The Predicting Media Memorability Task at MediaEval 2019

Mihai Gabriel Constantin[1], Bogdan Ionescu[1], Claire-Hélène Demarty[2],
Ngoc Q. K. Duong[2], Xavier Alameda-Pineda[3], Mats Sjöberg[4]
[1]University Politehnica of Bucharest, Romania
[2]InterDigital, France
[3]INRIA, France
[4]CSC, Finland

ABSTRACT

In this paper, we present the Predicting Media Memorability task, which is running for the second year at the MediaEval 2019 Benchmarking Initiative for Multimedia Evaluation. Participants are required to create systems that are able to automatically predict the memorability scores of a collection of videos, which should represent the "short-term" and "long-term" memorability of the samples. We will describe all the aspects of this task, including its main characteristics, a description of the development and test data sets, the ground truth, the evaluation metrics and the required runs.

1 INTRODUCTION

The latest developments in multimedia information processing have led to the development of systems and methods that can predict the way humans perceive and react to images and videos, i.e., infering interestingness, aesthetics, emotional content, etc. [7]. Such processing tools are gaining importance on media platforms, social networks and recommender systems considering that the amount of available data is continually growing so does the need to filter media content according to a wide variety of factors. Memorability is one of these factors. Furthermore, the analysis of video memorability is a domain of media processing with a wide array of possible applications such as content retrieval, education, summarization, advertising, content filtering, and recommendation systems. The study of memorability attracted different research communities, including psychologists, behavior specialists, and computer scientists. Early human-based studies on visual memory capabilities indicated a massive storage capacity for visual data [19, 22], also showing that, even in a long-term study, subjects are able to retain specific details of images, not just the general gist [3]. Also noteworthy are studies showing that memorability is an intrinsic property of images [13]. Computer vision scientists used these results and created methods for the prediction of image memorability [2, 9, 14, 15] and, more recently, video memorability [4, 6, 11, 21]. Recent studies also show that style transfer can be used to increase image memorability [23, 24]. However, in many of these examples, the authors used different datasets or different splits, thus making it hard to compare methods and draw a clear set of conclusions with regards to the accuracy of individual approaches [6]. The Predicting Media Memorability task addresses this problem, and, starting with last year's competition [5], creates a common benchmarking protocol and provides a dataset for short-term and long-term video memorability using common definitions. Details regarding the first edition of this task, including the methods used by all the participants and their results, can be found in the proceedings of the 2017 MediaEval workshop.[1]

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

2 TASK DESCRIPTION

The 2019 Predicting Media Memorability task is a continuation of last year's task [5]. Participants are required to create systems that can predict the memorability score for video samples. Just like in the previous settings of this task, ground truth data contains scores for both "short-term" and "long-term" memorability, created via memory performance tests. These two different objectives follow psychological and human subject studies, such as [16, 17], that analyze the effect that time has on visual memory. While short-time annotations measure the immediate retention of samples, long-time annotations measure retention after a longer period of time, usually ranging from hours to days [16, 18] and may be appropriate for different types of applications. Therefore two subtasks are proposed to participants:

- The prediction of **short-term memorability** - scores were measured a few minutes after the memorization process.
- The prediction of **long-term memorability** - scores were measured 24-72 hours after the memorization process.

3 DATA DESCRIPTION

The proposed dataset consists of 10,000 7-second videos without sound, split into 8,000 videos for the development set (devset) and 2,000 for the testing set (testset). Participants must train their systems on the devset and submit runs containing memorability scores for the testset. Ground truth scores and information regarding the number of annotators are provided for each video sample in the devset, for both subtasks.

We provided some pre-computed features that could help teams get their systems started and provide easier access to the task to a broader community of researchers. First, some frame-based features were extracted, for each video, analyzing the first, middle and last frames. Among these frame-based features are: Histogram of Oriented Gradients (HoG) [8], calculated on 32×32 windows for grayscale frames, Local Binary Patterns (LBP) [12], calculated for patches of 8×15 pixels, Color histogram in HSV space and ORB features [20]. Also, we extracted the output of the fc7 layer of InceptionV3 [25]. Another set of handcrafted features are the Aesthetic Visual Features (AVF) [10], representing color, texture and object-based descriptors, aggregated by the mean and median values extracted every 10 frames in a video. Second, we also extracted

[1]http://ceur-ws.org/Vol-2283/

video-level features representing the final category of visual descriptors. They have the role of motion or temporal descriptors that analyze the video as a whole and naturally represent the movement in these samples. We provide the Histogram of Motion Patterns (HMP) [1] and the output of the final classification layer of the convolutional neural network C3D model [26]. Finally, each video is accompanied by a short caption-like title or description text, that can be used if necessary as tag-like or textual features by the participants.

4 GROUND TRUTH AND ANNOTATION PROTOCOL

As we previously mentioned, memorability annotations are created via performance tests for both the short-term and long-term memorability subtasks and partially inspired by the work of [14]. The participants to these tests were shown a set of target samples (videos that did repeat after a certain time) and distractor samples (videos that did not repeat, having the role of fillers).

In the short-term phase, participants to these tests viewed 40 target videos that reappeared in the testing phase and 140 distractor videos that are played only once, adding up to a sequence of 180 total videos. In the long-term phase, after 24-72 hours, the same participants viewed 40 videos repeated from the previous distractor collection and another 120 new distractor videos, adding up to a sequence of 160 videos. The videos that repeat do so in a variable manner. Each repetition appears after a randomly chosen interval ranging from 45 to 100 videos. Participants were asked to press the space key each time they considered a repetition of a video sample occurred. Each sample from the dataset received between 13 and 38 annotations from the participants and in general more annotations were made for the short-term subtask, given that it proved difficult to collect data after an extended period from the first viewing. In order to assess the permanent attention of the annotators, control videos were repeated after a random number of videos between three and six.

We also applied specific correction protocols for the generation of the final memorability scores, inspired by the work of [15]. In the case of short-term annotation, in the first step, we calculated the percentage of memory test participants that correctly recognized the repetition of each sample, therefore obtaining an initial score in the interval [0,1]. However, given that these figures do not take into account the interval between the first viewing of the sample and its second appearance, a score normalization protocol, similar to the one presented in [15], was applied. The correlation between the repetition interval and memorability scores was previously studied in [14], where, in a paper on image memorability, the authors concluded that scores decrease when the interval grows, but that the ranks of the samples tend to remain unchanged. We confirmed this observation on our short-term memory tests too; indeed, a linear correlation existed between short-term memorability scores and the interval between the repetitions of the video sample, and therefore we decided to apply a linear correction to the initial scores. However, the same observation was not valid in the case of long-term memorability, where the second annotation was carried out 24 to 72 hours after the short-term stage of the experiment; therefore no correction was applied. More insights about the dataset, annotation protocol and some factors concerning video memorability can be found in [4].

Overall, the ground truth files are composed of the short-term and long-term memorability scores described above and the number of annotators for both subtasks, for each movie individually.

5 RUN DESCRIPTION

Teams are required to submit a run to each of the two subtasks, i.e., short-term memorability required run, and long-term memorability required run. In total, 10 runs can be submitted, 5 to each subtask.

For the two required runs, all information can be used in the development of the system, meaning provided features, ground-truth data, video sample titles, features extracted from the visual content and even external data. However, the only exception, in this case, is that the required short-term memorability run must not use long-term memorability score annotations and the required long-term memorability run must not use short-term memorability score annotations. For the rest of the runs, a maximum of 4 per subtask, everything is permitted, including using cross-annotations between the subtasks.

6 EVALUATION

Three classic metrics will be extracted from the submitted runs and returned to the participating teams: Spearman's rank correlation, Pearson correlation and Mean squared error; however, we will use the Spearman's rank correlation as the official metric. This choice comes from the desire to make comparisons between methods, allowing for the normalization of the output of different systems by taking into account monotonic relationships between ground truth and system output. Though primarily a prediction task, the use of Spearman's rank as the official metric will allow for the evaluation of the systems based on the ranking of different video samples from the testset.

7 CONCLUSIONS

In this paper we presented the 2019 Predicting Media Memorability task, running for its second yeat at the MediaEval Benchmarking Initiative. We created a framework that allows the comparative study of different approaches for predicting short-term and long-term memorability, based on a common video sample dataset, devset-testset split, annotations, and metric. Details regarding the methods employed by participants and their results can be found in the proceedings of the 2019 MediaEval workshop.

ACKNOWLEDGMENTS

We would like to thank Ricardo Manhães Savii (Federal University of São Paulo) for providing the features that accompany the dataset. 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).

REFERENCES

[1] Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. 2011. Comparison of video sequences with histograms of motion patterns. In *2011 18th IEEE International Conference on Image Processing*. 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 *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 491–495.

[3] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* 105, 38 (2008), 14325–14329.

[4] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, and Martin Engilberge. 2019. VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability. In *International Conference on Computer Vision (ICCV)*.

[5] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. Mediaeval 2018: Predicting media memorability task. In *Proceedings of the MediaEval 2017 Workshop*.

[6] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, 178–186.

[7] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. 2019. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR)* 52, 2 (2019), 25.

[8] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *n Computer Vision and Pattern Recognition, 2005. CVPR 2005*, Vol. 1. 886–893.

[9] 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.

[10] 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.

[11] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. 2014. Learning computational models of video memorability from fMRI brain imaging. *IEEE transactions on cybernetics* 45, 8 (2014), 1692–1703.

[12] 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.

[13] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*. 2429–2437.

[14] P Isola, Jianxiong Xiao, A Torralba, and A Oliva. 2011. What makes an image memorable?. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 145–152.

[15] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*. 2390–2398.

[16] James L McGaugh. 2000. Memory–a century of consolidation. *Science* 287, 5451 (2000), 248–251.

[17] Jaap MJ Murre and Joeri Dros. 2015. Replication and analysis of EbbinghausâĂŹ forgetting curve. *PloS one* 10, 7 (2015), e0120644.

[18] James S Nairne and Addie Dutta. 1992. Spatial and temporal uncertainty in long-term memory. *Journal of Memory and Language* 31, 3 (1992), 396–407.

[19] Ronald A Rensink, J Kevin O'Regan, and James J Clark. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological science* 8, 5 (1997), 368–373.

[20] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. 2011. ORB: An efficient alternative to SIFT or SURF.. In *ICCV*, Vol. 11. Citeseer, 2.

[21] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision*. 2730–2739.

[22] Roger N Shepard. 1967. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior* 6, 1 (1967), 156–163.

[23] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2017. How to Make an Image More Memorable? A Deep Style Transfer Approach. In *ACM International Conference on Multimedia Retrieval*.

[24] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2019. Increasing Image Memorability with Neural Style Transfer. *ACM Transactions on Multimedia Computing Communications and Applications* (2019).

[25] 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.

[26] 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*. 4489–4497.