Technicolor/INRIA/Imperial College London at the MediaEval 2012 Violent Scene Detection Task

Cédric Penet, Claire-Hélène Demarty, Mohammad Soleymani, Guillaume Gravier, Patrick Gros

Technicolor/INRIA/Imperial College London at the MediaEval 2012 Violent Scene Detection Task*

Cédric Penet, Claire-Hélène Demarty
Technicolor/INRIA Rennes & Technicolor
975, ave des champs blancs
35510 Cesson-Sévigné, France
firstname.lastname@technicolor.com

Mohammad Soleymani
Department of Computing
Imperial College London
SW7 2AZ London, United Kingdom
m.soleymani@imperial.ac.uk

Guillaume Gravier, Patrick Gros
CNRS/IRISA & INRIA Rennes
Campus de Beaulieu
35042 Rennes, France
guig@irisa.fr
patrick.gros@inria.fr

ABSTRACT

This paper presents the work done in Technicolor, INRIA and Imperial College London regarding the Affect Task at MediaEval 2012. This task aims at detecting violent shots in movies. Four different systems and a fusion of three of them are proposed in this paper.

1. INTRODUCTION

The MediaEval 2012 Affect Task is the continuation of the MediaEval 2011 Affect Task and aims at detecting violence in movies. A complete description of the task and datasets may be found in [3].

This paper is a joint effort between Technicolor, INRIA and the Imperial College of London. The different systems used are presented in section 2 and the results are discussed in section 3.

2. SYSTEMS DESCRIPTION

In this section, we briefly present the different systems behind each run for this year's task.

2.1 Run #1: Similarity

The idea that motivated the method for run #1 is to check whether it is feasible to achieve comparable results to those of the 2011 campaign using only a similarity measure between the training and test sets. Besides avoiding the use of machine learning techniques, this run is an attempt, given a new event, to classify it simply by measuring the similarity between this event and some violent and non violent events. The final decision is made using a knn scheme, measuring the shortest distance between the event and his neighbours and their labels. For each event, a set of only 4 video features was used: three color harmonization features (an harmonization template, its corresponding angle and a minimum energy value [2]) and a motion activity measure over each frame were computed and then agregated over one shot, by either taking a maximum voting scheme (harmonization template), or an average value over all frames (angle, energy or motion activity). We used a set of 10 movies from the learning database to create some frame clusters, using the lib-pq and yael software detailed in [5]. Each frame of the three test movies was then labeled according to its closest neighbour's label. The final decision at shot level was made by simply labelling any shots with at least one violent frame as violent.

2.2 Run #2: Bag of Audio Words

For this run, TF-IDF Bag of Audio Words (BoAW) were used as features. First, we extracted spectrum coefficients using a 24 filterbank along with the deltas and accelerations on 20 ms audio frames with 10 ms overlap. Then, spectrally coherent audio segments were extracted using [1], the silence segments were removed using Spro[1] and all 20ms audio frames lying within each segment were averaged. In order to produce audio words, the yael implementation of K-Means was used [5]. 128 clusters were extracted from the 15 development movies, and one cluster was added for silence segments. We finally extracted TF-IDF histograms [6] considering that each shot is a document. For classification, we used SVM with Histogram Intersection kernel (HIK) and Chi-squared kernel (χ^2):

$$K_{HIK}(x,y) = \sum_{k=0}^{D} min(x_k, y_k) \quad (1)$$

$$K_{\chi^2}(x,y) = \sum_{k=0}^{D} \frac{x_k * y_k}{2(x_k + y_k)^2} \quad (2)$$

where $x, y \in \Re^D$, and x_k is the k^{th} element of x. In order to obtain confidence scores to compute the Mean Average Precision at 100 (MAP@100), a sigmoid function was applied to the distance of each sample to the classification hyperplane. A grid search was applied on the SVM parameter C, and to cope with the imbalance data problem, the C parameter for violent samples was set to $10 * C$. In cross-validation over the development set, the best MAP@100 was obtained for the χ^2 kernel, with a weight of 10 for violent samples and $C = 2^{-6}$. The value reached for the MAP@100 is 44.77% and the standard deviation (STD) over all movies is 15.01%.

2.3 Run #3: Bayesian Networks

The method presented in [7] and developed for the MediaEval 2011 campaign was used again with more features

*This work was partly achieved as part of the Quaero Program, funded by OSEO, French State agency for innovation.
†The work of Soleymani is funded by European Research Council under the FP7 Marie Curie Intra-European Fellowship, EmoTag.

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

[1] https://gforge.inria.fr/projects/spro

and a new temporal filter. For the video modality, color coherence and color harmonisation [2] features were added, which brings the number of features from 4 up to 10. For the audio modality, only the roll-off at 90% was added.

We employed a majority vote post-processing filter that works over a window of size n (we used $n = 5$). For the confidence scores, we consider two cases, depending on whether the maximum vote is violent or non violent:

Violent The confidences of the non violent samples are set to $min(P(S_v))$ where $P(S_v)$ is the set of confidences of the violent samples within the window.

Non violent The confidences of the violent samples are set to $max(P(S_n v))$ where $P(S_n v)$ is the set of confidences of the non violent samples within the window.

As in [7], the parameters and configurations are chosen based on the system's performance in a one-movie-out cross-validation on the development set. The best system configuration, with a MAP@100 of 43.18% and a STD of 18.67%, was obtained using late fusion with the following configuration:

Audio K2-learned structure with non contextual features and the new temporality.

Video Naive structure with contextual features and the new temporality.

Late Fusion Averaging filter applied to the confidence scores.

2.4 Run #4: Imperial College

The system implemented by UniGe in Mediaeval 2011 [4] was re-implemented with a refined set of features. The results of text features on the training set were not satisfactory. Therefore, only audio and visual modalities were used after a weighted decision level fusion. Naïve Bayesian classifiers' confidence scores were added using the weights obtained based on the best results on one-movie-out cross validation on the development set. The weight for the audio modality was set to 0.95 and the visual modality weight was set to 0.05 respectively. Although the confidence scores of naïve Bayesian classifiers are not reliable in case of redundant features, feature reduction techniques such as Principal Component Analysis failed to improve the MAP on the development set.

2.5 Run #5: Classifier fusion

For this system, the results from runs #2, #3 and #4 were fused by simply multiplying the confidence scores with the effect of reducing the false alarms rate.

3. RESULTS AND DISCUSSION

Table 1 presents the results obtained on the test set. The first noticeable point from this table is the high average precision (AP) variation between the different movies. This first result highlights clearly the difficulty of the task, as well as the high variability of violence between movies. It is also interesting to note that for runs #1 and #2, the movie that have the worst result is the least violent movie (according to the definition), i.e., "Dead Poet Society". The inferior results on "Fight Club" compared to "Independence Day" can be due to the nature of the violent events present in these movies. Violent actions such as *fist fight* in "Fight Club" are under-represented in the training set whereas the violent events in "Independence Day", e.g., *explosions*, are more present in the training set. The fact that "Independence Day" systematically provides the best results can be as a result of the existence of similar genre of movies in the development set, i.e., movies depicting disasters, e.g., "Armageddon".

Table 1: Results for submitted runs (MAP: MAP@100, AP: average precision for "Deap Poet Society" (AP-1), "Fight Club" (AP-2) and "Independence Day" (AP-3), STD is the standard deviaton over the test movies and MC11: MediaEval Cost from 2011 campaign).

Run	**MAP (%)**	AP-1 (%)	AP-2 (%)	AP-3 (%)	STD (%)	MC11
#1	**13.89**	0.00	12.91	28.77	14.41	2.29
#2	**40.54**	10.85	52.98	57.77	25.82	2.50
#3	**61.82**	60.56	53.15	71.76	9.37	3.57
#4	**46.27**	40.03	22.97	75.82	26.97	3.64
#5	**57.47**	64.52	37.21	75.69	17.82	4.60

The best performing run is the third run using a Bayesian network with temporal integration post-processing. This further emphasizes the importance of multimodal approaches as well as the value of taking into account the temporal dimension. Nevertheless, this result might not generalize over a larger development set since the obtained MAP is almost 20% higher than the cross-validation MAP. Indeed, the standard deviation and mean might have more meaning if extracted on more test movies.

Finally, run #1 has proved that using only a similarity measure between such events is worth begin further tested. Although the MAP@100 value is not so good and it used a really small set of features, the MC11 is one of the best among all the participants.

4. REFERENCES

[1] R. Andre-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(1):29 –40, jan 1988.

[2] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu. Color harmonization. In *ACM SIGGRAPH 2006 Papers*, page 624Ű630, New York, NY, USA, 2006.

[3] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2012 Affect Task: Violent Scenes Detection. In *MediaEval 2012 Workshop*, 2012. ceur-ws.org.

[4] G. Gninkoun and M. Soleymani. Automatic Violence Scenes Detection: A multi-modal approach. In *MediaEval 2011 Workshop*, 2011. ceur-ws.org.

[5] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 33(1):117–128, jan 2011.

[6] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[7] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Multimodal information fusion and temporal integration for violence detection in movies. In *ICASSP*, Kyoto, Japon, Mar. 2012.