The MediaEval 2014 Affect Task: Violent Scenes Detection

Mats Sjöberg
Aalto University
Espoo, Finland
mats.sjoberg@aalto.fi

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

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

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

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

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

ABSTRACT

This paper provides a description of the MediaEval 2014 Affect Task: Violent Scenes Detection, which is running for the fourth year. The task originates from a use case at Technicolor[1]that aims to help users find suitable contents from a movie database. We provide insights on the use case, task challenges, data set and ground truth, required and optional participant runs and evaluation metrics.

1. INTRODUCTION

The Affect Task: Violent Scenes Detection is part of the MediaEval 2014 Benchmarking Initiative for Multimedia Evaluation. The objective of the task is to *automatically detect violent segments in movies*. This challenge is proposed for the fourth year in the MediaEval benchmark. It derives from a use case at Technicolor[1] that involves helping parents choosing movies that are suitable for their children with respect to their violence contents. Parents decide to select or reject movies after previewing the most violent parts of the movies.

In the literature, detection of violence in movies has been marginally addressed until recently [1, 2, 3]. As most of the proposed methods suffer from a lack of a consistent evaluation which usually requires the use of a constrained and closed dataset, the task's main objective is to propose a public common evaluation framework for the research in this area.

This year we concentrate on a subjective definition of violence that is closer to the considered use case than the more objective definition used in the previous editions. Another novelty is the addition of a new generalization task which transposes the detection to short web video footage. The idea is to assess how well approaches generalize to kinds of video material other than typical Hollywood movies. User-generated videos shared via on-line video platforms have been strongly gaining in popularity during the past couple of years. Taking such material into account is of vital interest for future research.

2. TASK DESCRIPTION

The task requires participants to deploy multimedia features to automatically detect movie segments that contain violent material. Segments are regarded as arbitrary length video time intervals, e.g., start – end frame. In contrast to previous years, video shot segmentation is no longer provided. Violence is being defined as content which "*one would not let an 8 years old child see in a movie because it contains physical violence*". To solve the task, participants are allowed to use either only features extracted from the original movie DVDs, or to use also additional external data, e.g., extracted from the web.

3. DATA DESCRIPTION

Two different data sets are proposed: (i) a set of 31 Hollywood movies whose DVDs must be purchased by the participants, for the main task and (ii) a set of 86 short YouTube[2] web videos under Creative Commons licenses that allow redistribution, for the generalization task.

3.1 Hollywood Movies

The proposed movies are of different genres and show different amounts of violence, from extremely violent movies to movies without violence. From the DVDs, participants can extract various information from different modalities, namely: visual (frames), audio (soundtracks) and text (subtitles and any additional metadata present in the DVDs).

From these 31 movies, 24 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"*, *"The Bourne Identity"*, *"The Wizard of Oz"*, *"Dead Poets Society"*, *"Fight Club"*, *"Independence Day"*, *"Fantastic Four 1"*, *"Fargo"*, *"Forrest Gump"*, *"Legally Blond"*, *"Pulp Fiction"*, *"The God Father 1"* and *"The Pianist"*. The remaining 7 movies, *"8 Mile"*, *"Braveheart"*, *"Desperado"*, *"Ghost in the Shell"*, *"Jumanji"*, *"Terminator 2"* and *"V for Vendetta"*, will serve as the test set for the actual benchmarking.

3.2 Web Videos

For the generalization task, we gathered 86 videos from YouTube, which are indicated by uploaders to fall under a Creative Commons license (total duration ca. 157 minutes). They vary in length between 6 seconds and 6 minutes. The dataset contains both violent and non-violent videos, from very diverse categories: video games, amateur videos of accidents, sport events, etc. Videos were retrieved with search queries reflecting violence, such as "killing video games" or

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

[1] http://www.technicolor.com/
[2] http://www.youtube.com/

"brutal accident". Results were then filtered for Creative Commons and short duration clips, and the final videos were manually selected from the remaining results. Along with the actual videos, we provide participants with a variety of metadata from YouTube, including YouTube-ID, upload date, title, description, keywords, duration, view counts, ratings, likes and dislikes.

This kind of video material is particularly challenging due to factors such as bad quality in general and worse quality than Hollywood movies, presence of different languages, overlay text, black framing of the actual frames, or other modifications of the raw video content.

4. GROUND TRUTH

This year ground truth (for test set and generalization task) was created by several human assessors[3] who followed the subjective definition of violence introduced in Section 2. The training data annotations (i.e., 24 movies) are the ones from the previous edition of the task [4]. This year's annotations consisted in the following protocol. Firstly, all the videos were annotated separately by two groups of annotators from two different countries. For each group, regular annotators labeled all the videos which were then reviewed by master annotators. Regular annotators were graduate students (typically single with no children) and master annotators were senior researchers (typically married with children). No discussions were held between annotators during the annotation process. Group 1 used 2 regular annotators and 1 master annotator. Group 2 used 5 regular annotators and 3 master annotators. Annotators labeled different sets of movies. In the end, each movie received 2 different annotations which were then merged by the master annotators. Secondly, the achieved annotations from the two groups were merged and reviewed once more by the task organizers. All the uncertain, e.g., borderline, cases were solved via panel discussions, involving different people from different countries, to avoid cultural bias in the annotations. A textual description was added to each segment to reflect the choices of the annotators. 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.

In addition to segments containing physical violence, annotations also include high-level concepts for the visual and audio modalities of the first 17 Hollywood movies in the training set. Seven visual concepts (*"presence of blood", "fights", "presence of fire", "presence of guns", "presence of cold weapons", "car chases"* and *"gory scenes"*) and three audio concepts (*"presence of screams", "gunshots"* and *"explosions"*) are provided. These are the concepts proposed in the previous editions of the task, see [4].

5. RUN DESCRIPTION

This year, there are two subtasks: the (i) main task, and the (ii) generalization task. In the main task participants are required to detect violence in the 7 Hollywood movies which serve as the test set. In the generalization task, participants are expected to use the same systems as for the main task, but this time to detect violence in the 86 YouTube videos provided by the organizers. The training data is the same for both subtasks.

Participants can submit two types of runs for each subtask: generated using official training data only, or using external sources (e.g., Internet). In all runs, participants are required to provide the violent segments by specifying the starting and ending time of each segment together with a confidence score (the higher the value, the more likely that the segment is violent).

6. EVALUATION CRITERIA

The official evaluation metric is the Mean Average Precision (MAP). In addition to this, for comparison reasons, metrics from the previous editions of the task will be computed as well, e.g., false alarm and miss detection rates, AED-precision and recall, the MediaEval cost, which is a function weighting false alarms (FA) and missed detections (MI), etc. To avoid evaluating systems only at a given operating point and enable full comparison of the pros and cons of each system, we use detection error trade-off (DET) curves, plotting the false reject rate as a function of the false positive rate, given a violence confidence score for each segment. The false reject and false positive rates are calculated on a per unit of time basis, i.e., durations of both references and detected segments are compared. Segments not in the output list are considered as non-violent.

7. CONCLUSIONS

The Affect Task: Violent Scenes Detection provides participants with a comparative and collaborative evaluation framework for violence detection in movies. This year in particular, the task explores also the generalization of such systems to web footage. Details on the methods and results of each individual team can be found in the working note papers of the participating teams in these proceedings.

Acknowledgments

This task is supported by the following projects: Academy of Finland grants 255745 and 251170, UEFISCDI SCOUTER grant 28DPST/30-08-2013, National Natural Science Foundation of China grants 61201387 and 61228205, China's National 973 Program 2010CB327900, Vietnam National University Ho Chi Minh City grant B2013-26-01, Austrian Science Fund P25655 and EU FP7-ICT-2011-9 project 601166.

8. REFERENCES

[1] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, and Patrick Gros, "Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies", IEEE ICASSP, 2012.

[2] Bogdan Ionescu, Jan Schlüter, Ionut Mironica, and Markus Schedl, "A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies", ACM ICMR, pp. 215-222, 2013.

[3] Esra Acar, Frank Hopfgartner, and Sahin Albayrak, "Violence Detection in Hollywood Movies by the Fusion of Visual and Mid-level Audio Cues", ACM Multimedia, pp. 717-720, 2013.

[4] Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Bogdan Ionescu, Vu Lam Quang, and Yu-Gang Jiang, "The MediaEval 2013 Affect Task: Violent Scenes Detection", CEUR-WS, Vol. 1043, `http://ceur-ws.org/Vol-1043/mediaeval2013_submission_4.pdf`, Spain, 2013.

[3] annotations were made available by Fudan University, Vietnam University of Science, and Technicolor. Any publication using these data should acknowledge these institutions.