LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words

Nadia Derbas, Bahjat Safadi and Georges Quénot
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217,
Grenoble, F-38041, France
FirstName.LastName@imag.fr

ABSTRACT

This paper describes the LIG participation to the MediaEval 2013 Affect Task on violent scenes detection in Hollywood movies. We submitted four runs at the shot level for each subtasks: objective violent scenes detection and subjective violent scenes detection. Our four runs are: hierarchical fusion of descriptors and classifier combinations, the same with joint audio-visual words, and the same two with reranking. Our reference run obtained with the official MAP@100 metric a performance of 69% for the subjective violence and 52% for the objective violence. The joint audio-visual words bring a slight improvement on the MAP@100 and they improve the precision in the head of the returned list while the temporal re-ranking improves the P@100.

1. INTRODUCTION

The MediaEval 2013 Affect Task: Violent Scenes Detection is fully described in [1]. It directly derives from a Technicolor use case which aims at easing a user's selection process from a movie database. This task therefore applies to movie content. This year, two different subtasks were proposed for violent segments, corresponding to objective violence and subjective violence. An objective violent scene is defined as "physical violence or accident resulting in human injury or pain", a subjective one is the scene "one would not let an 8 years old child see in a movie because it contains physical violence".

Our motivation is to test the performance of a new descriptor based on joint audio-visual bi-modal codewords on the violent scenes detection. As well, we aim to see how a generic system for general concept classification in video shots would perform compared to systems specifically designed for the task like [5]. Our system is a refined version of last year's system which was roughly a four-stage pipeline: descriptor extraction, descriptor optimization, classification and hierarchical late fusion. Besides using more descriptors, we proposed a new multi-modal feature, the "audio-visual words". Most of the stages have been optimized and specifically tuned on MediaEval development data.

2. SYSTEM DESCRIPTION

2.1 Descriptor Extraction

The descriptors were computed using audio, image and motion information. Six types of descriptors were used:

- color: a $4 \times 4 \times 4$ RGB color histogram (64-dim);
- texture: a 5-scale $\times$ 8-orientation Gabor transform (40-dim);
- SIFT: bag of SIFT descriptors computed using Koen van de Sande's software [6], 1000-bin histograms; four variants were used: Harris-Laplace filtering or dense sampling with both hard and fuzzy clustering;
- audio: bag of MFCCs, 4096-bin histograms;
- STIP: bag of HOFs, 4096-bin histograms;
- joint audio-visual BoW: bag of MFCCs and HOFs, 32768-bin histograms (see section 2.6).

2.2 Descriptor Optimization

Descriptor optimization was done using a method which combines a PCA-based dimensionality reduction with a power transformation [3]. The power transformation normalizes the distributions of the values, especially in the case of histogram components. A PCA is then performed for reducing the size (number of dimensions) of the descriptors while improving performance by removing noisy components.

2.3 Classification

Classification was done using two different learning methods, one based on multiple SVMs for a better handling of the class imbalance problem and one based on the k nearest neighbors.

2.4 Fusion

Classification was done separately for each classifier and each descriptor variant. The outputs of these individual classifiers are then merged at the level of normalized scores (late fusion). A linear combination of the scores is used with weights optimized on the MediaEval development set.

2.5 Temporal Re-ranking

As for our participation to MediaEval 2012 [2], we used a temporal re-ranking method. The method is based on the assumption that the violence will be more (or less) likely for a given shot if it appears within a movie with a high (or low) frequency of violent shots and/or if there are more (or less) violent shots in its temporal neighborhood. We have proposed to exploit this either at a global or at a local level by computing a detection score either at the video or at a

MediaEval 2013 Workshop October 18-19, 2013, Barcelona, Spain

	Objective			Subjective			All
Metric	MAP	MAP@100	P@100	MAP	MAP@100	P@100	MAP@100
LIG-hierarchicalFusion	0.501	0.514	0.381	**0.673**	**0.690**	0.584	0.602
LIG-hierarchicalFusionJoint	**0.505**	**0.520**	0.398	0.669	**0.690**	0.602	**0.605**
LIG-hierarchicalFusionReranking	0.443	0.502	0.412	0.627	0.685	0.624	0.593
LIG-hierarchicalFusionJointReranking	0.453	0.512	**0.418**	0.627	0.686	**0.635**	0.599

Table 1: Performance of the LIG system for the objective and the subjective violent scenes detection

neighborhood level and then re-evaluate the score of each shot according to this global or local score. The first step is done by a kind of temporal smoothing and the second one by a kind of averaging [4].

2.6 Audio-Visual Representation

For this year, we proposed a joint audio-visual representation in order to capture the dependence/relation between the audio and the visual information based on their simultaneous occurrence throughout the movies for a given concept, an idea inspired by [7]. In this approach the video content is modeled using the joint relations between the audio (MFCC) and the visual (HOF) modalities. The audio and visual features are first extracted from the movies separately. These two features are then normalized and joined by concatenating both feature vectors for each shot. Finally, the bi-modal descriptors are quantized into bi-modal words using a standard k-means clustering method, producing the "joint audio-visual" bag-of-words representation.

3. EXPERIMENTAL RESULTS

We submitted four runs at the shot level for both of the objective and subjective definitions. The hierarchical fusion of descriptors and classifier combinations and the same with the joint audio-visual words and/or with temporal re-ranking. The hierarchical fusion run is our baseline and the other three are contrastive ones. Table 1 shows the performance of the LIG system variants using different metrics, the Mean Average Precision (MAP), the Precision at 100 (P@100) and the official MediaEval metric for this task which is the MAP@100.

Considering these metrics, our system produces quite good results in the detection of the objective and subjective violent scenes in movies, with an average MAP@100 of about $60,50\%$. In general, our system provides better results for the subjective definition with a MAP@100 of about 69% and of about 52% for the objective definition . This could be related to the fact that the subjective definition is more related to the "basic violence" than the objective one. We observe that the hierarchical fusion with the joint audio-visual descriptor always improves the performance in terms of MAP@100 and especially in terms of P@100 (even if it is a slight improvement in some case). That is due to the fact that the bi-modal words consider the relation between audio and visual information while the other methods fuse them without exploiting their mutual dependence. Further, we notice that the re-ranking improves results just in terms of P@100 but it is slightly lowering the MAP@100 and even more the overal MAP.

4. CONCLUSIONS AND FUTURE WORK

We have participated in the MediaEval 2013 affect task with the same baseline system as for MediaEval 2012 but with different descriptors. This system includes a hierarchical fusion of classifiers' outputs using two different classification methods and a number of shot content descriptors. However, two new descriptors were added this year: the classical motion descriptor (STIP-HOF) and our proposed joint audio-visual descriptor. Four variants of the system were evaluated in which the joint audio-visual descriptor and the temporal re-ranking were added or not to the baseline. Our system obtained good results with a MAP@100 of about 69% for the subjective definition and of about 52% for the objective definition. The joint audio-visual descriptor always improves the MAP@100 and the P@100 while the re-ranking just improves the P@100.

In the future, we plan to extend our work on the joint audio-visual descriptor and focus on optimizing it and on testing it with more than just two features.

5. ACKNOWLEDGMENTS

This work was partly realized as part of the Quaero Program funded by OSEO, French State agency for innovation.

6. REFERENCES

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

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

[3] B. Safadi and G. Quénot. Descriptor optimization for multimedia indexing and retrieval. In *CBMI*, pages 65–71, Veszprém, Hungary, June 17-19 2013.

[4] B. Safadi and G. Quénot. Re-ranking for Multimedia Indexing and Retrieval. In *ECIR 2011: 33rd European Conference on Information Retrieval*, pages 708–711, Dublin, Ireland, April 18-21 2011.

[5] F. D. M. d. Souza, G. C. Chavez, E. A. d. Valle Jr., and A. d. 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.

[6] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[7] 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.