MULTI-MODAL EXTRACTION OF HIGHLIGHTS FROM TV FORMULA 1 PROGRAMS

Milan Petkovic[1], Vojkan Mihajlovic[2], Willem Jonker[1], S. Djordjevic-Kajan[2]

[1] Computer Science Department, University of Twente
PO BOX 217, 7500 AE Enschede, The Netherlands
{milan, jonker}@cs.utwente.nl

[2] Computer Science Department, University of Nis
Beogradska 14, 18000 Nis, Yugoslavia
{vojkan.m, sdjordjevic}@elfak.ni.ac.yu

ABSTRACT

As amounts of publicly available video data grow, the need to automatically infer semantics from raw video data becomes significant. In this paper, we focus on the use of Dynamic Bayesian Networks (DBNs) for that purpose, and demonstrate how they can be effectively applied for fusing the evidence obtained from different media information sources. The approach is validated in the particular domain of Formula 1 race videos. For that specific domain we introduce a robust audio-visual feature extraction scheme and a text recognition and detection method. Based on numerous experiments performed with DBNs, we give some recommendations with respect to the modeling of temporal and atemporal dependences within the network. Finally, we present the experimental results for the detection of excited speech and the extraction of highlights, as well as the advantageous query capabilities of our system.

1. INTRODUCTION

Numerous approaches presented in literature have shown that is now becoming possible to extract high-level semantic events from video. However, the majority of approaches including our previous work [1] uses the individual visual or audio cues, and is error-prone suffering from robustness problems due to detection errors. Fusing the evidence obtained from different sources should result in more robust and accurate systems. Furthermore, some events are naturally multi-modal demanding the gathering of evidence from different media sources.

On the other hand, the fusion of the multi-modal cues is quite challenging, since it has to deal with indications obtained from different media information sources, which might contradict each other. Only a few attempts to multi-modal analysis of audio-visual information have appeared recently, such as a probabilistic model for event detection in a classroom lecture environment [2], and a Bayesian approach for topic segmentation and classification in TV programs [3].

In this paper, we contribute by demonstrating how dynamic Bayesian networks can be effectively used for content-based video retrieval by fusing the evidence obtained from different media information sources. We validate our approach in the particular domain of Formula 1 race videos. For that specific domain we introduce a robust audio-visual feature extraction scheme and a text recognition and detection method. Based on numerous experiments performed for fusing extracted features in order to extract highlights, we give some recommendations with respect to the modeling of temporal and atemporal dependences in DBNs.

2. INFORMATION SOURCES

In this section, we briefly describe the extraction of multi-modal cues obtained from three different media components of the TV broadcasting Formula 1 program. In particular, we concentrate on audio, video, and text.

Audio plays a significant role in the detection and recognition of events in video. In our domain, the importance of the audio signal is even bigger, since it encapsulates the reporter's comment, which can be considered as a kind of the on-line human annotation. Furthermore, whenever something important happens the announcer raises his voice due to his excitement, which is a good indication for the highlights.

Based on a few experiments we select four audio features to be used for speech endpoint detection and extraction of excited speech. We chose Short Time Energy (STE), pitch, Mel-Frequency Cepstral Coefficients (MFCCs), and pause rate. A description of methods we developed for excited speech and speech endpoint detection can be found in [4]. For the recognition of specific keywords in announcer's speech we used a keyword-spotting tool based on a finite state grammar.

In our visual analysis, we use color, shape and motion features. First, the video is segmented into shots based on the differences of color histograms among several consecutive frames. Then, we calculate the amount of motion and apply semaphore, dust, sand, and replay detectors in order to characterize passing, start, and fly-out events, as well as to find replay scenes (for a description of these detectors see [4]).

The third information source we use is the text that is superimposed on the screen. This is another type of on-line annotation done by the TV program producer, which is intended to help viewers to better understand the video content. The superimposed text often brings some additional information that is difficult or even impossible to deduce solely by looking at the video signal. In order to speed up the detection and recognition of the superimposed text we modified the existing technique considering the properties of Formula 1 race videos [4].

3. PROBABILISTIC FUSION

As the majority of techniques for event detection, which relay solely on the one-media cues, showed to have robustness problems, we decided to base our analysis on the fusion of the evidence obtained from the aforementioned information sources. In order to find the most appropriate technique, we performed

0-7803-7304-9/02/$17.00 C2002 IEEE

numerous experiments and compare Bayesian Networks (BNs) versus Dynamic Bayesian Networks (DBNs), different network structures, temporal dependences, and learning algorithms.

A dynamic Bayesian network is a probabilistic network which is able to model stochastic temporal processes. It is a special case of singly connected Bayesian networks specifically aimed at time series modeling. A time-slice is used to represent each snapshot of the evolving temporal process. A DBN satisfies the first order Markov property. So, each state at time t may depend on one or more states at time $t-1$ and/or some states in the same time instant. The conditional probabilities between time-slices define the state evaluation model.

The parameters of a DBN can be learned from a training data set. As we work with DBNs that have hidden states, for this purpose we employ the Expectation Maximization learning algorithm. In the inferencing process, we use the modified Boyen-Koller algorithm for approximate inference [5]. For a detail description of both algorithms see [4].

3.1. Data set and audio-visual features

We digitized three Formula 1 races of the 2001 season, namely, the German, Belgian, and USA Grand Prix (GP). The average duration of these Formula 1 races was about 90 minutes or 135,000 frames for a PAL video. Videos were digitized as a quarter of the PAL standard resolution (384x288). Audio was sampled at 22kHz with 16 bits per audio sample.

Feature values, extracted from the audio and video signals, are represented as probabilistic values in range from zero to one. Since the parameters are calculated for each 0.1s, the length of feature vectors is ten times longer than the duration of the video measured in seconds. The features we extracted form a Formula 1 video are: keywords (f_1), pause rate (f_2), average values of STE (f_3), dynamic range of STE (f_4), maximum values of STE (f_5), average values of pitch (f_6), dynamic range of pitch (f_7), maximum values of pitch (f_8), average values of MFCCs (f_9), maximum values of MFCCs (f_{10}), part of the race (f_{11}), replay (f_{12}), color difference (f_{13}), semaphore (f_{14}), dust (f_{15}), sand (f_{16}), and motion (f_{17}). Since we also employed text detection and recognition algorithms, we were also able to extract text from the video. We decide to extract the names of Formula 1 drivers, and the semantic content of superimposed text (for example if it is a pit stop, or driver's classification is shown, etc.).

3.2. Excited speech

We decided to start our experiments by comparing the results that can be achieved by employing BNs versus DBNs for processing only audio cues to determine exited speech. We developed three different structures of BNs and corresponding DBN structures. The intention was to explore how different network structures can influence the inference step in this type of networks. The structures of BNs, which are also used for one time slice of DBNs, are depicted in Figure 1.

The query node is Excited Announcer (EA), since we want to determine if the announcer raises his voice due to an interesting event that is taking place in the race. The shaded nodes represent evidence nodes, which receive their values based on features extracted from the audio signal of the Formula 1 video.

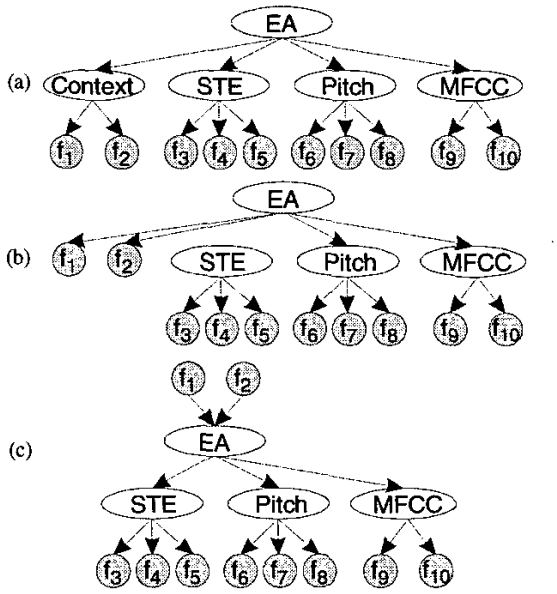


Figure 1. Different structures for processing of audio features: a) Fully parameterized structure; b) Structure with direct influence from evidence to query node; c) Input/output BN structure

The temporal dependencies between nodes from two consecutive time slices of DBNs were defined as in Figure 2. For learning and inference algorithms we considered all nodes from one time slice as belonging to the same cluster ("exact" inference end learning).

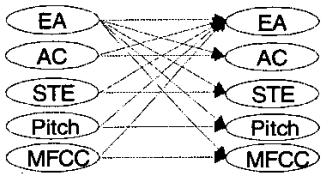
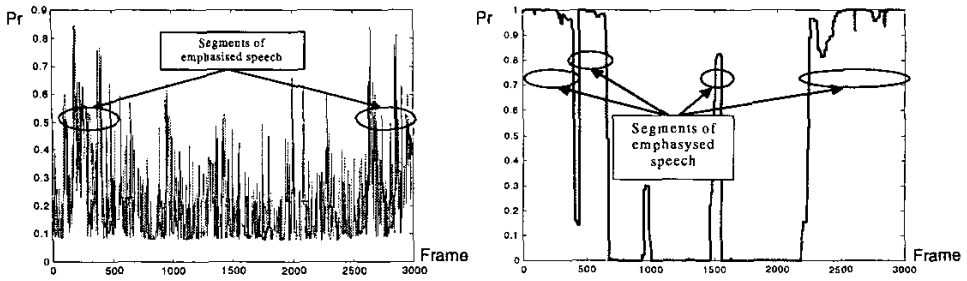


Figure 2. Temporal dependencies for the DBNs

We learned the BN parameters on a sequence of 300s, consisting of 3000 evidence values, extracted from the audio signal. For the DBNs, we used the same video sequence of 300s, which was divided into 12 segments with 25s duration each. The inference was performed on audio evidence extracted from the digitized German GP. For each network structure we computed precision and recall.

Note that we had to process the results obtained from BNs since the output values cannot be directly employed to distinguish the presence and time boundaries of the excited speech. This is shown in Figure 3a. Therefore, we accumulated values of a query node over time to make a conclusion whether the announcer is excited.

The results obtained from a dynamic Bayesian network were much smoother (see Figure 3b), and we did not have to process the output. The results from conducted experiments with previously described networks are shown in Table 1.



(a) (b)

Figure 3. Results of audio BN (a) and DBN (b) inference for 300s long "avi" file

Table 1. BNs and a DBN for detection of excited speech

Network structure	BN (Fig. 1a)	BN (Fig. 1b)	BN (Fig. 1c)	DBN (Fig. 1a, Fig. 2)
Precision	60 %	54 %	50 %	85 %
Recall	66 %	61 %	76 %	81 %

By comparing different BN structures we can see that there is no significant difference in precision and recall obtained from them (Table 1). The corresponding DBNs perform similarly, except for the DBN that corresponds to the BN with fully parameterized structure (Fig. 1a). It gives much better results than the other BN/DBN networks (last column in Table 1).

Next, we explored the influence that different temporal dependencies have on learning and inference procedures in DBNs. We developed three DBNs with the same structure of one time slice (Fig. 1a) but different temporal dependencies between two consecutive time slices: (1) the structure with emission query node (Fig. 4a), (2) one with collecting query node (Fig. 4b), and (3) one with dependencies as in Fig. 2. The evaluation showed that the last one significantly outperforms the first and slightly the second structure.

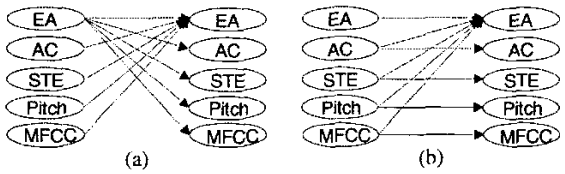


Figure 4. Temporal dependencies: emission (a), collecting (b)

Conclusions from experiments performed are twofold. From the first group of experiments we conclude that the DBN learning and inference procedures depend a lot on the selected DBN structure for one time slice. We can see that this is not the case when inference and learning are performed with BNs. These experiments also showed the advantages of the fully parameterized DBN structure over the other BN/DBN networks. Secondly, we conclude that chosen temporal dependencies between nodes of two consecutive time slices have strong influence on the results of DBN inference. The best result was obtained with temporal dependencies depicted in Fig. 2.

Finally, we selected the fully parameterized DBN structure, with one cluster for nodes in the same time slice, as the most powerful DBN structure for detection of the emphasized announcer speech. To evaluate the chosen network structure we employed it for detecting the emphasized speech in the audio signal of all three races (Table 2).

Table 2. Evaluation results for the audio DBN

Race	German GP	Belgian GP	USA GP
Precision	85 %	77 %	76 %
Recall	81 %	79 %	81 %

3.3. Highlight extraction

The audio DBN can only extract the segments of the Formula 1 race where the announcer raises his voice. Other interesting segments (highlights), which were missed by the announcer, could not be extracted. Therefore, the employment of the audio DBN for highlight extraction would lead to high precision, but low recall (if we count replay scenes, recall is about 50%).

To improve the results obtained solely from audio cues we developed an audio-visual DBN for highlight detection. The structure that represents one time slice of this network is depicted in Figure 5. The Highlight node was chosen to be the main query node, while we also queried nodes: Start, Fly Out, and Passing, in our experiments. We used the same kind of temporal dependencies as for the audio network.

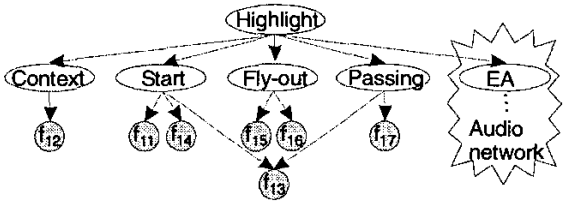


Figure 5. Audio-visual DBN for one time slice

We employed the learning algorithm on 6 sequences with 50s duration each. The results are shown in Table 3. Based on the value of the main query node (highlight), the values of the other query nodes are calculated. We calculated the most probable candidates during each "highlight" segment, and pronounce it as a start, fly out, or passing based on values of corresponding nodes. For segments longer than 15s we performed this operation every 5s to enable multiple selections.

The supplemental query nodes are incorporated in the scheme in order to classify different interesting events that take place in the Formula 1 race. We can see from Table 3 that for the German GP we gained high accuracy for highlights and start, while the most misclassifications were for fly out and passing events. Main reason for this is that we used very general and less powerful video cues for fly out, and especially passing.

Table 3. Evaluation results for audio-visual DBN

Audio/Video DBN		German	Belgian	USA[1]
Highlights	Precision	84 %	43 %	73 %
	Recall	86 %	53 %	76 %
Start	Precision	83 %	100 %	100 %
	Recall	100 %	67 %	50 %
Fly Out	Precision	64 %	100 %	0[2] %
	Recall	78 %	36 %	0 %
Passing	Precision	79 %	28 %	
	Recall	50 %	31 %	

For the Belgium and the USA GP we had a big decrement in our results, mostly because of the "passing" part of the network. Therefore, we simplified the overall audio-visual network, and excluded the "passing" sub-network. A significant difference in results obtained with (Belgian) and without the passing sub-network (USA) is presented in Table 3. The network with the passing sub-network worked fine in the case of the German GP, but failed with the other two races. The explanation for this is a different camera work in the German GP. This just confirms the fact that general low-level visual features might yield very poor results in the context of high-level concepts (to characterize passing we used motion). Obviously, more domain dependent features, which characterize the trajectories of Formula 1 cars, would be much robust and give a better result for the passing event, which is a direction for our future work.

4. CONTENT-BASED RETRIEVAL

Besides the excited speech, highlights, and the three events modeled by the DBN, our system can be used to query the Formula 1 videos based on recognized superimposed text, as well as based on audio-visual features directly. Results obtained from text recognition algorithm enable user to ask for the race winner, the classification in the *i*th lap, the position of a driver in the *i*th lap, relative positions of two drivers in the *i*th lap, the final lap, etc. To give an impression of the system capabilities, we list some query examples: (1) "Retrieve the video sequences with Michael Schumacher leading the race"; (2) "Retrieve the video sequences where Michael Schumacher is first, and Mika Hakkinen is second"; (3) "Retrieve the video sequences showing Barrichello in the pit stop"; (4) "Retrieve the sequences with the race leader crossing the finish line", etc.

Furthermore, our system benefits of combining the results obtained from Bayesian fusion and text recognition, and is capable to answer very detailed complex queries, such as: (1) "Retrieve all highlights showing the car of Michael Schumacher", (2) "Retrieve all fly outs of Mika Hakkinen in this season", (3) Retrieve all highlights at the pit line involving Juan Pablo Montoya, etc.

[1] These results are obtained by the audio-visual DBN that excludes the passing sub-network

[2] There were no fly-outs in the USA Grand Prix

5. CONCLUSIONS

This paper focuses on DBNs, and their use for content-based retrieval. We have conducted numerous experiments with different DBN and BN structures and demonstrated the advantage of DBNs over BNs for our application. Next, the influence of different atemporal and temporal connections within a DBN network has been explored. The chosen atemporal, but also temporal dependencies between nodes of two consecutive time slices, have strong influence on the results of DBN inference. The best result was obtained with the fully parameterized DBN structure and the direct temporal dependencies depicted in Fig. 1a and Fig. 2, respectively.

The approach has been validated for the extraction of highlights in the particular domain of the Formula 1 TV program. We have based our analysis on the fusion of the evidence obtained from different information sources (audio, video, and text). Consequently, a robust feature and text extraction schemes have been introduced for the audio-visual analysis of our particular domain.

The fusion of cues from the three different media has resulted in much better characterization of Formula 1 races. The audio DBN was able to detect a great number of segments where announcer raised his voice (recall 81%), which correspond to only 50% of all interesting segments, i.e. highlights in the race. The integrated audio-visual DBN was able to correct the result and detect about 80% of all interesting segments in the race. However, the audio part is still useful for the detection of the segments with the excited announcer speech, where it showed high recognition accuracy. By integrating the superimposed text, audio and video subsystems we have built a powerful tool for indexing the Formula 1 races videos, which can answer very detailed and specific queries.

6. REFERENCES

[1] M. Petkovic, W. Jonker, "Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events," IEEE Intl. Workshop on Detection and Recognition of Events in Video, Vancouver, Canada, 2001, pp. 75-82.

[2] T. Syeda-Mahmood, S. Srinivasan, "Detecting Topical Events in Digital Video", In *Proc. of ACM Multimedia*, Los Angeles, CA, 2000, pp. 85-94.

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

[4] V. Mihajlovic, M. Petkovic, "Automatic Annotation of Formula 1 Races for Content-based Video Retrieval", Technical Report, *TR-CTIT-01-41*, 2001.

[5] X. Boyen, D. Koller, "Tractable Inference for Complex Stochastic Processes," Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, 1998.