The MediaEval 2012 Affect Task: Violent Scenes Detection

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

The MediaEval 2012 Affect Task: Violent Scenes Detection*

Claire-Hélène Demarty, Cédric Penet
Technicolor, Rennes, France
claire-helene.demarty@technicolor.com
cedric.penet@technicolor.com

Guillaume Gravier
Irisa/CNRS
Rennes, France
guig@irisa.fr

Mohammad Soleymani
Imperial College London
United Kingdom
m.soleymani@imperial.ac.uk

ABSTRACT
This paper provides a description of the MediaEval 2012 Affect Task: Violent Scenes Detection. This task derives directly from a Technicolor use case which aims at easing a user's selection process from a movie database. This task will therefore apply to movie content. We provide some insight into the Technicolor use case, before giving details on the task itself. Dataset, annotations, and evaluation criteria as well as the two required and optional runs are described.

Keywords
Violence detection, Affect, Movies, Annotation, Benchmark

1. INTRODUCTION
The Affect Task - Violent Scenes Detection is part of the MediaEval 2012 benchmarking initiative for multimedia evaluation. It involves automatic detection of violent segments in movies. This challenge is a follow-up of last year's edition which served as a pilot, and therefore will only see slight modifications in 2012. It derives from a use case at Technicolor (http://www.technicolor.com). As a provider of services in multimedia entertainment, Technicolor is developing services connected to the management of movie databases, through content indexing and content discovery, for content creators. In that context, the company constantly seeks to help users select the most appropriate content, according to, for example, their profile or other constraints. Given this, a particular use case arises which involves helping users choose movies that are suitable for children in their family. The movies should be suitable in terms of their violent content, e.g., for viewing by users' families. Users select or reject movies by previewing parts of the movies (i.e., scenes or segments) that include the most violent moments. Despite its importance, there are only few published studies on the detection of violent scenes in videos. Among them, only a few use multimodal approaches [3, 1]. Moreover, these methods suffer from a lack of a common and consistent database, and usually use a limited developement set [2].

2. TASK DESCRIPTION
For 2012, the same task definition was kept: it still requires participants to deploy multimodal features to automatically detect portions of movies containing violent material. Defining the term 'Violence' is not an easy task, as this notion remains subjective and thus dependent on people. Since 2011, the chosen definition is the following: violence is defined as "physical violence or accident resulting in human injury or pain". Any features automatically extracted from the provided video, including the subtitles, may be used by participants. No external additional data such as metadata collected from the Internet can be used in this task. Only the content of the movie extractable from DVDs is allowed for feature extraction.

3. DATA DESCRIPTION
With respect to the use case, the dataset selected for the developed corpus is a set of 18 Hollywood movies that must be purchased as DVDs by the participants. The movies are of different genres (from extremely violent movies to movies without violence). The content extractable from DVDs consists of information from different modalities, namely, at least visual information, audio signals and subtitles. From these 18 movies, 15 are dedicated to the training process: *Armageddon, Billy Elliot, Eragon, Harry Potter 5, I am Legend, Leon, Midnight Express, Pirates of the Caribbean 1, Reservoir Dogs, Saving Private Ryan, The Sixth Sense, the Wicker Man, Kill Bill 1, The Bourne Identity, the Wizard of Oz.* The remaining 3 movies, *Dead Poets Society, Fight Club and Independance Day*, will serve as the evaluation set. As in 2011, we tried to respect the genre repartition (from extremely violent to non violent) both in the training and evaluation sets.

4. GROUNDTRUTH
The ground truth[1] was created by 9 human assessors. In addition to segments containing physical violence (with the above definition), annotations also include high-level concepts for the visual and the audio modalities. Each annotated violent segment contains only one action, whenever it is possible. In the cases where different actions are overlapping, the whole segment is proposed with different actions. This was indicated in the annotation files by adding the tag "multiple action scene". Each violent segment is annotated at frame level, i.e. it is defined by its starting and ending video frame numbers.

Seven visual and three audio concepts are provided: *presence of blood, fights, presence of fire, presence of guns, pres-*

*The work that went into MediaEval 2012 has been supported, in part, by the Quaero Program http://www.quaero.org/.

[1]The annotations, shot detections and key frames for this task were made available by Technicolor. Any publication using these data should acknowledge Technicolor's contribution.

ence of cold weapons, car chases and gory scenes (for the video modality); *presence of screams, gunshots and explosions* (for the audio modality). Participants should note that they are welcome to carry out detection of the high-level concepts. However, concept detection is not the goal of the task and these high-level concept annotations are only provided for training purposes and only on the training set. For the video concepts, each of them follows the same annotation format as for violent segments, i.e. starting and ending frame numbers and possibly some additional tags. Regarding blood annotations, a proportion of the amount of blood in each segment is provided, as the percentage of the image surface covered by blood. Four different types of fights are annotated: only two people fighting, a small group of people (roughly less than 10), large group of people (more than 10), distant attack (i.e. no real fight but somebody is shot or attacked at distance). As for the presence of fire, anything from big fires and explosions to fire coming out of a gun while shooting, a candle, a cigarette lighter, a cigarette, or sparks was annotated, e.g. a space shuttle taking off also generates fire and receives fire label. An additional tag may indicate special colors of the fire (i.e. not yellow or orange). If a segment of video showed the presence of firearms (respectively cold weapons) it was annotated by any type of (parts of) guns (respectively cold weapons) or assimilated arms. By "cold weapon", we mean any weapon that does not involve fire or explosions as a result from the use of gun powder or other explosive materials. Annotations of gory scenes are more delicate. In the present task, they are indicating graphic images of bloodletting and/or tissue damage. It includes horror or war representations. As this is also a subjective and difficult notion to define, some additional segments showing really disgusting mutants or creatures are annotated as gore. In this case, additional tags describing the event/scene are added. For the audio concepts, each temporal segment is annotated with its starting and ending times in seconds, and an additional tag corresponding to the type of event, chosen from the list: *nothing, gunshot, canon fire, scream, scream effort, explosion, multiple actions, multiple actions canon fire, multiple actions scream effort*. Automatically generated shot boundaries with their corresponding key frames are also provided with each movie. Shot segmentation was carried out by Technicolor's software.

5. RUN DESCRIPTION

Participants can submit two types of runs: the required run or shot-classification run and the optional run which is the segment-level run. For the shot-classification run, participants are required to provide a violent scene detection at the shot level, according to the provided shot boundaries. Each shot should be classified as violent or non violent, with possibly a confidence score. As for the segment-level run, participants are required to, independently of shot boundaries, provide violent segments for each test movie. Once again, confidence scores may be added for each segment. In both cases, confidence scores are optional. However, providing a list of segments that covers the entire duration of the videos enables plotting of detection error trade-off curves based on the scores which should be of great interest to analyze and compare the different techniques. Hence, we encourage participants to do so. Scores will in any case not be used for the official performance evaluations which will be based solely on the decisions provided in the submitted resulting file.

6. EVALUATION CRITERIA

Several performance measures will be used for diagnostic purposes (false alarm and miss detection rates, AED-precision and recall as defined in [4], mean average precision, etc.). The MediaEval cost, used in 2011 as a basis for systems comparison, will still be computed for the sake of comparison with last year's results. However, as it has proven to be highly biased towards high precision rate values, it will be replaced by the computation of the mean average precision at the 20 top ranked violent shots. We nevertheless recall the definition of the MediaEval cost, which is a function weighting false alarms and missed detections, according to

$$C = C_{fa}P_{fa} + C_{miss}P_{miss} \quad (1)$$

with the costs $C_{fa} = 1$ and $C_{miss} = 10$. P_{fa} and P_{miss} are the estimated probabilities of resp. false alarm (false positive) and missed detection (false negative) given the system's output and the ground truth. In the shot classification, the false alarm and miss probabilities are then calculated on a per shot basis while, in the segment level run, they are computed on a per unit of time basis, i.e. durations of both references and detected segments are compared. To avoid only evaluating systems at given operating points and enable full comparison of the pros and cons of each system, we will use detection error trade-off (DET) curves whenever possible, plotting P_{fa} as a function of P_{miss} given a segmentation and a score for each segment, where the higher the score, the more likely the violence. Note that in the segment level run, DET curves are possible only for systems returning a dense segmentation (a list of segments that spans the entire video): segments not in the output list will be considered as non violent for all thresholds.

7. CONCLUSIONS

The Affect Task on Violent Scenes Detection in the context of the MediaEval 2012 benchmarking initiative has been presented. Dataset and groundtruth, specifications of the expected runs and evaluation criteria were detailed to give an overview of this new challenge.

8. REFERENCES

[1] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su. Violence detection in movies. In *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, pages 119 –124, aug. 2011.

[2] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In S. Konstantopoulos et al., editor, *Artificial Intelligence: Theories, Models and Applications*, volume 6040 of *Lecture Notes in Computer Science*, pages 91–100. Springer Berlin / Heidelberg, 2010.

[3] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies. In *ICASSP - 37th International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japon, 2012.

[4] A. Temko, C. Nadeu, and J.-I. Biel. Acoustic Event Detection: SVM-Based System and Evaluation Setup in CLEAR'07. In R. Stiefelhagen et al., R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans*, volume 4625 of *Lecture Notes in Computer Science*, pages 354–363. Springer Berlin / Heidelberg, 2008.