LIG at MediaEval 2012 affect task: use of a generic method

Nadia Derbas, Franck Thollard, 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 2012 Affect Task on violent scenes' detection in Hollywood movies. We submitted four runs at the shot level: hierarchical fusion of descriptors and classifier combinations (LIG-4), the same with conceptual feedback (LIG-3), and the same two with reranking (LIG-2 and LIG-1). Our reference run obtained a performance slightly above the median with the official MAP@100 metric. The temporal re-ranking brings a significant improvement on the overall (classical) MAP but has almost no effect on the MAP@100. The conceptual feedback does not improve the overall MAP but it improves the precision in the head of the returned list (MAP@100 or P@100).

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 2012 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 [5]. Our system is an improved version of our last year system which was roughly a four-stage pipeline: descriptor extraction, descriptor optimization, classification and hierarchical late fusion. Besides using more descriptors and classifiers, two improvements have been introduced this year: a conceptual feedback and a temporal re-ranking. Most of the stages have been optimized for the TRECVID 2012 semantic indexing task [4] [2] but some parameters have been specifically tuned on MediaEval development data.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2. SYSTEM DESCRIPTION

2.1 Descriptor extraction

The descriptors were computed using audio and still image information (no motion). Four 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 hard or fuzzy clustering;
- audio: bag of MFCCs, 256-bin histograms; two variants were used: MFCCs only and MFCCs plus their first and second derivatives.

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 $x \leftarrow sign(x)|x|^{\alpha}$ 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 using two different classification methods, one based on the use of multiple SVMs for a better handling of the class imbalance problem and one based on the use of 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 weight optimized on the MediaEval development set.

2.5 Conceptual feedback

The conceptual feedback is the way we have chosen to make use of the annotations available on the 10 concepts other than *violentscenes*. We trained symmetrically classifiers on the 11 annotated concepts. We produced classification scores for them either directly on the test set or by cross-validation within the devleopment set. We built for each shot either in the development or in the test set a 11-component vector made of these scores and considered the produced vectors as a new descriptor on which we trained classifiers exactly as with the signal-based descriptors. Finally, the outputs of these classifiers were included in the hierarchical fusion.

2.6 Temporal re-ranking

Temporal re-ranking is based on the assumption that violence (or any other concept) appear more in some movies than in others and that it appears as "bursts" within a given one. In other words: 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 local level by computing a detection score either at the video or 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 [3].

2.7 Threshold selection

The threshold for the MediaEval cost metrics was tuned by cross-validation for optimizing a 1:3 cost. For the 1:10 version (last year metric), pour system was not able to do better that the "all positive" baseline.

3. EXPERIMENTAL RESULTS

Metric	MAP@100	1:2 cost	MAP	P@100
Best	0.6506	0.8225	0.3183	0.4833
LIG-1	0.3138	1.1295	0.1723	0.3167
LIG-2	0.3122	1.1009	0.1731	0.3034
LIG-3	0.3138	1.3534	0.1307	0.3166
LIG-4	0.3122	1.3734	0.1259	0.3033
Median	0.3122	1.2475	0.1249	0.2600

Table 1: Performance oghf the LIG system, lower is better for 1:2 cost and higher is better for MAP@100, MAP and P@100

We submitted four runs at the shot level: hierarchical fusion of descriptors and classifier combinations (LIG-4), the same with conceptual feedback (LIG-3), and the same two with reranking (LIG-2 and LIG-1). The runs are numbered according to theur expected performance, LIG-1 being the reference one and the other contrastive ones.

Table 1 shows the performance of the LIG system variants using four different metrics. The first one, MAP@100 is the official MediaEval metric for the task. The second metric displayed is the one with a 1:2 cost ratio. The third and fourth metrics are the classical MAP and Precision at 100 (P@100). Considering these metrics, our system variants performs slightly better than the median one with some differences. P@100 and MAP@100 are very correlated.

Considering the contrastive runs we made, the conceptual feedback significantly improves the MAP@100 and P@100 (head of the returned list) and the temporal re-ranking significantly improves the MAP and 1:2 cost but has almost no effect on the MAP@100 and P@100. The best combination always involves temporal re-ranking but the conceptual feedback should be used if the target metrics are MAP or 1:2 cost and should be avoided is the target metric is the MAP@100 and P@100.

4. CONCLUSIONS AND FUTURE WORK

We have participated to the MediaEval 2012 affect task with a system designed for general purpose concept detection in video shots. This system used audio and still image information but no motion information. The system includes hierarchical fusion of classifiers using two different classification methods and a number of shot content descriptors. Four variants of the system were evaluated, using or not conceptual feedback and temporal re-ranking. Our runs were generally above the median for the considered metrics (MAP@100, 1:2 cost, MAP and P@100). Temporal re-ranking always improve the result (1:2 cost and MAP) or has no significant effect (MAP@100 and P@100). The conceptual feedback has a negative impact for the 1:2 cost and MAP metrics and a positive impact on the MAP@100 and P@100 (head of the returned list).

In our future work, we plan to improve this system by including motion descriptors based on optical flow or on Space-Time Interest Points (STIP) and better audio descriptors.

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 2012 Affect Task: Violent Scenes Detection. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[2] B. Safadi, N. Derbas, A. Hamadi, F. thollard, and G. Quénot. Quaero at TRECVID 2012. In *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, November 26-28 2012.

[3] 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, apr 2011. Springer.

[4] 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.

[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, 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.