Overview of The MediaEval 2022 Predicting Video Memorability Task

Lorin Sweeney[1], Mihai Gabriel Constantin[2], Claire-Hélène Demarty[3], Camilo Fosco[4], Alba G. Seco de Herrera[5,*], Sebastian Halder[5], Graham Healy[1], Bogdan Ionescu[2], Ana Matran-Fernandez[5], Alan F. Smeaton[1] and Mushfika Sultana[5]

[1]*Dublin City University, Ireland*
[2]*University Politehnica of Bucharest, Romania*
[3]*InterDigital, France*
[4]*Massachusetts Institute of Technology Cambridge, USA*
[5]*University of Essex, UK*

Abstract

This paper describes the 5th edition of the *Predicting Video Memorability Task* as part of MediaEval2022. This year we have reorganised and simplified the task in order to lubricate a greater depth of inquiry. Similar to last year, two datasets are provided in order to facilitate generalisation, however, this year we have replaced the TRECVid2019 Video-to-Text dataset with the VideoMem dataset in order to remedy underlying data quality issues, and to prioritise short-term memorability prediction by elevating the Memento10k dataset as the primary dataset. Additionally, a fully fledged electroencephalography (EEG)-based prediction sub-task is introduced. In this paper, we outline the core facets of the task and its constituent sub-tasks; describing the datasets, evaluation metrics, and requirements for participant submissions.

1. Introduction

As the natural world unwinds in an endless cacophony of sensory threads, the human brain selectively spins it into intelligible spools—filtering out information it deems unnecessary and spinning the rest into an intelligible internal representation. The human brain is an equally masterful weaver as it is spinster; weaving a colourful tapestry of meaning from its spools of intelligible threads by deciding which threads should be stitched into the canvas of our mind—what should be remembered and what should not.

The question is, what criteria does it use to decide what should and should not be remembered? Unfortunately, a satiating answer presently remains out of reach, leaving "what it deems to be important" as our appetizer. Memorability—the likelihood that a given piece of content will be recognised upon subsequent viewing—can accordingly be viewed as a proxy for human importance, which is what ultimately motivates and brings meaning to its exploration. After all, what could be more important than a measure of importance itself?

Memorability is accordingly the quintessential media metric by virtue of its proximal nature to the bedrock of human experience. If a system can predict the memorability of incoming information, it can evaluate its utility, then discard, filter, or augment the scantily useful, and ultimately curate more meaningful media content.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online
*Corresponding author.
alba.garcia@essex.ac.uk (A. G. Seco de Herrera)

2. Related Work

The subject of memorability has seen an influx in interest since the likelihood of images being recognised upon subsequent viewing was found to be consistent across individuals [1]. Driven primarily by the MediaEval Media Memorability tasks [2, 3, 4, 5], recent research has extended beyond static images, pivoting to the more dynamic and multi-modal medium of video. In 2018, a video memorability annotation procedure was established, and the first ever large video memorability dataset VideoMem [4]—10,000 short soundless videos with both long-term and short-term memorability scores—was created. Additionally, the first ever analysis of human consistency and video memorability was conducted. In 2019, the task ran for a second time using the same dataset allowing participants to learn from the previous year's task and carry out comparative analysis of results from one year to the next. In 2020, a new smaller dataset was introduced which included audio for the first time [6]. In 2021, that dataset was extended, with a second large short-term dataset—Memento10k [7]—being released, short-term memorability was sub-categorised into *raw* and *normalised* scores, an optional generalisation sub-task was proposed, and a pilot EEG study [8] was conducted.

Over the course of those four tasks, we have learned that short-term video memorability is easier to predict than long-term memorability, simple image features, such as hue, saturation, or spatial frequency, have repeatedly been found not to correlate with memorability, properties such as aesthetics and interestingness likewise do not correlate with memorability, ensembles that combine different modalities provide the best results, combining deep visual features in conjunction with semantically rich features such as captions, emotions, or actions [9, 7, 10, 11] is a highly effective approach, dimensionality reduction improves prediction results, and certain semantic categories of objects or places are more memorable than others [4, 7].

3. Description

In this edition, the *Predicting Video Memorability* task challenges participants to develop systems that automatically predict short-term memorability scores for short form videos. Participants are provided with three datasets and offered three sub-tasks in which to participate.

3.1. Sub-task 1: How memorable is this video? - Video-based prediction

Using the Memento10k [7] dataset, participants are required to generate automatic systems that predict short-term memorability scores of new videos based on the given video dataset and their memorability scores.

3.2. Sub-task 2: How memorable is this video? - Generalisation (optional)

Sub-task 2 is a natural extension of sub-task 1, where participants can evaluate their systems from sub-task 1 (trained on Memento10k) on the VideoMem dataset. Alternatively, participants can also train a system on the VideoMem dataset and evaluate it on Memento10k.

3.3. Sub-task 3: Will this person remember this video? - EEG-based prediction (optional)

Participants are required to generate automatic systems that predict whether or not a given subject will recognise a given video upon subsequent viewing (N.B., this differs from memorability as it is subject specific and a binary prediction, rather than subject agnostic and a floating

point prediction) based on the provided EEG data. Participants may choose to use the provided EEG features in concert with sub-task 1's visual features or in isolation. However, they must use the EEG features in some capacity.

4. Dataset Details

In the interest of clarity, standardisation, and the facilitation of more directed inquiry, we have narrowed the scope of the tasks forgoing with raw and long-term memorability scores in favour of normalised short-term scores. Additionally, in order to address systemic data quality issues highlighted by a consistent disparity between participant systems trained on the TRECVid2019 dataset and the Memento10k dataset, we have opted to replace the TRECVid2019 dataset with VideoMem, and to elevate Memento10k to primary dataset status. Additionally, a fledged EEG dataset (EEGMem) is provided.

The following set of pre-extracted features are provided along with the Memento10k and VideoMem datasets:

- Image-level features: AlexNetFC7 [12], HOG [13], HSVHist, RGBHist, LBP [14], VGGFC7 [15], DenseNet121 [16], ResNet50 [17], EfficientNetB3 [18]
- Video-level features: C3D [19]

Three frames—the first, middle, and last—from each video were used to extract image-level features.

4.1. Memento10k

Memorability scores were collected through *Memento: The Video Memory Game*, a memorability experiment predicated the old-new recognition paradigm [1], where crowdworkers from Amazon's Mechanical Turk (AMT) watch a continuous stream of three-second video clips, and are asked to press the space bar when they see a repeated video. To maximise the pace and keep the experiment engaging, videos are shown as a continuous stream. When participants press their spacebar, they receive either a red (incorrect) or green (correct) flash as feedback. If a repeat is correctly identified, known as a "hit", the stream skips ahead to the next video; there is no feedback for missed repeats. Each level of the memory game contains on average 204 videos (with repeats) and lasts $\sim$ 9 minutes. The number of intervening videos between the first and second occurrence of a repeated video is known as the "lag". The game consists of "vigilance" repeats that occur at short lags of 2-3 videos and are used to filter out inattentive workers and "target" repeats at lags of 9-200 videos that provide memorability data.

The Memento10k dataset [7] consists of 10,000 three-second videos depicting in-the-wild scenes, each with associated short-term memorability scores, memorability decay values, action labels, and five human generated captions. The memorability scores were computed with an average of 90 annotations per video, and the videos were deafened before being shown to participants. 7,000 videos are released as part of the training set, and 1,500 are provided for validation. The remaining 1,500 videos are kept for the official test set.

4.2. VideoMem

The VideoMem dataset [4] consists of 10,000 soundless seven-second videos each with associated short-term and long-term memorability scores, however, long-term scores are omitted from this

year's task. Videos were extracted from cinematic raw stock footage and come with a caption. 7,000 videos are released as part of the training set, and 1,500 are provided for validation. The remaining 1,500 videos are kept for the official test set.

4.3. EEGMem

The EEGMem dataset comprises pre-extracted features from EEG recordings for 12 subjects captured while they watched a subset of the Memento10k [7] videos. Participants watched the same videos again through a custom-built online portal between 24–72 hours after the video-EEG recording session, where they were required to indicate for each video whether or not they recognised it, providing binary ground truth annotations[1].

5. Evaluation

A total of five runs can be submitted by each participant for each sub-task. For sub-task 1 all information relating to the Memento10k dataset, i.e., ground-truth data, annotation data, pre-extracted features, and features extracted from provided material, may be used to build the system. For sub-task 2, in similar fashion to sub-task 1, all information relating to the Memento10k and VideoMem datasets may be used to build the system, however, only one dataset may be used per run, and must be evaluated on the other dataset to assess generalisability. For sub-task 3 the only requirement is that EEG data must be, to some extent, included in the system.

Three standard metrics will be used to assess participant system performance for sub-tasks 1 and 2: Spearman's rank correlation, Pearson correlation, and mean squared error. However, similar to previous years, Spearman's rank correlation will be adopted as the official metric as it enables inter-method comparisons by taking into account monotonic relationships between ground-truth data and system output. Submissions for sub-task 3 will be evaluated using the Area Under the Receiver Operating Characteristic Curve.

6. Conclusions

This paper presents an overview of the fith edition of the MediaEval Predicting Video Memorability task. Similar to previous years, the task presents a framework to evaluate the prediction of the memorability of short form videos. This year the task focuses on short-term memorability and introduces a task based on EEG signals. Details regarding the participants' approaches and their results can be found in the proceedings of the 2022 MediaEval workshop[2].

Acknowledgements

Science Foundation Ireland under Grant Number SFI/12/RC/2289_P2, cofunded by the European Regional Development Fund. Financial support also provided by the University of Essex Faculty of Science and Health Research Innovation and Support Fund. Financial support also provided under project AI4Media, a European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911.

[1]Further details on the EEGMem dataset and data collection protocol are available at: https://bit.ly/3BTstj7
[2]See CEUR Workshop Proceedings (CEUR-WS.org).

References

[1] P. Isola, J. Xiao, A. Torralba, A. Oliva, What makes an image memorable, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 145–152.

[2] R. S. Kiziltepe, M. G. Constantin, C.-H. Demarty, G. Healy, C. Fosco, A. García Seco de Herrera, S. Halder, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, L. Sweeney, Overview of the MediaEval 2021 predicting media memorability task, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2021. URL: http://ceur-ws.org/Vol-3181/.

[3] A. García Seco de Herrera, R. Savran Kiziltepe, J. Chamberlain, M. G. Constantin, C.-H. Demarty, F. Doctor, B. Ionescu, A. F. Smeaton, Overview of MediaEval 2020 predicting media memorability task: What makes a video memorable?, in: Working Notes Proceedings of the MediaEval 2020 Workshop, 2020. URL: http://ceur-ws.org/Vol-2882/.

[4] 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 International Conference on Computer Vision, 2019, pp. 2531–2540.

[5] R. Cohendet, K. Yadati, N. Q. Duong, C.-H. Demarty, Annotating, understanding, and predicting long-term video memorability, in: Proceedings of the 2018 ACM International Conference on Multimedia Retrieval, 2018, pp. 178–186.

[6] R. S. Kiziltepe, L. Sweeney, M. G. Constantin, F. Doctor, A. G. S. de Herrera, C.-H. Demarty, G. Healy, B. Ionescu, A. F. Smeaton, An annotated video dataset for computing video memorability, Data in Brief 39 (2021) 107671.

[7] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 223–240.

[8] L. Sweeney, A. Matran-Fernandez, S. Halder, A. G. S. de Herrera, A. F. Smeaton, G. Healy, Overview of the EEG pilot subtask at MediaEval 2021: predicting media memorability, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2021. URL: http://ceur-ws.org/Vol-3181/.

[9] D. Azcona, E. Moreu, F. Hu, T. Ward, A. F. Smeaton, Predicting media memorability using ensemble models, in: Proceedings of MediaEval 2019, Sophia Antipolis, France, CEUR Workshop Proceedings, 2019. URL: http://ceur-ws.org/Vol-2670/.

[10] L. Sweeney, G. Healy, A. F. Smeaton, The influence of audio on video memorability with an audio gestalt regulated video memorability system, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2021. URL: http://ceur-ws.org/Vol-3181/.

[11] T. Zhao, I. Fang, J. Kim, G. Friedland, Multi-modal ensemble models for predicting video memorability, in: Proceedings of the MediaEval 2020 Workshop, CEUR Workshop Proceedings, 2020. URL: http://ceur-ws.org/Vol-2882/.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60 (2017) 84–90.

[13] 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.

[14] D.-C. He, L. Wang, Texture unit, texture spectrum, and texture analysis, IEEE Transactions on Geoscience and Remote Sensing 28 (1990) 509–512.

[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[16] 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 Recognitionn, 2017, pp. 4700–4708.

[17] 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.

[18] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[19] 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.