VIDEO SCOUTING: AN ARCHITECTURE AND SYSTEM FOR THE INTEGRATION OF MULTIMEDIA INFORMATION IN PERSONAL TV APPLICATIONS

R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, and J. Zimmerman

Philips Research, 345 Scarborough Road, Briarcliff Manor, N.Y., 10510

ABSTRACT

Currently available Personal Video Recorders find and store whole TV programs. Our system, Video Scouting, not only finds and stores programs; it automatically segments and indexes story segments from the programs according to viewers' profiles. The extracted descriptions serve the viewers' content information requests for program segment selection, e.g. play the three minute interview with Hillary Clinton. To achieve this, the system combines information from the audio, visual, and transcript domains in a probabilistic framework based on Bayesian networks. In this paper we describe the overall architecture, a system implementation, and discuss some experimental results.

1. INTRODUCTION

Personal entertainment devices such as TiVo and ReplayTV, Personal Video Recorders (PVRs), have begun to change the way people consume media. Using metadata descriptions of content, PVRs find and record whole TV programs. Our system, Video Scouting, not only finds and stores programs; it offers sub-program segment selection and recording. For example, users watching a talk show can choose to watch only the specific interview segment with Michelle Pffeifer. Video Scouting advances current PVRs that use only metadata for Electronic Program Guides (EPGs) and personal program profile (PPP) information. Video Scouting allows users to input specific content requests via the content personal program profile (CPPP), extending the PPP. Using information from CPPP, our system segments and indexes TV program according to content. In particular, Video Scouting allows segmentation and indexing of TV programs according to high level, i.e., semantic information (e.g., celebrities and financial topics). Users should feel more rewarded because they can interact with these systems in terms of simple high level information and related commands.

This paper is organized as follows. In Section 2 related work to Video Scouting is described. In Section 3 the architecture is described. In Section 4, the system implementation is described. Finally, in Section 5, we draw conclusions.

2. RELATED WORK

Literature abounds with methods for video segmentation, retrieval by similarity, video classification, and automatic content description [1, 6]. However, the focus so far has been mostly on algorithms for retrieval in large video archives. There are several (research) systems that combine multiple aspects of audio and visual processing. Notable examples are Query-By-Image-and Video-Content (QBIC)[7], VisualGrep [8], DVL of AT&T, InforMedia [11], VideoQ [13], MoCA [12], Vibe [9] and CONIVAS [10]. In particular, the InforMedia, MoCA, and VideoQ systems are more related to Video Scouting. The InforMedia project is a digital video library system containing methods to create a short synopsis of each video primarily based on speech recognition, natural language understanding, and caption text. The MoCA project is designed to provide content-based access to a movie database. Besides segmenting movies into salient shots and generating an abstract of the movie, the system detects and recognizes title credits and performs audio analysis. The VideoQ system classifies videos using compressed domain analysis consisting of three modules: parsing, visualization and authoring. The main advantage of Video Scouting over the previously mentioned systems is that it uses audio (A), visual (V), and transcript (T) processing for story segmentation and classification as opposed to indexing video based on low level features. In addition, Video Scouting uses multimodal processing for consumer applications, which have been ignored by previously described systems.

3. VIDEO SCOUTING: ARCHITECTURE

First, we discuss general properties of the VS architecture, and second, we present a probabilistic framework for the representation and integration of VS features.

3.1. General Properties

The VS architecture is shown in Figures 1 and 2. Figure 1 shows the three main modules of VS: Video Pre-Processing, Segmentation and Indexing, and User Interface with Storage. Figure 2 expands the segmentation and indexing modules of Figure 1. We can see in Figure 2 the three layers, low, mid, and high-level. Next we explain the functional part described by these two figures.

Figure 1 shows three types of input: (i) the video (marked as "Video In" over the vertical arrow), (ii) the EPG, PPP, and CPPP (shown at the right hand side of the horizontal arrow), and (iii) user input (shown under the vertical arrow pointing towards the User Interface).

In the Video Pre-Processing module, TV programs are demultiplexed and then decoded into the A, V, and T streams. These three streams are then sent to the Segmentation and Indexing module for processing. In addition to the A, V and T streams, the Segmentation and Indexing module uses the EPG, PPP, and CPPP. EPG is a standard table with program information such as program title, broadcast time, actors, and genre. The PPP is a file derived from user specified program preferences such as action movies, or programs about economical issues. This file could be dynamically updated [2] to reflect changes in user preferences. Using the PPP the system additionally generates a list of viewing/recording

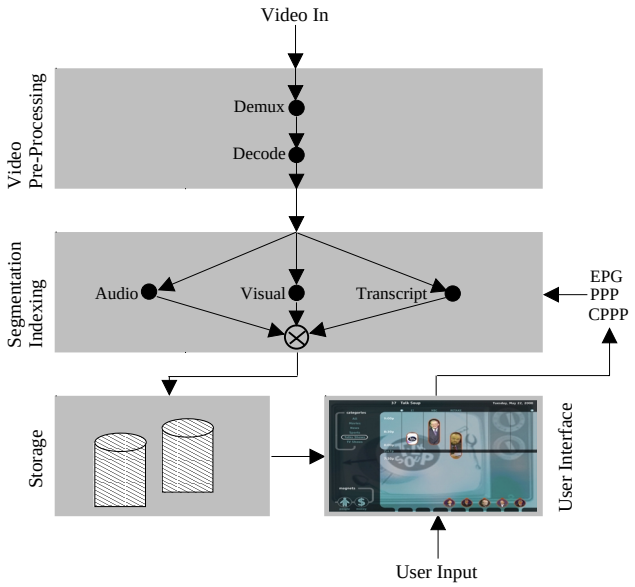

Fig. 1. The overall Video Scouting architecture.

recommendations. Both, the EPG and the PPP are standard features of PVRs, such as TiVo. The CPPP file differentiates Video Scouting. This file contains user information/preferences in terms of company names, celebrity names, topics, and keywords. The CPPP can also evolve dynamically according to user's program content preference changes. This is shown in Figure 1 by the arrow going from the "User Input" area, and this is triggered by the "User Input" input arrow. The Segmentation and Indexing module illustrates the integration part of this module where the A, V, and T circles are linked by arrows converging onto the ⊗ integration symbol. Figure 2 displays the Segmentation and Indexing including the multi-modal integration. The details of how this module combines the EPG, PPP, and CPPP information with the processed content is described in Section 4. Finally, there is a storage module that stores the video segments which were generated by the Segmentation and Indexing module. The UI of the retrieval application handles the users' requests and generates desired output using a retrieval back-end engine.

Figure 2 displays the core elements of Video Scouting in terms of the Low, Mid, and High-Level layers. In these layers video features are segmented and indexed. Within each layer, and in particular in the High-level one, multimodal integration is performed. The input is the demultiplexed and decoded video stream of the TV program selected for analysis. Note that certain feature extraction algorithms (e.g. edges for videotext detection) in the system use spatial domain analysis while others use compressed domain analysis (e.g. color in keyframe extraction) [2]. Different features are extracted at the low-level layer from each stream. In the current implementation of Video Scouting, the following features are extracted. In the visual domain, the system extracts color, edge, and shape. In the audio domain the system extracts 143 different parameters, including average energy, bandwidth, pitch, Mel-Frequency cepstral coefficients (MFCC), linear predictive coding coefficients (LPC), zero-crossings (ZCR). In the transcription domain, twenty different categories are extracted such as politics, economy, and sports.

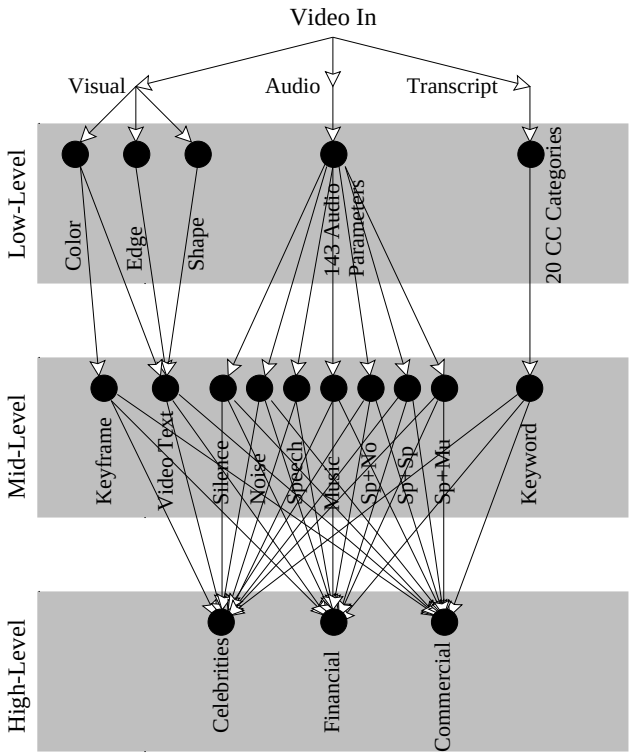

Fig. 2. The three layered Segmentation and Indexing module.

Mid and high-level features are generated via integration of information. This is especially true for the elements of the high-level layer, which describe semantic content. This multimodal integration uses inter-domain or intra-domain information. The intra-domain integration is done separately for each domain, while the inter-domain integration is done across domains. Mid-level features consist of keyframes and videotext in the visual domain; silence, noise, speech, music, speech with background noise (Sp+No), speech with speech (Sp+Sp), and speech with music (Sp+Mu) in the audio domain; and keywords in the transcript domain [15].

High-level features correspond to the division between the actual program, described here by the celebrities and financial topics, and the non-program, i.e., the commercials. The system uses multimodal integration to perform this commercial detection. Next, we describe the probabilistic framework used to implement the feature segmentation and indexing of Figure 2.

3.2. Probabilistic Framework

We chose a probabilistic framework because of its precise handling of uncertainty information. This is a general framework for integration of multimodal information and for the recursive updating of information. Figure 2 illustrates the different types of representation and granularity for each layer. In addition, each layer can have its own set of granularities. In this section we describe a framework for the probabilistic representation based on Bayesian networks – called Bayesian Engine (BE). We chose a Bayesian network framework because it can automatically encode the conditional dependency between the various elements within each layer and/or between each layer. There are many detailed descriptions

of Bayesian networks such as [4]. Here we give a brief definition of the terminology and present the main concepts.

According to [4], Bayesian networks are directed acyclical graphs (DAG) in which: (i) the nodes correspond to (stochastic) variables, (ii) the arcs describe a direct causal relationship between the linked variables, and (iii) the strength of these links is given by conditional probability distributions (cpds). Let the set $\Omega(x_1, \ldots x_N)$ of N variables define a DAG. For each variable there exists a sub-set of variables of Ω, Π_{x_i}, the parents set of x_i, i.e., the predecessors of x_i in the DAG, such that $P(x_i \mid \Pi_{x_i}) = P(x_i \mid x_1, \ldots, x_{i-1})$, where $P(\cdot|\cdot)$ is a cpd, strictly positive. Now, given the joint probability density function (pdf) $P(x_1, \ldots, x_N)$, using the chain rule [4], we get that $P(x_1, \ldots, x_N) = P(x_N \mid x_{N-1}, \ldots, x_1) \times \cdots \times P(x_2 \mid x_1) P(x_1)$. According to this equation, the parent set Π_{x_i} has the property that x_i and $\{x_1, \ldots, x_N\} \setminus \Pi_{x_i}$ are conditionally independent given Π_{x_i}.

In Figure 2 the flow diagram of the BE has the structure of a DAG made up of three layers. In each layer, each element corresponds to a node in the DAG. The directed arcs join one node in a given layer with one or more nodes of the proceeding layer. Basically, two sets of arcs join the elements of the three layers. For a given layer, and for a given element, we compute a joint pdf as previously described. More precisely, for an element (node) $i^{(l)}$ associated with the lth layer, the joint pdf is:

$$\begin{aligned} & P^{(l)}(x_{i^{(l)}}^{(l)}, \Pi^{(l-1)}, \ldots, \Pi^{(2)}) = P(x_{i^{(l)}}^{(l)} \mid \Pi^{(l)}) \\ \times \quad & \{P(x_1^{(l-1)} \mid \Pi_1^{(l-1)}) \ldots P(x_{N^{(l-1)}}^{(l-1)} \mid \Pi_{N^{(l-1)}}^{(l-1)})\} \ldots \\ \times \quad & \{P(x_1^2 \mid \Pi_1^2) \ldots P(x_{N^2}^2 \mid \Pi_{N^2}^{(2)})\}, \end{aligned} \tag{1}$$

where for each element $x_i^{(l)}$ there exists a parent set $\Pi_i^{(l)}$; the union of the parent sets for a given level l, i.e., $\Pi^{(l)} \equiv \sum_{i=1}^{N^{(l)}} \Pi_i^{(l)}$. There can exist an overlap between the different parent sets for each level.

4. VIDEO SCOUTING: A SYSTEM IMPLEMENTATION

In this section we present issues related to the Video Scouting system implementation based on the architecture described in Section 3. The visual and transcript processing runs in on a TriMedia™ Tricodec card. Keyframe extraction presented in [5] uses macroblock level information from the DCT coefficients to determine frame differences. Videotext extraction consists of the following sequence of operations: edge detection, thresholding, region merging, and character shape extraction. In the current implementation, we only look for the presence or absence of text characters. ASCII characters are extracted from the decoded closed caption packets. A file containing a set of keywords gives category information. This file is generated as follows.

According to Figure 2 a set of twenty closed caption categories are extracted at the low-level layer. These are: weather, international (affairs), crime, sports, movie, fashion, tech stock, music, automobile, war, economy, energy, stock, violence, financial, national (affairs), biotech, disaster, art, and politics. For each of these categories we have a knowledge base: an association table of keywords and categories. After statistical processing the categorization is performed using category vote histograms. If a word in the closed caption file matches a knowledge base keyword, then the corresponding category gets a vote. Audio features are extracted on a PC. Seven mid-level features (silence, music, speech, noise and speech with background music, speech, and noise) are extracted by examining the 143 low-level features. All resulting mid-level features have an associated set of probabilities [15].

The following are the BE's main steps for the segmentation and indexing of TV programs according to users' requests. Inputs to the BE are output files from visual, audio, and transcript feature processing. These files contain either numbers, e.g., probability values, or characters, e.g., closed caption text. Based on these content descriptions files, the BE segments the TV program into: (i) commercials, and (ii) program content. Second, it classifies the program part into sub-parts according to financial or celebrity segments. In the current implementation, Video Scouting handles story segmentation in structured program genres (financial, celebrities) that include a lot of conversations. The closed captions contain most of the relevant information for the BE classification task. The closed caption data drives the initial program segmentation into sub-program parts. The other cues, i.e., visual and audio, are used jointly with the closed caption data for the inference part (see the High-Level layer in Figure 2). In order to realize all this, the BE:

1. Reads in the content description files and a set of predefined threshold values.
2. For each (video) successive frame, the BE segments the video into the commercial and program parts.
   (a) Computes the commercial joint probability, according to Equation 1. This is given by the product of the probabilities for keywords, commercial detection given keywords, videotext, commercial detection given videotext, keyframes, commercial detection given keyframes, audio segmentation, and commercial detection given audio segmentation.
   (b) Determines if the result corresponds to a commercial or a program part based on CC information. It also verifies if the joint probability for the commercial is larger than a pre-defined threshold.
3. For each (video) frame corresponding to a program part the BE:
   (a) Computes the financial and celebrities joint probabilities. For the celebrity topic this is the product of the probabilities of keywords, celebrity given keywords, audio segmentation, celebrities given audio segmentation, celebrity categorization. An identical combination of probabilities defines the joint probability that it is a financial segment.
   (b) Compares the financial and celebrity joint probabilities with two corresponding thresholds. If one of them is larger than the threshold, then a vote of 1 is given to that frame. In the opposite case a zero vote is given.

Each frame receives either a 1 or 0 vote per segment genre. For each sub-program segment the total number of votes as computed from the joint financial and celebrity probabilities is counted and divided by the total number of frames. The resulting number describes the frequency or probability of a segment being associated with a financial or celebrity program.

Video Scouting has been applied to a set of nine different TV programs. Below we show samples of input and output files for the financial TV program "Wall Street Week."
1. a sample of the (input) probabilities for the set of seven audio categories, were the first two columns indicate the start and end times in frames, and the remaining columns display the probabilities for these categories:

000 025 1.00 0.00 0.00 0.00 0.00 0.00 0.00
026 118 0.00 1.00 0.00 0.00 0.00 0.00 0.00
119 167 0.00 0.67 0.00 0.33 0.00 0.00 0.00

2. a sample of the (input) probabilities for the set of twenty CC categories, in the order defined above, with the first two columns corresponding to the start and end times of each segment in frames:

12136 13086 0 0 0 0.09 0.09 0 0 0.09 0 0 0.36 0 0.27 0 0.09 0 0 0 0 0
13086 14466 0 0 0 0.09 0.09 0 0 0 0 0 0.45 0 0.18 0 0.09 0.09 0 0 0 0
14466 15599 0.08 0 0.08 0.08 0.08 0 0 0.08 0 0 0.33 0 0.17 0 0.08 0 0 0 0 0

As a sample of outputs, we show,

1. a sample of segmented and indexed program segments:

12136 13086 economy ; economy rate interest stock going ;
13086 14466 economy ; indicator rate interest stock Basis Point raise may points ;
14466 15599 economy ; bond rate stock Start july. ;

2. probabililities for the two topic-based program segment classification. The two probabilities, for financial and celebrity topics, correspond to joint probabilities. Out of the probabilities only the probabilities corresponding to the CC and audio categories vary; the other ones are fixed to one. Now, each of these two probabilities are obtained, separatelly, by chosing the largest of a sub-set of CC or audio probabilities. These sub-sets are: 1. financial: 1.1 audio: speech and speech plus music, 1.2 CC: stock, tech-stock, financial, politics, energy, economy, biotech, international, national, automobile, and war; 2. celebrities: 2.1 audio: noise, speech, and speech plus music, 2.2 CC: movie, music, fashion, sport, and art. Therefore, a sample result for the segment between frames 13086 and 14466:

(a) audio category probabilities:

pSIL = 0.050, pNOI = 0.000, pSP = 0.931, pMUS = 0.000,
pSPNOI = 0.000, pSPSP = 0.000, pSPMUS = 0.366.

(b) CC category probabilities:

pMUS = 0.000, pAUT = 0.000, pSTK = 0.000, pTSTK = 0.000,
pVIL = 0.000, pFIN = 1.000, pPOL = 0.000, pFASH = 0.000,
pWAR = 1.000, pENER = 1.000, pECON = 0.000, pBIOT = 0.000,
pWEAT = 0.000, pINT = 0.000, pCAT = 0.000, pSPO = 0.000,
pNAT = 0.000, pDIS = 0.000, pART = 0.000.

Out of these probabilities, the resulting program topic (joint) probabilities are: pFINANCIAL-TOPICS = 0.931, pCELEBRITIES-TOPICS = 0.000.

For this particular program used in this example, 30 out of 31 program segments were classified as financial. For the other three financial programs used, they had 23 out 23, 10 out 11, and 25 out of 28 program segments correctly classified as financial.

5. CONCLUSIONS

Video Scouting segments TV programs based of high-level inferences of content information. The distinguishing elements are: (i) video content segmentation and indexing; (ii) multimodal integration of audio, visual, and transcript information; (iii) a probabilistic framework for representation, processing, and integration of multimodal information; (iv) a generalization of current personal TV technology allowing selection of sub-program video clips according to content.

Video Scouting provides sub-program level segmentation based on user specified content. The system can distinguish between program segments and non-program segments (TV commercials). In addition, Video Scouting offers increased granularity of description. Within a program segment it can identify specific topics. For example: the system can find a several minute story on Philips within Financial news program. Video Scouting advances current video indexing systems and current commercial Personal Video Recorders through its processing of audio, visual, and transcript streams within its Bayesian Engine.

Future work includes: (i) the augmentation of visual, audio, and transcript features, e.g., visual faces, color histograms; (ii) the increase in the number of topics covered by VS; (iii) multimedia genre information; (iv) content augmentation information.

6. REFERENCES

[1] P. Aigrain, H. Zhang, and D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review," Multimedia Tools and Applications, Vol. 3, 179-202, November 1996.

[2] N. Dimitrova, T. McGee, L. Agnihotri, S. Dagtas, R. Jasinschi, "On Selective Video Content Analysis and Filtering," SPIE Storage and Retrieval for Media Databases, Jan. 2000.

[3] www.tivo.com.

[4] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference," Morgan Kaufmann Publishers, Inc., San Mateo, California, 1988.

[5] N. Dimitrova, T. McGee and H. Elenbaas, "Video Keyframe Extraction and Filtering: A Keyframe is not a Keyframe to Everyone," ACM Conf. Info. Knowledge Mgmt, 1997.

[6] N. Dimitrova, "Multimedia Content Analysis for Indexing and Retrieval Applications," ACM Journal of Info. Sci, 1999.

[7] W. Niblack, J.L. Hafner, T. Breuel, D. Ponceleon, "Updates to the QBIC System," SPIE, vol. 3312, pp. 150-161, 1997.

[8] A. Gupta and R. Jain, "Visual Information Retrieval," Communications of the ACM, vol. 40, pp. 71-79, May 1997.

[9] J.-Y. Chen, C. Taskiran, A. Albiol, E. J. Delp, and C. A. Bouman, "Vibe: A Compressed Video Database Structured for Active Browsing and Search," Purdue University 1999.

[10] M. Abdel-Mottaleb, N. Dimitrova, R. Desai, J. Martino, "CONIVAS: CONtent-based Image and Video Access System," ACM Multimedia, Boston, 1996.

[11] A. Hauptmann and M. Smith, "Text, speech, and vision for video segmentation: The Informedia project," AAAI Symp. on Comp. Models for Integrating Lang. and Vision, 1995.

[12] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting Digital Movies Automatically," Journal on Visual Comm. Image Repres., vol. 7, pp. 345-353, 1996.

[13] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An Automated Content Based Video Search System Using Visual Cues," ACM Multimedia, 1997.

[14] N. Vasconcelos and A. Lippman, "Bayesian Representations and Learning Mechanisms for Content Based Image Retrieval", SPIE Storage & Retrieval for Media DB, San Jose, 2000.

[15] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of General Audio Data for Content-Based Retrieval," Pattern Recognition Letters 2000