MediaEval 2017 Predicting Media Interestingness Task

Claire-Hélène Demarty[1], Mats Sjöberg2, Bogdan Ionescu[3], Thanh-Toan Do[4],
Michael Gygli[5], Ngoc Q. K. Duong[1]
[1]Technicolor, Rennes, France
[2]HIIT, University of Helsinki, Finland
[3]LAPI, University Politehnica of Bucharest, Romania
[4]University of Science, Vietnam, University of Adelaide, Australia
[5]ETH Zurich, Switzerland & Gifs.com, US

ABSTRACT

NOT CHANGED - copy-past from last year This paper provides an overview of the Predicting Media Interestingness task that is organized as part of the MediaEval 2017 Benchmarking Initiative for Multimedia Evaluation. The task, which is running for the second year, expects participants to create systems that automatically select images and video segments that are considered to be the most interesting for a common viewer. In this paper, we present the task use case and challenges, the proposed data set and ground truth, the required participant runs and the evaluation metrics.

1 INTRODUCTION

Bogdan, could you help for this?

2 TASK DESCRIPTION

NOT REALLY CHANGED - This comes from the wiki. Ngoc, if you want to rephrase, please help.

The Predicting Media Interestingness Task was proposed for the first time last year. This year's edition is a follow-up which builds incrementally upon the previous experience. The task requires participants to automatically select images and/or video segments which are considered to be the most interesting for a common viewer. Interestingness of media is to be judged based on visual appearance, audio information and text accompanying the data, including movie metadata. To solve the task, participants are strongly encouraged to deploy multimodal approaches.

As in 2016, interestingness should be assessed according to a practical use case in industry, in particular at Technicolor[1], which involves helping professionals to illustrate a Video on Demand (VOD) web site by selecting some interesting frames and/or video excerpts for the movies. The frames and excerpts should be suitable in terms of helping a user to make his/her decision about whether he/she is interested in watching the whole movie. Once again, two subtasks will be offered to participants, which correspond to two types of available media content, namely images and videos. Participants are encouraged to submit to both subtasks. In both cases, the task is a binary classification task and prediction will be carried out on a per movie basis.

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

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

- *Predicting Image Interestingness* Given a set of key-frames extracted from a certain movie, the task involves automatically identifying those images that viewers report to be interesting. To solve the task, participants can make use of visual content as well as accompanying metadata, e.g., Internet data about the movie, social media information, etc.
- *Predicting Video Interestingness* Given a set of video segments extracted from a certain movie, the task involves automatically identifying the segments that viewers report to be interesting. To solve the task, participants can make use of visual and audio data as well as accompanying metadata, e.g., subtitles, Internet data about the movie, etc.

3 DATA DESCRIPTION

The data is extracted from Creative Commons licensed Hollywood-like videos: 103 movie trailers and 4 continuous extracts of ca. 15min from full-length movies.

For the video interestingness subtask, the data consists of video segments obtained after a manual segmentation. These segments correspond to shots (video shots are the continuous frame sequences recorded between a camera turn on and off) for all videos but four. Their average duration is of one second. The four last videos which correspond to the full-length movie extracts cited above, were manually segmented into longer segments with an average duration of ***, to better take into account the a certain unity of meaning and the audio information of the resulting segments. For the image subtask, the data consists of collections of key-frames extracted from the video segments used for the video subtask (one key-frame per segment). This will allow comparing results from both subtasks. The extracted key-frame corresponds to the frame in the middle of each video segment. In total, 7,396 video segments and 7,396 key-frames are released in the development set, whereas the test set consists of *** video segments and the same number of key-frames.

*** Here we could just copy-past what was said for the provided features last year or rewrite everything. Should we talk about the bug that was discovered in the face-tracking feature and the advice not to take into account negative values ? I am not sure of this. I am copying text from last year below. If we agree that this needs rephrasing, Michael, can you help for this part?***

To facilitate participation from various communities, we also provide some pre-computed content descriptors, namely: *low level features — dense SIFT* (Scale Invariant Feature Transform) which are computed following the original work in [5], except that the local

frame patches are densely sampled instead of using interest point detectors. A codebook of 300 codewords is used in the quantization process with a spatial pyramid of three layers [4]; *HoG descriptors* (Histograms of Oriented Gradients) [1] are computed over densely sampled patches. Following [8], HoG descriptors in a 2×2 neighborhood are concatenated to form a descriptor of higher dimension; *LBP* (Local Binary Patterns) [6]; *GIST* are computed based on the output energy of several Gabor-like filters (8 orientations and 4 scales) over a dense frame grid like in [7]; *color histogram* computed in the HSV space (Hue-Saturation-Value); *MFCC* (Mel-Frequency Cepstral Coefficients) computed over 32ms time-windows with 50% overlap. The cepstral vectors are concatenated with their first and second derivatives; *fc7 layer* (4,096 dimensions) and *prob layer* (1,000 dimensions) of AlexNet [3]; *mid level face detection and tracking related features*[2] — obtained by face tracking-by-detection in each video shot with a HoG detector [1] and the correlation tracker proposed in [2].

*** Michael, could you add a few words about your C3D feature? Your reference is also missing.***

4 GROUND TRUTH

Mats, could you help here? ** Do not forget to say that the question has changed in the web tool **

Otherwise, i think it is exactly the same protocol as last year, except that we compute the final BTL on all the new data with an initialization from last year ranking. 1/ This was decided because of the cheating process to help remove the false data 2/ We checked visually that the rankings were ok and even a little better for some of the videos compared to what was obtained in 2016 (but very similar to what would have been obtained this year by following the same iterative process to compute the BTL) => this validates both ways of computing the BTL (one against the other) + the fact that the increased number of iterations did improve the ranking at least for the images.

Should we talk about the anti-cheating measures?

Plus add a few figures as last year about the annotators population.

5 RUN DESCRIPTION

Every team can submit up to 10 runs, 5 per subtask. For each subtask, a required run is defined: *Image subtask - required run:* classification is to be achieved with the use of the visual information. External data is allowed. *Video subtask - required run:* classification is to be achieved with the use of *both* audio and visual information. External data is allowed.

Apart from these required runs, any additional run for each subtask will be considered as a general run, i.e., anything is allowed, both from the method point of view and the information sources.

6 EVALUATION

For both subtasks, the official evaluation metric will be the mean average precision at 10 (MAP@10) computed over all videos, and over the top 10 best ranked images/video shots. MAP@10 is selected because it reflects the VOD use case, where the goal is to select a small set of the most interesting images or video segments for each movie. To provide a large overview of the systems' performances, other common metrics will also be provided. All metrics will be computed by using the `trec_eval` tool from NIST[3].

7 CONCLUSIONS

NOT CHANGED - copy-past from last year - Dont think we really need to change it. If yes, Bogdan can you do it? The 2017 Predicting Media Interestingness task provides participants with a comparative and collaborative evaluation framework for predicting content interestingness with explicit focus on multimedia approaches. Details on the methods and results of each individual participant team can be found in the working note papers of the MediaEval 2017 workshop proceedings.

ACKNOWLEDGMENTS

We would like to thank Yu-Gang Jiang and Baohan Xu from the Fudan University, China, Hervé Bredin, from LIMSI, France, and Michael Gygli for providing the features that accompany the released data. Part of the task was funded under research grant PN-III-P2-2.1-PED-2016-1065, agreement 30PED/2017, project SPOTTER.

REFERENCES

[1] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE CVPR Conference on Computer Vision and Pattern Recognition*.

[2] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference*.

[3] Yu-Gang Jiang, Qi Dai, Tao Mei, Yong Rui, and Shih-Fu Chang. 2015. Super Fast Event Recognition in Internet Videos. *IEEE Transactions on Multimedia* 177, 8 (2015), 1–13.

[4] S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE CVPR Conference on Computer Vision and Pattern Recognition*. 2169–2178.

[5] D. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision* 60 (2004), 91–110.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7) (2002), 971–987.

[7] A. Oliva and A. Torralba. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision* 42 (2001), 145–175.

[8] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE CVPR Conference on Computer Vision and Pattern Recognition*. 3485–3492.

[2] http://multimediaeval.org/mediaeval2016/persondiscovery/

[3] http://trec.nist.gov/trec_eval/