A PROBABILISTIC LAYERED FRAMEWORK FOR INTEGRATING MULTIMEDIA CONTENT AND CONTEXT INFORMATION

R. S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, D. Li, and J. Louie
Philips Research USA, 345 Scarborough Road, Briarcliff Manor, 10510, U.S.A.

ABSTRACT

Automatic indexing of large collections of multimedia data is important for enabling retrieval functions. Current approaches mostly draw on a single or dual modality of video content analysis. Here we describe a framework for the integration of *multimedia content* and *context* information, which generalizes and systematizes current methods. *Content* information in the visual, audio, and text domains, is described at different levels of granularity and abstraction. *Context* describes the underlying structural information that can be used to constrain the possible number of interpretations. We introduce a probabilistic framework that combines (a) Bayesian networks that describe both content and context and (b) hierarchical priors that describe the integration of content and context. We present an application that uses this framework to segment and index TV programs. We discuss experimental results on segment classification on six and a half hours of broadcast video. In our experiments we used audio context information. Classification results for financial segments yield 83% and for celebrity segments 89%.

1. INTRODUCTION

Powerful search mechanisms and personalized access to massive collections of multimedia data are essential for the future home multimedia archives. Also, the need for good quality visual communication systems under restricted bandwidth conditions, such as, video conferencing, require the integration of visual and audio information. In order to address these issues different methods for content-based video analysis and indexing have been developed. These methods rely basically on the processing of multimedia content information, and they create complex, and in most cases, restricted and brittle results.

In this paper we describe a framework that proposes a systematic extension of current methods for multimedia information processing. This framework includes a systematic description of multimedia content and context information. A probabilistic framework that describes the representation and integration of multimedia content and context information is presented. This framework combines Bayesian networks with hierarchical priors.

The experimental work is performed within the Video Scout [1,2,3] prototype system that aims at generalizing the functionalities of personal video recorders (PVR)[4]. The outputs of Video Scout are TV program sub-segments that are indexed and classified according to high-level topics. We benchmark the multimedia integration framework within Video Scout on 6 ½ hours of recorded broadcast material and we show, among other things, that the integration of content and context information in TV program sub-segment improves classification tasks considerably.

Related work on multimodal integration includes the one by Vasconcelos and Lippman [5] that proposes a statistical model for video shot detection and semantic characterization. Here they only use audio and visual information, however they do not use textual information. Also, another approach based on the observation that semantic concepts in videos interact and appear in context is proposed by Naphade [6]. This approach uses a probabilistic graphical network. In both these cited references they do not have a layered organization of context and content information. In [7] Chen discusses methods for the bi-modal integration of visual (lip) and audio (speech) information.

Section 2 discusses conceptual properties of content and context information. In Section 3 we describe Bayesian networks and hierarchical priors. In Section 4 we describe experiments and their results implemented in Video Scout. Finally in Section 5 we draw conclusions.

2. CONTENT AND CONTEXT INFORMATION

Multimedia information spans three domains: visual, audio, and transcript. By examining these three domains we can detect objects – content information, and infer the environment or situation in which these objects exist – context information. Next, we will discuss in detail the general properties of multimedia content and context information.

2.1. Multimedia Content

A content object can be a specific shot, a particular frame in a shot, a line of dialogue, a face, etc. In order to systematize the definition of content we need the following two concepts: *granularity* and *abstraction*. Granularity describes the *levels of detail*. These levels depend on spatial and/or temporal scales. For each domain and for each attribute we employ a granularity scale. For example, in the visual domain spatial-granularity can be defined as local (pixels, voxels), regional, or global. Abstraction is used to describe semantic information. This information is especially valuable when dealing with semantically rich content such as TV programs and movies. Semantic information is created based on the relationships among different objects. Objects are related within each level of granularity, producing different levels of abstraction. We define three levels of abstraction: low, mid, and high. The low level focuses on individual objects. The mid-level relates objects and infers events. The high-level relates the events to understand the overall story. This characterization of content information in terms of granularity and abstraction allows for an efficient representation of content for access and integration purposes.

2.2. Multimedia Context

Context information corresponds to a signature, pattern, or underlying structure in the multimedia information. Context information can be used to constrain the content information, reducing the number of possible interpretations. We divide context into visual, audio, and transcript. We determine visual and audio context by looking for specific patterns. The transcript context focuses on high-level aspects of the text, such as story, motivation, and symbolism. One important element in determining context is the idea of "global pattern matching". By training the system on large amounts of content, patterns that help determine context emerge, revealing "regularity". We show in Section 4.2 how to extract and use audio patterns in the process of multimodal integration.

3. PROBABILISTIC FRAMEWORK

Next, we discuss the Multimodal Integration Engine, and Bayesian networks and hierarchical priors.

3.1. Scout's Multimodal Integration Engine

The Multimodal Integration Engine (MIE) is at the core of Video Scout's system [1,2,3]. It describes the multimodal integration of content and context information in a probabilistic form. The MIE is organized in two layers*:* the content and context layers. The input to these two layers is a TV program.

The content layer receives a visual, audio, and text stream from the decoded video. This layer contains three successive levels: the low, mid, and high as depicted in Figure 1. At the low level the visual information is given by pixel-based color, edge, and shape attributes. Audio information employs 20 different signal processing parameters such as MEL cepstral coefficients, and average energy. Transcript information is extracted from the closed-captions (CC) text. Each of these features are computed "locally" and separately from each other. The granularity is high. At the mid-level the visual features include keyframes, faces, and videotext, the audio features correspond to seven mid-level categories computed out of the 20 low-level features. These audio categories are silence, noise, speech, music, speech plus noise, speech plus speech, and speech plus music. The CC features are given by 20 different categories [1,2]. Each category has an associated set of keywords. Finally, at the high-level, TV program segments are classified according to topics.

3.2. Bayesian Network

Bayesian networks are a natural way of processing probabilistically content information. Bayesian networks [9] combine a directed acyclical graph (DAG) with probabilities. A DAG is made up of nodes and directed arcs or arrows between the nodes. In our framework each node represents a feature such as color, edge, and shape. The directed arcs represent relationships between these nodes. The nodes and the directed arcs are associated with conditional probability densities (cpd). The formal structure of Bayesian networks as applied to this work is described in [1,2]. Figure 1 shows the content layer of the MIE as a Bayesian network where each node is represented by an ellipse and each directed arrow represents a relationship between nodes.

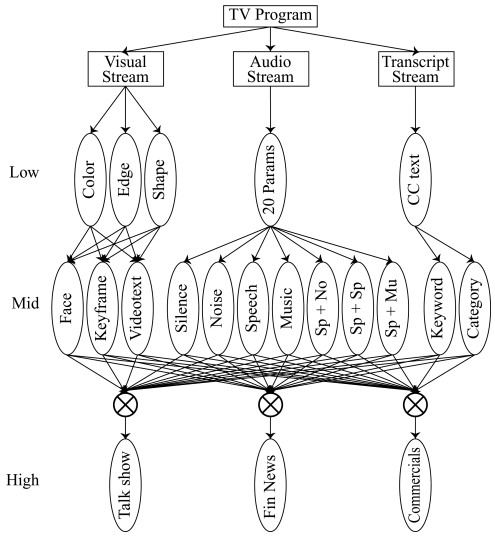

Figure 1: *MIE's content layer as a Bayesian network.*

3.3. Hierarchical Priors

Hierarchical priors [10] model context information and its relationship to content information. In the MIE this is described by the content and context layers. Each layer is made up of nodes distributed in three successive levels that are connected via directed arcs; a Bayesian network represents each layer. Hierarchical priors model the directed arcs between nodes in the two layers. Figure 2 shows a general diagram of the content and context layers.

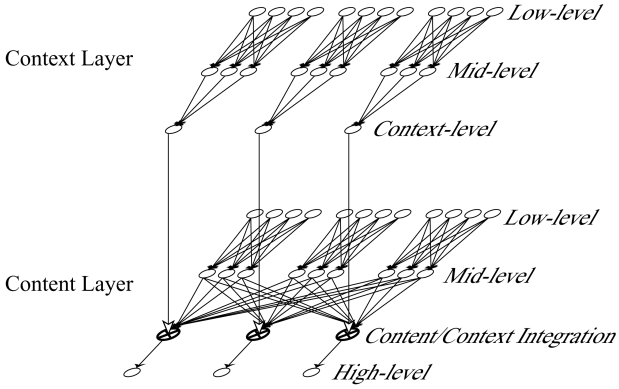

Figure 2: *Content and context Layers.*

In Figure 2, the content and context layers each contain three levels: low, mid and high-level. Each node x_i describes a given feature, such as a "face" in Figure 1. In general, the i[th] node in a given level is related to nodes in a previous level according to $P(x_i \mid \Pi_{x_i}) = P(x_i \mid x_1, \cdots, x_{i-1})$, where Π_{x_i} is the parent set of x_i and $P(\cdot \mid \cdot)$ is a conditional probability density (cpd). Out of these cpds we can build joint probability densities (jpd) as $P(x_1, \cdots, x_N) = P(x_N \mid x_{N-1}, \cdots, x_1) \times \cdots \times P(x_2 \mid x_1) \times P(x_1)$. These jpds are shown by converging arrows in Figure 2.

The vertical arrows from the context layer to the content layer, which converge to the nodes marked by a $\otimes$ symbol are described by hierarchical priors. As an example, let us assume that node x_1 = "face" and x_2 = "color"; we want to estimate x_1 given x_2 by maximizing the cpd $P(x_1 \mid x_2)$. Now, if we know that x_2 also depends on another variable, say λ_1 = "TV program is a talk show", then $P(x_1 \mid x_2)$ has to be replaced by $P(x_1 \mid \lambda_1, x_2) \times P(\lambda_1 \mid x_2)$, where the cpd $P(\lambda_1 \mid x_2)$ plays the role of a (conditional) prior on λ_1. This is formalized by the equation $P(x_1 \mid x_2) = \int d\lambda_1 P(x_1 \mid \lambda_1, x_2) \times P(\lambda_1 \mid x_2)$ called the Chapman-Kolmogorov [8] equation. The term $P(\lambda_1 \mid x_2)$ represents the prior terms of λ_1-- context layer -- and $P(x_1 \mid \lambda_1, x_2)$ is the cpd defined in the content layer. Typically, a prior, as described by Bayes' theorem [16] determines the additional information about a given variable independently of its "measurement". The product in the Chapman-Kolmogorov corresponds to the $\otimes$ symbol and it represents the integration of content and context information. This can be generalized to an arbitrary number of new λ variables by the recursive use of this equation. For example, we can write iterate the use of the Chapman-Kolmogorov equation twice and get as a result of this $P(x_1 \mid x_2) = \int d\lambda_1 \int d\lambda_2 P(x_1 \mid \lambda_1, x_2) \times P(\lambda_1 \mid \lambda_2, x_2) \times P(\lambda_2 \mid x_2)$. Using this formalism we can add new layers of information above the context layer.

4. EXPERIMENTS

The goals of the experiments were to (i) detect TV commercials, (ii) segment and index the program part into sub-segments; and (iii) classify these program sub-segments as either financial news or talk shows. In all, we extracted all features and probability values as presented on Figure 1 on six and a half hours of video. The extracted context information is also derived from the same data set.

4.1. Content-Based Segmentation

As described in 3.1 the MIE processes content information in three stages: low, mid, and high-level. At each level, and for each feature we associate a probability value. In the current implementation we use probabilities for the mid and high levels.

Video Scout performs the initial segmentation and indexing using closed-captions CC information (transcript information.) The symbol ">>" marks the change of speakers and ">>>" the change of topics. We associate the text between ">>" to a CC unit, and associate it with a given CC category. In order to index these units as financial news or talk show, the MIE performs two tasks. First, it combines the context information with the mid-level content information, producing a probability for financial news and another for talk shows, for audio, visual and transcript domains. Second, it computes a joint probability for financial news and for talk show by taking the product of each probability for each domain. This is motivated by the fact that if all individual probabilities are used then the final product will be an ultra small number (less than 10^{-30}). To avoid this small number we compute the joint probabilities within each domain separately. This also includes the combination of content and audio context domain. The end result is two joint probabilities: one for financial news and one for talk show. The final decision is based on the largest of these two probabilities.

4.2. Context Extraction: Audio Patterns

Audio context is defined as the *combination* of probabilities, one for each mid-level category. This combination forms a pattern that is used in the integration with content information. In order to extract this audio pattern we have to perform a "learning" process for TV program segments and commercials. This process consists of "combining" the probability values in time across different program segments for the same program and across different programs of the same genre. The overall elements of this process are described in [11]. The resulting audio patterns for news, talk shows, and commercials are shown in Figure 3.

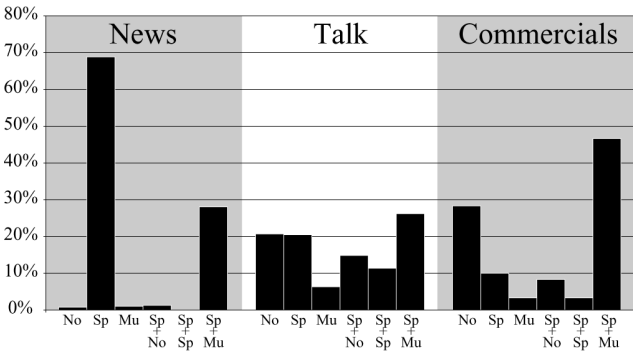

Figure 3: *Audio context patterns*

We can notice that the dominant categories for commercials are noise (sound effects), speech, and speech plus music; for news they are speech and speech plus music, and for talk shows noise, speech, and speech plus music. Noise represents all audio information that is not encountered in the other five categories such as crowd cheering, laughing, etc.

4.3. Experimental Results

The following figures display results of the experiments on 9 different TV programs. These include the talk shows hosted by David Letterman and Jay Leno, and the business programs Wall Street Journal Report (WSJR), Wall Street Weekly (WSW), Marketwatch and Moneyline (ML). These results are divided into two classes: one that uses the integration of content and context information via hierarchical priors and Bayesian networks and the other one that uses exclusively Bayesian networks. In Figures 4 and 5 we show the results with just Bayesian networks, without content and context integration. Here the classification results are obtained with a single content layer. For this case the audio joint probability was the product of all the non-zero probabilities associated with the mid-level audio categories, as prescribed by Bayesian networks. The majority of TV program segments are not correctly classified and many of them are unclassified. Figures 6 and 7 display the results of the integration of content and context. Most of the TV program segments are correctly classified. Here the audio context results are used as explained in Section 4.2. Classification results for financial segments yield 83% and for celebrity segments 89%.

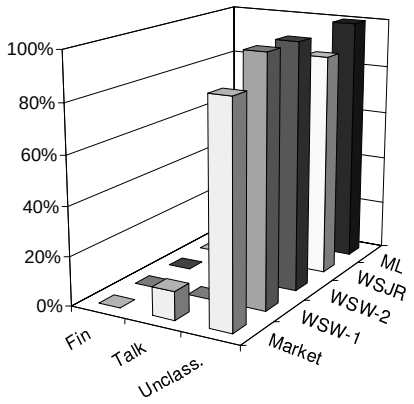

Figure 4: *Financial TV program segmentation using only content layer.*

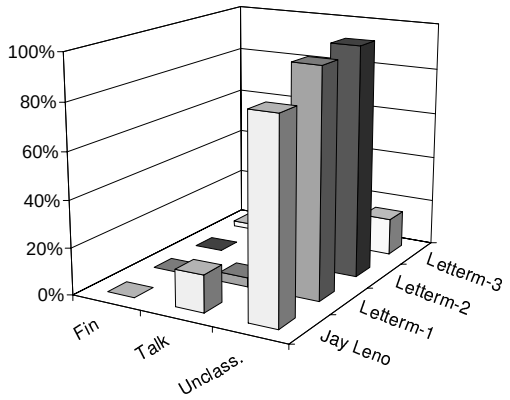

Figure 5: *Talk show TV program segmentation using only content layer.*

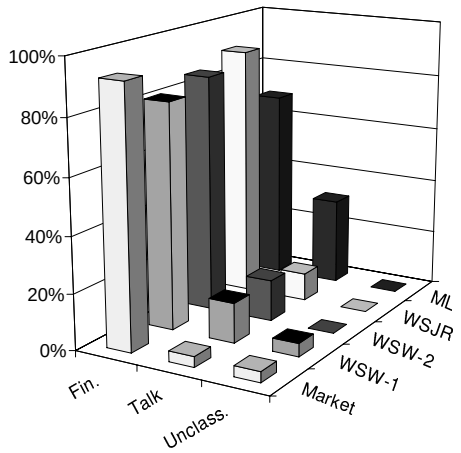

Figure 6: *Financial TV program segment classification based on the integration of content and context information.*

Clearly, without context information the classification results are poor. Moreover, these results point to the importance of audio for multimodal integration.

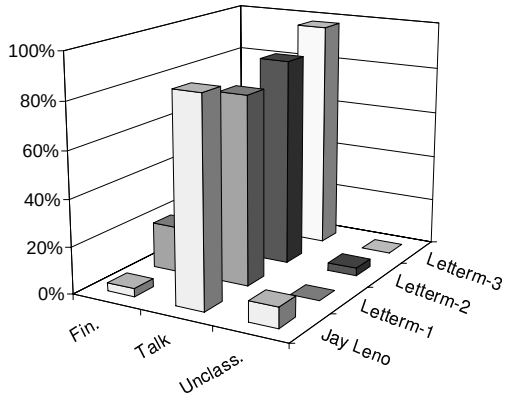

Figure 7: *Talk show TV program segment classification with the integration of content and context information.*

5. CONCLUSION

We introduced a layered probabilistic representation for processing video information that includes a content layer and a context layer. Within each layer the representation is based on Bayesian networks. Hierarchical priors provide the connection between the two layers. The novelty of our framework is in the information layering and the connection via hierarchical priors. The strength of this framework is demonstrated in an end-to-end system called Video Scout that selects, indexes, and stores TV program segments based on topic classification. We show that, by running Video Scout on real video of TV programs, the classification task is highly improved by combining context and content information.

6. References.

[1] R. S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, and J. Zimmerman, "*Video Scouting: An Application and System for the Integration of Multimedia Information in Personal TV Applications*", Proc. IEEE ICASSP 2001, Salt Lake City, Utah, May 2001.

[2] R. S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, and D. Li, "*Integrated Multimedia Processing for Topic Segmentation and Classification*", Proc. of IEEE ICIP, Oct. 2001.

[3] N. Dimitrova, R. S. Jasinschi, L. Agnihotri, J. Zimmerman, T. McGee, and D. Li, "*Personalizing Video Recorders using Video Processing and Integration*", Proc. ACM Multimedia, pp. 564-567, Oct. 2001.

[4] *www.tivo.com*.

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

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

[7] T. Chen and R. R. Rao, "*Audio-Visual Integration in Multimodal Communication*", Proc. of IEEE, Vol. 86, 5, 837-852, 1998.

[8] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, Mc-Graw Hill, New York, 1984.

[9] J. Pearl, "*Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*", Morgan Kaufmann, San Mateo, CA, 1988.

[10] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.

[11] R. S. Jasinschi and J. Louie, "*Automatic TV Program Genre Classification Based on Audio Patterns*", 370-375, Proc. of the 27th EUROMICRO Conference, Warsaw, Poland, 4-6 September, 2001.