The MediaEval 2017 Emotional Impact of Movies Task

Emmanuel Dellandréa[1], Martijn Huigsloot[2], Liming Chen[1], Yoann Baveye[3] and Mats Sjöberg[4]
[1]Ecole Centrale de Lyon, France, {emmanuel.dellandrea, liming.chen}@ec-lyon.fr
[2]NICAM, Netherlands, Huigsloot@nicam.nl
[3]Université de Nantes, France, yoann.baveye@univ-nantes.fr
[4]HIIT, University of Helsinki, Finland, mats.sjoberg@helsinki.fi

ABSTRACT

This paper provides a description of the MediaEval 2017 "Emotional Impact of Movies task". It continues to build on previous years' editions. 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, arousal and fear. Here we provide a description of 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 [3] or video indexing [13], and efficient movie visualization and browsing [14]. Beyond the analysis of existing video material, affective computing techniques can also be used to generate new content, e.g., movie summarization [9], or personalized soundtrack recommendation to make user-generated videos more attractive [11]. Affective techniques can also be used to enhance the user engagement with advertising content by optimizing the way ads are inserted inside videos [12].

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 to build on previous years' editions [5] with a first subtask, which is a mix of last year's tasks related to valence and arousal prediction, and a new subtask dedicated to fear prediction.

2 TASK DESCRIPTION

The task requires participants to deploy multimedia features and models to automatically predict the emotional impact of movies. This emotional impact is considered here to be the prediction of the expected emotion. The expected emotion is the emotion that the majority of the audience feels in response to the same content. In other words, the expected emotion is the expected value of experienced (i.e. induced) emotion in a population. While the induced emotion is subjective and context dependent, the expected emotion can be considered objective, as it reflects the more-or-less unanimous response of a general audience to a given stimulus [8].

This year, two new scenarios are proposed as subtasks. In both cases, long movies are considered and the emotional impact has to be predicted for consecutive 10-second segments sliding over the whole movie with a shift of 5 seconds:

(1) Valence/Arousal prediction: participants' systems are supposed to predict a score of expected valence and arousal for each consecutive 10-second segments. 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 [10];
(2) Fear prediction: the purpose here is to predict for each consecutive 10-second segments whether they are likely to induce fear or not. The targeted use case is the prediction of frightening scenes to help systems protecting children from potentially harmful video content. This subtask is complementary to the valence/arousal prediction task in the sense that the mapping of discrete emotions into the 2D valence/arousal space is often overlapped (for instance, fear, disgust and anger are overlapped since they are characterized with very negative valence and high arousal) [7].

3 DATA DESCRIPTION

The dataset used in this task is the LIRIS-ACCEDE dataset[1]. It contains videos from a set of 160 professionally made and amateur movies, shared under Creative Commons licenses that allow redistribution [2]. 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 continuous part of LIRIS-ACCEDE [1] is used as the development test for both subtasks. It consists of a selection of 30 movies. 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.

The test set consists of a selection of 14 movies other than the selection of the 160 original movies. They are between 210 and 6,260 seconds long (mean = 2045.2sec ± 2450.1sec SD). The total length of the 14 selected movies is 7 hours, 57 minutes and 13 seconds.

In addition to the video data, participants are also provided with general purpose audio and visual content features. To compute audio features, movies have first been processed to extract consecutive 10-second segments sliding over the whole movie with a shift of 5 seconds. Then, audio features have been extracted from these

Copyright held by the owner/author(s).
MediaEval'17, 13-15 September 2017, Dublin, Ireland

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

segments using openSmile toolbox[2] [6]. The default configuration named "emobase2010.conf" was used. It allows the computation of 1,582 features, which result from a base of 34 low-level descriptors (LLD) with 34 corresponding delta coefficients appended, and 21 functionals applied to each of these 68 LLD contours (1 428 features). In addition, 19 functionals are applied to the 4 pitch-based LLD and their four delta coefficient contours (152 features). Finally the number of pitch onsets (pseudo syllables) and the total duration of the input are appended (2 features).

Beyond audio features, for each movie, image frames were extracted every one second. For each of these images, several general purpose visual features have been provided. They have been computed using LIRE library[3], except CNN features (VGG16 fc6 layer) that have been extracted using Matlab Neural Networks toolbox[4]. The visual features are the following: Auto Color Correlogram, Color and Edge Directivity Descriptor, Color Layout, Edge Histogram, Fuzzy Color and Texture Histogram, Gabor, Joint descriptor joining CEDD and FCTH in one histogram, Scalable Color, Tamura, Local Binary Patterns, VGG16 fc6 layer.

4 GROUND TRUTH

Annotations are provided to participants for the 30 movies from the development set. Thus, for each movie, a first file contains valence and arousal values for consecutive 10-second segments sliding over the whole movie with a shift of 5 seconds, and a second file contains the indication whether these segments are supposed to induce fear (value 1) or not (value 0).

4.1 Ground Truth for the first 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 [4] 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 considering the induced valence and the other subset considering 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 a shift of 1 second overlap (i.e., 1 value per second) in order to remove the noise due to unintended movements 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 [1]. For the purpose of the first subtask, these values have been averaged to obtain a single value of valence and a single value of arousal for every consecutive 10-second segments sliding over the whole movie with a shift of 5 seconds.

4.2 Ground Truth for the second subtask

Fear annotations for the second subtask were generated using a tool specifically designed for the classification of audio-visual media allowing to perform annotation while watching the movie (at the same time). The annotations have been realized by two well experienced team members of NICAM[5] both of them trained in classification of media. Each movie has been annotated by 1 annotator reporting the start and stop times of each sequence in the movie exptected to induce fear. From this information, the 10-second segments sliding over the whole movie with a shift of 5 seconds have been labeled as fear (value 1) if they intersect one of the fear sequences and as not fear (value 0) otherwise.

5 RUN DESCRIPTION

Participants can submit up to 5 runs for each of the two subtasks, so 10 runs in total. Models can rely on the features provided by the organizers or any other external data.

6 EVALUATION CRITERIA

Standard evaluation metrics are used to assess systems performance. The first subtask can be considered as a regression problem (estimation of expected valence and arousal scores) while the second subtask can be seen as a binary classification problem (the video segment is supposed to induce/not induce fear).

For the first subtask, the official metric is the Mean Square Error (MSE), which is the common measure generally used to evaluate regression models. However, to allow a deeper understanding of systems' performance, we also consider Pearson's Correlation Coefficient. Indeed, MSE 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.

For the second subtask, the official metric is the Mean Average Precision (MAP). Moreover, Accuracy, Precision, Recall and F1-score are also considered to provide insights into systems behaviours.

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, arousal and fear. The LIRIS-ACCEDE dataset has been used as development and test sets. Details on the methods and results of each individual team can be found in the papers of the participating teams in the MediaEval 2017 workshop proceedings.

ACKNOWLEDGMENTS

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

[2] http://audeering.com/technology/opensmile/

[3] http://www.lire-project.net/

[4] https://www.mathworks.com/products/neural-network.html

[5] http://www.kijkwijzer.nl/nicam

REFERENCES

[1] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. 2015. Deep Learning vs. Kernel Methods: Performance for Emotion Prediction in Videos. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII).*

[2] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. 2015. LIRIS-ACCEDE: A Video Database for Affective Content Analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[3] L. Canini, S. Benini, and R. Leonardi. 2013. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology* 23, 4 (2013), 636–647.

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

[5] E. Dellandréa, L. Chen, Y. Baveye, M. Sjöberg, and C. Chamaret. 2016. The MediaEval 2016 Emotional Impact of Movies Task. In *MediaEval 2016 Workshop.*

[6] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *ACM Multimedia (MM), Barcelona, Spain.*

[7] L.-A. Feldman. 1995. Valence focus and arousal focus: Individual differences in the structure of affective experience. 69 (1995), 153–166.

[8] A. Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine* (2006).

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

[10] J. A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review* (2003).

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

[12] K. Yadati, H. Katti, and M. Kankanhalli. 2014. Cavva: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia* 16, 1 (2014), 15–23.

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

[14] S. Zhao, H. Yao, X. Sun, X. Jiang, and P. Xu. 2013. Flexible presentation of videos based on affective content analysis. *Advances in Multimedia Modeling* 7732 (2013).