



An Annotated Video Dataset for Computing Video Memorability

Rukiye Savran **Kiziltepe**^a, Lorin **Sweeney**^b, Mihai Gabriel **Constantin**^c, Faiyaz **Doctor**^a, Alba **García Seco de Herrera**^a, Claire-Hélène **Demarty**^d, Graham **Healy**^b, Bogdan **Ionescu**^c, Alan F. **Smeaton**^{b,*}

^a*University of Essex, UK*

^b*Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland*

^c*University Politehnica of Bucharest, Romania*

^d*InterDigital, R&I, France*

ARTICLE INFO

Keywords:

Video memorability

Machine learning

Human memory

MediaEval Benchmark

ABSTRACT

Using a collection of publicly available links to short form video clips of an average of 6 seconds duration each, 1,275 users manually annotated each video multiple times to indicate both long-term and short-term memorability of the videos. The annotations were gathered as part of an online memory game and measured a participant's ability to recall having seen the video previously when shown a collection of videos. The recognition tasks were performed on videos seen within the previous few minutes for short-term memorability and within the previous 24 to 72 hours for long-term memorability. Data includes the reaction times for each recognition of each video. Associated with each video are text descriptions (captions) as well as a collection of image-level features applied to 3 frames extracted from each video (start, middle and end). Video-level features are also provided. The dataset was used in the Video Memorability task as part of the MediaEval benchmark in 2020.

Specifications Table

*Corresponding author.

e-mail: Alan.Smeaton@DCU.ie (Alan F. Smeaton)

Subject	Computer Vision and Pattern Recognition
Specific subject area	Ground truth data (videos, video features plus annotations) needed to build and train systems for the automatic computation of the memorability of short video clips
Type of data	Text files (csv)
How data were acquired	Raw videos are already publicly available online. Low level features were extracted automatically from videos and annotation data was collected through crowdsourcing using a video memorability game with the participation of both volunteers and paid workers on Amazon Mechanical Turk.
Data format	Raw Analyzed
Parameters for data collection	The maximum false alarm rate (short-term): 30% The maximum false alarm rate (long-term): 40% The minimum recognition rate of vigilance fillers (short-term): 70% The minimum recognition rate (long-term): 15% The false alarm rate must be lower than the recognition rate (long-term).
Description of data collection	1,500 short videos selected from the Vimeo Creative Commons (V3C1) dataset and used in the TRECVID 2019 Video-to-Text task were divided into three non-overlapping subsets: training, development, and testing. Multiple manual memorability annotations for each video were collected via a video memorability game, which displays a series of short videos and requires users to press the spacebar when they recall a video previously seen by them. The game consists of two parts: in the first part where videos are repeated within a few minutes, the user interaction with a repeated video was collected to calculate short-term memorability scores. The second part took place between 24 and 72 hours after initial viewing of videos, and this time the participants' responses to previously seen videos from the first part were collected to acquire long-term memorability scores. After analysing the collected annotations, the short-term and the long-term memorability scores of each video were calculated as a percentage of correctly recalled videos, respectively. Each video memorability annotation is accompanied by the video timestamp offsets at which it was recalled by users, response times of the users, the key pressed when watching each video, and textual captions describing each video from the TRECVID benchmark. The Media Memorability 2020 dataset is included here with memorability annotations on 590 videos as part of the training set and 410 additional videos as part of the development set. In this dataset we provide memorability annotations for the development and training set videos but not the test set as this is used in future MediaEval memorability benchmark tasks.
Data source location	Primary data sources: TRECVID 2019 Video-to-Text dataset [1], available from: https://www-nplir.nist.gov/projects/trecvid/trecvid.data.html Institution: National Institute of Standards and Technology (NIST) City: Gaithersburg Country: USA
Data accessibility	Repository name: Figshare Direct URL to data: https://doi.org/10.6084/m9.figshare.15105867.v2 Source code used to process this is adapted from [2] and is available at https://github.com/InterDigitalInc/VideoMemAnnotationProtocol/
Related research article	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? MediaEval Workshop, online 14-15 December, 2020 [3] DOI:

Value of the Data

- Media platforms such as social networks, media advertisement, information retrieval and recommendation systems deal with exponentially growing data day after day. Enhancing the relevance of multimedia data – including video – in our everyday lives requires new ways to analyse, index and organise such data. In particular it requires us to be able to

discover, find and retrieve digital content like video clips and that means automatically analysing video so that it can be found. Much work in the computer vision community has concentrated on analysing video in terms of its content, identifying objects in the video or activities taking place in the video but video has other characteristics such as aesthetics or interestingness or memorability. Video memorability refers to how easy it is for a person to remember seeing a video and video memorability can be regarded as useful for a system to make a choice between competing videos on which video to present to a user when that user is searching for video clips. Video memorability will also be useful in areas like online advertising or video production where the memorability of a video clip will be important. The data provided here can be used to train a machine learning system to automatically calculate the likely memorability of a short form video clip.

- Researchers will find this data interesting if they work in the areas of human perception and scene understanding, such as image and video interestingness, memorability, attractiveness, aesthetics prediction, event detection, multimedia affect and perceptual analysis, multimedia content analysis, or machine learning.
- The dataset provides links to publicly available short form video clips, each of 6 seconds duration, features which describe those videos and annotations as to the memorability of those videos. This is all the data needed to train and evaluate the accuracy of machine learning classifier to predict video memorability.
- A huge amount of video material is now available to us at our fingertips, including from video sharing platforms like YouTube and Vimeo, video streaming platforms like Netflix and Amazon Prime, videos shared on social media platforms and even the video clips we ourselves generate on our smartphones. Unlike searching text documents on the WWW, searching through all this video content in order to find a clip you may have seen previously or a clip you think might exist but you are not sure and you would like to find it, such information search is not currently supported. Eventually technology companies will catch up with the growth in the amount of available video content and as they do, the intrinsic memorability of a video clip will be a characteristic of a video clip that will be important in deciding whether to retrieve it for a user. This means that video search will give us search results which will be with the better, more memorable videos more highly ranked.
- The specific use cases of creating video commercials or creating educational content re-

quires videos which people will remember. Because the impact of different forms of visual multimedia content – images or videos – on human memory is unequal, the capability of predicting or computing the likely memorability of a clip of video content is obviously of high importance for professionals in the fields of advertising and education.

- Beyond advertising and educational applications, other areas such as filmmaking will find use for methods which calculate the memorability of video clips. We may see film and documentary makers creating videos in such a way that the key moments in a movie or documentary will be created in ways so as to maximise their likely memorability by the viewer and that opens up new ways for creating video material.

1. Data Description

The Media Memorability 2020 dataset contains a subset of short videos selected from the TRECVID 2019 Video-to-Text dataset [1] and a sample of frames from some of these is shown in Figure 1. The dataset contains links to, as well as features describing and annotations on, 590 videos as part of the training set and 410 videos as part of development set. It also contains links to, and features describing, 500 videos used as test videos for the MediaEval Video Memorability benchmark in 2020.



Fig. 1. A sample of frames from some of the videos in the *TRECVID 2019 Video-to-Text dataset*.

Each video in the training and development sets is distributed with both short-term and long-term memorability ground truth scores and several automatically calculated features. The collected annotations for each video are also published along with the overall memorability scores. We collected a minimum of 14 and a mean of 22 annotations in the short-term memorability step, and a minimum of 3 and a mean of 7 annotations in the long-term memorability step for

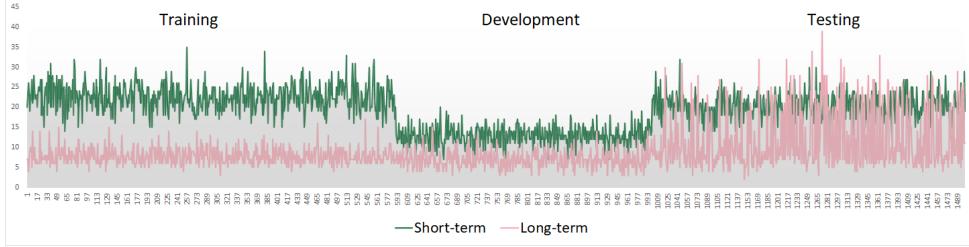


Fig. 2. The number of annotations in the training, development, and test sets.

the training set. The development set has similar annotation numbers in the long-term step with a minimum of 3 and a mean of 7 annotations, however, the number of annotations for the development set in the short-term step is lower than in the training set with a minimum of 6 and a mean of 12. Figure 2 shows the distribution of the number of annotations for short-term and long-term memorability. We will continue improving and update the development set with more annotations in the near future.

Five files are released for each of the training and development sets as presented in Table 2.

Training Set	Development Set
video_urls.csv	dev_video_urls.csv
short_term_annotations.csv	dev_short_term_annotations.csv
long_term_annotations.csv	dev_long_term_annotations.csv
scores.csv	dev_scores.csv
text_descriptions.csv	dev_text_descriptions.csv

Table 2. Text files in the training and development sets of the MediaEval2020 Predicting Media Memorability dataset

The dataset also contains the same features and structure in the training and development sets. Table 3 presents the text files with the features and their descriptions.

Table 3: The text files in the training and development sets of the MediaEval2020 Predicting Media Memorability dataset

Text File	Feature Name	Description
Video URLs	video.id	the unique video id
	video.url	the video url
Short-term annotations	video.id	the unique video id
	video.url	the video url
	user.id	the id number of the user performing the annotation
	rt	response time in milliseconds for the second occurrence of the video (-1 for no response given by the user)
	key_press	the key code pressed by the user for the second occurrence of the video (32 is for spacebar, -1 for no response)
	video.position.first	the position of video seen first time in the current stream (1-180)
	video.position.second	the position of video seen second time in the current stream (1-180)
Long-term annotations	correct	1 is for the correct response 0 is for incorrect response
	video.id	the unique video id
	video.url	the video url
	user.id	the id number of the user performing the annotation
	rt	response time in milliseconds for the occurrence of the video (-1 for no response given by the user)
	key_press	the key code pressed by the user for the second occurrence of the video (32 is for spacebar, -1 for no response)
	video.position	the position of target video seen in the current stream (1-180)
Text Descriptions	correct	1 is for the correct response 0 is for incorrect response
	video.id	the unique video id
	video.url	the video url
Scores	description	text description for the video
	video.id	the unique video id
	video.url	the video url
	ann_1	the number of annotations for short-term memorability
	ann_2	the number of annotations for long-term memorability
	part_1_scores	short-term memorability score
	part_2_scores	long-term memorability score

Additional pre-computed features are provided in individual folders per feature type and in individual csv files per sample, which are available in the data repository. There are seven folders containing the seven features for each of the training, development and test sets as follows:

- AlexNetFC7 (image-level feature) [4]
- HOG (image-level feature) [5]
- HSVHist (image-level feature)
- RGBHist (image-level feature)
- LBP (image-level feature) [6]
- VGGFC7 (image-level feature) [7]
- C3D (video-level feature) [8]

For image-level features we extract features from 3 frames for each video, each one in an individual file, where the filenames are composed as follows: <video_id>-<frame_no>.csv. The 3

frames per video represent the first, the middle and the last frames in the video clip. For example, for video_id 8 we extract the following AlexNet feature-files:

- AlexNetFC7/00008-000.csv : AlexNetFC7 feature for video.id = 8, frame_no = 0 (first frame)
- AlexNetFC7/00008-098.csv : AlexNetFC7 feature for video.id = 8, frame_no = 98 (middle frame)
- AlexNetFC7/00008-195.csv : AlexNetFC7 feature for video.id = 8, frame_no = 195 (last frame)

For video-level features we extract 1 feature for each video, where the filenames are composed as follows: <video_id>.mp4.csv. Using the same video_id 8 as an example, we extract the following C3D feature-file:

- C3D/00008.mp4.csv : C3D features for video.id = 8

Figure 3 shows the minimum and maximum reaction times for the annotations for short-term memorability for each of the 590 videos in the training set while Figure 4 shows the same for long-term memorability. The figures reads from left to right, with each column being the vertical continuation of the preceding column. Reaction times are sorted greatest to shortest difference between minimum and maximum reaction time, x-axis is the reaction time in milliseconds, numbers on the y-axes refer to video_id. The figure illustrates a large range of min-to-max reaction times, those appearing later in the graph (rightmost column, towards the bottom) appear to be universally memorable to all annotators while those at the other end of the graph are memorable to some as soon as video playback commences, and less memorable to others. The positioning of the blue dots in the graph indicates that all videos have at least some annotators who remember the video early during the playback, in many instances almost as soon as video playback commences. The differences between short- and long-term memorability annotations indicate long-term recall happens sooner, i.e., earlier during video playback.

2. Experimental Design, Materials and Methods

Each video has two associated scores of memorability that refer to its probability to be remembered after two different durations of memory retention. 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 using a video memorability game protocol proposed by Cohendet et al. [2]. In a first step (short-term memorization), participants watched 180 videos, among which 40 target videos are repeated after a few minutes to collect short-term memorability



Fig. 3. Comparison of the short-term minimum and maximum reaction times for 510 videos.

labels. The task is basically to press the space bar once a participant recognises a previously seen video, which enables to determine videos recognised and not recognised by them. 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 between 24 and 72 hours, the same participants attend the second step for collecting long-term memorability labels. During this second step, they each watch 120 videos comprised of 40 target videos chosen randomly from among non-vigilance fillers from the first step and 80 fillers selected randomly from new videos which 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 recognitions for each video, by the participants.

The experimental protocol was written in PhP and JavaScript (a modified version of the JavaScript library in [9] was used) and interacts with a MySQL database. The interaction with Amazon mechanical turk was performed through JavaScript code. The optimisation problem for generating positions was written in Matlab.

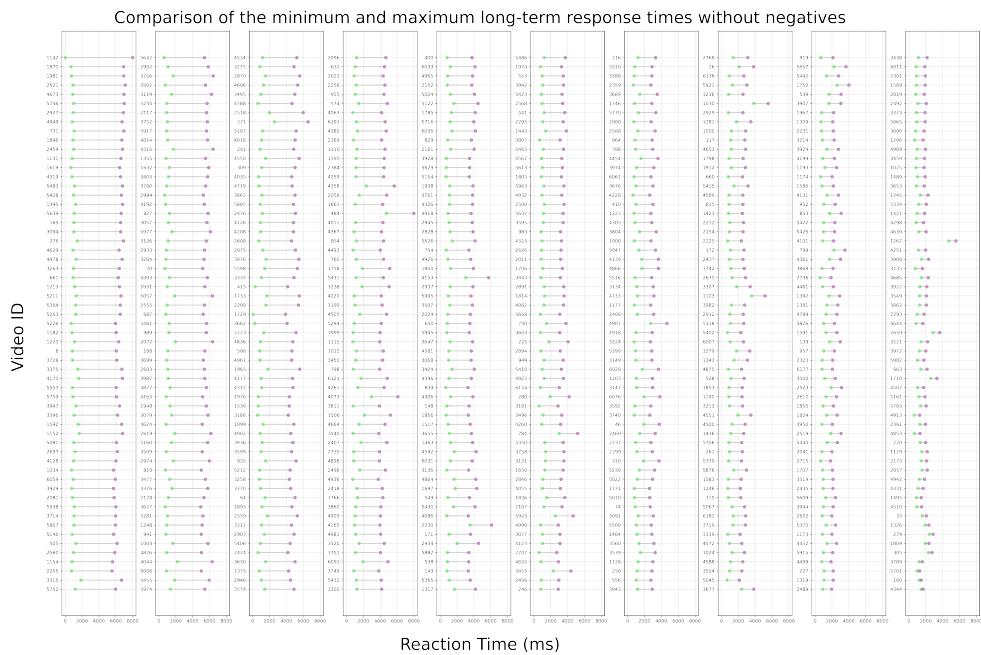


Fig. 4. Comparison of the long-term minimum and maximum reaction times for 510 videos.

A participant could participate only once in the study. The order of videos was randomly assigned, using an algorithm that randomly selects from among the last 1,000 least annotated videos, and which generates random positions from 45 to 100 videos (i.e., 4 to 9 minutes).

Several vigilance tests were settled up upon the results on an in-lab test and only participants that met the controls were retained for the analysis:

1. 20 vigilance fillers were added in the short-term step and we expected a recognition rate of those fillers of 70%.
 2. a minimal recognition rate of 15% in the long-term step.
 3. a maximal false alarm rate of 30% for short-term and 40% for long-term.
 4. a false alarm rate lower than the recognition rate for long-term.

Two versions of the memorability game using three language options: English, Spanish and Turkish were published for different audiences and in different contexts. One was published on Amazon Mechanical Turk (AMT) and another was issued for general use among an audience essentially made up of students. A total of 1,275 different users participated in the short-term memorability step while 602 participated in the long-term memorability step. Only about 48%

of the participants who completed the short-term step came back to participate in the long-term step.

Ethics Statement

Institutional ethical approval for eliciting human participation for the memorability game was granted by the University of Essex, with protocol number ETH1920-1049. Anonymity of participants was maintained, and informed consent was obtained, and only data from consenting participants was used in constructing the memorability dataset. With respect to the use of Amazon Mechanical Turk (AMT), the design of the game and nature of information captured from participants ensured security and confidentiality concerns were minimal.

CRediT Author Statement

Rukiye Savran Kiziltepe: Conceptualization, Methodology, Software Programming, software development, Validation, Formal analysis, Writing - Original Draft; **Lorin Sweeney**: Formal analysis, Writing - Original Draft, Visualization; **Mihai Gabriel Constantin**: Conceptualization, Methodology, Validation; **Faiyaz Doctor**: Methodology, Writing - Original Draft; **Alba García Seco de Herrera**: Conceptualization, Methodology, Validation, Writing - Original Draft; **Claire-Hélène Demarty**: Conceptualization, Methodology, Software Programming, software development, Validation, Writing - Original Draft, **Graham Healy**: Methodology, Validation; **Bogdan Ionescu**: Conceptualization, Methodology, Validation; **Alan F. Smeaton**: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgements

The work of Mihai Gabriel Constantin and Bogdan Ionescu was supported by the project AI4Media, A European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911. The work of Graham Healy, Alan Smeaton and Lorin Sweeney is partly

supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund. The work of Rukiye Savran Kızıltepe is partially funded by the Turkish Ministry of National Education. Funding for the annotation of videos was provided through an award from NIST No. 60NANB19D155. We thank Cohendet et al. [2] for sharing their source code.

References

- [1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, A. Smeaton, Y. Graham, W. Kraaij, G. Quenot, TRECVID 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval, in: 2019 TREC Video Retrieval Evaluation Notebook Papers and Slides, 2020.
- [2] 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.
- [3] A. García Seco de Herrera, R. Savran Kızıltepe, 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: MediaEval 2020, volume 2882, CEUR-WS.org, 2020.
- [4] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
- [5] 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.
- [6] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on pattern analysis and machine intelligence 24 (2002) 971–987.
- [7] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
- [8] 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.
- [9] J. R. De Leeuw, jspsych: A javascript library for creating behavioral experiments in a web browser, Behavior research methods 47 (2015) 1–12.