MediaEval 2016 Predicting Media Interestingness Task

Claire-Hélène Demarty[1], Mats Sjöberg[2], Bogdan Ionescu[3], Thanh-Toan Do[4], Hanli Wang[5], Ngoc Q. K. Duong[1], Frédéric Lefebvre[1]
[1]Technicolor, Rennes, France
[2]HIIT, University of Helsinki, Finland
[3]LAPI, University Politehnica of Bucharest, Romania
[4]Singapore University of Technology and Design, Singapore & University of Science, Vietnam
[5]Tongji University, China
{claire-helene.demarty, quang-khanh-ngoc.duong, frederic.lefebvre}@technicolor.com
mats.sjoberg@helsinki.fi, bionescu@imag.pub.ro, thanhtoan_do@sutd.edu.sg, hanliwang@tongji.edu.cn

ABSTRACT

This paper provides an overview of the Predicting Media Interestingness task that is organized as part of the MediaEval 2016 Benchmarking Initiative for Multimedia Evaluation. The task, which is running for the first 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

The ability of multimedia data to attract and keep people's interest for long periods of time is gaining more and more importance in the field of multimedia, where concepts such as memorability [4], aesthetics [9], interestingness [13, 11], attractiveness [14], affective value [25], are intensely studied, especially in the context of the ever growing market value of social media and advertising. In particular, although interestingness has been studied for a long time in the psychology community [21, 2, 22] and more recently but actively in the image processing community [23, 5, 1, 6, 10, 18], no common definition exists in the literature. Moreover datasets publicly available are only a few and no benchmark exists for the evaluation of what makes a media interesting.

In this paper we introduce the 2016 MediaEval[1] Predicting Media Interestingness Task, which is a pioneer benchmarking initiative for automatic prediction of image and video interestingness. The task which is in its first year derives from a practical use case at Technicolor[2]. It involves helping professionals to illustrate a Video on Demand (VOD) web site by selecting some interesting frames and/or video excerpts for the posted 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 underlying movie. The data in this task is therefore adapted to this particular context which provides a more focused definition for interestingness.

[1]http://multimediaeval.org/
[2]http://www.technicolor.com/

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

2. TASK DESCRIPTION

The task requires participants to deploy algorithms that automatically select images and video segments of Hollywood-like movies which are considered to be the most interesting for a common viewer. Interestingness of the media is judged based on the visual appearance, audio information and text accompanying the data. Therefore, the multimodal facet of interestingness prediction can be investigated.

Two different subtasks are provided, which correspond to the two types of available media content, namely:

- *predicting image interestingness* — given a set of key-frames extracted from a movie, the task requires to automatically identify those images for the given movie that viewers report to be the most interesting. To solve the task, participants can make use of visual content as well as external metadata, e.g., Internet data about the movie, social media information, etc;
- *predicting video interestingness* — given the video shots of a movie, the task requires to automatically identify those shots that viewers report to be the most interesting in the given movie. To solve the task, participants can make use of visual and audio data as well as external data, e.g., subtitles, Internet data, etc.

In both cases, the task is a binary classification task, thus participants are expected to label the provided data as being interesting or not (Note that prediction will be carried out on a per movie basis). However, a confidence value is required for the provided prediction.

3. DATA DESCRIPTION

The 2016 data is extracted from Creative Commons licensed trailers of Hollywood-like movies. It consists of a *development data* intended for designing and training the methods (information is extracted from 52 trailers) and a *testing data* which is used for the final benchmarking (with information from 26 trailers). The choice of using trailers instead of full movies is driven by the need to find data, both freely distributable and still representative in content and quality of Hollywood movies. Trailers are the results of some manual filtering of movies to keep interesting scenes, but also less attractive shots to balance their content. We therefore believe they are still representative for the task.

For the *predicting video interestingness* subtask, the data consists of the video shots obtained after the manual segmentation of the videos (video shots are the continuous

frame sequences recorded between a camera turn on and off), i.e., 5,054 shots for the development data, and 2,342 shots for the test data.

For the *predicting image interestingness* subtask, the data consists of collections of key-frames extracted from the video shots used for the video subtask. One single key-frame is extracted per shot, therefore leading to 5,054 key-frames for the development set and 2,342 for the test set. This single key-frame is chosen as the middle frame, as it is highly likely to capture the most important information of the shot.

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 [17], 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 [15]; *HoG descriptors* (Histograms of Oriented Gradients) [7] are computed over densely sampled patches. Following [24], HoG descriptors in a 2×2 neighborhood are concatenated to form a descriptor of higher dimension; *LBP* (Local Binary Patterns) [19]; *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 [20]; *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 [12]; *mid level face detection and tracking related features*[3] — obtained by face tracking-by-detection in each video shot with a HoG detector [7] and the correlation tracker proposed in [8].

4. GROUND TRUTH

All data was manually annotated in terms of interestingness by human assessors. A dedicated web-based tool was developed to assist the annotation process. Overall, more than 312 annotators participated to the annotation for the video data and 100 for the images. The cultural distribution is over 29 different countries in the world.

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 of the content is more interesting for them. The process is repeated by scanning the whole dataset. As an exhaustive comparison of all possible pairs is basically impossible due to the required human resources, a boosting selection was used instead, i.e., a modified version of the adaptive square design method [16], for which several annotators participate to each iteration.

To achieve the final ground truth, pair-based annotations are aggregated with the Bradley-Terry-Luce (BTL) model computation [3] resulting in an interestingness degree for each image/shot. The final binary decisions are obtained after the following processing steps: (i) the interestingness values are ranked in increasing order and normalized between 0 and 1; (ii) the resulting curve is smoothed with a short averaging window, and the second derivative is computed; (iii) for both shots and images, and for all videos, a threshold empirically set to 0.01 is applied on the second derivative to find the first point whose value is above the threshold. This position corresponds to the limit between non interesting and interesting shots/images. The underlying motivation for this empirical rule is the following: the non interesting population has rather similar interestingness values which increase slowly, while a gap happens when one switches from this non interesting population to the population of more interesting samples. The second derivative was chosen preferably to the first derivative, as it allowed to select those gaps more precisely.

Ground truth is provided in binary format, i.e., 1 for interesting and 0 for non interesting, for each image and video in the two subtasks.

5. RUN DESCRIPTION

Each participating team is expected to submit up to 5 runs for both subtasks altogether. Among these 5 runs, two runs are *required*, one per subtask: for the *predicting image interestingness subtask*, the required run is built on visual information only and no external data is allowed; for the *predicting video interestingness subtask* only audio and visual information is allowed (no external data) for the required run.

Note that in this context, *external data* can be understood as: (i) additional datasets and annotations dedicated to interestingness classification; (ii) pre-trained models, features, detectors obtained from such dedicated additional datasets; and (iii) additional metadata that could be found on the Internet on the provided content (e.g., from IMDB[4]).

On the contrary, CNN features trained on generic datasets such as ImageNet (typically the provided CNN features) are allowed for use in the required runs. By generic datasets, we mean datasets that were not designed to support research in the task area, i.e., for the classification/study of image and video interestingness.

6. EVALUATION

The official evaluation metric is the *mean average precision* (MAP) over the interesting class, i.e., the mean over the average precision scores computed for each trailer. This metric, adapted to retrieval tasks, fits perfectly the chosen use case in which we want to help a user choose between different samples by providing him a list of suggestions, ranked according to interestingness. For assessing the performance, we use the `trec_eval` tool provided by NIST[5]. In addition to MAP, other commonly used metrics such as precision and recall will be provided to participants.

7. CONCLUSIONS

The 2016 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 2016 workshop proceedings.

8. ACKNOWLEDGMENTS

We would like to thank Yu-Gang Jiang and Baohan Xu from the Fudan University, China, and Hervé Bredin, from LIMSI, France for providing the features that accompany the released data, and Alexey Ozerov and Vincent Demoulin for their valuable inputs to the task definition.

[3] `http://multimediaeval.org/mediaeval2016/persondiscovery/`

[4] `http://www.imdb.com/`

[5] `http://trec.nist.gov/trec\_eval/`

9. REFERENCES

[1] X. Amengual, A. Bosch, and J. L. de la Rosa. *Review of Methods to Predict Social Image Interestingness and Memorability*, pages 64–76. Springer, 2015.

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

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: the method of paired comparisons. *Biometrika*, (39 (3-4)):324–345, 1952.

[4] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, (116):165–178, 2015.

[5] C. Chamaret, C.-H. Demarty, V. Demoulin, and G. Marquant. Experiencing the interestingness concept within and between pictures. In *Proceeding of SPIE*, Human Vision and Electronic Imaging, 2016.

[6] S. L. Chu, E. Fedorovskaya, F. Quek, and J. Snyder. The effect of familiarity on perceived interestingness of images. volume 8651, pages 86511C–86511C–12, 2013.

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

[8] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference*, 2014.

[9] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[10] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 1017–1026, New York, NY, USA, 2013.

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

[12] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang. Super fast event recognition in internet videos. *IEEE Transactions on Multimedia*, 177(8):1–13, 2015.

[13] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yan. Understanding and predicting interestingness of videos. In *AAAI Conference on Artificial Intelligence*, 2013.

[14] S. Kalayci, H. K. Ekenel, and H. Gunes. Automatic analysis of facial attractiveness from video. In *IEEE ICIP International Conference on Image Processing*, pages 4191–4195, 2014.

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

[16] J. Li, M. Barkowsky, and P. L. Callet. Boosting paired comparison methodology in measuring visual discomfort of 3dtv: performances of three different designs. In *SPIE Electronic Imaging, Stereoscopic Displays and Applications*, volume 8648, 2013.

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

[18] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 83–92, New York, NY, USA, 2010.

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

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

[21] P. J. Silvia. *Exploring the psychology of interest.* Oxford University Press, 2006.

[22] C. Smith and P. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838, 1985.

[23] M. Soleymani. The quest for visual interest. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 919–922, New York, NY, USA, 2015.

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

[25] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi. Multimedia content analysis for emotional characterization of music video clips. *EURASIP Journal on Image and Video Processing*, (26), 2013.