Overview of MediaEval 2020 Predicting Media Memorability Task: What Makes a Video Memorable?

Alba G. Seco de Herrera[1], Rukiye Savran Kiziltepe[1], Jon Chamberlain[1], Mihai Gabriel Constantin[2], Claire-Hélène Demarty[3], Faiyaz Doctor[1], Bogdan Ionescu[2], Alan F. Smeaton[4]

[1]University of Essex, UK
[2]University Politehnica of Bucharest, Romania
[3]InterDigital, R&I, France
[4]Dublin City University, Ireland.
alba.garcia@essex.ac.uk

ABSTRACT

This paper describes the MediaEval 2020 *Predicting Media Memorability* task. After first being proposed at MediaEval 2018, the *Predicting Media Memorability* task is in its 3rd edition this year, as the prediction of short-term and long-term video memorability (VM) remains a challenging task. In 2020, the format remained the same as in previous editions. This year the videos are a subset of the TRECVid 2019 Video-to-Text dataset, containing more action rich video content as compared with the 2019 task. In this paper a description of some aspects of this task is provided, including its main characteristics, a description of the collection, the ground truth dataset, evaluation metrics and the requirements for participants' run submissions.

1 INTRODUCTION

Media platforms such as social networks, media advertisements, information retrieval and recommendation systems deal with exponential growth. Enhancing the relevance of multimedia occurrences in our everyday lives requires new ways to organise – in particular, to retrieve – digital content. Like other video metrics of importance, such as aesthetics or interestingness, memorability can be regarded as a useful aspect to help make a choice between competing videos. This is even truer when one considers specific use cases of creating commercials or educational content. Because the impact of different multimedia content, images or videos, on human memory is unequal, the capability of predicting the memorability of a given piece of video content is of high importance for professionals in the field of advertising and other fields. Beyond advertising, other applications, such as film-making, education, content retrieval, etc., may also be influenced by this task.

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], 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 understanding of video memorability follows on from the study of image memorability prediction, which has attracted increasing attention since the seminal work of Isola et al. [7]. Models have achieved very good results at predicting image memorability [8, 15] and we have recently started to see the use of techniques like style transfer to improve image memorability [13] thus illustrating that we have now moved from just measuring memorability, to using memorability as an evaluation metric.

In contrast, research on visual memorability (VM) from a computer science point of view is in its early stage. Recently we have seen other work on video memorability [11] with a particular focus on short term, but the scarcity of studies on VM can be explained by several reasons. Firstly, there is no publicly available data set to train and test models, though the VideoMem [12] and the Memento10k [11] datasets are recent additions. The second point, closely related to the previous one, is the lack of a common definition for VM. Regarding modelling, previous attempts at predicting VM [3, 12] have highlighted several features which contribute to the prediction of VM, such as semantic, saliency and colour features, but the work is far from complete and our capacity to propose effective computational models will help to meet the challenge of VM prediction.

The goal of this task is to participate in the harmonisation and the advancement of this emerging multimedia field. Furthermore, in contrast to previous work on image memorability prediction, where memorability was measured a few minutes after memorisation, we propose a dataset with longer term memorability annotations. We expect the predictions of the models trained on this data to be more representative of long-term memory, which is used preferably in numerous applications.

3 TASK DESCRIPTION

The *Predicting Media Memorability* task requires participants to automatically predict memorability scores for short form videos, that reflect the probability for a video to be remembered. Participants were provided with a dataset of videos with short-term and long-term memorability annotations, related information, and pre-extracted state-of-the-art visual features. Therefore, two subtasks were proposed to participants:

- **Short-term VM prediction** - scores were measured a few minutes after the memorisation process;
- **Long-term VM prediction** - scores were measured 24-72 hours after the memorisation process.

MediaEval'20, 14-15 December 2020, Online

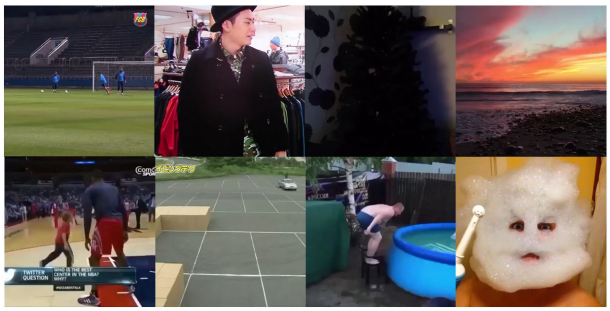

Figure 1: A sample of frames of the videos in the *TRECVid 2019 Video-to-Text dataset.*

4 COLLECTION

The dataset is composed of a subset of short videos selected from the TRECVid 2019 Video-to-Text dataset [1] (see Figure 1). These videos are shared under Creative Commons licenses that allow their redistribution. The TRECVid videos have much more action happening in them compared with those in the 2019 VM task, and thus they correspond to more generic use cases.

Each video consists of a coherent unit in terms of meaning and is associated with two scores of memorability that refer to its probability to be remembered after two different time durations of memory retention. A set of pre-extracted features are also distributed:

- image-level features: AlexNetFC7 [9], HOG [5], HSVHist, RGBHist, LBP [6], VGGFC7 [14];
- video-level feature: C3D [16].

The image-level features were extracted from 3 frames for each video: the first, the middle and the last frame. In addition, each TRECVid video is accompanied by two textual captions describing the activity. Additional information on the annotation was also provided to allow further investigation of the user interaction for memorability. Hence, the annotations collected from participants were provided including the first appearance position and the second appearance position of each target video along with the response time of the user and the key pressed when watching each video.

The TRECVid 2019 Video-to-Text dataset [1] contains 6,000 videos. In 2020, three subsets were distributed as part of the MediaEval Predicting Media Memorability task. The training set contained 590 videos, the development set 410 videos and the test set 500 videos. Each video was annotated by at least 16 annotators for their short term memorability. However, there are fewer long term annotations.

Similar to previous editions of the task [2, 4], memorability has been measured using recognition tests, i.e., through an objective measure, a few minutes after the memorisation of the videos (short term), and then 24 to 72 hours later (long term). The ground truth dataset was collected by using a video memorability game protocol proposed by Cohendet et al. [3]. Two versions of the memorability game were published. One was published on Amazon Mechanical Turk (AMT) and another one was issued for general use with following three language options: English, Spanish and Turkish.

In the game of video memorability, participants are expected to watch 180 and 120 videos in short-term and long-term memorisation steps, respectively. The task is basically to press the space bar once the participants recognise a previously seen video, which enables to determine videos recognised and not recognised by them. In the first step of the game, 40 target videos are repeated after a few minutes to collect short-term memorability labels. As for filler videos in the first step, 60 non-vigilance filler videos are displayed once. 20 vigilance filler videos are repeated after a few seconds to check participants' attention to the task. After 24 hours to 72 hours, the same participants are expected to attend the second step for collecting long-term memorability labels. This time, 40 target videos chosen randomly from among non-vigilance fillers of the first step and 80 fillers selected randomly from new videos are displayed to measure long-term memorability scores for those target videos. Both short-term and long-term memorability scores are calculated as the percentage of correct recognition for each video by the participants. Relevant screenshots and label collection procedures are demonstrated on the MediaEval task web page [10].

5 SUBMISSION AND EVALUATION

As in previous editions of the task, each team is required to predict both short and long term memorability. In total, 10 runs can be submitted, 5 for each. 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. The only exception, in this case, is that the required short-term and long-term memorability runs must not use each other's score annotations. For the rest of the runs, a maximum of 4 per subtask, everything is permitted, including using cross-annotations between the subtasks.

The outputs of the prediction models – i.e., the predicted memorability scores for the videos – will be compared with ground truth memorability scores using classic evaluation metrics (e.g., Spearman's rank correlation).

6 DISCUSSION AND OUTLOOK

In this paper we presented the third edition of the Predicting Media Memorability at the MediaEval 2020 Benchmarking initiative. The task provides a framework that allows a comparative study of different state of the art Machine Learning approaches aiming to predict short and long-term memorability. A collection of videos is provided as well as memorability annotations and a common evaluation metric. In addition, related information has been provided to help participants in developing their approaches. Details regarding the methods employed by participants and their results can be found in the proceedings of the 2020 MediaEval workshop[1].

ACKNOWLEDGMENTS

This work was part-funded by NIST Award No. 60NANB19D155, by Science Foundation Ireland under grant number SFI/12/RC/2289_P2 and under project AI4Media, A European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant 951911.

[1] See CEUR Workshop Proceedings (CEUR-WS.org).

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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] MediaEval. 2020. MediaEval 2020: Predicting Media Memorability. (2020). https://multimediaeval.github.io/editions/2020/tasks/memorability/ Accessed: 2020-11-26.

[11] 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.

[12] 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.

[13] 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

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

[15] 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.

[16] 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.