MTM at MediaEval 2014 Violence Detection Task

Bruno do Nascimento Teixeira
Universidade Fedederal de Minas Gerais
Belo Horizonte, Brazil
bruno.texeira@dcc.ufmg.br

ABSTRACT

This paper describes the team MTM participation in Violent Scenes Detection (VSD) task of the MediaEval 2014 campaign. We propose an approach to the problem of detecting violence, which is based on probabilistic graphical models using Mel-frequency cepstral coefficients (MFCCs) as audio feature. In our approach, we employ Dynamic Bayesian Networks (DBNs) to represent a violent scene as an dynamic system.

1. INTRODUCTION

The goal of the Violent Scenes Detection (VSD) task of the MediaEval 2014 benchmarking campaign is to detect violence in movies [5]. This year the organizers of the VSD task released two datasets: (i) a set of 31 Hollywood movies, where 24 are used for training and 7 for the testing (our focus); (ii) Youtube set, composed of 86 violent and non-violent videos. Violence is defined as *"one would not let an 8 years old child see in a movie because it contains physical violence"*. A model based on the variable-duration hidden Markov model is proposed to detect complex events using latent variables in Internet videos [6]. The authors of [1] propose an audio-visual approach to video genre classification using content descriptors that exploit audio, color, temporal, and contour information and demonstrated good results over other existing approaches by using a combination of these descriptors in genre classification. In [2], temporal structure of broadcast tennis video is recovered from HMMs. This trained HMM is used to analyze the temporal interleaving shots.

We propose to model video based on temporal structure and principle of causality using Dynamic Bayesian Networks (DBN).

2. METHOD

For this year's benchmark, we have developed an acoustic system based on temporal data (MFCC vector). The main idea behind this approach is to represent a violent scene as a dynamic system.

2.1 Dynamic Bayesian Network

A DBN (see Figure 1) is a state-space model of random variable V_t [3]:

$$V_t = (U_t, X_t, Y_t), \tag{1}$$

where U_t represents the hidden, X_t the input and Y_t the output variable. A pair (B_1, B_2) defines a DBN, where B_1 and B_2 are BNs. The two-slice temporal Bayes net B_2 (DBN unrolled for 2 slices), defines $P(V_t|V_{t-1})$:

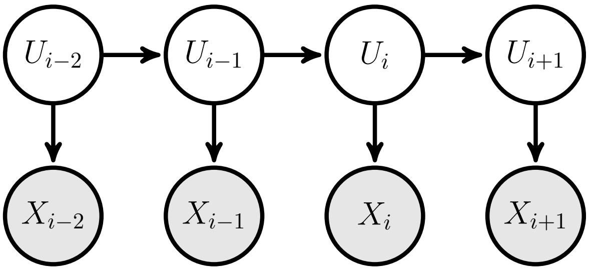

Figure 1: A graphical-model view of an DBN unrolled for 4 slices with hidden state sequence U and a observed node X.

$$P(V_t|V_{t-1}) = \prod_{i=1}^{N} P(V_i^t|Pa(V_i^t)), \tag{2}$$

where $Pa(V_i^t)$ are the parents in the net. Next, our acoustic feature detector is described.

2.2 Acoustic Feature Detector

Our audio concept detector is based on MFCCs. The audio signal is segmented into acoustic frames with overlapping. Acoustic frames are used to group samples using a window with fixed length. We split the audio signal into frames of 40ms length, with 20ms overlap, and apply a Hamming window to each frame. The Hamming function is given by:

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}). \tag{3}$$

For each audio frame, 12 MFCCs (range 133Hz-6855Hz) and their first and second derivates are computed to build an acoustic vector y^j:

$$y^j = (y_1^j, y_2^j, ..., y_{36}^j). \tag{4}$$

2.3 Bag of Audio Words representation

After the feature extraction, a way of representing audio is through a feature vector model using Bag of Audio Words (BoAW). In this representation, each vector has the size of the vocabulary, where each vocabulary word represents a position vector. The i^{th} vector value for a n audio segment equals the number of occurrences of that word i in the audio segment.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: Performance of DBNs for the violence detection task at MediaEval 2014.

Source	run #1 DBN			run #2 DBN BoAW		
	Mean Average Precision (MAP)	*Mean Average Precision 2014 (MAP2014)*	*Mean Average Precision at 100 (MAP@100)*	*Mean Average Precision (MAP)*	*Mean Average Precision 2014 (MAP2104)*	*Mean Average Precision at 100 (MAP@100)*
8 MILE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
BRAVEHEART	0.0429	0.0029	0.0369	0.0572	0.0149	0.2977
DESPERADO	0.1875	0.0159	0.1407	0.2165	0.0173	0.1635
GHOST IN THE SHELL	0.1018	0.0125	0.0458	0.1401	0.0423	0.1970
JUMANJI	0.0480	0.0235	0.1000	0.0443	0.0180	0.0307
TERMINATOR 2	0.1974	0.0518	0.1993	0.1113	0.0133	0.0295
V FOR VENDETTA	0.1201	0.0364	0.1432	0.0985	0.0794	0.4311

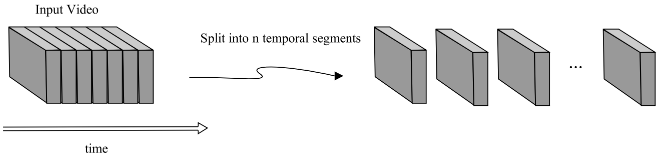

Figure 2: Given a video, we split into segments and build BoAW histograms for each segment.

3. SUBMITTED RUNS

For each run, a naive DBN is trained using two different observed vectors Y_t: (i) acoustic vector y^j, and (i) BoAW by^j with 128 audio words (see Figure 2). The likelihood of a model M, $P(y_{1:T}|M)$, is used to assign a sequence $y_{1:T}$ to non-violent or violent label as follows:

$$M^*(y_{1:T}) = arg \max_M P(y_{1:T}|M)P(M). \qquad (5)$$

The Bayes Net Toolbox for Matlab (BNT) [4] is used to train the dynamic networks.

Table 2: Global results for the violence detection task at MediaEval 2014.

Run	*MAP@100*	*MAP2014*
#1 (MFCC-DBN)	9.51 %	2.04 %
#2 (MFCC-BoAW-DBN)	16.51 %	2.64 %

4. RESULTS AND DISCUSSION

Table 1 shows the *Mean Average Precision* (*MAP*): *MAP2014* and *MAP@100* for the test movies. DBN with BoAW and DBN without have similar performances. Both approaches (run #1 and run #2) fail at detecting of violent scenes in the movie "8 Mile". The run #2 results are higher in the movies "BRAVEHEART", "DESPERADO", "GHOST IN THE SHELL" and "V FOR VENDETTA", but lower for the movies "TERMINATOR 2" and "JUMANJI" in comparisom with run #1 (using *MAP@100* and *MAP2014* metrics). Run #2 uses BoAW representation, that has less observations (temporal segments) than run #1 approach, which uses directly the acoustic feature vector built from MFCCs. Our best result is 16.51% (*MAP@100*) or 2.64 % (*MAP2014*) for run #2 (see Table 2). We investigated the results and came to the presumption that BoAW removes noisy observations,while reducing the number of observations per segment. It might be related with the observation "grouping" when the BoAW is computed for the temporal segment (see Figure 2). Thus, BoAW removes data noise and builds a better representation for a scene (model observation). However, the results are still very poor. We suppose it could be due to features, only MFCCs seems not capable of distinguishing all violence and non-violence segments and generalize the violence concept. Further work directions relies in capture the causality in violence segments using different structures and other feature modalities (feature selection).

5. ACKNOWLEDGMENTS

This work was supported in part by two grants from CAPES and CNPq.

6. REFERENCES

[1] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, and P. Lambert. Video genre categorization and representation using audio-visual information. *Journal of Electronic Imaging*, 21(2):023017–1–023017–17, 2012.

[2] E. Kijak, L. Oisel, and P. Gros. Temporal structure analysis of broadcast tennis video using hidden markov models. In M. M. Yeung, R. Lienhart, and C.-S. Li, editors, *Storage and Retrieval for Media Databases*, volume 5021 of *SPIE Proceedings*, pages 289–299. SPIE, 2003.

[3] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, July 2002.

[4] K. P. Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 33, 2001.

[5] M. Sjöberg, B. Ionescu, Y. Jiang, V. Quang, M. Schedl, and C. Demarty. The MediaEval 2014 Affect Task: Violent Scenes Detection. In *MediaEval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[6] K. Tang. Learning latent temporal structure for complex event detection. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '12, pages 1250–1257, Washington, DC, USA, 2012. IEEE Computer Society.