

Overview of The MediaEval 2023 Predicting Video Memorability Task

Mihai Gabriel Constantin¹, Claire-Hélène Demarty², Camilo Fosco³, Alba García Seco de Herrera⁴, Sebastian Halder⁴, Graham Healy⁵, Bogdan Ionescu¹, Ana Matran-Fernandez⁴, Rukiye Savran Kiziltepe⁶, Alan F. Smeaton⁵ and Lorin Sweeney⁵

¹University Politehnica of Bucharest, Romania

²InterDigital, France

³Massachusetts Institute of Technology Cambridge, USA

⁴University of Essex, UK

⁵Dublin City University, Ireland

⁶Karadeniz Technical University, Turkey

Abstract

This paper describes the sixth edition of the *Predicting Video Memorability* task, part of the MediaEval¹ multimedia evaluation benchmark initiative. Similar to previous editions, we use video data and annotations from two datasets, the Memento10k, and the VideoMem datasets. In light of the consistent performance plateau observed in previous iterations of the prediction task, in which participants were required to train and test on the same dataset, we have taken the decision to drop the prediction task from this year's competition. This modification allows participants the opportunity to redirect their efforts toward more challenging tasks. Therefore, for this edition we propose two tasks: the generalization task, where participants are required to train on one dataset and test their results on a different dataset, and the EEG task, where participants are required to predict memorability using EEG-related data. In this paper we present the main aspects of the 2023 Predicting Video Memorability task, exploring the proposed tasks, the datasets, evaluation methods and metrics, as well as the requirements for participants.

1. Introduction

Multimedia processing systems bear the formidable task of accurately predicting and correlating a vast array of media content with the intricacies of the human cognitive process. This role places them at the heart of media retrieval and media recommendation systems, where the fusion of computer vision, deep learning and cognitive sciences is of paramount importance in providing useful and insightful results. In this context, memorability is one of the most important aspects of human cognition that is explored by researchers from various domains. Defined as the likelihood that a certain piece of multimedia content will be remembered and recognized on subsequent viewing, memorability and the question “what makes a video memorable?” is still an open research question.

The 2023 MediaEval Predicting Video Memorability task attempts to answer some of these questions, proposing a common evaluation benchmark for models that target memorability prediction for videos. This represents the sixth edition of this task, following the success and learning from the patterns and lessons discovered in previous editions of the memorability task. Thus, this task has continually evolved and adapted throughout the editions, taking into


¹<https://multimediaeval.github.io/>

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online

✉ mihai.constantin84@upb.ro (M. G. Constantin)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

account the general trends of results, observations with regards to the data, annotations and ground truth, as well as valuable participant feedback.

2. Related work

Several key works in the study of human perception of multimedia data have shown not only an astonishing capacity for memorization from human viewers [1], but also that people tend to retain very specific characteristics and details of the visual samples they are shown [2]. In this context, numerous works have analyzed memorability from a computer vision standpoint, targeting images [3, 4] and videos [5, 6]. Important developments in this domain also target the use of physiological data like fMRI [5] and EEG [7]. Furthermore, researchers have studied different sets of low- and high-level human understandable attributes and their correlation with memorability, including but not limited to the presence of certain objects [8], photographic quality and emotions [9], and natural scene types [10].

The Predicting Video Memorability task builds upon these findings and initial ideas, and, in the previous five editions [11] has featured short- and long-term video memorability tasks, multiple datasets including Memento10k [12], VideoMem [13], and a memorability annotated subset of the 2019 TRECVID Video-to-Text dataset [14], each dataset featuring various modalities, including visual, audio, and textual. Multiple facets of memorability prediction are studied throughout the editions, manifested as three different subtasks: (i) a prediction subtask, which asks participants to train their models on the training and validation subsets of one dataset, and submit their runs for testing on the same dataset, (ii) the generalization subtask, where participants train and validate their models on one dataset, and test generalization properties on the testing subset of another dataset, and (iii) an EEG-based subtask where participants must use EEG data in order to infer whether a certain viewer will memorize or not a given video. Results thus far show some interesting trends. For the prediction subtask, results seem to reach a plateau around Spearman’s rank correlation values of 0.7, making us theorize that the maximum possible performance or values very close to that maximum potential have been reached. On the other hand, results for the generalization task show lower performance. This leaves a lot of room for development in this area of memorability, showing the need for researching less dataset-specific systems. Finally, the EEG task, while it only has one full edition in 2022 and a pilot edition in 2021, has shown some promising initial results.

3. Task description

Given the performance plateau registered on the prediction task, and the problems participants’ systems had on the generalization task, for this edition we propose to drop the prediction subtask, thus allowing participants to focus the Generalization of memorability predictor systems (Sub-task 1). We also continue the EEG-based prediction task (Sub-task 2), given its encouraging start in the previous edition of MediaEval.

3.1. Subtask 1: Generalization

Sub-task 1 deals with the Generalization of memorability predictor systems, thus testing them in a challenging scenario, but a scenario that would be closer to real-world applications. Participants are asked to train and validate their systems on the training and devset sections of the Memento10k dataset, and submit their runs and predictions on the testset split of VideoMem. Participants are allowed a maximum of 5 runs for this task. One of the runs must consist of

systems trained only with Memento10k data, while the other four can augment the training dataset in any way the participants feel is necessary, as long as they do not use VideoMem data.

3.2. Subtask 2: EEG-based prediction

Participants must create systems that can automatically predict whether a given human subject will remember a certain video or not on subsequent viewing, starting from the provided EEG data. For each video, in addition to the specific EEG features, we also provide the identifier of the volunteer, the label, and the id of the video that was being watched, so features from the video available for the other subtasks can be used. There is an obligation however to include EEG data in each system the participants develop for this task.

4. Datasets

This edition of the memorability task uses three datasets for its two subtasks. In the generalization subtask, the Memento10k dataset is provided and used for system training and validation, while the VideoMem dataset is used for system testing. On the other hand, the EEG subtask uses human physiological data from the EEGMem dataset, that consists of human subjects' EEG responses recorded while watching videos from the Memento10k dataset. This Section presents these datasets, the data they encompass, annotation protocols, and the features we provide associated with each dataset.

The following set of pre-extracted features are provided along with the Memento10k and VideoMem datasets, namely: (i) image-level features: AlexNetFC7 [15], HOG [16], HSVHist, RGBHist, LBP [17], VGGFC7 [18], DenseNet121 [19], ResNet50 [20], EfficientNetB3 [21]; and (ii) video-level features: C3D [22].

Given the different nature and modality of the EEG data, a different set of features is computed and provided for this data: ERPs (i.e., EEG amplitudes at the start of the video), ERSPs (features in the time-frequency domain, spanning the whole duration of the video), and images (also conveying time-frequency information, but appropriate for feeding into a CNN or some other sort of computer vision system).

4.1. Memento10k

The Memento10k dataset is an extensive and comprehensive dataset for investigating and analysing video memorability. The dataset consists of a collection of 10,000 three-second real-world video clips sourced from the Internet. Each video is accompanied by corresponding short-term memorability scores, memorability decay values, action labels, and five human-annotated captions. This dataset comprehensively encompasses the concept of memorability throughout a range of presentation delays, from seconds to minutes. This provides valuable insights into the temporal dynamics of memorability and how it changes over time. The short-term memorability scores are derived from "Memento: The Video Memory Game" experimental approach [9], involving crowdworkers tasked with identifying repeated videos, and are based on their responses. On average, each video clip has been annotated with 90 annotations and the dataset has a high level of human consistency, as indicated by a Spearman's rank correlation coefficient of 0.73.

From the Memento10k dataset [12] we will provide the training (7000 video samples) and validation (1500 video samples) sets, which will be used as the official training and validation (or development) sets of MediaEval 2023 Predicting Video Memorability task.

4.2. VideoMem

The VideoMem is a large-scale dataset composed of 10,000 soundless seven-second videos created to predict short-term and long-term video memorability. The video clips were obtained from a collection of cinematic raw stock footage, including different scenes from animals, food, nature, people, and transportation. Every video is accompanied by a caption or its original title with short-term and long-term memorability scores. The dataset aims to facilitate research focused on understanding the memorability of videos and assessing methodologies for predicting multimedia content memorability. A novel annotation protocol is demonstrated and both short-term and long-term memorability performances are measured via recognition tests conducted shortly after viewing the videos and 24-72 hours later, respectively [13].

From the VideoMem dataset the testing set (2000 video samples) will be provided and used as the official training set for the competition.

4.3. EEGMem

The EEGMem [7] dataset is composed of EEG data collected from 12 subjects while watching a subset of the Memento10k [12] dataset. The subjects are then asked to watch the same videos through a custom-built online portal between 24–72 hours after the video-EEG recording session, indicating whether they have recognised a video.

5. Evaluation

Two different metrics will be used as the main metrics for the proposed subtasks. Subtask 1 - generalization will use three metrics, namely Spearman’s rank correlation, Pearson correlation, and mean squared error. Similar to the previous editions of the Memorability task, Spearman’s rank correlation will be used as the official main metric for subtask 1. Subtask 2 - EEG will use the Area Under the Receiver Operating Characteristic Curve as the official metric.

6. Conclusions

This paper presents the sixth edition of the MediaEval Predicting Video Memorability task. This year’s edition proposes two subtasks, one based on the generalization of memorability prediction systems, and another one based on EEG data generated through the analysis of human subjects.

Acknowledgements

Financial support provided under project AI4Media, a European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911.

References

- [1] R. N. Shepard, Recognition memory for words, sentences, and pictures, *Journal of Verbal Learning and Verbal Behavior* 6 (1967) 156–163.
- [2] T. F. Brady, T. Konkle, G. A. Alvarez, A. Oliva, Visual long-term memory has a massive storage capacity for object details, *Proceedings of the National Academy of Sciences* 105 (2008) 14325–14329.

- [3] Y. Baveye, R. Cohendet, M. Perreira Da Silva, P. Le Callet, Deep learning for image memorability prediction: The emotional bias, in: *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 491–495.
- [4] J. Fajtl, V. Argyriou, D. Monekosso, P. Remagnino, Amnet: Memorability estimation with attention. arxiv 2018, arXiv preprint arXiv:1804.03115 (1804).
- [5] J. Han, C. Chen, L. Shao, X. Hu, J. Han, T. Liu, Learning computational models of video memorability from fMRI brain imaging, *IEEE Transactions on Cybernetics* 45 (2014) 1692–1703.
- [6] S. Shekhar, D. Singal, H. Singh, M. Kedia, A. Shetty, Show and recall: Learning what makes videos memorable, in: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2730–2739.
- [7] L. Sweeney, A. Matran-Fernandez, S. Halder, A. G. S. de Herrera, A. Smeaton, G. Healy, Overview of the EEG pilot subtask at MediaEval 2021: predicting media memorability, arXiv preprint arXiv:2201.00620 (2021).
- [8] M. A. Kramer, M. N. Hebart, C. I. Baker, W. A. Bainbridge, The features underlying the memorability of objects, *Science Advances* 9 (2023) eadd2981.
- [9] P. Isola, D. Parikh, A. Torralba, A. Oliva, Understanding the intrinsic memorability of images, *Advances in Neural Information Processing Systems* 24 (2011).
- [10] J. Lu, M. Xu, R. Yang, Z. Wang, Understanding and predicting the memorability of outdoor natural scenes, *IEEE Transactions on Image Processing* 29 (2020) 4927–4941.
- [11] R. Savran Kiziltepe, M. G. Constantin, C.-H. Demarty, G. Healy, C. Fosco, A. Garcia Seco De Herrera, S. Halder, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, et al., Overview of the MediaEval 2021 predicting media memorability task, in: *MediaEval Workshop 2021, CEUR Workshop Proceedings*, volume 3181, 2021.
- [12] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, Springer, 2020, pp. 223–240.
- [13] R. Cohendet, C.-H. Demarty, N. Q. Duong, M. Engilberge, Videomem: Constructing, analyzing, predicting short-term and long-term video memorability, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2531–2540.
- [14] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, et al., Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval, arXiv preprint arXiv:2009.09984 (2020).
- [15] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Communications of the ACM* 60 (2017) 84–90.
- [16] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 1, IEEE, 2005, pp. 886–893.
- [17] D.-C. He, L. Wang, Texture unit, texture spectrum, and texture analysis, *IEEE Transactions on Geoscience and Remote Sensing* 28 (1990) 509–512.
- [18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
- [19] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [21] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114.
- [22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.