

MediaEval 2018: Predicting Media Memorability

Romain Cohendet¹, Claire-Hélène Demarty¹, Ngoc Q. K. Duong¹, Mats Sjöberg², Bogdan Ionescu³,
Thanh-Toan Do⁴

¹Technicolor, Rennes, France

²Aalto University, Finland

³Multimedia Lab, University Politehnica of Bucharest, Romania

⁴ARC Center of Excellence for Robotic Vision, University of Adelaide, Australia

ABSTRACT

In this paper, we present the Predicting Media Memorability task, which is proposed as part of the MediaEval 2018 Benchmarking Initiative for Multimedia Evaluation. Participants are expected to design systems that automatically predict memorability scores for videos, which reflect the probability of a video being remembered. In contrast to previous work in image memorability prediction, where memorability was measured a few minutes after memorization, the proposed dataset comes with "short-term" and "long-term" memorability annotations. All task characteristics are described, namely: the task's challenges and breakthrough, the released data set and ground truth, the required runs and the evaluation metrics.

1 INTRODUCTION

Following the rapid expansion of the research field of image memorability prediction [2, 5, 10, 11], the challenge has recently been extended to videos [3, 7, 15]. An important motivation for video memorability (VM) prediction derives from the need for new techniques that can help to organize and retrieve digital content, to make it more useful in our daily lives. The problem is a pressing one since media platforms, such as social networks, search engines, and recommender systems deal with growing amounts of content data day after day. Like other cues of video importance, such as aesthetics or interestingness, memorability can be regarded as useful to help make a choice between otherwise comparable videos. Consequently, a large number of applications, e.g., education and learning, content retrieval and search, content summarization, storytelling, targeted advertising, content recommendation and filtering, would benefit from models capable of ranking videos according to their memorability. Despite its potential of being an active area of research in the computer vision community, VM prediction suffers from two main obstacles that were described in [3]. Firstly, among the previous attempts at predicting VM [3, 7, 15] no clear definition of VM has been established, nor does a common and unified protocol for its measurement exist, contrary to what can be found in the literature for image memorability. Secondly, no large-scale dataset is available, for the community to build its models. The purpose of this task is therefore to propose a public benchmark to assess the memorability of videos, based on a publicly released large-scale dataset and on an objective and clear measurement protocol.

2 TASK DESCRIPTION

The Predicting Media Memorability Task requires participants to build systems that are capable of predicting how memorable a video

is, by computing for each video a memorability score. Participants will be provided with an extensive dataset of videos with memorability annotations. The ground truth has been collected through recognition tests, and, for this reason, reflects objective measures of memory performance. In contrast to previous work on image memorability prediction [10, 11], where memorability was measured a few minutes after memorization, the dataset comes with both "short-term" and "long-term" memorability annotations. Because memories continue to evolve in long-term memory [12], in particular during the first day following memorization (see e.g., the forgetting curve in the seminal work of Ebbinghaus [13]), we expect long-term memorability annotations to be more representative of long-term memory performance, which is more relevant in many applications. Participants will be required to train computational models capable of inferring video memorability from visual content. Optionally, descriptive titles attached to the videos may be used. Two subtasks will be offered to participants: **Short-term Memorability Subtask**: the task involves predicting a short-term memorability score for a given video. **Long-term Memorability Subtask**: the task involves predicting a long-term memorability score for a given video. For the two subtasks, depending on the runs, participants will be allowed to use external data.

3 DATA DESCRIPTION

The dataset is composed of 10,000 short soundless videos shared under a license that allows their use and redistribution in the context of MediaEval 2018. These 10,000 videos were split into 8,000 videos for the development set and 2,000 videos for the test set. They were extracted from raw footage used by professionals when creating content. Of 7s-duration each, they are varied and contain different scenes types. Each video also comes with its original title. These titles can often be interpreted as a list of tags (textual metadata) that might be useful to infer the memorability of the videos. Participants are free to use them or not. To facilitate participation from various communities, we also provide some pre-computed content descriptors. Two of them are video-dedicated features: *C3D spatio-temporal visual features* that are obtained by extracting the output of the final classification layer of the C3D model, a 3-dimensional convolutional network proposed for generic video analysis [17], and *HMP* [1], the histogram of motion patterns for each video. Additional frame-based features are provided that were extracted on three key-frames (first, middle and last frames) for each video: *HoG descriptors* (Histograms of Oriented Gradients) [4] are calculated on 32x32 windows on a grey scale version of each frame; *LBP* (Local Binary Patterns) [8] are calculated for patches of 8x15 pixels; *InceptionV3 features* [16] correspond to the output of the fc7 layer of the InceptionV3 deep network; *ORB features* [14] result from a fusion of

FAST keypoint detector and BRIEF descriptor and *Color histograms* are computed in the HSV space. Additionally, following the work in [6], a set of *Aesthetic visual features*, composed of color, texture and object based descriptors, aggregated through the computation of their mean and median values, are extracted for each 10-frame of one single video.

4 GROUND TRUTH

4.1 Protocol to measure video memorability

To collect memorability annotations, we introduced a new protocol to measure human long-term memory performance for videos, partly inspired by [10]. The protocol consists of two parts, and is based upon recognition tests for our memorability scores to reflect objective measures of memory performance. The first part consists of interlaced viewing and recognition tasks. Participants viewed a sequence of 180 various videos they were unfamiliar with, 40 of which being *targets* videos, i.e., repeated videos, and the other being fillers, i.e., videos that occurred only once. Their task was to press the space bar whenever they detected a repetition. After 24 to 72 hours, participants viewed a new sequence of videos consisting of 40 targets, which were videos randomly chosen from the fillers of the first part, and 120 new fillers. In contrast to previous work [10, 11], where memorability was measured a few minutes after memorization, memory performance was therefore measured twice: a few minutes after memorization and again (on different items) 24-72 hours later. Thus, the dataset comes with both short-term and long-term annotations. These two scores will allow a comparison of the participants' systems for both short and long term memorability prediction. However, because of the difficulty to collect data after a long delay through crowdsourcing, the number of annotations is bigger for short-term memorability scores than for long-term ones. On average, each video received 38 and 13 annotations in the short-term and long-term recognition task, respectively. For each video in the development set, we provide the number of annotations for both tasks.

4.2 Memorability scores calculation

We assigned an initial memorability score to each video, defined as the percentage of correct detections by participants, for both short-term and long-term memory performances. The percentage scores are presented as floats in the interval [0,1].

The short-term raw scores are further refined by applying a linear transformation that takes into account the memory retention duration to correct/normalize the scores. Indeed, in our measurement protocol, the second occurrence (i.e., repetition) of a video happens after variable time intervals (i.e., each video is repeated after a variable number of other videos randomly chosen in the range of [45;100] videos). In [9], using a similar approach for images, it has been shown that memorability scores change as a function of the time interval between repeats while memorability ranks are largely conserved. We were able to prove the same relation for videos. Thus, as in [11], we use this information to apply a correction to our raw memorability scores to explicitly account for the difference in interval lengths, with the objective for our short-term memorability scores to be the most representative of the typical memory performance after the max interval (i.e., 100 videos). Because we observed that memorability decreases linearly

when the retention duration increases, we decided to apply a linear correction. Nevertheless, note that the applied correction only has a little effect on the memorability scores both in term of absolute and relative values. On the contrary, we did not apply any correction for long-term scores. Indeed, we observed no specific relationship between retention duration and long-term memorability from our collected scores. This was expected from what can be found in the literature: according to our protocol, the second measure was carried out 24 to 72 hours after the first measure. After such a long retention duration, it is expected that the memory performance is no more subjected to substantial decrease due to the retention duration.

5 RUN DESCRIPTION

Every team can submit up to 10 runs, 5 per subtask. For each subtask, a required run is defined. **Short-term memorability subtask – required run:** Any information (extracted from the content, the provided features, the short-term memorability scores or external data) is allowed to build the systems, except the use of the long-term memorability scores which is not allowed. **Long-term memorability subtask – required run:** Any information (extracted from the content, the provided features, the long-term memorability scores or external data) is allowed to build the systems, except the use of the short-term memorability scores which is not 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 Spearman's rank correlation between the predicted memorability scores and the ground-truth memorability scores computed over all test videos. Although the task remains a prediction task, only the ranking of the different videos will be evaluated by the official metric. The choice of the Spearman's rank correlation as official measure indeed corresponds to a desire of normalizing the output of the different systems and making the comparison easier. For this reason, participants are encouraged to really consider the task as a prediction task. Other classic metrics (i.e., Pearson correlation and Mean squared error) will also be computed and provided to the participants for the sake of comparison between the different runs and systems.

7 CONCLUSIONS

A complete and comparative framework for the evaluation of video memorability is proposed. Details on the methods and results of each individual participant team can be found in the working note papers of the MediaEval 2018 workshop proceedings.

ACKNOWLEDGMENTS

We would like to thank Ricardo Manhaes Savii (Federal University of São Paulo) and Mihai Gabriel Constantin (University Politehnica of Bucharest) for providing the features that accompany the released data. This work was partially supported by the Romanian Ministry of Innovation and Research (UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002), the Academy of Finland (project 313988) and the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No 780069).

REFERENCES

- [1] Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. 2011. Comparison of video sequences with histograms of motion patterns. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 3673–3676.
- [2] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. 2016. Deep Learning for Image Memorability Prediction: the Emotional Bias. In *Proc. ACM Int. Conf. on Multimedia (ACMM)*. 491–495.
- [3] Romain Cohendet, Karthik Yadati, Ngoc Q.K. Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proc. of the ICMR 2018 Workshop, Yokohama, Japan, June 11-14*.
- [4] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 886–893.
- [5] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. AMNet: Memorability Estimation with Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6363–6372.
- [6] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, and others. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3 (2015), e1390.
- [7] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. 2015. Learning computational models of video memorability from fMRI brain imaging. *IEEE Transactions on Cybernetics* 45, 8 (2015), 1692–1703.
- [8] Dong-Chen He and Li Wang. 1990. Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing* 28, 4 (1990), 509–512.
- [9] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1469–1482.
- [10] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 145–152.
- [11] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. 2390–2398.
- [12] James L McGaugh. 2000. Memory—a century of consolidation. *Science* 287, 5451 (2000), 248–251.
- [13] Jaap MJ Murre and Joeri Dros. 2015. Replication and analysis of Ebbinghaus forgetting curve. *PLoS one* 10, 7 (2015), e0120644.
- [14] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2564–2571.
- [15] Sumit Shekhar, Dhruv Singal, Harvaneet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2730–2739.
- [16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.
- [17] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. 4489–4497.