LIG at MediaEval 2011 affect task: use of a generic method

Bahjat Safadi
UJF-Grenoble 1 / UPMF-Grenoble 2 /
Grenoble INP / CNRS, LIG UMR 5217,
Grenoble, F-38041, France
Bahjat.Safadi@imag.fr

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

ABSTRACT

This paper describes the LIG participation to the MediaEval 2011 Affect Task on violent scenes' detection in Hollywood movies. We submitted only the required run (shot classification run) with a minimal system using only the visual information. Color, texture and SIFT descriptors were extracted from key frames. The performance of our system was below the performance of the systems using both audio and visual information but it appeared quite good in precision.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation

Keywords

Violence detection, Affect, Video Annotation, Benchmark

1. INTRODUCTION

The MediaEval 2011 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.

Our motivation was to see how a generic system for general concept classificationn in video shots would perform compared to systems specifically designed for the task like [4]. Our system is roughly a four-stage pipeline: descriptor extraction, descriptor optimization, classification and fusion. Most of the stages have been optimized for the TRECVID 2011 semantic indexing task [3] [2] but some parameters have been specifically tuned on MediaEval development data.

2. SYSTEM DESCRIPTION

2.1 Descriptor extraction

The descriptors were computed only on the visual information (no audio) and even only on the key frames (no motion). Three 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 [5], 1000-bin histograms, four variants were used: Harris-Laplace filtering or dense sampling with hard or fuzzy clustering.

2.2 Descriptor optimization

The descriptor optimization consists of two steps:

- power transformation: its goal is to normalize the distributions of the values, especially in the case of histogram components. It simply consists in applying an $x \leftarrow x^{\alpha}$ tranformation on all components individually. The optimal value of alpha can be optimized by cross-validation and is often close to 0.5 for histogram-based descriptors.
- PCA reduction: its goal is both to reduce the size (number of dimensions) of the descriptors and to improve performance by removing noisy components. For color and texture, the optimal number of dimension is close to half of the original one. For the SIFT-based descriptors, it is in the 150-250 range.

2.3 Classification

The classification was done here using a kNN-based classifier. It is a bit less efficient than an SVM one but it is much faster.

2.4 Fusion

Classification was done separately with one kNN for 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 weight optimized on the MediaEval development set. It finally appeared that, for the MediaEval task, the SIFT descriptors did not help, compared to color and texture alone; this was not the case in the general context of TRECVID.

3. EXPERIMENTAL RESULTS

Figure 1 shows the false alarms' rate versus miss rates for participants' best runs. It is obtained by the application of a varying threshold on the scores provided by the participants. The LIG system performs less well than other systems using both audio and visual information. However, it appears to be as good as all of them in the area of the low false alarm

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

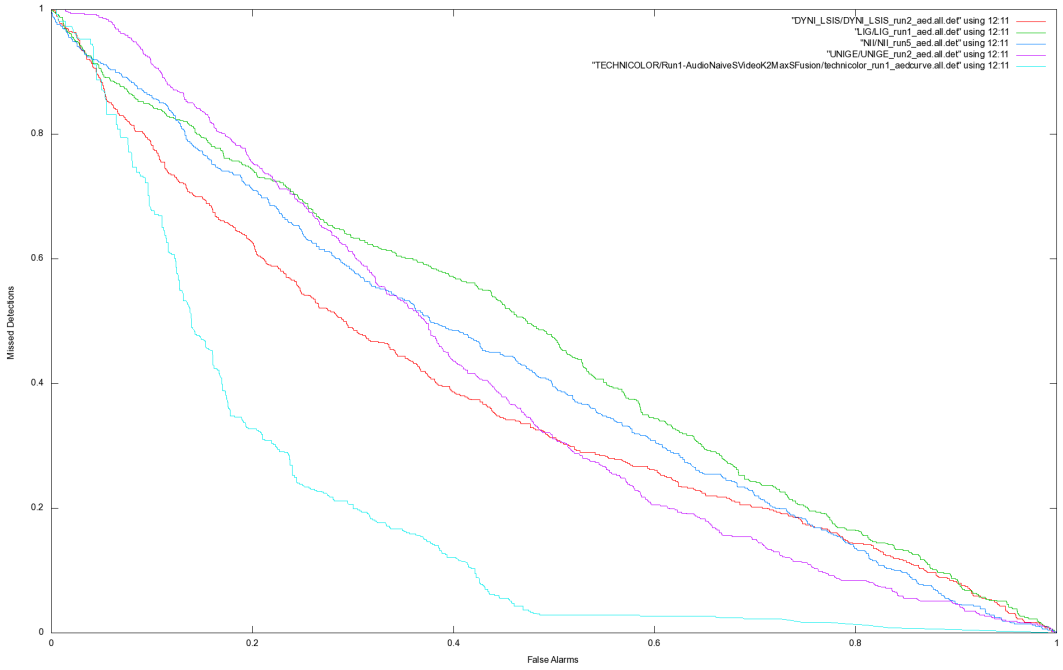

Figure 1: False alarms' rate versus miss rate for the participants' best runs

rates. This means that the LIG system is able to find with a good confidence a fraction of the shots containing physical violence but beyond these, it fails to detect others, probably because the audio and/or motion modalities are necessary for them.

	F-measure	MediaEval cost
Kill bill	0.19	8.58
The Bourne Identity	0.24	6.07
The wizard of Oz	0.00	10.1
All	0.20	7.94

Table 1: Performance of the LIG system

Table 1 shows the performance of the LIG system using the AED F-Measure (common in information retrieval) and the official MediaEval Cost. The MediaEval cost is highly biased towards recall and while the threshold of our system was also biased in this direction it was not biased enough for being optimal for this measure.

While the performance of the system is consistent on *Kill Bill* and *The Bourne Identity*, it is very bad for *The wizard of Oz*. The system did not found any of the 46 violent shots though it predicted 60 positives (all false) in a total of 908 shots. This seems to be worse than random.

4. CONCLUSIONS AND FUTURE WORK

We have participated to the MediaEval 2011 affect task with a basic system designed for general purpose concept detection in video shots. This system used only the information available in the key frames (no audio or motion). This system was initially intended to be used as a baseline and specific extensions were considered but they could not be finalized in time. Also, concerning the target measure, the threshold was biased a bit toward recall but not enough for an optimal result with the same ranking.

In our future work, we plan to improve this baseline system by using a better classifier (SVM-based) and include motion descriptors based on optical flow and audio descriptors based on MFCC.

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, 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.

[2] B. Safadi, N. Derbas, A. Hamadi, F. thollard, and G. Quénot. LIG at TRECVID 2011. In *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, December 5-7 2011.

[3] A. F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In A. Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.

[4] 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, 2010.

[5] 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.