MediaEval 2017 Predicting Media Interestingness Task

Claire-Hélène Demarty[1], Mats Sjöberg[2], Bogdan Ionescu[3], Thanh-Toan Do[4], Michael Gygli[5], Ngoc Q. K. Duong[1]
[1]Technicolor, Rennes, France
[2]Dept. of Computer Science and Helsinki Institute for Information Technology 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

In this paper, the Predicting Media Interestingness task which is running for the second year as part of the MediaEval 2017 Benchmarking Initiative for Multimedia Evaluation, is presented. For the task, participants are expected to create systems that automatically select images and video segments that are considered to be the most interesting for a common viewer. All task characteristics are described, namely the task use case and challenges, the released data set and ground truth, the required participant runs and the evaluation metrics.

1 INTRODUCTION

Predicting the interestingness of media content has been an active axis of research in the computer vision community for several years now [1, 7, 8, 10] and it has even been studied earlier in the psychological community [2, 16, 17]. But, definitions of interestingness were multiple, datasets publicly available were only a few and, until last year, no benchmark did exist to assess the interestingness of content. In 2016, a task for the Prediction of Media Interestingness was proposed in the MediaEval 2016 Benchmarking Initiative for Multimedia Evaluation. This task was the opportunity to propose a clear definition of interestingness, compliant to a use case in industry, and in particular at Technicolor[1]. The 2017 edition of the MediaEval benchmark includes a following of the Predicting Media Interestingness Task. This paper gives an overview of the task description in its second year, together with a description of the data and ground truth. The required runs and chosen evaluation metrics are also detailed. In all cases, changes that happen in this year's edition are highlighted compared to last year's edition.

2 TASK DESCRIPTION

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.

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

MediaEval'17, 13-15 September 2017, Dublin, Ireland

As in 2016, interestingness should be assessed according to a practical use case at Technicolor, 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 will be considered as a binary classification AND a regression task. Prediction will be carried out on a per movie basis. The two taskes are:

Predicting Image Interestingness Given a set of key-frames extracted from a certain movie, the task involves automatically (1) identifying those images that viewers report to be interesting and (2) ranking all images according to their level of interestingness. 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 (1) identifying the segments that viewers report to be interesting and (2) ranking all segments according to their level of interestingness. 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 (243) with an average duration of 11.4s, to better take into account 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 2435 video segments and the same number of key-frames.

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 [13], 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 [11]; *HoG descriptors* (Histograms of Oriented Gradients) [4] are computed over densely sampled patches. Following [19], HoG descriptors in a 2×2 neighborhood are concatenated to form a descriptor of higher dimension; *LBP* (Local Binary Patterns) [14]; *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 [15]; *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 [9]; *mid level face detection and tracking related features*[2] — obtained by face tracking-by-detection in each video shot with a HoG detector [4] and the correlation tracker proposed in [5]. In addition to these frame-based features, we provide C3D [18] features, which were extracted from *fc6 layer* (4,096 dimensions) and averaged on a segment level.

4 GROUND TRUTH

Both video and image data was manually and independently annotated in terms of interestingness by human assessors, to allow to study the correlation between the two subtasks. A dedicated web-based annotation tool was developed by the organising team for the previous edition of the task [6]. This year we made some improvements and released it as free and open source software[3]. Overall, more than 202 annotators participated in the annotation for the video data and 144 for the images. The cultural distribution is over 20 different countries in the world.

As in last year's setup we use a pair-wise comparison protocol [3] where annotators are provided with a pair of images/shots at a time and asked to tag which one in the pair is the more interesting for them. As a change from last year, we now ask the question in a way more directly connected to the commercial application: "Which image/video makes you more interested in watching the whole movie?", with the intent to make the annotators' decision clearer. As an exhaustive annotation of all possible pairs is practically impossible due to the required human resources, a boosting selection was used instead. In particular, we used a modified version of the adaptive square design method [12], in which several annotators participated in each iteration. In this method the number of comparisons for each iteration is reduced from all possible pairs $n(n-1)/2 \sim O(n^2)$ to a subset of pairs $n(\sqrt{n}-1) \sim O(n^{\frac{3}{2}})$, where n is the number of segments or images. For the development set, we started from iteration 6, as we could reuse the annotations done last year. To achieve the ranking used as the basis for the next round, the pair-based annotations are aggregated with the Bradley-Terry-Luce (BTL) model computation [3] resulting in an interestingness degree for each image/shot. Previously the same procedure was also used to get the final interestingness values. This year we used an alternative method, which took into account all pair comparisons from all rounds done this year into a single large BTL calculation. This was done mainly because we discovered afterwards that some annotations from earlier rounds had to be discarded, because of some cheating annotators. These annotators occasionally switched to cheating, where they simply always selected the first, or the second item as the most interesting one without actually assessing the media contents. In the development set as much as 10% of the annotations were marked as invalid and not included in the final BTL calculation. We added some heuristic anti-cheating measures to the system, although it is not possible to avoid all cheating processes. Unfortunately, in the iterative approach, we could only discard annotations from the most recent round, as it would be based on the previous round's BTL output, which is why we developed another solution to compute the final BTL ranking. The final binary decisions are obtained using a thresholding scheme that tries to detect the boundary where interestingness values make the "jump" between the underlying distributions of the non interesting and interesting populations. See last year's overview paper for a more detailed description [6].

[2]http://multimediaeval.org/mediaeval2016/persondiscovery/
[3]https://github.com/mvsjober/pair-annotate

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[4].

7 CONCLUSIONS

Thanks to the 2017 Predicting Media Interestingness task, participants are proposed a complete and comparative framework for the evaluation of content interestingness. 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.

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

REFERENCES

[1] Xesca Amengual, Anna Bosch, and Josep Lluís de la Rosa. 2015. *Review of Methods to Predict Social Image Interestingness and Memorability.* Springer, 64–76. https://doi.org/10.1007/978-3-319-23192-1_6

[2] Daniel E. Berlyne. 1960. *Conflict, arousal and curiosity.* Mc-Graw-Hill.

[3] R. A. Bradley and M. E. Terry. 1952. Rank Analysis of Incomplete Block Designs: the method of paired comparisons. *Biometrika* 39 (3-4) (1952), 324–345.

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

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

[6] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang, Ngoc Q.K. Duong, and Frédéric Lefebvre. 2016. MediaEval 2016 Predicting Media Interestingness Task. In *Proceedings of the MediaEval 2016 Workshop*. Hilversum, Netherlands.

[7] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *IEEE International Conference on Computer Vision and Pattern Recognition.*

[8] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. van Gool. 2013. The Interestingness of Images. In *ICCV International Conference on Computer Vision.*

[9] 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.

[10] Y-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yan. 2013. Understanding and Predicting Interestingness of Videos. In *AAAI Conference on Artificial Intelligence.*

[11] 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.

[12] Jing Li, Marcus Barkowsky, and Patrick Le Callet. 2013. Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. In *SPIE Electronic Imaging, Stereoscopic Displays and Applications*, Vol. 8648.

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

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

[15] 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.

[16] Paul J. Silvia. 2006. *Exploring the psychology of interest.* Oxford University Press.

[17] Craig Smith and Phoebe Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology* 48, 4 (1985), 813–838.

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision.*

[19] 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.