The MediaEval 2016 Emotional Impact of Movies Task

Emmanuel Dellandréa[1], Liming Chen[1], Yoann Baveye[2], Mats Sjöberg[3] and Christel Chamaret[4]
[1]Ecole Centrale de Lyon, France, {emmanuel.dellandrea, liming.chen}@ec-lyon.fr
[2]Université de Nantes, France, yoann.baveye@univ-nantes.fr
[3]HIIT, University of Helsinki, Finland, mats.sjoberg@helsinki.fi
[4]Technicolor, France, christel.chamaret@technicolor.com

ABSTRACT

This paper provides a description of the MediaEval 2016 "Emotional Impact of Movies" task. It continues builds on previous years' editions of the Affect in Multimedia Task: Violent Scenes Detection. However, in this year's task, participants are expected to create systems that automatically predict the emotional impact that video content will have on viewers, in terms of valence and arousal scores. Here we provide insights on the use case, task challenges, dataset and ground truth, task run requirements and evaluation metrics.

1. INTRODUCTION

Affective video content analysis aims at the automatic recognition of emotions elicited by videos. It has a large number of applications, including mood based personalized content recommendation [5] or video indexing [12], and efficient movie visualization and browsing [13]. Beyond the analysis of existing video material, affective computing techniques can also be used to generate new content, e.g., movie summarization [7], or personalized soundtrack recommendation to make user-generated videos more attractive [9]. Affective techniques can also be used to enhance the user engagement with advertising content by optimizing the way ads are inserted inside videos [11].

While major progress has been achieved in computer vision for visual object detection, scene understanding and high-level concept recognition, a natural further step is the modeling and recognition of affective concepts. This has recently received increasing interest from research communities, e.g., computer vision, machine learning, with an overall goal of endowing computers with human-like perception capabilities. Thus, this task is proposed to offer researchers a place to compare their approaches for the prediction of the emotional impact of movies. It continues builds on previous years' editions of the Affect in Multimedia Task: Violent Scenes Detection [10].

2. TASK DESCRIPTION

The task requires participants to deploy multimedia features to automatically predict the emotional impact of movies. We are focusing on felt emotion, i.e., the actual emotion of the viewer when watching the video, rather than for example what the viewer believes that he or she is expected to feel. The emotion is considered in terms of valence and arousal [8]. Valence is defined as a continuous scale from most negative to most positive emotions, while arousal is defined continuously from calmest to most active emotions. Two subtasks are considered:

1. Global emotion prediction: given a short video clip (around 10 seconds), participants' systems are expected to predict a score of induced valence (negative-positive) and induced arousal (calm-excited) for the whole clip;

2. Continuous emotion prediction: as an emotion felt during a scene may be influenced by the emotions felt during the previous ones, the purpose here is to consider longer videos, and to predict the valence and arousal continuously along the video. Thus, a score of induced valence and arousal should be provided for each 1s-segment of the video.

3. DATA DESCRIPTION

The development dataset used in this task is the LIRIS-ACCEDE dataset (liris-accede.ec-lyon.fr) [3]. It is composed of two subsets. The first one, used for the first subtask (global emotion prediction), contains 9,800 video clips extracted from 160 professionally made and amateur movies, with different genres, and shared under Creative Commons licenses that allows to freely use and distribute videos without copyright issues as long as the original creator is credited. The segmented video clips last between 8 and 12 seconds and are representative enough to conduct experiments. Indeed, the length of extracted segments is large enough to get consistent excerpts allowing the viewer to feel emotions and is also small enough to make the viewer feel only one emotion per excerpt. A robust shot and fade in/out detection has been implemented using to make sure that each extracted video clip start and end with a shot or a fade. Several movie genres are represented in this collection of movies such as horror, comedy, drama, action and so on. Languages are mainly English with a small set of Italian, Spanish, French and others subtitled in English.

The second part of LIRIS-ACCEDE dataset is used for the second subtask (continuous emotion prediction). It consists in a selection of movies among the 160 ones used to extract the 9,800 video clips mentioned previously. The total length of the selected movies was the only constraint. It had to be smaller than eight hours to create an experiment of acceptable duration. The selection process ended with the choice

Copyright is held by the author/owner(s).
MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

of 30 movies so that their genre, content, language and duration are diverse enough to be representative of the original LIRIS-ACCEDE dataset. The selected videos are between 117 and 4,566 seconds long (mean = 884.2sec ± 766.7sec SD). The total length of the 30 selected movies is 7 hours, 22 minutes and 5 seconds.

In addition to the development set, a test set is also provided to assess participants' methods performance. 49 new movies under Creative Commons licenses have been considered. With the same protocol as the one used for the development set, 1,200 additional short video clips have been extracted for the first subtask (between 8 and 12 seconds), and 10 long movies (from 25 minutes to 1 hour and 35 minutes) have been selected for the second subtask (for a total duration of 11.48 hours).

In solving the task, participants are expected to exploit the provided resources. Use of external resources (e.g., Internet data) will be however allowed as specific runs.

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

4. GROUND TRUTH

4.1 Ground Truth for the first subtask

The 9,800 video clips included in the first part of the LIRIS-ACCEDE dataset are ranked along the felt valence and arousal axes by using a crowdsourcing protocol [3]. To make reliable annotations as simple as possible, pairwise comparisons were generated using the quicksort algorithm and presented to crowdworkers who had to select the video inducing the calmest emotion or the most positive emotion.

To cross-validate the annotations gathered from various uncontrolled environments using crowdsourcing, another experiment has been created to collect ratings for a subset of the database in a controlled environment. In this controlled experiment, 28 volunteers were asked to rate a subset of the database carefully selected using the 5-point discrete Self-Assessment-Manikin scales for valence and arousal [4]. 20 excerpts per axis that are regularly distributed have been selected in order to get enough excerpts to represent the whole database while being relatively few to create an experiment of acceptable duration.

From the original ranks and these ratings, absolute affective scores for valence and arousal have been estimated for each of the 9,800 video clips using Gaussian process regression models as described in [1].

To obtain ground truth for the test subset, each of the 1,200 additional video clips has first been ranked according to the 9,800 video clips from the original dataset. Then, its valence and arousal ranks have been converted into a valence and arousal score using the regression models mentioned previously.

4.2 Ground Truth for the second subtask

In order to collect continuous valence and arousal annotations, 16 French participants had to continuously indicate their level of arousal while watching the movies using a modified version of the GTrace annotation tool [6] and a joystick (10 participants for the development set and 6 for the test set). Movies have been divided into two subsets. Each annotator continuously annotated one subset along the induced valence and the other into the induced arousal. Thus, each movie has been continuously annotated by five annotators for the development set, and three for the test set.

Then, the continuous valence and arousal annotations from the participants have been down-sampled by averaging the annotations over windows of 10 seconds with 1 second overlap (i.e., 1 value per second) in order to remove the noise due to unintended moves of the joystick. Finally, these post-processed continuous annotations have been averaged in order to create a continuous mean signal of the valence and arousal self-assessments. The details of this processing are given in [2].

5. RUN DESCRIPTION

Participants can submit up to 5 runs for the first subtask (global emotion prediction). For the second subtask (continuous emotion prediction), there can be 2 types of run submissions: full runs that concerns the whole test set (the 10 movies, total duration: 11.48 hours) and light runs that concern a subset of the test set (5 movies, total duration: 4.82 hours). In each case (light and full), up to 5 runs can be submitted. Moreover, each subtask has a required run which uses no externale 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

Standard evaluation metrics (Mean Square Error and Pearson's Correlation Coefficient) are used to assess systems performance. Indeed, the common measure generally used to evaluate regression models is the Mean Square Error (MSE). However, this measure is not always sufficient to analyze models efficiency and the correlation may be required to obtain a deeper performance analysis. As an example, if a large portion of the data is neutral (i.e., its valence score is close to 0.5) or is distributed around the neutral score, a uniform model that always outputs 0.5 will result in good MSE performance (low MSE). In this case, the lack of accuracy of the model will be brought to the fore by the correlation between the predicted values and the ground truth that will be also very low.

7. CONCLUSIONS

The Emotional Impact of Movies Task provides participants with a comparative and collaborative evaluation framework for emotional detection in movies, in terms of valence and arousal scores. The LIRIS-ACCEDE dataset [1] has been used as development set, and additional movies under Creative Commons licenses and ground truth annotations have been provided as test set. Details on the methods and results of each individual team can be found in the papers of the participating teams in the MediaEval 2016 workshop proceedings.

8. ACKNOWLEDGMENTS

This task is supported by the CHIST-ERA Visen project ANR-12-CHRI-0002-04.

[1] http://liris-accede.ec-lyon.fr

9. REFERENCES

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

[2] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.

[3] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 2015.

[4] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 1994.

[5] L. Canini, S. Benini, and R. Leonardi. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013.

[6] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton. Gtrace: General trace program compatible with emotionml. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013.

[7] H. Katti, K. Yadati, M. Kankanhalli, and C. TatSeng. Affective video summarization and story board generation using pupillary dilation and eye gaze. In *IEEE International Symposium on Multimedia (ISM)*, 2011.

[8] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 2003.

[9] R. R. Shah, Y. Yu, and R. Zimmermann. Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *ACM International Conference on Multimedia*, 2014.

[10] M. Sjöberg, Y. Baveye, H. Wang, V. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The mediaeval 2015 affective impact of movies task. In *MediaEval 2015 Workshop*, 2015.

[11] K. Yadati, H. Katti, and M. Kankanhalli. Cavva: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia*, 2014.

[12] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia*, 2010.

[13] S. Zhao, H. Yao, X. Sun, X. Jiang, and P. Xu. Flexible presentation of videos based on affective content analysis. *Advances in Multimedia Modeling*, 2013.