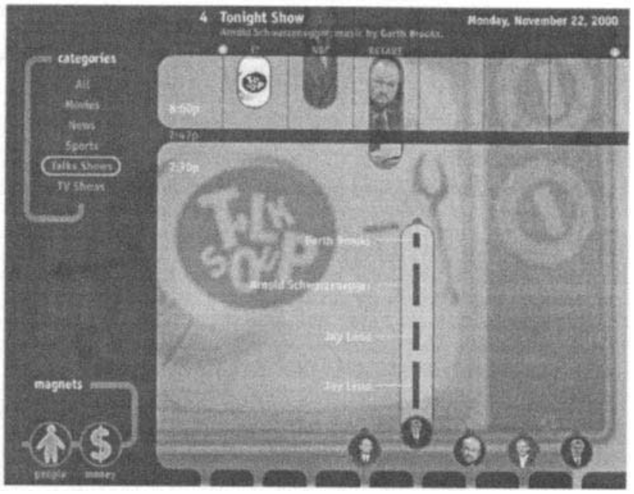

Figure 2. Screen dump of the video scouting system.