Joint Audio-Visual Words for Violent Scenes Detection in Movies

Nadia Derbas Georges Quénot
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
{Nadia.Derbas, Georges.Quenot}@imag.fr

ABSTRACT

This paper presents an audio-visual data representation for violent scenes detection in movies. Existing works in this field consider either the audio or the visual information; or their shallow fusion. None has yet explored their joint dependence for violent scenes detection. We propose a feature which provides strong multi-modal audio and visual cues by first joining the audio and the visual features and then revealing statistically the joint multi-modal patterns. Experimental validation was conducted in the context of the Violent Scenes Detection task of the MediaEval 2013 Multimedia benchmark. The obtained results show the potential of the proposed approach in comparison to methods using audio and visual features separately and other fusion methods.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, Indexing methods*

General Terms

Algorithms, Experimentation

Keywords

Semantic Indexing, Content Analysis, Audio Visual Fusion, Multimedia, MediaEval.

1. INTRODUCTION

The movie industry produces thousands of movies each year. In order to manage this vast amount of multimedia material, many automatic movie content analysis techniques were implemented such as the genre classification and scene characterization. Some of these techniques aim at selecting appropriate movies for different user profiles or audiences. Typically, a large number of movies are not suitable for children, especially those containing violent scenes. Therefore, some approaches have already been proposed to address the problem of automatic violence detection in movies to help preventing children from watching such movies.

ICMR '14, April 01 - 04 2014, Glasgow, United Kingdom.

http://dx.doi.org/10.1145/2578726.2578799.

Defining the term "violence" is not an easy task due to its ambiguity and its subjectivity, some viewers may see violence where others may not. Each researcher has to clarify its own definition of violence. We can find literal ones such as "physical violence or accident resulting in human injury or pain" [7]. There are also the more technical ones where violence is defined by specific visual-auditory indicators, for instance high-speed movements or fast-paced music [12].

In this paper, we propose an effective method for the automatic violent scenes detection in the context of movies. Our approach relies on an audio-visual Bag-of-Words (BoW) representation. This BoW describes the video contents based on the joint relations between the audio and the visual modalities. The remainder of the paper is organized as follows: in section 2 we discuss the related works. We describe the proposed method in section 3. In section 4, we evaluate the proposed feature using the MediaEval 2013 benchmark and we report the obtained results. We draw conclusions and discuss future works in section 5.

2. RELATED WORK

The literature related to the detection of violent content is limited and usually based on visual or spatio-temporal features [5, 3, 6]. Other methods focus on using only the audio features [10]. Furthermore, Penet *et al.* [18] proposes a novel use of the well-known audio words representations by describing each segment by one or several audio words obtained by applying product quantization to standard features.

Violent scenes in movies often occur with specific audio events (e.g. gun shots) and they typically show consistent audio-visual patterns. We therefore believe that successful violent scenes detection methods should exploit both the audio and visual modalities. Most popular audio-visual analysis techniques are based on multi-modal fusion. We can distinguish between early fusion, late fusion and kernel fusion methods [20]. In early fusion the audio and the visual features are combined before classification [12] while in late fusion the classification scores from the individual feature models are combined [9, 11, 16]. The kernel fusion can be considered as an intermediate fusion, the audio and the visual features at the kernel level are merge before performing the classification [1, 17]. These methods fuse the audio and visual modalities without considering their correlations. We can however find some works which jointly analyze the audio and visual content in other related domains. In video event

detection, Ye *et al.* [21] model the relation between audio and visual modalities within a bipartite graph followed by a partitioning over this graph to reveal the joint patterns. In the event recognition in video surveillance cameras, some authors propose a method which integrates audio and visual information [4]. They compute an "audio visual concurrence matrix" to detect and segment audio visual events. In the object tracking domain, Beal *et al.* [2] decided to exploit the statistical structure of audio and visual data along with their mutual dependencies. They model it into a single probabilistic graphical model. In general video concept classification, Jiang *et al.* [13, 14] studied the statistical temporal causality between audio and visual code words to represent the video content as audio-visual patterns.

3. THE JOINT AUDIO-VISUAL REPRESENTATION

This section describes the proposed joint audio-visual representation for violent scenes detection. Its goal is to exploit the strong correlations between the audio and the visual information in order to discover specific audio-visual patterns able to identify violent scenes. For instance, the presence of blood (detected with a color based visual feature) with persons shouting (detected with audio feature), usually illustrates a violent scene. Audio-visual patterns are expected to give better results than the simple fusion (either early or late) of the audio and the visual modalities without considering their correlations. The method is composed of three steps. Firstly, the local audio and visual features are extracted separately. Secondly, the bi-modal patterns or, in other terms, the bi-modal words, are generated. Thirdly, the bi-modal BoW representations are built using these words. The framework is illustrated by Figure 1.

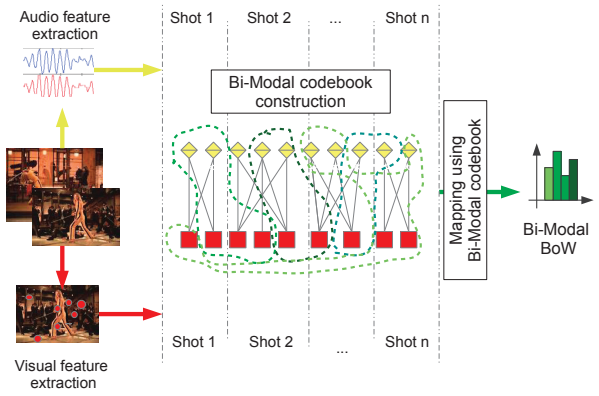

Figure 1: The overall process for generating our audio-visual BoW

We consider a video collection decomposed into n_s video shots. Local audio and video descriptors are extracted for each video shot. For a video shot s_i, these can be noted a_{ij} and v_{ik} with $1 \leq i \leq n_s$, $1 \leq j \leq n_{ai}$ and $1 \leq k \leq n_{vi}$, and with n_{ai} and n_{vi} being respectively the number of local audio descriptors and the number of local video descriptors in the video shot s_i. All the a_{ij} (resp. v_{ik}) are vectors of fixed dimension d_a (resp. d_v) and the number of local descriptors n_{ai} and n_{vi} generally depends upon the content of the video shot s_i. In the classical and per modality bag of word approach, the a_{ij} (resp. v_{ik}) are aggregated by histogramming according to clusters previously computed on all available local descriptors of the same type on the whole collection. The clusters form a dictionary of a predefined fixed size which also corresponds to the size of the aggregated representation.

For the joint audio-visual representation, we consider the set of $m_{ijk} = a_{ij} \oplus v_{ik}$ vectors with $1 \leq i \leq n_s$, $1 \leq j \leq n_{ai}$ and $1 \leq k \leq n_{vi}$, and where $a_{ij} \oplus v_{ik}$ is simply the concatenation of the a_{ij} and v_{ik}. The m_{ijk} therefore all have the same fixed dimension $d_a + d_v$ and the number of local descriptors $n_{ai} \times n_{vi}$ generally depends upon the content of the video shot s_i. Before concatenating the audio and visual descriptors, normalization and possibly weighting can be performed. The normalization can be made so that the average distance between two local descriptors is equal to 1. If a weighting is performed, it can be done according to the relative performance of the audio and visual descriptors taken separately and evaluated by cross-validation within a development set.

Since there may be a lot of local audio and a lot of visual descriptors for each video shot and also a lot of video shots in a collection, this may lead to a very large number of joint audio-visual descriptors. Even though such a representation does not need to be stored, the aggregation has to be performed on it. In order to make the approach generalizable to the case in which more than two local descriptors are fused in that way, we propose to bound the number of considered audio-visual combinations to a given value n_{max}. In this case, the local joint audio-visual representation will be a set of m_{il} with $1 \leq i \leq n_s$ and $1 \leq l \leq n_{ml}$ with n_{ml} being the number of local audio-visual descriptors in the video shot s_i. In the case in which $n_{ai} \times n_{vi} \leq n_{max}$, we take $n_{ml} = n_{ai} \times n_{vi}$ and all the $a_{ij} \oplus v_{ik}$ are considered. In the case in which $n_{ai} \times n_{vi} > n_{max}$, we take $n_{ml} = n_{max}$ and only n_{max} randomly selected $a_{ij} \oplus v_{ik}$ out of the $n_{ai} \times n_{vi}$ are considered. Finally, the bag of word type of aggregation is performed exactly in the same way on the m_{il} local descriptors as it is performed on the a_{ij} and v_{ik} ones.

4. EXPERIMENTS AND RESULTS

4.1 Dataset

We measured the effectiveness of our joint audio-visual representation in the context of the Violent Scenes Detection task of MediaEval 2013. This task defines two kinds of violence: objective and subjective. Objective violence is defined as "physical violence or accident resulting in human injury or pain". Subjective violence is defined as "scenes that you would not let an 8 years old child see in a movie because they contain physical violence" [7]. The data collection consists of two sets: a training set and a test set. The training set contains 18 annotated Hollywood movies: *Kill Bill, The sixth sense, Armageddon ...* The test set contains 7 other Hollywood movies. The samples of the training set are annotated as containing or not (subjective and objective) violent scenes shot by shot in addition to ten other concepts: *blood, fire, screams, car chase, firearms, gore, cold arms, explosions, gun shots, fights.* These ten concepts could be used to detect the violent scenes.

4.2 Parameters tuning

To generate the proposed joint audio-visual feature, we used a standard Mel Frequency Cepstral Coefficients (MFCC)

representation for the local audio descriptors and Spatio-Temporal Interest Points (STIP) based representation for the local visual descriptors [15], as motion is very important for violence detection. Before the generation of the joint audio-visual representation, these two descriptors were optimized separately on the twelve annotated concepts for avoiding over-fitting for the two target concepts (objective and subjective violence). Classification was done using two different learning methods, one based on multiple SVMs [19] (MSVM) and one based on the search of the k nearest neighbors.

We used Ivan Laptev's tool for computing the Spatio-Temporal Interest Points (STIP) [15]. A 90-dimensional Histogram of Optical Flow (HOF) vector is produced for each detected STIP. Aggregation is performed on the shot duration.

We used Guillaume Gravier's spro tool[1] for computing MFCC descriptors. One 13-dimensional vector is produced every 10ms. The window minimal duration impacts the bag of MFCCs performance. Therefore, we optimized it by cross-validation within the training set. The performance with different window durations is shown in Figure 2. Aggregation is performed over the shot segment extended before and/or after up to the minimun duration if necessary.

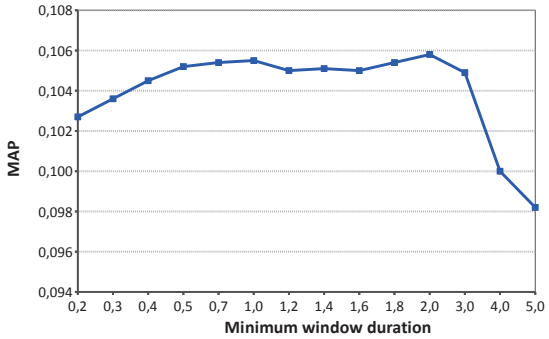

Figure 2: MFCCs performance with different window duration.

The number of clusters computed on the extracted local descriptors (dictionary size) used for the BoW generation also influence directly the descriptor performance. We optimized as well the dictionary size by cross-validation within the training set. Figure 3 shows the influence of the number of clusters on the global system performance. 4096 clusters gave the best results, thus all aggregations, either by modality or joint, are made using a BoW representation with a dictionary size of 4096 words.

Finally, we also tuned the number of considered audio-visual combinations (n_{max}) experimentally by cross-validation within the training set. We tested for different n_{max}, the best result was obtained with 32,768 as we can see in Figure 4.

4.3 Results

The official metric for the task is the Average Precision at 100 (AP@100). We compare the performance of the different visual/audio features. First, we compare for each feature separately. Then, we oppose them with the late fusion by

[1] `http://www.irisa.fr/metiss/guig/spro/`

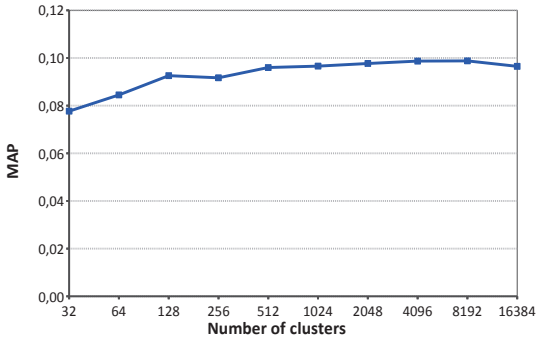

Figure 3: MFCCs performance with different number of clusters.

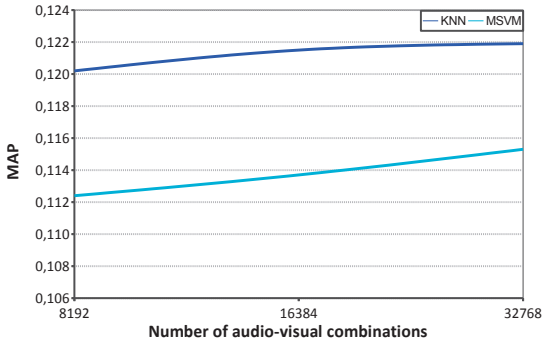

Figure 4: Influence of n_{max} on the performances.

averaging the output scores of the classifiers trained independently on each feature. Finally, we contrast them with the proposed joint MFCC-HOF feature and the fusion of the joint audio-visual feature with the two originals features (BoW of MFCC and BoW of HOF). We report in Figure 5 the AP@100 for objective violent scenes detection for 6 of the 7 films in the test set[2]. Results show that the joint audio visual feature outperforms the individual visual HOF and the audio MFCC. The joint audio visual feature and the MFCC-HOF late fusion perform comparably overall, each having different advantages over different films. Indeed, the joint audio visual feature outperforms the different visual/audio features for the movies containing good consistency between the image content and the audio signal for instance *Fantastic four1* and *Forrest gump*. This represents the interest of the proposed joint audio-visual feature.

For our official participation to the MediaEval 2013 Violent Scenes Detection task, we added two other features (opponent SIFT and color-texture) in addition to a post processing step, detailed in [8]. In Table 1, we report the obtained Mean Average Precision at 100 (MAP@100) for our official submission, for the best submission and for the median one. Our submission includes the fusion of features with the joint audio-visual feature (Submission with jointAV). This submission ranked us first over five participating groups for the subjective violence definition (69%) and second over nine participating groups for the objective one (52%).

[2] As the seventh film (Legally blonde) does not contain any violent scenes, we did not include it.

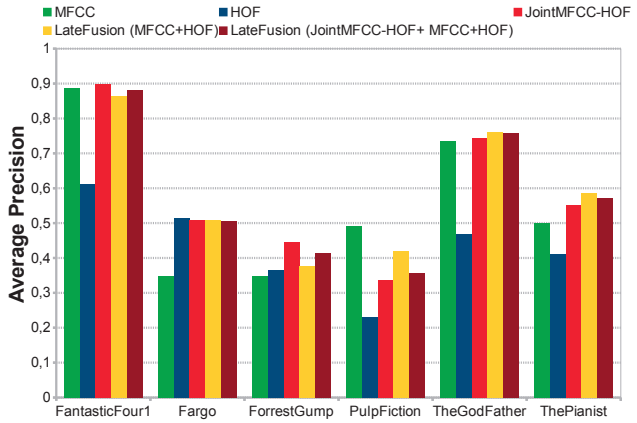

Figure 5: AP@100 for the different BoW representations and their late fusion for the objective Violent Scenes Detection task in MediaEval 2013.

	Objective	Subjective	Both
Best Submission	0.550	0.690	0.620
Submission with jointAV	0.520	0.690	0.605
Median Submission	0.400	0.570	0.485

Table 1: MAP@100 of our submission in MediaEval 2013 Violent Scenes Detection task in comparison to the best run and the median one.

5. CONCLUSION AND FUTURE WORKS

In this paper, a new method for a joint audio-visual content representation in the context of automated violent scenes detection has been proposed. It exploits the correlations between audio and visual information by building a joint audio-visual codebook in order to discover specific audio-visual patterns. Comparatively to other fusion approaches, it can be seen as an "early-early" one since fusion is done before the aggregation step while in the classical early one it is done after this aggregation step (and before the classification step).

Experimental validation on real Hollywood movies (MediaEval 2013 data collection) has shown that the joint audio-visual fusion produced results as good as a late fusion. Our future work will explore the joint representation for different kind of dynamic concepts, other than violence. Also, we plan to explore the use of joint audio-visual features at a time scale smaller than the whole shot one since the useful correlation might be better captured with a stronger time locality. Finally, this work can be extended to use more than just the two original features, MFCC and STIP-HOF.

Acknowledgements

This work was partly realized as part of the Quaero Program and Camomile project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency).

6. REFERENCES

[1] S. Ayache, G. Quénot, and J. Gensel. Classifier fusion for svm-based multimedia semantic indexing. In *Advances in Information Retrieval*, pages 494–504. Springer, 2007.

[2] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):828–836, 2003.

[3] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns*, pages 332–339. Springer Berlin Heidelberg, 2011.

[4] M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *Multimedia, IEEE Transactions on*, 9(2):257–267, 2007.

[5] A. Datta, M. Shah, and N. da Vitoria Lobo. Person-on-person violence detection in video data. In *Pattern Recognition*, volume 1, pages 433–438 vol.1, 2002.

[6] F. de Souza, G. ChaÌĄvez, E. do Valle, and A. de A Araujo. Violence detection in video using spatio-temporal features. In *Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, pages 224–230, Washington, DC, USA, August 30-Septembre 3 2010.

[7] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. L. Quang, and Y.-G. Jiang. The MediaEval 2013 Affect Task: Violent Scenes Detection. In *MediaEval Workshop*, Barcelona, Spain, October 18-19 2013.

[8] N. Derbas, B. Safadi, and G. Quénot. Lig at mediaeval 2013 affect task: Use of a generic method and joint audio-visual words. In *MediaEval Workshop*, Barcelona, Spain, October 18-19 2013.

[9] N. Derbas, F. Thollard, B. Safadi, and G. Quénot. Lig at mediaeval 2012 affect task: Use of a generic method. In *MediaEval Workshop*, Pisa, Italy, October 4-5 2012.

[10] T. Giannakopoulos, D. I. Kosmopoulos, A. Aristidou, and S. Theodoridis. Violence content classification using audio features. In *SETN*, pages 502–507, 2006.

[11] 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*, pages 91–100. Springer Berlin Heidelberg, 2010.

[12] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao. Detecting violent scenes in movies by auditory and visual cues. In *Advances in Multimedia Information Processing - PCM 2008*, pages 317–326. Springer Berlin Heidelberg, 2008.

[13] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. Loui. Short-term audio-visual atoms for generic video concept classification. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 5–14. ACM, 2009.

[14] W. Jiang and A. C. Loui. Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. In *ACM Multimedia*, pages 123–132, 2011.

[15] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, Sept. 2005.

[16] J. Lin and W. Wang. Weakly-supervised violence detection in movies with audio and video based co-training. In *Advances in Multimedia Information Processing - PCM 2009*, pages 930–935. Springer Berlin Heidelberg, 2009.

[17] M. Mühling, R. Ewerth, J. Zhou, and B. Freisleben. Multimodal video concept detection via bag of auditory words and multiple kernel learning. In *Advances in Multimedia Modeling*, pages 40–50. Springer, 2012.

[18] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Audio event detection in movies using multiple audio words and contextual bayesian networks. In *Workshop on Content-Based Multimedia Indexing*, pages 17–22, 2013.

[19] B. Safadi and G. Quénot. Evaluations of multi-learner approaches for concept indexing in video documents. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 88–91, 2010.

[20] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM International Conference on Multimedia*, pages 399–402, 2005.

[21] G. Ye, I.-H. Jhuo, D. Liu, Y.-G. Jiang, D. Lee, and S.-F. Chang. Joint audio-visual bi-modal codewords for video event detection. In *ACM International Conference on Multimedia Retrieval (ICMR)*, Hong Kong, June 5-8 2012.