The MediaEval 2013 Affect Task: Violent Scenes Detection
Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Ionescu Bogdan, Vu Lam Quang, Yu-Gang Jiang

To cite this version:

HAL Id: hal-00932551
https://hal.archives-ouvertes.fr/hal-00932551
Submitted on 21 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The MediaEval 2013 Affect Task: Violent Scenes Detection

Claire-Hélène Demarty
Technicolor
Rennes, France
claire-helene.demarty@technicolor.com

Cédric Penet
Technicolor
Rennes, France
cedric.penet@technicolor.com

Markus Schedl
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

Bogdan Ionescu
University Polytechnica of Bucharest
Romania
bionescu@imag.pub.ro

Vu Lam Quang
University of Science, VNU-HCMC
Vietnam
lамquangvu@gmail.com

Yu-Gang Jiang
Fudan University
China
yugang.jiang@gmail.com

ABSTRACT
This paper provides a description of the MediaEval 2013 Affect Task Violent Scenes Detection. This task, which is proposed for the third year to the research community, 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, which has seen some changes in 2013. Dataset, annotations, and evaluation criteria as well as the required and optional runs are described.

1. INTRODUCTION
The Affect Task Violent Scenes Detection is part of the MediaEval 2013 benchmarking initiative for multimedia evaluation. The objective is to automatically detect violent segments in movies. This challenge is proposed for the third year in the MediaEval benchmark. It derives from a use case at Technicolor (http://www.technicolor.com), 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. In the literature, the detection of violence was not a lot studied [2, 1, 3], until recently when it has gained interest. As most of the proposed methods suffer from a lack of a common and consistent database, and usually use a limited development set, the task was launched to propose a public and common framework for the research community. This year, among other changes, two definitions of violence will be studied, an objective one and a subjective one (see below). The addition of a subjective definition was motivated by the fact that the one from 2012 has proven to lead to annotations which do not correspond to the use case.

2. TASK DESCRIPTION
The task requires participants to deploy multimodal features to automatically detect portions of movies containing violent material. For 2013, two definitions of violence are studied.

3. DATA DESCRIPTION
With respect to the use case, the dataset selected for the developed corpus is a set of 25 Hollywood movies that must be purchased as DVDs by the participants. The movies are of different genres and show different amounts of violence (from extremely violent movies to movies without violence). The content extractable from DVDs consists of information from different modalities, namely, visual information, audio signals and subtitles, and any additional metadata present in the DVDs. From these 25 movies, 18 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, Dead Poets Society, Fight Club and Independence Day. The remaining 7 movies, Fantastic Four, Fargo, Forrest Gump, Legally Blond, Pulp Fiction, The God Father 1 and The Pianist, will serve as the evaluation set.

4. GROUND TRUTH
The ground truth was created by several human assessors. In addition to segments containing physical violence (with the two above definitions), 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, presence 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 themselves. 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 by the following tags: unnoticeable, low, medium and high. 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 thus receives a 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 (or cold weapons) it was annotated by any type of (parts of) guns (or 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 indicated by graphic images of bloodletting and/or tissue damage. This 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.

As a first step towards a qualitative evaluation, participants are encouraged to present at the MediaEval workshop a video summary of the most violent scenes found by their algorithms. This will not be evaluated by the organizers this year, but it will serve as a first basis for future evolution of the task.

6. EVALUATION CRITERIA

As in 2012, the official evaluation metric will be the mean average precision at the N top ranked violent shots. Several performance measures will be used for diagnostic purposes (false alarm and miss detection rates, AED-precision and recall as defined in [4], the MediaEval cost, which is a function weighting false alarms (FA) and missed detections (MI), etc.). 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, 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. $P_{FA}$ and $P_{miss}$ are respectively the FA and MI rates given the system’s output and the reference annotation. In the shot classification, the FA and MI rates were calculated on a per shot basis while, in the segment level run, they were computed on a per unit of time basis, i.e., durations of both references and detected segments are compared. 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. REFERENCES