STRUCTURE LEARNING IN A BAYESIAN NETWORK-BASED VIDEO INDEXING FRAMEWORK

Siwar Baghdadi
Thomson R&D France
1 av Belle Fontaine-CS 17616
35576 Cesson-Sévigné. France.
siwar.baghdadi@thomson.net

Guillaume Gravier
IRISA/CNRS
Campus de Beaulieu
35042 Rennes Cedex. France.
guillaume.gravier@irisa.fr

Claire-Hélène Demarty
Thomson R&D France
1 av Belle Fontaine-CS 17616
35576 Cesson-Sévigné. France.
claire-helene.demarty@thomson.net

Patrick Gros
IRISA/INRIA
Campus de Beaulieu
35042 Rennes Cedex. France.
patrick.gros@irisa.fr

ABSTRACT

Several stochastic models provide an effective framework to identify the temporal structure of audiovisual data. Most of them need as input a first video structure, *i.e.* connections between features and video events. Provided that this structure is given as input, the parameters are then estimated from training data. Bayesian networks offer an additional feature, namely structure learning, which allows the automatic construction of the model structure from training data. Structure learning obviously leads to an increased generality of the model building process. This paper investigates the trade-off between the increase of generality and the quality of the results in video analysis. We model video data using dynamic Bayesian networks (DBNs) where the static part of the network accounts for the correlations between low-level features extracted from the raw data and between these features and the events considered. It is precisely this part of the network whose structure is automatically constructed from training data. Experimental results on a commercial detection case study application show that, even though the model structure is determined in a non supervised manner, the resulting model is effective for the detection of commercial segments in video data.

1. INTRODUCTION

A huge amount of video is produced everyday. For efficient access and retrieval of this digital data, tools that can automatically understand the semantic content of a video are becoming compulsory.

Recently, rule-based approaches have been widely used. Announcer's excited speech or sequence replay have been used to detect important events in sport videos [1]. Monochrome frames and silence have been used to detect commercials in videos [2]. Statistical approaches have also been used for video indexing. Among them, Hidden Markov Models (HMMs)-based systems have been widely used for their good properties and modeling capabilities. In [3], the authors used HMMs in order to do video structuring in tennis video. DBNs, which are a generalization of HMMs, have also been used recently. Many research works [4, 5, 6] have shown that DBNs are a powerful tool for semantic video analysis. Compared to HMMs, DBNs allow to use a set of random variables instead of only one hidden state node at each time instant. The work presented in [6] has also shown that DBNs are an effective fusion tool. In this work, they allow fusion of different modalities (visual and audio) in order to extract highlights from Formula 1 race videos.

However, designing a rule-based or a statistical-based video parser is highly domain-dependent. It requires a highly knowledgeable expert in order to manually choose relevant features and define (or ignore) the relations between them. In addition to the fact that this is a manual and hence costly process, it also restricts significantly the applications where rule-based or statistical based video parsers can be used.

To overcome this problem, and in addition to classical parameters learning, we propose in this paper to use structure learning in order to automatically construct the static part of a DBN. This automatic procedure aims at reflecting the different kinds of interactions that exist between the model‘s variables. Our approach is presented and evaluated on a commercial detection application.

The paper is organized as follows. The global DBN based scheme is introduced in Section 2. The proposed approach for DBN structure learning is presented in Section 3. The commercial case study is then presented in Section 4. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

2. DYNAMIC BAYESIAN NETWORK MODEL

Bayesian networks are the result of a marriage between probability theory and graph theory. A Bayesian network is a directed acyclic graph (DAG) whose nodes represent random variables and whose edges represent statistical dependence relations among the variables. In a Bayesian network, an edge from node A to B (A is a parent of B) can be informally interpreted as indicating that A *causes* B. A conditional probability distribution (CPD) is associated to each variable. These

CPDs are the parameters of the network.

Bayesian networks also allow encoding time evolution by way of their dynamic version. In DBNs, the time flow is discretized, and a static Bayesian network is assigned to each temporal slice. Variables of different time-slices are connected through directed edges, which explicitly represent the time flow in terms of conditional dependence.

We choose to use DBNs in order to construct a video analysis system. The model takes into account the characteristics of the video: local characteristics (audio visual features and their relations with the events) and global temporal characteristics (event's duration and time correlation between events).

The static part of the DBN allows to deal with the local characteristics of the video. In fact, the nodes that compose the static part correspond to the extracted audio-visual features and the events of interest. The structure of the static part of the DBN reflects the local interactions that exist firstly between the features and secondly between the features and the events. The way of constructing the static part of the model will be described in Section 3.

DBNs handle the temporal correlations that exist in audiovisual signals through the connections that exist between nodes of different time slices. In order to integrate the static structure into a DBN, we assume that the features of different time slices are conditionally independent given the event variables, that is, features from different time slices are not connected. Hence the information from different time slices goes around through the event nodes, which are connected from one slice to another.

3. BUILDING STATIC BAYESIAN NETWORK

Building a Bayesian network can be divided in two steps: construction of the structure with the determination of the topology of the network and estimation of the parameters with the computation of the conditional probabilities using statistical methods. A special focus is put in this section on the structure construction procedure. Structure encodes the dependencies between the system variables. In all the frameworks for video indexing based on Bayesian networks that can be found in the literature, the structure of the model is designed manually. It is done using expert knowledges on the different interactions between the variables used. Unfortunately, this knowledge is not always available. Moreover it is expensive and tedious to get, in particular for complex problems of semantic video analysis. Bayesian networks provide interesting opportunities for automatically learning the model's structure from the training data,which has already been used with success in the medical domain [7].

Structure learning has a first advantage to help the system designer by constructing automatically a model constructed from the training data. It also allows the system designer to get a better understanding of the system under study, as it produces an easily readable network. On the network learned, the existence of an edge between two variables points out a causal relation between them. Moreover, thanks to structure learning, the constructed video indexing scheme has the power to discard non relevant features in the system, as it is shown in Section 4. Hence, a set of all available features can be used during the structure learning process. Only relevant features for event detection will be kept in the constructed structure, and non relevant features will be disconnected.

Different solutions are proposed for structure learning of Bayesian networks. The most widely used is the K2 algorithm [8]. This algorithm is a typical search and scoring method. It starts by assuming that a node lacks parents, after which, in every step, it adds incrementally the parent whose addition most increases the probability of the resulting structure given the data. The K2 algorithm stops adding parents to the nodes when the addition of a single parent cannot increase the probability. However the K2 algorithm has the drawback to be very sensitive to initialization. In order to bypass this disadvantage, we use a tree generated by the Maximum Weight Spanning Tree algorithm [9] as initialization for the K2 algorithm.

4. COMMERCIAL DETECTION: A CASE STUDY

We have chosen the commercial detection application as a case study for our structure learning scheme. Motivations behind this choice are twofold. First, this application has been extensively studied and is very important to many services like set-top-box applications. Second, the evaluation of the results obtained is straightforward. This is not always the case as some semantic video analysis applications require manual and subjective evaluation.

In this particular case study, low-level features have been selected based on the observation that commercial segments tend to be more appealing than program segments. Advertising producers use motion, color, text and audio to emphasize the appealing character of their commercials spots.

We use temporal properties of color histogram to detect cuts and segment videos into shots. For each shot, we compute a set of multi-modal features which are described in the following section.

4.1. Feature extraction

1. *Action features*: Fast-moving objects, frequent hard cuts and many zooms result in an impression of action [2]. Therefore, we use motion intensity and shot length as features describing the amount of action in the video.
2. *Color coherence*: The coherence of color between commercial shots is low. A color histogram is computed for each shot. A Bhattacharya distance is then used to estimate the color similarity of the histograms for consecutive shots.
3. *Text caption*: The excessive use of overlay text during the commercials is a way of both capturing the user attention and conveying the information. Text captions are tracked in each shot and we use their number and surface as features.

4. *Low short time energy ratio (LSTER)*: This feature has been proposed in [10] in order to perform audio segmentation. It represents the variation of short-time energy (*STE*). The *LSTER* is defined as the ratio of the number of frames whose *STE* are less then 0.5 time the average *STE* in a w_s window. *LSTER* is a good discriminator between speech and music.
5. *Silence/Monochrome frames*: When available, silence shots and monochrome frames are considered as strong indicators of commercials. We will not always use this feature in the experiments of Section 5, but when used, it will consist in binary shot taging: 1 for shots containing monochrome frames and with no audio activity and 0 otherwise.

4.2. Model construction

As explained in Section 3 and unlike the existing Bayesian-based frameworks for video indexing, the structure of the static part of the model is automatically generated through the structure learning process provided by Bayesian networks. Therefore, before learning the CDPs parameters, we use the K2 algorithm [8] in order to find the optimal structure of the static Bayesian network in relation with the training data.

For the application of commercial detection, the structure obtained is depicted in Figure1. It shows the different kinds of interaction that exist between features. Note: we did not include the Silence/Monochrome frames feature in this first test.

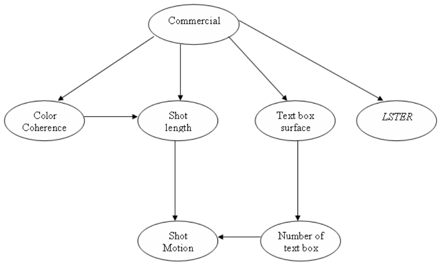

Fig. 1. Static structure of the DBN model

We find that the commercial node is directly related to the color coherence, shot length, *LSTER* and text nodes. We can consider that these variables are the most characteristics features that distinguish commercial from non commercial segments. In this model, a dependence edge also appears between the shot length variable and shot motion node. The existence of this edge can be explained by the fact that within high level action scenes, the shot detector we use, generates short-length over-segmented shots. Therefore shot motion and shot length are highly correlated variables. The structure obtained is consistent with our knowledge. Structure learning provides user designer with a tool for a better understanding of the studied domain.

In order to test the ability of the structure learning procedure to discard non-pertinent variables, we have introduced a variable that represents the existence of green color in the shot. Intuitively, the green color variable is not relevant to the problem of commercial detection since it is equally present in the video. In the Bayesian network resulting from the structure learning step, the green color variable has not been related to any other variable. This test reenforces the fact that the system designer should be able to introduce all the available features, without taking care of whether these variables are relevant or not, since the constructed model will only incorporate pertinent features to the problem. Structure learning is therefore a tool for selecting relevant features for a particular problem.

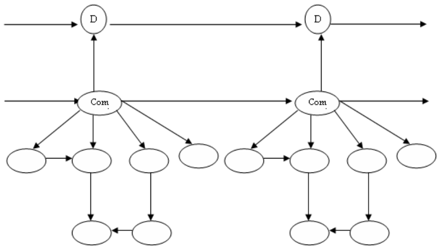

Fig. 2. DBN Model for Commercial Detection

Once the structure of the static Bayesian network is obtained and its parameters are estimated, the static network is used in the global DBN. Using this structure is possible since we suppose that features from different time slices are conditionally independent given the event nodes. In case of commercial detection, we use the global DBN illustrated in Figure 2. Usually durations of commercial segments are not randomly set. In order to use the knowledge about duration of commercial segments, we have introduced the variable 'D'. At the beginning of a commercial segment this variable takes the predicted time length of the commercial segment. Then it indicates the remaining time steps before the end of the commercial segment.

5. EXPERIMENTAL RESULTS

We have evaluated our approach on a data set of 30 hours of video taken from 3 different general interest French TV channels. Program types include: news, series, movies, entertainment and some integrated programs (weather forecasts, short programs). 20 hours of the data set are used for training and 10 hours for testing. We have used recall, precision and F-measure metrics in order to evaluate the performance of our system. These metrics are calculated at the shot level.

In order to illustrate the benefit of using structure learning to handle the interaction between low level features, we have compared our automatically constructed Bayesian network with a multi-stream HMM coupled with a duration model where only the parameters are learned automatically from the data. This last model uses the same features as our constructed model. At this step, we are not using the silence and monochrome frames feature. Results are illustrated in first part of Table 1.

Table 1. Evaluation of the contribution of structure learning in building a Bayesian network for commercial detection. (a) without the silence/monochrome feature, (b) with the silence/monochrome feature.

	Recall (%)	Precision (%)	F-mesure (%)
Manually Constructed Model	88	76	81
Automatically Constructed Model (a)	93	80	86
Automatically Constructed Model (b)	93	90	91

We can notice an improvement of the detection results when using structure learning. In fact unlike a standard HMM model, where no knowledge about different features interactions can be learned from training data, our model incorporates automatically this kind of knowledge. It takes advantage of the different kinds of interactions between low level features through the learned structure of the static model. However as in HMM, our model includes the global knowledge of the problem by simulating the duration of commercial segments. Some false alarms persist however. After examining these false alarms, we have found that they are very similar to commercials in terms of visual content, motion activity and duration. One example of these false alarms is the opening credit of the serie ”ER” which is designed to be attractive, with text caption, high shot rate and motion, as in commercials segments.

In the second part of our experiments, we have followed the same protocol as in the first part and we have additionally included the silence/monochrome frames feature. The results obtained are comparable to existing commercial detection system [11, 2]. The precision has been improved to 90%, that is, an F-measure of 91%. This improvement is mainly due to the fact that the silence and monochrome frames are known to be hardly occurring in non commercial segment. In the constructed model, the silence/monochrome frames node is directly related to the commercial node, which indicates that this feature has a direct effect on the commercial variable.

6. CONCLUSION

Our first experimental tests have proved that structure learning in DBNs may be used to automate the task of defining a statistical model of the video: we did bypass the need for expert knowledge, while still accurately representing the domain of interest, in our application of commercial detection. The experiments also showed that in addition to being easy to construct, the model has similar performance with existing commercial detection techniques. We may conclude that the DBNs' specific feature of structure learning is promising. It needs to be further validated in more complex applications, with a richer database of features and events. This is what we aim at for our future work. Another perspective will be to extend structure learning to the dynamic part of the network and see if it still correctly models the video.

7. REFERENCES

[1] Y. Rui, A. Gupta, and A. Acero, “Automatically extracting highlights for TV baseball programs,” in *Proc. of ACM*, 2000, pp. 105–115.

[2] P. Lienhart, C. Kuhmunch, and W. Effelsberg, “On the detection and recognition of television commercials,” in *Proc. IEEE Conference on Multimedia Computing and Systems*, 1997, pp. 509–516.

[3] E. Kijak, G. Gravier, L. Oisel, and P. Gros, “Audiovisual integration for sport broadcast structuring,” *Multimedia Tools and Applications*, vol. 30, pp. 289–312, 2006.

[4] N. Oliver and E. Horvitz, “A comparison of HMMs and dynamic Bayesian networks for recognizing office activities,” in *Proc. of International Conference on User Modeling*, 2005, pp. 199–209.

[5] F. Wang, Y. Ma, H. Zhang, and J. Li, “A generic framework for semantic sports video analysis using dynamic bayesian networks,” in *Proc. International Multimedia Modelling Conference*, 2005, pp. 115–122.

[6] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, “Multi-modal extraction of highlights from TV formula 1 programs,” in *Proc. IEEE International Conference on Multimedia and Expo*, 2002, pp. 817–820.

[7] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, “Using bayesian network to analyze expression data,” *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[8] G. F. Cooper and E. Herskovits, “A bayesian method for the induction of probabilistic networks from data,” *Machine learning*, vol. 9, no. 2, pp. 309–347, 1992.

[9] C. Chow and C. Liu, “Approximating discrete probability distributions with dependence trees,” *IEEE Transaction on Information Theory*, vol. 11, no. 3, pp. 462–467, 1968.

[10] L. Lu, H. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[11] L. Lu, H. Zhang, and H. Jiang, “Robust learning-based TV commercial detection,” in *IEEE International Conference on Multimedia and Expo*, 2005, pp. 4–7.