INTEGRATED MULTIMEDIA PROCESSING FOR TOPIC SEGMENTATION AND CLASSIFICATION

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

Philips Research, 345 Scarborough Road, Briarcliff Manor, NY 10510

ABSTRACT

In this paper we describe integrated multimedia processing for Video Scout, a system that segments and indexes TV programs according to their audio, visual, and transcript information. Video Scout represents a future direction for personal video recorders. In addition to using electronic program guide metadata and a user profile, Scout allows the users to request specific topics within a program. For example, users can request the video clip of the President speaking from a half-hour news program.

Video Scout has three modules: (i) Video Pre-Processing, (ii) Segmentation and Indexing, and (iii) Storage and User Interface. Segmentation and Indexing, the core of the system, incorporates a Bayesian framework that integrates information from the audio, visual, and transcript (closed captions) domains. This framework uses three layers to process low, mid, and high-level multimedia information. The high-level layer generates semantic information about TV program topics. This paper describes the elements of the system and presents results from running Video Scout on real TV programs.

1. INTRODUCTION

Personal Video Recorders (PVRs), such as TiVo [3], use electronic program guide metadata to automatically select and store whole TV programs based on users' profiles. Video Scout represents a future direction for PVRs. Scout analyzes the audio, visual, and transcript (closed captions) content in order to segment and index TV programs. In addition, Scout allows users to request topic level information. For example: users can request speeches by the President, and when the nightly news is broadcast, Scout can extract the President's speech from the half-hour news broadcast. This paper describes the elements of the system and presents results from running Video Scout on real TV programs.

Video Scout employs three interconnected modules: (i) Video Pre-Processing; (ii) Segmentation and Indexing; and (iii) Storage and User Interface. Segmentation and Indexing, the core of the system, uses a three layered probabilistic system, called the Bayesian Engine (BE). The BE integrates information from the audio, visual, and transcript domains in order to generate semantically indexed information from TV program segments. We implemented Video Scout as an end-to-end system, and in its current state it segments, indexes, and stores TV programs for financial news and for talk shows.

Section 2 offers a brief overview of related work. Section 3 gives a system overview of Scout. Section 4 describes the Video Pre-Processing module. Section 5 presents the Segmentation and Indexing module. Section 6 describes topic segmentation and classification. Section 7 shares our experimental results. And in Section 8 we draw conclusions from this work.

2. RELATED WORK

Content based video analysis and processing has been an active topic of research [5]. However, there are very few initial approaches to multimodal processing of audio, visual, and transcript information. The majority of content-based systems include Query-By-Image-and-Video-Content (QBIC), VisualGrep, DVL of AT&T, InforMedia, VideoQ, MoCA, Vibe, and CONIVAS. In particular, the InforMedia, MoCA, and VideoQ systems are more related to Video Scout.

However, there are very few initial approaches to multimodal processing of audio visual and transcript information in video analysis. Vasconcelos and Lippman [6] propose a statistical model that uses video content structure information to perform video shot detection and semantic characterization. Rui et al. [9] present an approach that uses low level features in audio to detect excited speech and a baseball hit within a probabilistic framework for automatic highlight extraction. Syeda-Mahmood et al. [8] present event detection in multimedia presentations from teaching and training videos. They use a probabilistic model that exploits the co-occurrence of visual and audio events. Another approach based on the observation that semantic concepts in videos interact and appear in context is proposed by Naphade [1]. To model this contextual interaction explicitly, a probabilistic graphical network of multijects or a multinet is proposed.

The main advantage of our system over the previously mentioned approaches is the use of multilevel audio, visual, and transcript processing for story segmentation and topic classification as opposed to indexing video based on low-level features. In addition, our system uses multimodal processing for consumer applications, which have been ignored by previously described systems.

3. SYSTEM OVERVIEW

Figure 1 illustrates Video Scout's three modules. Video Pre-Processing, receives raw video that has been, de-muxed and decoded as necessary. This generates three streams: audio, visual, and transcript. Decoding is optional and is only implemented when further video processing is done in the uncompressed domain.

Segmentation and Indexing, takes these three streams as input. It also receives electronic program guide metadata that describes the TV programs. This metadata contains general information such as channel, time, title, cast, etc. In addition, Segmentation and Indexing has input from program profiles (PP) and the content program profiles (CPP).

Users specify their program preferences via the PP, which can be used to generate recommendations about TV programs. The CPP allows users to input topic level preferences. These preferences, such as, athlete, movie star, company stock, etc., go beyond

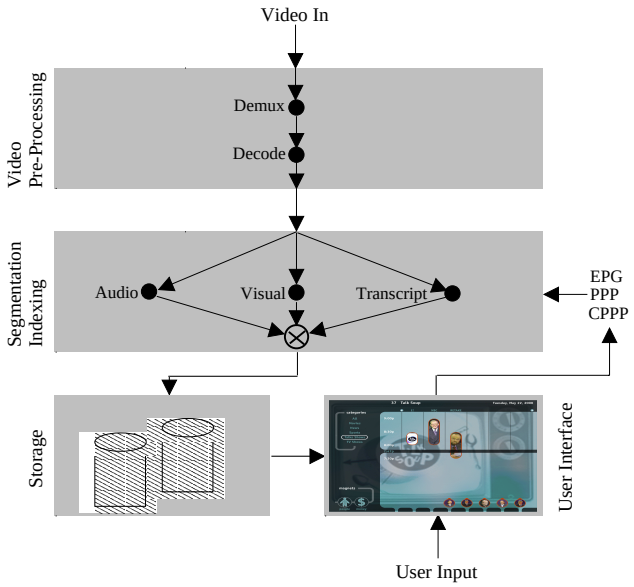

Fig. 1. The three modules of Video Scout.

what is described by the metadata and the PP in that they contain references to content information within a TV program. The central feature of the Segmentation and Indexing module is its ability to integrate information within and/or across the audio, visual, and, transcript domains.

Storage and User Interface offers users access and playback to stored programs and stored video clips. In addition, the user interface allows users to modify their PP and their CPP.

4. VIDEO PRE-PROCESSING

Video pre-processing is done by the Analysis Engine (AE) to combine video pre-processing and feature analysis. The AE takes MPEG-2 input and extracts the closed captions while performing a visual analysis for feature extraction. We selected a Philips TriMedia^{TM} Tricodec card for this task. This card has a TM1000 processor and 8Mb of memory. The TriMedia runs using the hosted mode in a 600 MHz PIII with a WinNT operating system. "C" code gets downloaded to the TriMedia where analysis is performed.

The AE performs shot detection first by extracting a new keyframe when it detects a significant difference between sequential I-frames. It employs two DCT based implementations for the frame differencing: histogram and macroblock. Unicolor keyframes or frames that appear similar to previously extracted keyframes get filtered out using a one-byte frame signature. This keyframe extraction produces an uncompressed image and a list giving the keyframe number and the probability of a cut. The AE bases this probability on the relative amount above the threshold using the differences between the sequential I-frames. The system then passes keyframes on for videotext detection. The AE looks for videotext using an edge-based method on the uncompressed image and tags each keyframe for the presence or absence of text. These keyframes are also analyzed for the presence of faces.

In addition to extracting the visual features, the AE also extracts closed captions from the user data field of the MPEG-2 video. The extraction process produces a complete time-stamped program transcript. We use time-stamps in order to align the transcript data with the related video.

Twenty low-level audio parameters are extracted using .wav files on a PC. The outputs of the Video Pre-Processing are used in the Segmentation and Indexing.

5. SEGMENTATION AND INDEXING

Figure 2 contains an illustration of the BE, the core of Segmentation and Indexing. We organized this probabilistic framework in three consecutive layers: low, mid and high-level. Each layer has nodes with associated probabilities. Arrows between the nodes indicate a causal relationship.

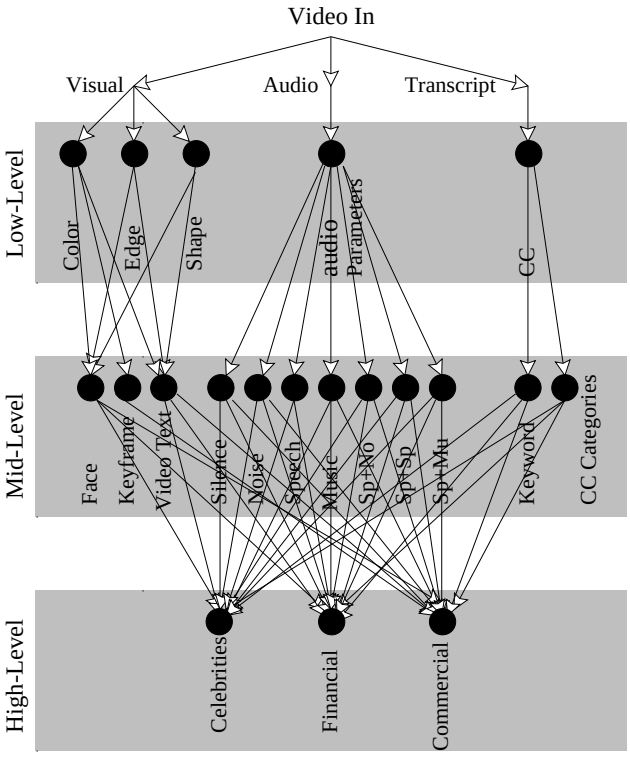

Fig. 2. The three layers of the Segmentation and Indexing module of Video Scout.

The low-level layer describes signal-processing parameters. In the current implementation parameters include: the visual features color, edge, and shape; twenty audio parameters [7] such as average energy, bandwidth, pitch, mel-frequency cepstral coefficients, linear prediction coding coefficients, and zero-crossings; and the transcript (text pulled from the ASCII characters of the closed-captions).

Arrows indicate the combinations of low-level features that create the mid-level features. Mid-level features are associated with whole frames or collections of frames while low-level features are associated with pixels or short time intervals. Keyframes (first frame of a shot), faces, and video text are mid-level visual features; silence, noise, speech, music, speech plus noise, speech plus speech, and speech plus music are mid-level audio features; and keywords and the twenty transcript categories make up the mid-level transcript features.

High-level features describe semantic video content obtained through the integration of mid-level features across the different modalities. In the current implementation Scout classifies segments as either part of a talk show, financial news, or a commercial.

We based the formal structure of the BE on Bayesian networks [4]. We chose this approach because probabilistic frameworks are designed to deal with uncertain information, and they are appropriate for representing the integration of information. The BE's probabilistic integration employs either intra or inter-modalities. Intra-modality integration refers to intergration of features within a single domain. For example: integration of color, edge, and shape information for videotext represents intra-modality integration because it all takes place in the visual domain. Integration of mid-level audio categories with the visual categories face and videotext offers an example of inter-modalities.

Bayesian networks are directed acyclical graphs (DAG) in which the nodes correspond to (stochastic) variables. The arcs describe a direct causal relationship between the linked variables. The strength of these links is given by conditional probability distributions (cpds). More formally, 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, 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 preceding layer. 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 l-th 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., $\Pi^{(l)} \stackrel{(df)}{=} \sum_{i=1}^{N^{(l)}} \Pi_i^{(l)}$. There can exist an overlap between the different parent sets for each level.

6. TOPIC SEGMENTION AND CLASSIFICATION

Topic segmentation and classification performed by BE is shown in the third layer (high-level) of Figure 2. The complex nature of multimedia content requires integration across multiple domains. We use the comprehensive set of data from the audio, visual, and transcript domains.

In the BE structure, Figure 2, for each of the three layers, each node and arrow is associated to a cpd. In the low-level layer the cpd's are assigned by the AE as described above. For the mid-level layer, twenty closed captions categories are generated: weather, international, crime, sports, movie, fashion, tech stock, music, automobile, war, economy, energy, stock, violence, financial, national (affairs), biotech, disaster, art, and politics. We use a knowledge tree for each category made up of an association table of keywords and categories. After a statistical processing, the system performs categorization using category vote histograms. If a word in the closed captions file matches a knowledge base keyword, then the corresponding category gets a vote. The probability, for each category, is given by the ratio between the total number of votes per keyword and the total number of votes for a closed captions paragraph.

Video Scout performs segmentation and indexing of TV programs according to the users' requests. It performs this task by (i) reading users' requests from the PP and CPP files, the metadata and the input data; (ii) segmenting the TV program into commercial vs. non-commercial parts; (iii) classifying the non-commercial parts into segments based on two high-level categories: financial news and talk shows (performed by the BE).

Initial segmentation and indexing is done using closed caption data to divide the video into program and commercial segments. Next the closed captions of the program segments are analyzed for single, double, and triple arrows. Double arrows indicate a speaker change. The system marks text between successive double arrows with a start and end time in order to use it as an atomic *closed captions unit*. Scout uses these units as the indexing building blocks. In order to determine a segment's high-level indexing (whether it is financial news or a talk show) Scout computes two joint probabilities. These are defined as:

p-FIN-TOPIC = p-VTEXT * p-KWORDS * p-FACE * p-AUDIO-FIN * p-CC-FIN * p-FACETEXT-FIN (2),

p-TALK-TOPIC = p-VTEXT * p-KWORDS * p-FACE * p-AUDIO-TALK * p-CC- TALK * p-FACETEXT- TALK (3).

The audio probabilities p-AUDIO-FIN for financial news and p-AUDIO-TALK for talk shows are created by the combination of different individual audio category probabilities. The closed captions probabilities p-CC-FIN for financial news and p-CC-TALK for talk shows are chosen as the largest probability out of the list of twenty probabilities. The face and videotext probabilities p-FACETEXT-FIN and p-FACETEXT-TALK are obtained by comparing the face and videotext probabilities p-FACE and p-TEXT which determine, for each individual closed caption unit, the probability of face and text occurrence. One heuristic used builds on the fact that talk shows are dominated by faces while financial news has both faces and text. The high-level indexing is done on each closed captions unit by computing in a new pair of probabilities: p-FIN-TOPIC and p-TALK-TOPIC. The highest value dictates the classification of the segment as either financial news or talk show.

7. EXPERIMENTAL RESULTS

We used Scout to analyze seven TV programs. We looked at the half-hour financial news programs Marketwatch, Wall Street Week (WSW), and Wall Street Journal Report (WSJR) as well as the one-hour talk shows hosted by Jay Leno and David Letterman. The total video analyzed was about 6 hours. The results of our experiments are shown in following tables. For each table "Num. CC Units" indicates the total number of closed captions units for the given TV program. "Fin. News" and "Talk Shows" indicate the number of closed captions units (see Section 6 above) classified as

either financial news or as a talk show. "Unclassified" indicates the number of closed captions units that could not be classified.

In order to better understand the role audio, visual, and transcript in the integration, we performed experiments using two subsets of features. For example: Tables 1 and 2 display results using only visual and transcript information. For this we used the joint probabilities p-FIN-TOPIC = p-VTEXT * p-KWORDS * p-FACE * p-CC-FIN * p-FACETEXT-FIN and p-TALK-TOPIC = p-VTEXT * p-KWORDS * p-FACE * p-CC- TALK * p-FACETEXT-TALK. Using these features, Scout can reasonably classify talk shows, but it cannot find financial news.

Table 1: Financial News Using Visual and Transcript

Show	CC Units	Fin	Talk	Unclassified
Marketwatch	26	0	7	19
WSW-1	21	0	16	5
WSW-2	27	0	20	7
WSJR	10	0	5	5

Table 2: Talk Shows Using Visual and Transcript

Show	CC Units	Fin	Talk	Unclassified
Jay Leno	57	0	46	11
Letterman-1	51	0	44	7
Letterman-2	60	0	53	7

Tables 3 and 4 display results using only audio and transcript information. For this we used the joint probabilities p-FIN-TOPIC = p-AUDIO-FIN * p-KWORDS * p-CC-FIN and p-TALK-TOPIC = p-AUDIO-TALK * p-KWORDS * p-CC- TALK. With the visual features turned off Scout can easily classify financial news program, but it has trouble with talk shows. However, finding talk shows without using visual data is easier than finding financial news without using audio data.

Table 3: Financial News Using Audio and Transcript

Show	CC Units	Fin	Talk	Unclassified
Marketwatch	26	25	0	1
WSW-1	21	17	1	3
WSW-2	27	24	1	2
WSJR	10	9	0	1

Table 4: Talk Shows Using Audio and Transcript

Show	CC Units	Fin	Talk	Unclassified
Jay Leno	57	4	48	5
Letterman-1	51	18	31	2
Letterman-2	60	18	38	4

Tables 5 and 6 display the main results of our integration effort. When we employ all features, Scout can accurately classify both financial news and talk shows. Comparing Table 3 to Table 5, Scout loses some precision in classifying financial news when all features are combined. However, this slight loss in accuracy is more than made up for in Scout's increased ability to accurately classify talk shows (Table 4 compared to Table 6).

Table 5: Financial News Using All Features

Show	CC Units	Fin	Talk	Unclassified
Marketwatch	26	24	1	1
WSW-1	21	17	3	1
WSW-2	27	23	4	0
WSJR	10	9	1	0

Table 6: Talk Shows Using All Features

Show	CC Units	Fin	Talk	Unclassified
Jay Leno	57	2	50	5
Letterman-1	51	10	41	0
Letterman-2	60	5	53	2

Integration is clearly necessary to accurately discriminate between the two program genres. Without using audio, visual, and transcript data it is easy to misclassify entire genres.

8. CONCLUSION

This paper describes our work on Video Scout, a segmentation and indexing system for TV programs. The most important contributions from this work are: (i) integration of information from the audio, visual, and transcript domains; and (ii) high-level classification of TV program segments based on a probabilistic framework. In contrast to the majority of related multimodal integration research, our work on Video Scout comprehensively employs multimedia information from all three domains. Experimental results reveal that using audio, visual, and transcript information significantly improves classification accuracy across different TV program genres. Future work includes the introduction of other multimedia clues, such as color, and the use of more high-level information and reasoning.

9. REFERENCES

[1] M. Naphade, T. Huang, "A Probabilistic Framework for Semantic Indexing and Retrieval in Video", IEEE International Conference on Multimedia and Expo, New York, 31 July-2 August 2000.

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

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

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

[8] T. Syeda-Mahmood and S. Srinivasan, "Detecting Topical Events in Digital Video," Proc. of ACM Multimedia 2000, Marina Del Rey, November 2000, pp. 85-94.

[9] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," Proc. of ACM Multimedia 2000, Marina Del Rey, November 2000, pp. 105-115.