The MediaEval 2015 Affective Impact of Movies Task

Mats Sjöberg[1], Yoann Baveye[2], Hanli Wang[3], Vu Lam Quang[4], Bogdan Ionescu[5], Emmanuel Dellandréa[6], Markus Schedl[7], Claire-Hélène Demarty[2], and Liming Chen[6]

[1]Helsinki Institute for Information Technology HIIT, University of Helsinki, Finland, mats.sjoberg@helsinki.fi

[2]Technicolor, France, [yoann.baveye,claire-helene.demarty]@technicolor.com

[3]Tongji University, China, hanliwang@tongji.edu.cn

[4]University of Science, VNU-HCMC, Vietnam, lamquangvu@gmail.com

[5]University Politehnica of Bucharest, Romania, bionescu@imag.pub.ro

[6]Ecole Centrale de Lyon, France, emmanuel.dellandrea@ec-lyon.fr, liming.chen@liris.cnrs.fr

[7]Johannes Kepler University, Linz, Austria, markus.schedl@jku.at

ABSTRACT

This paper provides a description of the MediaEval 2015 "Affective Impact of Movies Task", which is running for the fifth year, previously under the name "Violent Scenes Detection". In this year's task, participants are expected to create systems that automatically detect video content that depicts violence, or predict the affective impact that video content will have on viewers. Here we provide insights on the use case, task challenges, data set and ground truth, task run requirements and evaluation metrics.

1. INTRODUCTION

The Affective Impact of Movies Task is part of the MediaEval 2015 Benchmarking Initiative. The overall use case scenario of the task is to design a video search system that uses automatic tools to help users find videos that fit their particular mood, age or preferences. To address this, we present two subtasks:

- *Induced affect detection*: the emotional impact of a video or movie can be a strong indicator for search or recommendation;
- *Violence detection*: detecting violent content is an important aspect of filtering video content based on age.

This task builds on the experiences from previous years' editions of the Affect in Multimedia Task: Violent Scenes Detection. However, this year, we introduce a completely new subtask for detecting the emotional impact of movies. In addition, we are introducing to MediaEval a newly extended data set consisting of 10,900 short video clips extracted from 199 Creative Commons-licensed movies.

In the literature, detection of violence in movies has been marginally addressed until recently [8, 6, 1]. Similarly, in affective video content analysis it has been repeatedly claimed that the field would highly benefit from a standardised evaluation data set [5, 9]. Most of the previously proposed methods for affective impact or violence detection suffer from a lack of a consistent evaluation, which usually requires the use of a constrained and closed data set [4, 7, 10]. Hence, the task's main objective is to propose a public common evaluation framework for the research in these closely-related areas.

2. TASK DESCRIPTION

The task requires participants to deploy multimedia features to automatically detect violent content and emotional impact of short movie clips. In contrast to previous years, the task no longer considers arbitrary starting and ending points of detected segments, but instead the short video clips are considered as single units for detection purposes with a single judgement per clip. This year, there are two subtasks: (i) induced affect detection, and (ii) violence detection. Both tasks use the same videos for training and testing.

For the induced affect detection task, participants are expected to predict, for each video, its valence class (*i.e.*, into one of negative, neutral or positive) and arousal class (*i.e.*, into one of calm, neutral or active). In this task, we are focusing on *felt emotion*, *i.e.*, the actual emotion of the viewer when watching the video clip, rather than for example what the viewer believes that he or she is expected to feel. Valence is defined as a continuous scale from most negative to most positive emotion, while arousal is defined continuously from most calm to most active emotion. However, to keep the two subtasks compatible and enable participants to use similar systems for both tasks, we have here opted to discretise the two scales into three classes as follows:

- valence: *negative*, *neutral*, and *positive*,
- arousal: *calm*, *neutral*, and *active*.

For the violence detection task, participants are expected to classify each video as violent or non-violent. Violence is defined as content that "*one would not let an 8 years old child see in a movie because it contains physical violence*".

To solve the task, participants are only allowed to use features extracted from the original video files, or metadata provided by the organisers. In addition, there is a possibility to use external data for runs which are specifically marked, however, at least one run for each subtask must be without any external data.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

3. DATA DESCRIPTION

This year a single data set is proposed: 10,900 short video clips extracted from 199 Creative Commons-licensed movies of various genres. The movies are split into a development set – intended for training and validation – and a test set as 100, respectively 99 movies, resulting in 6,144 respectively 4,756 extracted short video clips.

The proposed data set is actually an extension of the LIRIS-ACCEDE data set originally composed of 9,800 excerpts extracted from 160 movies [3]. For this task, 1,100 additional video clips have been extracted from 39 new movies and included in the test set. The selected feature films and short films can be considered professionally made or amateur movies but almost all are indexed on video platforms referencing best free-to-share movies or have been screened during film festivals. Since these movies are shared under Creative Commons licenses, the excerpts can also be shared and downloaded along with the annotations without infringing copyright. The excerpts have been extracted from the movies so that they last between 8 and 12 seconds and start and end with a cut or a fade.

Along with the video material and the annotations, features extracted from each video clip are also provided by the organisers. They correspond to the audiovisual features described in [3].

4. GROUND TRUTH

For each of the 10,900 video clips, the ground truth consists of: a binary value to indicate the presence of violence, the class of the excerpt for felt arousal (calm-neutral-active), and the class for felt valence (negative-neutral-positive). Before the evaluation, participants are provided only with the annotations for the development set, while those for the test set are held back to be used for benchmarking the submitted results.

The original video clips included in the LIRIS-ACCEDE data set were all already ranked along the felt valence and arousal axes by using a crowdsourcing protocol [3]. Pairwise comparisons were generated using the quicksort algorithm and presented to crowdworkers who had to select the video inducing the calmer emotion or the more positive emotion. In [2] the crowdsourced ranks were converted into absolute affective scores ranging from -1 to 1, which have been used to define the three classes for each affective axis for the MediaEval task. The negative and calm classes correspond respectively to the video clips with a valence or arousal score smaller than -0.15, the neutral class for both axes is assigned to the videos with an affective score between -0.15 and 0.15, and the positive and active classes are assigned to the videos with an affective score higher than 0.15. These limits have been defined empirically taking into account the distribution of the data set in the valence-arousal space.

For the 2015 MediaEval evaluation the test set was extended with an additional 1,100 video clips. Due to time and resource constraints, these were annotated using a simplified scheme which takes advantage of the fact that we do not need a full ranking of the new video clips, but only to separate them into three classes for each affect axis. Two pivot videos were selected for each axis, which had absolute scores very close to the -0.15 and 0.15 class boundaries. The annotation task could then be formulated as comparing each video clip to these pivot videos, and thus place them in their correct class. In total 17 annotators were involved from five different countries, and three judgements were collected for each pivot/affect dimension pair. Out of these three judgements the majority vote was selected.

For the violence detection the annotation process was similar to previous years' protocol. Firstly, all the videos were annotated separately by two groups of annotators from two different countries. For each group, regular annotators labelled 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 12 regular and 2 master annotators, while Group 2 used 5 regular and 2 master annotators. Within each group, each video received 2 different annotations which were then merged by the master annotators into the final annotation for the group. Finally, the achieved annotations from the two groups were merged and reviewed once more by the task organisers.

5. RUN DESCRIPTION

Participants can submit up to 5 runs for each subtask: induced affect detection and violence detection. Each subtask has a required run which uses no external training data, only the provided development data is allowed. Also any features that can be automatically extracted from the video are allowed. Both tasks also have the possibility for optional runs in which any external data can be used, such as Internet sources, as long as they are marked as "external data" runs.

6. EVALUATION CRITERIA

For the induced affect detection subtask the official evaluation measure is *global accuracy*, calculated separately for valence and arousal dimensions. Global accuracy is the proportion of the returned video clips that have been assigned to the correct class (out of the three classes).

The official evaluation metric for the violence detection subtask is *average precision*, which is calculated using the `trec_eval` tool provided by NIST[1]. This tool also produces a set of commonly used metrics such as precision and recall, which may be used for comparison purposes.

7. CONCLUSIONS

The Affective Impact of Movies Task provides participants with a comparative and collaborative evaluation framework for violence and emotion detection in movies. The introduction of the induced affect detection subtask is a new effort for this year. In addition, we have started fresh with a data set not used in MediaEval before, which consists of short Creative Commons-licensed video clips, which enables legally sharing the data directly with participants. Details on the methods and results of each individual team can be found in the papers of the participating teams in these proceedings.

Acknowledgments

This task is supported by the following projects: ERA-NET CHIST-ERA grant ANR-12-CHRI-0002-04, UEFISCDI SCOUTER grant 28DPST/30-08-2013, Vietnam National University Ho Chi Minh City grant B2013-26-01, Austrian Science Fund P25655, and EU FP7-ICT-2011-9 project 601166.

[1] `http://trec.nist.gov/trec_eval/`

8. REFERENCES

[1] E. Acar, F. Hopfgartner, and S. Albayrak. Violence detection in hollywood movies by the fusion of visual and mid-level audio cues. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 717–720. ACM, 2013.

[2] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. From crowdsourced rankings to affective ratings. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, July 2014.

[3] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, Jan 2015.

[4] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, Feb. 2005.

[5] M. Horvat, S. Popovic, and K. Cosic. Multimedia stimuli databases usage patterns: a survey report. In *Proceedings of the 36nd International ICT Convention MIPRO*, pages 993–997, 2013.

[6] B. Ionescu, J. Schlüter, I. Mironica, and M. Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *ICMR*, pages 215–222, 2013.

[7] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *IEEE Transactions on Multimedia*, 12(6):523–535, Oct. 2010.

[8] 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, 2012.

[9] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic. Corpus development for affective video indexing. *IEEE Transactions on Multimedia*, 16(4):1075–1089, June 2014.

[10] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia*, 12(6):510–522, Oct. 2010.