MULTIMODAL INFORMATION FUSION AND TEMPORAL INTEGRATION FOR VIOLENCE DETECTION IN MOVIES

Cédric Penet[1,3] *Claire-Hélène Demarty*[1] *Guillaume Gravier*[2] *Patrick Gros*[3]

[1] Technicolor Cesson-Sévigné, FRANCE
[2] CNRS/IRISA, FRANCE
[3] INRIA Rennes, FRANCE

ABSTRACT

This paper presents a violent shots detection system that studies several methods for introducing temporal and multimodal information in the framework. It also investigates different kinds of Bayesian network structure learning algorithms for modelling these problems. The system is trained and tested using the MediaEval 2011 Affect Task corpus, which comprises of 15 Hollywood movies. It is experimentally shown that both multimodality and temporality add interesting information into the system. Moreover, the analysis of the links between the variables of the resulting graphs yields important observations about the quality of the structure learning algorithms. Overall, our best system achieved 50% false alarms and 3% missed detection, which is among the best submissions in the MediaEval campaign.

Index Terms— Bayesian networks, structure learning, violence detection, multimodal fusion, temporal integration

1. INTRODUCTION

Violence detection in movies is a use case at Technicolor, which involves helping users choose movies that are suitable for children in their families by showing them the most violent segments.

This use case raises the issue of defining violence. Violence is a very subjective notion and may have several forms, such as physical violence or verbal violence. People perceive it differently from one person to another. Furthemore, in the context of movies, it may be spread over several modalities, and each event duration is variable. These characteristics highlight the complexity of the task.

This also appears in the small amount of literature on this subject over the past years. Until recently, it focused mainly on monomodal and static systems such as in [1], where the authors developed a system based on support vector machines and visual features to detect action scenes in movies. The work in [2] is an audio-based concept detection system. To our knowledge, it is also the only paper to introduce temporal information in the system through dynamic programming. More recently, the same authors combined audio and visual information in [3]. They added a visual activity detector and integrated it into their previous system through a nearest neighbour classifier to infer the presence of violence. In that paper, they no longer use any dynamic programming. To summarise, state-of-the-art systems use either multimodal information or temporal information but not both, and to the best of our knowledge, these characteristics have not been extensively studied.

In our opinion, the structure of the events requires the systems to integrate **both** multimodal and temporal information. For temporal integration, we propose to use contextual features as in [4] and to study their effect. We also propose to compare different types of multimodal fusion methods and we will conclude on the interest of both these subjects for detecting violent shots in movies. A direct comparison of our system with prior works is difficult as they are not built using the same dataset and the same definition of violence. Hence, it was tested in the context of the MediaEval 2011 Affect Task [5], which aims at comparing methods for detecting violent shots in movies using a common definition of violent events and a common dataset. A common definition, the task considers sequences in which an action or accident resulted in human pain or injury. It attempts to narrow the scope of the task, therefore reducing its subjectivity. The corpus is composed of 15 movies, hence increasing the size of the dataset compared to the literature. Through this common framework, the MediaEval 2011 Affect Task provided comparison of our technique with 5 different state-of-the-art systems.

The system developed is presented in section 2. In section 3.1, we present the dataset and evaluation metric. Finally, the results and their analysis are presented in section 3.2 and section 3.3.

2. SYSTEM DESCRIPTION

The system we developed aims at detecting violent video shots in movies. It contains four different parts: the features

THIS WORK WAS PARTLY ACHIEVED AS PART OF THE QUAERO PROGRAM, FUNDED BY OSEO, FRENCH STATE AGENCY FOR INNOVATION.

WE WOULD LIKE TO ACKNOWLEDGE THE MEDIAEVAL MULTIMEDIA BENCHMARK HTTP://WWW.MULTIMEDIAEVAL.ORG/ AND IN PARTICULAR THE AFFECT TASK 2011 FOR PROVIDING THE DATA USED IN THIS RESEARCH.

extraction, the classification step, the temporal integration and the fusion of modalities.

2.1. Features extraction

Movies have several components that convey different information, namely audio, video, subtitles... We chose to extract features from the audio and video modalities. The audio features are the energy, the centroid, the asymmetry, the zero crossing rate (ZCR) and the flatness, which are classical state-of-the-art low level features. They were extracted from 40 msec frames with 20 msec overlap and then averaged over a shot. The video features for each shot are the shot duration, the average number of blood-like color pixels in the HSV space, the average activity and the number of flashes (high luminance variation over three frames[1]). The features histograms were then rank-normalised over 22 values for each movie.

2.2. Classification

We chose to use Bayesian networks (BN) [6] as a probability distribution modelisation process. BN exhibit interesting features: in particular, they link the variables in a graph whose structure may be learned. On the downside, they consider each feature independently. During the inference stage, in the case of classification, the algorithm outputs the marginal probability for the sample of being violent. The decision is then taken by thresholding this probability.

The huge number of possible graphs is another drawback of BN, which is why structure learning algorithms were developed. They have been successfully used in the literature for soccer action detection [4], however their use in multimodal event detection appears to be quite recent. There exists two families of structure learning algorithms. The first one tries to find relations between variables using an independence test. The main drawback of these methods is that they are computationally intensive, especially when the number of variables increases. The second family groups score-based algorithms. They scan the different possible graphs and optimise a score function that is a trade-off between the likelihood of the data and the complexity of the graph. They mostly differ in the heuristics used to scan the possible graphs. Both these families are supervised.

For this paper, we focused on two score-based structure learning algorithms[2], which we compare to the naive case:

- The naive Bayesian network is the simplest network. It links the features to the classification node[3] assuming that they are all independent with respect to this node.
- The forest augmented naive Bayesian network (FAN) is introduced in [7]. It relaxes the features independence assumption by learning a forest between the features before connecting them to the classification node.
- The K2 [8] algorithm is a state-of-the-art score-based greedy search algorithm. It requires a node ordering to decide for a given node which nodes may be its children, hence reducing the number of graphs to test.

[1]Similar to `http://users.iit.demokritos.gr/~amakris/Violence.html`

[2]We used the Bayes Net Toolbox: http://code.google.com/p/bnt/.

[3]The classification node is the variable node corresponding to what we want to model.

2.3. Temporal integration

In order to account for the temporal structure of movies, we first used contextual features. However, contrary to what is done in [4], we did not restrict the contextual features to the five next video shots and also used the features from the five previous video shots. Hence, considering $X_t = [x_1, \cdots, x_K]$ the feature vector for sample at time t, the contextual features vector becomes $X_t^c = [X_{t-n}, \cdots, X_t, \cdots, X_{t+n}]$ with $n = 5$. We also considered two types of temporal filtering over 5 samples, that are used for smoothing decisions:

- Taking the maximum vote decision over a few samples, after probability thresholding.
- Averaging the samples probabilities of being violent, before probability thresholding.

2.4. Modalities fusion

As for multimodal fusion, two cases were considered: late (LF) and early fusion (EF). For early fusion, we simply concatenated the features from both modalities before learning, while for late fusion, we fused the probability of both modalities for the ith shot s_i using:

$$P_{fused}^{s_i}(P_{v_a}^{s_i}, P_{v_v}^{s_i}) = \begin{cases} max(P_{v_a}^{s_i}, P_{v_v}^{s_i}) & \text{if both are violent} \\ min(P_{v_a}^{s_i}, P_{v_v}^{s_i}) & \text{if both are non violent} \\ P_{v_a}^{s_i} * P_{v_v}^{s_i} & \text{otherwise} \end{cases} \quad (1)$$

where $P_{v_a}^{s_i}$ (respectively $P_{v_v}^{s_i}$) is the probability of being violent for the audio (respectively video) modality for the ith shot. This rule gives high scores when both audio and video find a violent segment, a low score if they both do not and an intermediate score if only one answers yes.

3. EXPERIMENTAL RESULTS

3.1. Protocol

The MediaEval 2011 Affect Task [5] corpus used for experimenting the system is composed of 12 Hollywood movies for training and 3 movies for testing. These movies were calibrated to cover a wide range of violence types in both sets. The total amount of available data is about 30 hours of video, which corresponds to 21,608 shots for the development data and 4,500 shots for the test data. A detailed description may be found in [5].

For evaluation, the metric used for the MediaEval 2011 Affect Task is a cost function (MC) of the shot false alarms (FA) and missed detections (M) rates:

$$MC = 10 * \%_M + 1 * \%_{FA} \quad (2)$$

Modality	Temporal integration			Structure learning		MediaEval cost
Late Fusion	C	A: Me	V: Ma	A: N	V: K2	**0.761**
		A: Me	V: Me	A: N	V: K2	**0.774**
Video	C	Ma		K2		**0.784**
		Me		K2		**0.840**
	-	Ma		N		**0.950**
		Ma		FAN		**1.009**
Audio	C	Me		K2		**0.805**
		Me		N		**0.843**
		Ma		K2		**0.943**
	-	-		K2		**0.967**
Early Fusion	C	Ma		K2		**0.892**
		-		K2		**0.998**

Table 1: Selected experiments results ordered by modalities used (MC: MediaEval cost, C: contextual, N: naive BN, Ma: max decision vote, Me: mean probability).

This cost function reflects the fact that missing a violent shot has much bigger consequences than having a false alarm. However, it implies that best score is mostly obtained when each shot is classified as violent, meaning a FA rate of 1 and a M rate of 0. The problem boils down to a FA reduction problem. For the sake of analysis, we also studied the FA and M curves which are proposed in Figure 1.

3.2. Results

Table 1 presents the results for 12 selected experiments chosen according to the MC metric and the false alarms vs. missed detections curves among all the possible combinations of the system. For the audio and the video experiments, the two best experiments according to each metric were chosen, while for the multimodal ones (namely early and late fusion), only the best ones are reported.

The first thing to notice is that most of the obtained scores have values lower than 1 which is better than the simple case where each sample is classified as violent, i.e. FA=100% and M=0%.

As most of the selected runs use temporal information, it seems clear that the introduction of temporality in the systems provides better results, especially for contextual data, which confirms the importance of the temporal structure of movies. The depth used for contextual data has been chosen arbitrarily, however considering other depths could be interesting. On the downside, it seems that these results depend on the algorithm used for learning the BN structures: FAN seems to work better with non contextual data contrary to the K2 algorithm. This is confirmed by the analysis of the additional experiments not presented in the paper.

The effect of multimodal fusion may be seen on Figure 1 and Table 1. It seems that EF provides worse results than LF and all the other experiments, selected or not selected, even though we thought it would improve the results. The analysis of the produced graphs (see Figure 2) yields that for non contextual data the algorithm only links the video features (red nodes) to the violence node (blue node), while the contextual data graphs are messy. Both these results indicate that the features from the two modalities are not correlated. In our opinion, it is mainly due to the semantic level of the features used, as the audio ones are low level features and the video ones mid level features, and they cannot be compared as such.

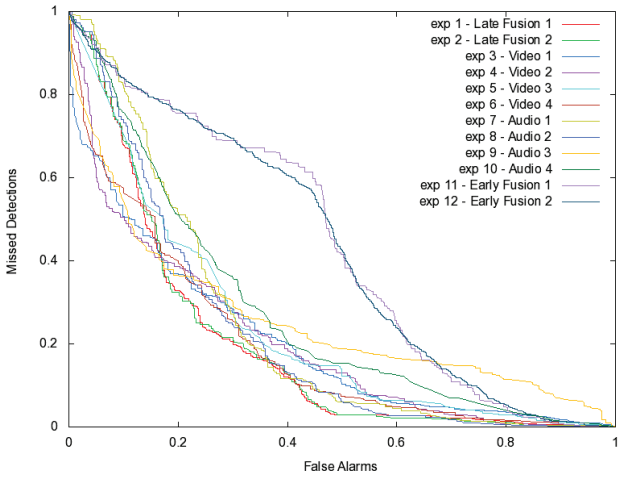

Fig. 1: False alarms vs. missed detection for each selected experiment. The legend is organised as in table 1.

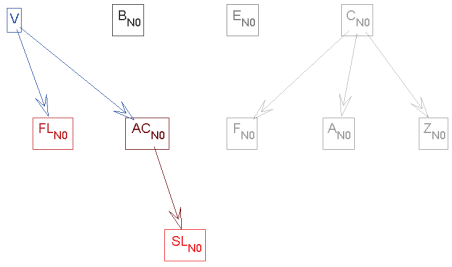

Fig. 2: Example of an early fusion graph using non contextual variables and learned with the K2 algorithm (V: violence, A: asymmetry, F: flatness, E: energy, C: centroid, ZCR: zero crossing rate, B: blood, AC: activity, SL: shot lenght, FL: flashes). The red color nodes are the video modality nodes connected to the violent node (in blue).

3.3. Graph analysis

Graph analysis produces very interesting results, particularly the analysis of the K2 graphs, as the approach is more generic than the others. A good example of the quality of the structure learning algorithm is proposed in Figure 3. Firstly, the links between features may be easily interpreted: the activity is linked to the shot length as the shot detector used tends to oversegment when the activity is high and blood has been found uncorrelated to violence which is legitimate considering the definition of violence adopted and the fact that the presence of blood highly differs from one movie to another. Furthermore, it seems that the algorithms produce a strict tem-

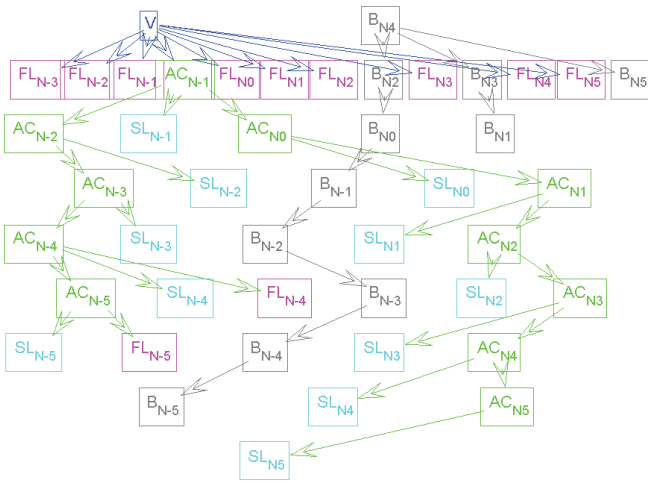

Fig. 3: Video graph using contextual variables and K2 algorithm. Each color corresponds to one group of contextual features.

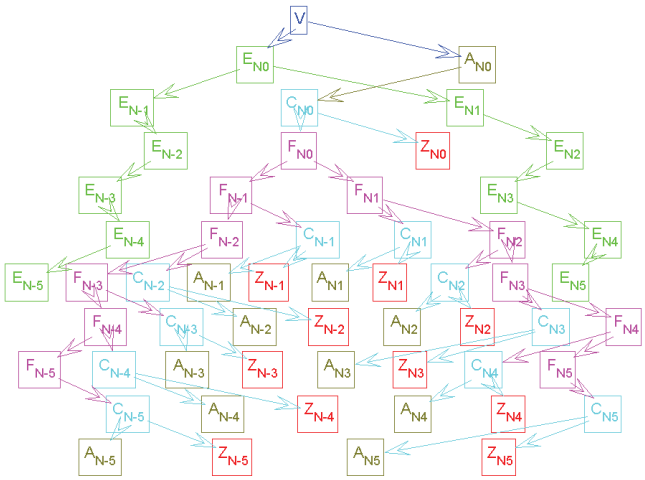

Fig. 4: Audio graph using contextual variables and K2 algorithm. Each color corresponds to one group of contextual features.

poral structure, i.e. the features from time $t = n$ are linked together and not to features from different times unless they are in temporal chains. There are two temporal chains in these graphs: activity and blood. It is easy to see that these features have a temporal coherency. On the other hand, the flash feature is only connected to the violence node and forms no chain, which is again logical as the flash feature only detects high luminance variations, and has therefore no well defined temporal structure. Comparable conclusions may be drawn from an example of an audio graph (see Figure 4).

On the downside, these results depend on the movie the system is used on: the best results were obtained for the Bourne identity, then Kill Bill and finally the Wizard of Oz. We think that it is mainly due to the low number of variables considered (4 for the video modality and 5 for the audio modality), which is probably not sufficient for the system to model each type of violence and generalise well. Again, further investigation is needed.

4. CONCLUSION

This paper presented a Bayesian network framework for temporal integration and multimodal information fusion. First, it was shown experimentally that the introduction of temporality in the framework through context and/or temporal smoothing helped improving the results. It was also shown that early fusion with features that have different natures lead to poor results, while late fusion seemed to be more promising. Finally, the structure learning algorithm output logical graphs with respect to the provided data: they were able to capture the links between features, detect the most significant variables and provide a coherent temporal structure.

This work provides a promising baseline for future work on the subject. We have several improvement ideas. Firstly, we will investigate further the contextual data and test other structure learning algorithms. Secondly, we want to add features from the text modality, as we think it also contains important information on the violent nature of the video shots.

5. REFERENCES

[1] C. Liang-Hua, S. Chih-Wen, W. Chi-Feng, and M. L. Hong-Yuan, "Action Scene Detection With Support Vector Machines," *Journal of Multimedia*, vol. 4, pp. 248–253, 2009.

[2] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "Gunshot detection in audio streams from movies by means of dynamic programming and Bayesian networks," in *Int. Conf. on Accoustic, Speech and Signal Processing*, 2008, pp. 21–24.

[3] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-Visual Fusion for Detecting Violent Scenes in Videos," in *Artificial Intelligence: Theories, Models and Applications*, LCNS, pp. 91–100. Springer Berlin / Heidelberg, 2010.

[4] S. Baghdadi, C-H. Demarty, G. Gravier, and P. Gros, "Apprentissage de structure dans les réseaux bayésiens pour la détection d'évènement vidéo," in *Traitement et analyse de l'information : méthodes et applications*, Hammamet, Tunisie, May 2009.

[5] C-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, "The MediaEval 2011 Affect Task: Violent Scenes Detection in Hollywood Movies," in *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.

[6] D. Heckerman, "A Tutorial on Learning with Bayesian Networks," Tech. Rep., Microsoft Research, 1995.

[7] P. Lucas, "Restricted Bayesian Network Structure Learning," in *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, 2002, pp. 217–232.

[8] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–347, 1992.