Overview of The MediaEval 2021 Predicting Media Memorability Task

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

[1]University of Essex, UK

[2]University Politehnica of Bucharest, Romania

[3]InterDigital, France

[4]Dublin City University, Ireland

[5]Massachusetts Institute of Technology Cambridge, Massachusetts, USA.

alba.garcia@essex.ac.uk

ABSTRACT

This paper describes the MediaEval 2021 *Predicting Media Memorability* task, which is in its 4th edition this year, as the prediction of short-term and long-term video memorability remains a challenging task. In 2021, two datasets of videos are used: first, a subset of the TRECVid 2019 Video-to-Text dataset; second, the Memento10K dataset in order to provide opportunities to explore cross-dataset generalisation. In addition, an Electroencephalography (EEG)-based prediction pilot subtask is introduced. In this paper, we outline the main aspects of the task and describe the datasets, evaluation metrics, and requirements for participants' submissions.

1 INTRODUCTION

Information retrieval and recommendation systems deal with exponential growth in media platforms such as social networks and media marketing. New methods of organising and retrieving digital material are needed in order to increase the usefulness of multimedia events in our daily lives. Memorability, like other important video properties, such as aesthetics or interestingness, can be viewed as useful to contribute in the selection of competing videos especially when developing advertising or instructional material. In advertising, predicting the memorability of a video is important since multimedia materials have varying effects on human memory. In addition to advertising, this task may have an impact on other fields such as film making, education, and content retrieval.

The *Predicting Media Memorability* task addresses this problem. The task is part of the MediaEval benchmark and, following the success of previous editions [2, 4, 6, 15], creates a common benchmarking protocol and provides a ground truth dataset for short-term and long-term memorability using common definitions.

2 RELATED WORK

The computational study of video memorability is a natural extension of research into image memorability prediction, which has gained increasing attention in the years after Isola et al.'s work [11]. Models have reached remarkable predictive accuracy for image memorability [12, 19], and we have just begun to see the application of approaches such as style transfer to enhance image memorability [17], demonstrating that we have progressed from simply measuring memorability to using it as an evaluation aspect.

In contrast, computer science research on visual memorability (VM) is still in its early stages. Recent studies on video memorability have focused on the short-term [14], but the lack of studies on VM can be explained by a number of factors. To begin with, there are currently not enough publicly available data sets for training and testing models.The second problem is the absence of a standardised definition for VM. In terms of modeling, previous attempts at VM prediction [3, 16] have identified several features that contribute to VM prediction, including semantic, saliency, and colour features. However, the work is far from complete, and our ability to propose effective computational models will aid in meeting the challenge of VM prediction.

The purpose of this task is to contribute to the harmonisation and advancement of this rapidly growing multimedia field. Additionally, in contrast to prior work on image memorability prediction in which memorability was tested only a few minutes after memorisation, we present a dataset containing long-term memorability annotations. We expect that models trained on this will produce predictions that are more indicative of long-term memorability, which is preferred in a wide variety of applications. This year we also distribute an external dataset for generalisation purposes and propose a new pilot task which is EEG-based video memorability.

3 TASK DESCRIPTION

The *Predicting Media Memorability* task asks participants to develop automatic systems that predict short-term and long-term memorability scores from short videos. Participants were given a dataset of short videos with short-term and long-term memorability scores, raw annotations, and extracted features. Participants were assigned three sub-tasks:

- **Video-based prediction**: Participants are required to generate automatic systems that predict short-term and long-term memorability scores of new videos based on the given video dataset and their memorability scores.
- **Generalization (optional)**: Participants will train their system on one of the two sources of data we provide and will test them on the other source of data. This is an optional sub-task.

MediaEval'21, December 13-15 2021, Online

- **Electroencephalography (EEG)-based prediction (pilot)**: Participants are required to generate automatic systems that predict short-term memorability scores of new videos based on the given EEG data. This is a pilot sub-task and details for it can be found in [20].

4 COLLECTION

This task utilises a subset of the TRECVID 2019 Video-to-Text video dataset [1]. The dataset contains Twitter Vine videos where various actions are performed. This year, the dataset has been expanded and normalised short-term memorability scores are provided with memory alpha decay values. Additionally, we open the Memento10K [14] dataset to participants. Apart from traditional video information like metadata and extracted visual features, part of the data will be accompanied by Electroencephalography (EEG) recordings that would allow to explore the physical reactions of users.

A set of pre-extracted features are also distributed as follows:

- image-level features: AlexNetFC7 [13], HOG [5], HSVHist, RGBHist, LBP [7], VGGFC7 [18], DenseNet121 [10], ResNet50 [8], EfficientNet b3 [21];
- video-level feature: C3D [22];
- audio-level feature: VGGish [9].

Three frames from each video were used to extract image-level features: the first, the middle, and the last frame. Additionally, each TRECVid video includes at least two textual captions summarising the action, whereas Memento10K includes five. The annotations acquired from participants included the first and second appearance positions of each target video, as well as participants' response times and the keys pressed while watching each video.

4.1 TRECVid 2019 Video-to-Text dataset

The TRECVid 2019 Video-to-Text dataset [1] contains 6,000 videos. In 2021, three subsets were distributed as part of the MediaEval Predicting Media Memorability task. The training set contained 588 videos, the development set 1,116 videos and the test set 500 videos. Each video has two associated memorability scores indicating its likelihood of being remembered after two distinct periods of memory retention. Similar to previous editions of the task [2, 4], memorability was measured twice using recognition test: a few minutes after the videos were shown (short-term) and 24-72 hours later (long-term). The videos are released under Creative Commons licences that allow their redistribution.

The ground truth dataset was generated using a video memorability game protocol proposed by Cohendet et al. [3]. The memorability game was formed in two versions. One was made available on Amazon Mechanical Turk (AMT), and another was made available for general use in three languages: English, Spanish, and Turkish.

In the video memorability game protocol, participants were expected to watch 180 and 120 videos in short-term and long-term memorisation steps, respectively. The goal was essentially to press the space bar whenever participants recognise a previously seen video, which allows for the determination of which videos they do and do not recognise. The game begins with the repetition of 40 target videos after a few minutes to accumulate short-term memorability labels. Regarding the first step's filler videos, 60 non-vigilance filler videos are shown once. After a few seconds, 20 vigilance filler videos are repeated to ensure that participants are paying attention to the task. After 24 to 72 hours, the same individuals are anticipated to return for the second step, which involves collecting labels for long-term memorability. This time, 40 target videos chosen at random from the non-vigilance fillers in the first stage and 80 fillers chosen at random from new videos are displayed to determine the target videos' long-term memorability scores. The percentage of correct recognition for each video is used to calculate short-term and long-term memorability scores.

4.2 Memento10K dataset

The Memento10K dataset [14] contains 10,000 three-second videos depicting in-the-wild scenes, with their associated short-term memorability scores, memorability decay values, action labels, and 5 accompanying captions. The scores were computed with 90 annotations per video on average, and the videos were shown to participants without sound. 7,000 videos were released as a training set, and 1,500 were provided for validation. The last 1,500 videos were used as the test set for scoring submissions.

5 SUBMISSION AND EVALUATION

As it is with previous tasks, each team is expected to submit both short-term and long-term memorability predictions. A total of ten runs, five for each, can be submitted for video-based prediction. In addition, participants can submit five runs per optional subtask (generalisation and EEG-based prediction). All information, including given features, ground truth data, video sample titles, features extracted from visual material, and even external data, may be used to build the system. Short-term and long-term annotation memorability runs must be submitted separately and must not include each others.

Classic evaluation metrics (including Spearman's rank correlation) are used to compare the predicted memorability scores for the videos with the ground truth memorability scores.

6 DISCUSSION AND OUTLOOK

In this paper we introduced the 4th edition of the Predicting Media Memorability at the MediaEval 2021 Benchmarking initiative. With this task, a comparative assessment of current state-of-the-art machine learning techniques to predict short- and long-term memorability can be conducted. A dataset containing short videos is distributed with memorability annotations and external data is provided for generalisation purposes. Moreover, EEG annotations are also provided for a pilot study. Related information has also been made available to participants so they can refine their strategies. The 2021 MediaEval workshop proceedings presents details on the participants' approaches to the task including methodologies used and findings.

ACKNOWLEDGMENTS

MGC and BI's contribution is supported under project AI4Media, a European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911. The work of RSK is partially funded by the Turkish Ministry of National Education. This work was part-funded by NIST Award No. 60NANB19D155 and by Science Foundation Ireland under grant number SFI/12/RC/2289_P2.

REFERENCES

[1] George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, and others. 2019. TRECVID 2019: An Evaluation Campaign to Benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & Retrieval. (2019).

[2] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. MediaEval 2018: Predicting media memorability task. In *Working Notes Proceedings of the MediaEval 2018 Workshop*. Sophia Antipolis, France.

[3] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. 2019. VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability. In *Proceedings of the IEEE International Conference on Computer Vision*. 2531–2540.

[4] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc QK Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. Predicting Media Memorability Task at MediaEval 2019. In *Working Notes Proceedings of the MediaEval 2019 Workshop*. Sophia Antipolis, France.

[5] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 886–893.

[6] Alba García Seco de Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-Hélène Demarty, Faiyaz Doctor, Bogdan Ionescu, and Alan F. Smeaton. 2020. Overview of MediaEval 2020 Predicting Media Memorability Task: What Makes a Video Memorable?. In *Working Notes Proceedings of the MediaEval 2020 Workshop*.

[7] 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.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[11] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2013. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2013), 1469–1482.

[12] 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.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[14] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 2020. Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 223–240.

[15] Rukiye Savran Kiziltepe, Lorin Sweeney, Mihai Gabriel Constantin, Faiyaz Doctor, Alba García Seco de Herrera, Claire-Hélène Demarty, Graham Healy, Bogdan Ionescu, and Alan F. Smeaton. 2021. An Annotated Video Dataset for Computing Video Memorability. *Data in Brief* (2021), 107671. https://doi.org/10.1016/j.dib.2021.107671

[16] 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 Workshops*. 2730–2739.

[17] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2019. Increasing Image Memorability with Neural Style Transfer. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 2, Article 42 (June 2019), 22 pages. https://doi.org/10.1145/3311781

[18] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[19] Hammad Squalli-Houssaini, Ngoc QK Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2018. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2371–2375.

[20] Lorin Sweeney, Ana Matran-Fernandez, Sebastian Halder, Alba García Seco de Herrera, Alan F. Smeaton, and Graham Healy. 2021. Overview of the EEG Pilot Subtask at MediaEval 2021: Predicting Media Memorability. *Working Notes Proceedings of the MediaEval 2021 Workshop*.

[21] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[22] 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.