Assessing the difficulty of predicting media memorability

Mihai Gabriel Constantin
mihai.constantin84@upb.ro
University Politehnica of Bucharest
Bucharest, Romania

Mihai Dogariu
University Politehnica of Bucharest
Bucharest, Romania

Andrei-Cosmin Jitaru
University Politehnica of Bucharest
Bucharest, Romania

Bogdan Ionescu
University Politehnica of Bucharest
Bucharest, Romania

ABSTRACT

Memorability is a critical aspect of human cognition that has been studied extensively in various fields, including psychology, education, and computer vision. The ability to remember information and experiences over time is essential for learning, decision-making, and creating lasting impressions. While the number of computer vision works that attempt to predict the memorability score of videos has recently seen a significant boost, thanks to several benchmarking tasks and datasets, some questions related to the performance of automated systems on certain types of videos are still largely unexplored. Given this, we are interested in discerning what makes a video sample easy or hard to classify or predict from a memorability standpoint. In this paper, we use a large set of runs, created and submitted by the participants to the MediaEval Predicting Video Memorability task, and, using their results and a set of visual, object, and annotator-based features and analyses, we attempt to find and define common traits that make the memorability scores of videos hard or easy to predict.

CCS CONCEPTS

• **Computing methodologies → Computer vision**; **Machine learning**.

KEYWORDS

media memorability, benchmarking, prediction analysis, perception

ACM Reference Format:
Mihai Gabriel Constantin, Mihai Dogariu, Andrei-Cosmin Jitaru, and Bogdan Ionescu. 2023. Assessing the difficulty of predicting media memorability. In *Proceedings of 20th International Conference on Content-based Multimedia Indexing (CBMI '23).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The study of memory and its capacity for filtering the large quantity of information humans are constantly bombarded with has been a subject of interest in research for a significant period of time. This domain attracted researchers from many domains like psychology and physiology [20, 23], machine learning and computer vision [19, 22], and benckmarking tasks, datasets and human-based surveys [16, 17]. Interestingly, memorability is shown to be an intrinsic property of images, that reflects "the extent to which the image can be remembered by the human mind" [12, 13].

While a substantial number of papers have been dedicated to studying methods of determining the memorability score of an image or video sample [16, 22], interest in this domain has significantly grown, thanks to the development of benchmarking tasks that target the creation of automated methods for the prediction of media memorability. Under the umbrella of the MediaEval Multimedia Evaluation Benchmark competition[1], the Predicting Video Memorability (PVM) task is now at its fifth edition [24], bringing at least 7 participating teams during each of its editions, and a total of 43 participating teams over the five editions. This not only greatly increases the amount of interest around the subject of media memorability, but also creates a significant baseline with regards to determining which methods, processing steps and learners may positively contribute to the precision of memorability predictor systems.

2 RELATED WORK

Previous work on memorability and how it is influenced by various factors saw studies that target the impact of colors, hues and brightness [5, 22], as well as high-level attributes like the presence and salience of certain objects [14]. On the other hand, memorability has been studied along side other concepts like image aesthetics and social interestingness [9], authors looking for positive or negative correlations with other concepts related to the subjective perception of multimedia items.

Our work is built upon the aforementioned 2022 MediaEval Predicting Video Memorability task [24]. This edition of the benchmarking task attracted 10 participating teams, that submitted a total of 33 runs. Participants are asked to develop and train their video memorability prediction systems on the training data, and submit their runs and predictions on the testing data. A wide variety of machine learning methods and models have been submitted, ranging from methods that use adapted language models [8] and convolutional and deep features [21] to methods that employ transformer networks [2] and ensembles of various networks [1]. Participants also starting delving into using electroencephalogram (EEG) data for inferring memorability, starting from data collected during a

CBMI '23, September 20–22, 2023, Orleans, France

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1] https://multimediaeval.github.io/

preliminary EEG study in the 2021 edition [25]. It is obvious that a large body of work is dedicated to the prediction of media memorability, using a vast diversity of methods, predictors and processing algorithms. However, to the best of our knowledge, no study has yet been made with regards to the correlations between the performance of these methods and the aspect, features and qualities of the videos themselves, while also trying to understand what makes a video hard or easy to predict from a computer vision point of view. Such a study would, in our opinion, contribute to the understanding and explainability of the performances of memorability prediction systems. On the other hand, by providing a list of visual samples that may be harder to classify, robustness may be increased, by providing a list of visual samples that have to be thoroughly analyzed, or augmented via traditional transformation methods, or even generative approaches. In this context, our paper proposes the following contributions over the current state of the art:

- We gather and process a large number of runs submitted by participants to the 2022 MediaEval Predicting Video Memorability task, and process them in order to understand which of the videos in the Memorability task are harder to automatically predict;
- We identify and test a large number of video features that are able to provide an easy to understand overview of what makes a video harder or easier to classify with regards to memorability;
- We analyze the correlation between video ground truth data, based on human annotator assessment, and how well the participant methods perform on the given testset.

3 METHODOLOGY

Starting from the 33 runs submitted by the participants to the 2022 edition of the Predicting Video Memorability task, included in the Prediction subtask (subtask 1), we propose the creation of a metric that could accurately measure how easy it is for the methods and models represented by the 33 runs to accurately classify the videos for this task. This section will present the systems we will work with, some pre-processing considerations, the main principles of the difficulty metric, and video categorization principles.

3.1 The 2022 Predicting Video Memorability dataset

The 2022 PMV task and dataset [24] proposes a dataset composed of 10,000 short three-second videos annotated for short-term memorability and extracted from the Memento10k [19] dataset. It is divided into a training set composed of 7,000 videos, a validation set of 1,500 videos, with the remaining 1,500 videos being alloted to the testing set. Participants must use the training and validation data in order to design, create and train their systems, and must submit their prediction runs on the final testing set. In this paper, we propose using the 33 systems submitted by participants and analyze their performance in order to better understand what types of videos are harder to classify with regards to their memorability.

A preliminary pre-processing step involves checking the runs submitted by participants. Runs with missing samples were accepted to the competition as valid, as this would affect the final performance of the systems, but it would not create a situation where it would be impossible to continue the experiments. However, in our case missing samples may significantly affect the computation of the Distance metric (see Section 3.2) and may change a video's ranking. For these reasons we had to drop two of the submissions, ending up with a total of 31 valid runs.

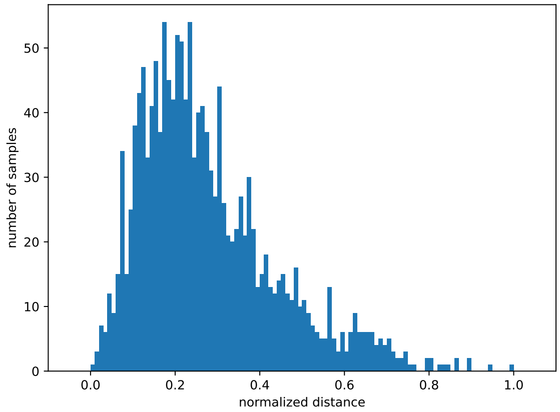

Figure 1: Distribution of the video samples in PVM according to the Distance metric.

3.2 Distance metric

The MediaEval PVM task used the same official metric for measuring the performance of individual systems throughout all its five editions, namely the Spearman's rank correlation metric:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{1}$$

where N is the number of samples in the two collections that are compared, and d_i represents the difference between the two ranks for sample i, with $i \in [1, N]$. Therefore, for the given data and metric, the difference from the ground truth data ranks $G = \{g_1, g_2, ..., g_n\}$ is defined by the d_i distance. Given a set of M participant runs, $R = \{R_1, R_2, ..., R_M\}$, each run having the ranks of its predictions for the N samples in the dataset $R_j = \{p_{1,j}, p_{2,j}, ..., p_{N,j}\}$, the distance for a video i can be expressed as:

$$D_i = \sum_{j=1}^{M} \left| p_{i,j} - g_i \right| \tag{2}$$

In the final step, we propose normalizing the set of distances for the N videos, obtaining a set of distances $\hat{D} = \{\hat{D_1}, \hat{D_2}, ..., \hat{D_N}\}$ with values $\hat{D_i} \in [0, 1]$.

3.3 Video categories

The result of the process described in Section 3.2 is presented in Figure 1. It is obvious that results are skewed towards lower values of the distance metric $\hat{D}$, indicating more easy to predict videos, with a median value of 0.2391 and an average value of 0.2774. We create two separate groups of video categories. The first one splits the videos into equal quartiles, denoted $Q1$, $Q2$, $Q3$, and $Q4$, where $Q1$ represents the easiest 25% of videos to predict according to

Figure 2: Randomly chosen samples from each video category, representing the four quartiles and four thresholds.

the distance $\hat{D}_i$ for each video, $Q2$ the second easiest set of videos, and so on. These sets are all composed of 375 videos. The next class divides the videos according to a set of thresholds, creating splits denoted as $T1$, $T2$, $T3$, and $T4$. In this case, the $T1$ category joins videos with $0 \leq \hat{D}_i < 0.25$, the $T2$ category videos with $0.25 \leq \hat{D}_i < 0.5$ and so on. The sizes of these four collections are, in order: 787, 558, 130, and 25, with the majority of videos going into the easiest to predict categories, namely $T1$ and $T2$. Some visual examples of these splits are presented in Figure 2.

4 PROPOSED ANALYSIS METHODS

Starting with these collections of video categories, we propose applying a set of feature computation methods that are able to describe each video in the entire testset, and the video categories themselves, as averages over the set of videos they are composed of. We will use three sets of features, namely: (i) a global visual feature set, computing a variety of visual attributes, (ii) an object-based feature set, that analyzes attributes related to the presence of objects in the videos, and (iii) an annotator-based computation, that analyzes if there are any correlations between the way human assessors annotated the videos and how well predictors perform.

4.1 Visual features

We implement a number of visual feature computation schemes, aiming at understanding the visual characteristics that make a video's memorability easy (i.e., belonging to $Q1$ or $T1$) or hard (i.e., belonging to $Q4$ or $T4$) to predict by machine learning methods. Considering the nature of the videos, we take three frames from each video and apply each feature computation function to each of these frames. The final value of the function for a video is expressed as the average over the three frames, while for an entire category of videos the value is computed as the average over the entirety of videos. While these features represent a traditional approach to processing videos, they provide a measurable, understandable, and explainable quantity for each chosen concept. They are chosen based on their previous use in many studies that seek to analyze the visual qualities of a visual sample [3, 7, 15, 18].

We start by computing the sharpness via the Laplacian (f_1) operator [26] and Canny (f_2) operator. Following this, we apply a metric for computing the colorfulness of images (f_3), based on the "psychophysical" experiments described in [10], and a contrast feature (f_4) that computes the contrast of images in RGB color space [15]. We then average the pixels of the images transformed to HSL color space, resulting in a hue (f_5), saturation (f_6) and brightness (f_7).

The final visual feature uses a video-level type of descriptor, analyzing the dynamism across the entire video (f_8). We compute this by summing up the absolute values of the movement magnitude vectors using a dense optical flow function computed via the Farneback method [6]. This feature uses all the frames in the video.

4.2 Object-based features

We theorize, based on previous experiments on the correlation between the presence of objects in images and subjective concepts like interestingness [3], that certain objects may help some prediction methods to perform better, while others, or the lack of objects may introduce uncertainty, not only in the way algorithms work, but also in the way human annotators perform for these videos. We use the architecture presented in [4] for automatically annotating the videos, an architecture based on the popular Mask R-CNN approach [11]. Our interest in this case is in exploring the most common objects present in the video categories. Concretely, this feature (f_9) explores the top-5 most common objects, as well as the percentage of images from a collection they appear in. Finally, using the masks of the detected objects, we compute the percentage of the frame that is covered with detectable objects, thus creating a second object-based feature (f_{10}).

4.3 Annotator-based feature

Finally we wish to analyze the correlation between the created video categories and the ground truth memorability values assigned to the data, according to the results registered by human assessors. Starting from the four quartiles ($Q1$, $Q2$, $Q3$, and $Q4$) that have an equal amount of videos assigned to them, we compute and analyze a set of histograms (f_{11}) that measure the distribution of each video category given the ground truth memorability of the videos according to their Spearman's coefficient.

5 EXPERIMENTAL RESULTS

Given this set of 11 features, we will present and analyze the results they show for the chosen video categories, either expressing these results as differences between categories or by analysing the data generated by each feature. It may be important to remember that, while the quartile sets of videos have the same number of elements in their composition, for the threshold categories, especially for $T4$, there may be few samples in the set and this may affect the results.

5.1 Visual features results

The results of the 8 visual features (f_1 - f_8) are presented in Table 1, where we present the differences between the categories and a baseline composed of $Q1$ or $T1$. It is interesting to note that in the majority of cases, the differences are all either positive or negative, indicating a clear tendency. Starting with the color analysis, the tendency seems to show that videos with higher colorfulness (f_3) and lower color saturation (f_6), and those with higher brightness (f_7) seem to be easier to correctly predict. On the other hand, the results for hue values (f_5) seem to be mostly constant, with the exception of $T4$. Also, when analyzing the sharpness features (f_1 and f_2) and contrasts (f_4) it would seem that the easiest to classify

Table 1: Percentage change between the harder quartiles ($Q2$ - $Q4$) and the easiest quartile ($Q1$), and between the harder threshold intervals ($T2$ - $T4$) and the easiest interval ($T1$) for the visual features (f_1 - f_8), the top 5 objects overall (with no objects added for a top 6), and object coverage feature (f_{10}).

Feature	Q2	Q3	Q4	T2	T3	T4
f_1	30.58%	40.82%	30.12%	20.67%	18.11%	32.48%
f_2	16.67%	18.59%	11.36%	5.67%	4.99%	31.25%
f_3	-6.54%	-4.25%	-4.11%	-3.62%	-1.53%	3.76%
f_4	15.54%	17.44%	18.65%	9.35%	3.27%	5.01%
f_5	-1.05%	-0.21%	-0.39%	-0.35%	2.42%	10.97%
f_6	0.54%	6.51%	2.59%	3.32%	1.55%	5.14%
f_7	-6.84%	-6.21%	-6.43%	-3.55%	-0.66%	-1.14%
f_8	-1.67%	-10.91%	-10.01%	-5.51%	-18.38%	-38.13%
$f_9 - pers$	5.63%	2.63%	1.12%	-1.99%	-3.55%	9.11%
$f_9 - none$	27.69%	19.94%	47.97%	-3.77%	32.67%	-27.1%
$f_9 - chair$	-4.21%	-8.28%	-12.5%	5.73%	-29.67%	-
$f_9 - car$	-21.1%	-31.55%	-10.65%	-7.41%	-35.73%	-
$f_9 - table$	92.24%	199.46%	99.46%	25.36%	45.77%	94.4%
$f_9 - bird$	62.91%	138.02%	62.91%	92.14%	74.64%	-
f_{10}	-7.78%	-12.08%	-10.56%	-7.44%	-11.09%	-12.34%

videos have lower average values, however they also have more dynamism, as shown by the f_8 feature.

5.2 Object-based results

The following objects are listed in the top 5 most present objects overall: "person", "chair", "car", "dining table", and "bird". We add a special class for videos with no objects detected, and present the results in Table 1. Regarding the top-5 most common objects, overall the person class is present in 72.6% of all the videos, chair in 8.94%, car in 5.86%, dining table 4.26%, and bird in 3.93%, while 10.66% of the videos do not have any detected objects in them. One of the most interesting observations in Table 1 seems to be related to videos without discernible objects in them ("none"), where there are clear differences between the four quartiles. There are also significant differences in the car, table, and bird classes of detections, however, considering their relative low representation overall in the entire collection of videos, this may not represent a significant change in video content. Another constant result is represented by the f_{10} feature, which measures the coverage of detectable objects in a video. Overall, both $Q1$ and $T1$ have more object coverage, while objects in the categories representing harder prediction samples have lower object coverage. The object coverage for the $Q1$ and $T1$ categories are 38.15% and 36.69%, while object coverage for $Q4$ and $T4$ are 34.12% and 32.16% respectively.

5.3 Annotator-based results

The results of the final experiment are presented in Figure 3. The histogram is grouped in memorability score intervals of 0.02. We only consider the quartile categories in this case, as these are the categories with equal number of samples in them, and therefore make this comparison fair towards all the categories. Visually, two observations stand out in this analysis: (i) the first one, containing samples that have average performance with regards to memorability ground truth scores are mostly dominated by samples belonging to the $Q2$ or $Q3$ categories, (ii) the samples that have either high or low memorability scores tend to belong to the $Q1$ or $Q2$ categories. Furthermore, most of the hardest to predict movies belong to the first category, being placed between memorability values of 0.7 – 0.9.

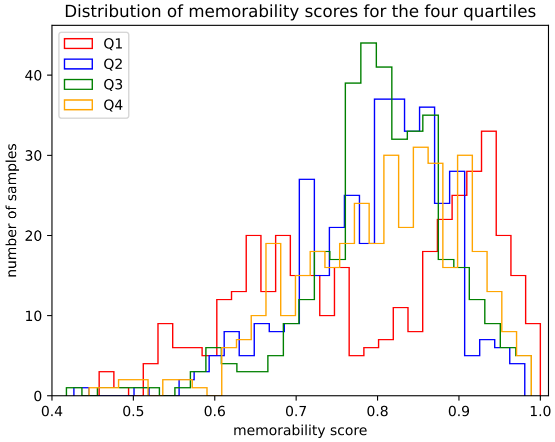

Figure 3: Distribution of the video samples in the four quatiles, according to their memorability score as annotated by human assessors, grouped in score intervals of 0.02.

6 CONCLUSIONS

This paper presents an analysis of typical video features and attributes, and their correlation with how easy or hard it is to classify video samples according to their memorability scores. We gathered runs submitted by participants during the 2022 MediaEval Predicting Video Memorability task, and created categories of videos according to how well the runs performed for each video in the testset. Feature analysis shows that videos that are hard or at least harder to classify have the following characteristics: (i) they have higher contrast and sharpness, but lower dynamism; (ii) they have lower brightness and colorfulness, and a higher saturation; (iii) they have fewer significant objects in them and the coverage of these objects is smaller; and (iv) they tend to have a mid-level memorability score. We believe that this analysis may be useful for future research, as movies in a dataset that have these particularities may be somehow augmented at training time in an attempt to improve classifier performance for hard to classify samples in particular, thus positively impacting overall performance.

ACKNOWLEDGMENTS

The results presented in this article have been funded by the Ministry of Investments and European Projects through the Human Capital Sectoral Operational Program 2014-2020, Contract no. 62461/03.06.2022, SMIS code 153735, and under project AI4Media, a European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911.

REFERENCES

[1] Muhammad Mustafa Ali Usmani, Sumaiyah Zahid, and Muhammad Atif Tahir. 2023. Modelling of Video Memorability using Ensemble Learning and Transformers. In *Working Notes Proceedings of the MediaEval 2022 Workshop*.

[2] Mihai Gabriel Constantin and Bogdan Ionescu. 2023. AIMultimediaLab at MediaEval 2022: Predicting Media Memorability Using Video Vision Transformers and Augmented Memorable Moments. In *Working Notes Proceedings of the MediaEval 2022 Workshop*.

[3] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*. IEEE, 1657–1664.

[4] Mihai Dogariu, Liviu-Daniel Stefan, Mihai Gabriel Constantin, and Bogdan Ionescu. 2020. Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios. In *2020 13th International Conference on Communications (COMM)*. IEEE, 157–160.

[5] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. 2015. What makes an object memorable?. In *Proceedings of the ieee international conference on computer vision*. 1089–1097.

[6] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 363–370.

[7] Helmut Grabner, Fabian Nater, Michel Druey, and Luc Van Gool. 2013. Visual interestingness in image sequences. In *Proceedings of the 21st ACM international conference on Multimedia*. 1017–1026.

[8] Camille Guinaudeau and Andreu Girbau Xalabarder. 2023. Textual Analysis for Video Memorability Prediction. In *Working Notes Proceedings of the MediaEval 2022 Workshop*.

[9] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The interestingness of images. In *Proceedings of the IEEE international conference on computer vision*. 1633–1640.

[10] David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, Vol. 5007. SPIE, 87–95.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[12] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the intrinsic memorability of images. *Advances in neural information processing systems* 24 (2011).

[13] 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.

[14] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *CVPR 2011*. IEEE, 145–152.

[15] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1. IEEE, 419–426.

[16] 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.

[17] Rukiye Savran Kiziltepe, Lorin Sweeney, Mihai Gabriel Constantin, Faiyaz Doctor, Alba García Seco de Herrera, Claire-Héléne Demarty, Graham Healy, Bogdan Ionescu, and Alan F Smeaton. 2021. An annotated video dataset for computing video memorability. *Data in Brief* 39 (2021), 107671.

[18] Michal Kucer, Alexander C Loui, and David W Messinger. 2018. Leveraging expert feature knowledge for predicting image aesthetics. *IEEE Transactions on Image Processing* 27, 10 (2018), 5100–5112.

[19] 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: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 223–240.

[20] WA Phillips. 1974. On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics* 16 (1974), 283–290.

[21] R. Gokul Prakash, Jayaraman Bhuvana, Eeswara Anvesh Chodisetty, Arjun Mukesh, and T.T. Mirnalinee. 2023. Multi-model Estimators and Ensemble-based Regressors for Predicting Video Memorability. In *Working Notes Proceedings of the MediaEval 2022 Workshop*.

[22] 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.

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

[24] Lorin Sweeney, Mihai Gabriel Constantin, Claire-Hélène Demarty, Camilo Fosco, Alba G. Seco de Herrera, Sebastian Halder, Graham Healy, Bogdan Ionescu, Ana Matran-Fernandez, Alan F. Smeaton, and Mushfika Sultana. 2023. Overview of The MediaEval 2022 Predicting Video Memorability Task. In *Working Notes Proceedings of the MediaEval 2022 Workshop*.

[25] Lorin Sweeney, Ana Matran-Fernandez, Sebastian Halder, Alba G Seco de Herrera, Alan 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* (2021).

[26] Jing Wan, Xiaofu He, and Pengfei Shi. 2007. An Iris Image Quality Assessment Method Based on Laplacian of Gaussian Operation.. In *MVA*. 248–251.

Received 24 April 2023; revised 06 July 2023; accepted xx xxxxx 2023