Annotating, understanding, and predicting long-term video memorability*

Extended Abstract†

Romain Cohendet
Technicolor
Rennes, France
romain.cohendet@technicolor.com

Karthik Yadati
University of Delft
Delft
N.K.Yadati@tudelft.nl

Ngoc Khanh Duong
Technicolor
Rennes, France
Quang-Khanh-Ngoc.Duong@technicolor.com

Claire-Helene Demarty
Technicolor
Rennes, France
Claire-Helene.Demarty@technicolor.com

ABSTRACT

Following the interest of these last years for image memorability prediction, video memorability has recently attracted attention of researchers. The growing number of video shared everyday forces us to find new way to deal with them to make the most relevant their occurrence in our everyday lives.

Recently, two studies have tried to predict video memorability ; however, the immaturity of this work is obvious. In particular, none of these studies on images nor videos tried to measure 'very' long-term memory: they just measured the memory of items some minutes after the memory encoding. In this paper, we propose a dataset of 700 videos annotated with memorability scores corresponding to real long-term memory. We discuss in details the method and the proponents. We then propose a model to predict memorability of videos and test it on the available dataset of images and videos. The material – videos plus theirs memorability scores – is made accessible for researchers.

KEYWORDS

Video memorability, Long-term memory, Measurement protocol, Memorability scores, Deep learning, Multimedia information retrieval

ACM Reference Format:
Romain Cohendet, Karthik Yadati, Ngoc Khanh Duong, and Claire-Helene Demarty. 2018. Annotating, understanding, and predicting long-term video memorability: Extended Abstract. In *Proceedings of ACM Yokohama conference (ICMR 2018)*. ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

*Produces the permission block, and copyright information
†The full version of the author's guide is available as `acmart.pdf` document

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Opening. To deal with the constant increase in shared videos, we must imagine new ways to organize – in particular, to retrieve – videos with the aim of improving the relevance of their occurrences in our every day lives. Contrary to other metrics used to quantify image of video importance, such as aesthetics or interestingness, whose corresponding ground truths corresponds to subjective judgments, memorability can be objectively measured by a memory test. As such, memorability can be regarded as a particularly relevant element to help us picking a video among several ones.

The field of video memorability is very immature. The image memorability prediction challenge has attracted increasing attention from the seminal work of Isola *et al.* released in 2001 [4]. Recently, models achieved very good results with the introduction of deep learning to address the challenge [1, 5]As a result of this success, the search field has been extended to videos. Video memorability prediction is however a very new field, and there are only two studies, to our knowledge, that address this issue [2, 6].

At least two important problems could explained this scarcity of studies. Firstly, protocols used in the existing studies are not smart enough, but are easily critisizeable. Secondly, authors didn't make their data accessible for community, so no database exist.

1.1 Problem 1: The problem of the weakness of the studies

Critics of Han et al. The previous attempt at creating memorability video dataset first approach to model video memorability had been propose by [2]. This first attempt is a "strange" adaptation of [4] protocols.

- Firstly, the authors call their task a "memory game" as Isola *et al.*. If it were justified for Isola et al., 2011, la tÃćche dure pour 1 unique participant environ 12 heures. The authors do not provide this time, but it cans be deduced from the by the following calculation:

Thenks to this maneer, participants obtained a dataset of videos with only 20 participants (we donť know if they are student or payed).

rÃĺpatit sur 10 jours difffÃĺrens (5 pour les phases dapprentissage + 5 pour les phase de mesure de la mÃĺmoire).

uen autre aberration e

In [6], - des questions pas forcÃľment aussi faciles/diffilec entre les vidÃľos, qui peuvent expliquer leur

Enfin, dans les deux Ãľtudes prÃľcÃľdentes, la congruence inter-humaine, minimum nÃľcessaire pour s'assurer de la cohÃľrence des donnÃľes et avoir une base de comparaison pour les modÃĺles, n'a pas ÃľtÃľ fournie.

on comprend les auteur d'avoir agit ainsi pour consituter une base de grand ampleur (Khosla et al.) It is long-term memory a proprement parlÃľ because. but memories evolove in l-g mem => Ebbinghaus curve So memorabilities of souvenirs can change after a certain delays

Problem 2: No existing database. => Dataset non disponible pour les deux Ãľtudes.

The fact that there exist no dataset to modeling video memorability is the most serious obstacle which prevent video memorability prediction to take off. This should be the very first goal of pioneers of this new research field: to provide data to community. For images, the pioneers created and made downloadable such a database [4], which enable to made image memorability research flourishing. The second released database, far larger, launched new possibility for deep learning and enable to obtained the aforementioned really good results [5]. Nous pouvons parier qu'il en sera de mÃłme pour le cham de la mÃľmorabilitÃľ des vidÃľos.

The establishment of a dataset, in particular of the first dataset of this kind, may not be self-evident.

Elle nÃľcessite de dÃľfinir les points cardinaux qui vont guider âĂŞ et contraindre âĂŞ les chercheurs qui utiliseront ces premiÃĺres donnÃľes disponibles.

Pour preuve, l'influence considÃľrable qu'Isola et al. (puis Khosla et al., de la mÃłme Ãľquipe, avec leur base de donnÃľes beaucoup plus consÃľquente rendue disponible en 2015) ont eu sur les Ãľtudes rÃľalisÃľes dans leur sillage : jusqu'Ãă aujourd'hui, la quasi-totalitÃľ des Ãľtudes ayant portÃľ sur la mÃľmorabilitÃľ des images ont adoptÃľ âĂŞ avec les bases de donnÃľes de ces auteurs âĂŞ leurs dÃľfinitions de la mÃľmorabilitÃľ.

Or, la maniÃĺre dont les auteurs ont mesurÃľ la mÃľmorabilitÃľ des images, qui n'est pratiquement jamais rediscutÃľe, n'est qu'une des maniÃĺres de procÃľder.

Aussi, si les modÃĺles de prÃľdiction de la mÃľmorabilitÃľ des images ont obtenu d'excellents rÃľsultats (trÃĺs proches de ceux qu'on obtiendrait en infÃľrant des rÃľsultats d'un groupe d'observateur les rÃľsultats d'un autre groupe d'observateurs), qu'est-ce que ces modÃĺles prÃľdisent si bien, en fait ?

Videos are more complex than images. Contrary to images, videos do not constitute clearly defined units, but have supplementary dimensions – sound and movement – that makes difficult the definition of what a video is.

Aussi cette Ãľtude veut participer Ãă rÃľflÃľchir aux protocoles qui pourraient Ãłtre employÃľe pour contituer Ãă l'avenir Ãă un protocole de plus grande envergure.

Goals of our study. These two reasons are linked. They give us our priority: build a dataset and define a protocol to do that. This dataset – and a fortiori the protocol – should avoid the drawbacks of previous work in video and image memorability.

=> To propose a new protocol to collect data/for annotation => To propose a dataset => The dataset will have "lasting" long-term memorability annotations We want our work would benefit to video but also to image memorability.

Secondary goals. il semble des derniÃĺre Ãľtudes qu'il serait intÃľressant de prendre en temps le dans le calcul de la mÃľmorabilitÃľ [LaMem + papier sur la VM oÃź ils prennent en compte le temps dans leurs scores]

The semantization that can affect long-term memorability of videos

2 CONSTRUCTION OF THE DATASET

In this section, we describe the new protocol we propose to measure video memorability in order to collect the long-lasting memorability of the videos. The protocol is also lighter than ordinarily since there is no need of learning step to encode the material before the recognition task, as we measure the participants' memory established prior the experiment.

The memory task involve three materials: videos sequences, questionnaire and participants.

2.1 Memory task

To measure the "real" long-term memory, we designed the recognition task presented in the figure ...

The 10-sec video sequences are displayed one after another, separated by an inter-stimuli interval (ISI).

Les vidÃľos cibles sont mÃłlÃľes aux vidÃľos de remplissage, et le participant doit dire, pour chaque vidÃľo qui lui est prÃľsentÃľe, sâĂŹil lâĂŹa dÃľjÃă vue ou non. To answer, he must press the spacebar during the displaying of the video sequences (than answer when they arrived during the 1-s inter-stimuli interval). movies they have seen (the targets) and never seen (the fillers). Even if the participant answers, the video continue to run until its end (to avoid the temptation of ansewering to faster finish the experiment).

"Respond ONLY if you remember the sequence AND NOT THE movie from which it comes".

2.2 The video sequences

A list of 100 movies (available with the downloadable data) was first established, for which we mixed their popularity and their genres. The movies quality was HD 720-1080, except for ??? dvdrip movies. orginal-english version (without subtitles) (occidental movies, except for Slumdog millionaire)

To constitute our dataset, we manually selected 7 video sequences of 10 seconds from each movies (for a total of 700 videos). The sequences had to meet several criteria to be selected. First, not to be part of the official trailers, to be sure that, if a participant report not having seen the movie, he won't have seen the sequence presented for the recognition in a trailer. Second, we wanted our sequences to constitute "logical" units, that is to say to "protect the sens" of the videos. Indeed, semantics is linked to memorability of images [3], from which we can imagine that it is also the case for videos. But, contrary to images that are clearly cut units, sequences memorability depends on the cut. To maintain a relative semantic consistency intra-sequence, the cutters were provided with the instruction to avoid to cut in the middle of a sentence, of to agglomerate very different plans together. The idea was to increase the coherence of the sequences to diminish the difficulty of he computational models to learn without impairing ther capacity to generalize.

During the manual selection of the sequence, we chose 127 sequences (among the 700) we called "generic" by contrast to "typical" ones. Generic sequences are parts of the movie which do not contain any elements that would enable someone to easily guess that the sequences belong to the particular movie. The list of undesirable includes but is not limited to: scenes with recognizable famous actors (e.g. in Kill Bill, avoid the scenes with Uma Thurman, Lucy Liu, David CarradineâĂę); scenes famous for their music, gesture... (e.g. in Kill Bill, avoid the scenes with music that enable to easily guess that the sequence comes from Kill Bill); scenes with typical elements (e.g. a spaceship or an alien from which you can easily guess that you are watching a Star Wars movie). The cutters were provided with the instructions to ask themselves the following question when choosing a generic sequence in a movie: âĂIJCould I easily guess this sequence belongs to the movie I extracted it from?âĂİ – if the answer was yes, so donâĂŹt select this sequence. In some movies (e.g. Kill Bill 1, Star wars, 2001 A Space Odyssey, etc.), just a few 10-sec sequences of no sequence at all were found to meet these criteria (e.g. The lord of the ring); for these movies, we don't have generic sequences. By contrast, non-generic sequences were namely two 10-sec parts of the movie that you think are very representative of the movie (you can choose any part of the movie without restriction). These two types of sequences should be considered in link with the instruction of the task written above (even if you know that a sequence comes from a movie, just answer if you recognize THE sequence): it is just a supplementary control (qui jour un rÃťle symÃľtrique au taux de FA pour la prise de risque).

In Isola et al., 2014, the authors showed that memorability directly correlated with color. To explain thi result, they advanced the hypothesis that the difference in memorability between indoor and outdoor scenes was due to the fact that the later tend to be less memorable that the older, and that the color tend to linked with a differences in colors, could explain this result. To validate this hypothesis, we manually annotated our 700 sequences as consisting of an indoor or outdoor scene.

2.3 Questionnaire

The targets and fillers that constitute the items of the recognition task were different for each participant. Indeed, the video sequences viewed by a participant were selected based on response to a questionnaire presented in the figure ??? Questionnaire with the 100 movies.

Le nom du film, suivi de lâĂŹaffiche, du nom du rÃľalisateur et des acteurs principaux âĂŞ qui permettent de sâĂŹassurer que le film que le participant a en tÃłte est bien celui qui lui est prÃľsentÃľ âĂŞ est prÃľsentÃľ au participant.

For each movie, the participant had first to answer the following question: "Do you remember watching this movie?" – with the fact that was madde clear that the film must been seen EN ENTIER.

If he answered "yes" Please let us know how confident you are of having seen this movie (I am not/slightly/50%/considerably/100% confident) Please let us know when did you last see this movie (<1 month; <1 year; <5 years; < 10 years; > 10 years) Please let us know how many times did you watch this movie (once; 2-4 times; 5-9 times; 10-19 times; >20 times)

If he answered "no" Please let us know how confident you are of NOT having seen this movie (I am not/slightly/50%/considerably/100% confident)

The goal of the confidence question was to be sure that we select video sequences from movie , to increase our certitude that target/filler were really targets/fillers.

The other two questions were used for the analysis of the results to assess factors a priori susceptible to influence video memorability. The question about the number of times the movie had been seen for us to evaluate the effect of repetition on memorization, and the question about the last time tha movie had been seen to evaluate the effect of time passage on long-term memorability, that continue to be affected in longterm memory

2.4 Participants, facilities and apparatus

??? (105) participants ($22 - 58$ years of age; $mean = 37.1$; $SD = 10.4$; 26% of them female) employees of [removed for double-blind review] participated in he experiment on a volunteer basis. All participants have either normal or corrected-to-normal visual acuity.

The images were displayed on a 40 inch monitor (TV SONY Bravia ???) with a display resolution of $1{,}920 \times 1{,}080$. The participants were comfortably seated at a distance of 150 centimeters from the screen (three times the screen height). The $1{,}024 \times 768$ images were centered on a black background; at a viewing distance of 150 cm, the stimuli subtended 18.85 degrees of vertical visual angle.

2.5 Procedure

Participants first answered the questionnaire. During the whole time, the experimenter was near the participant in case he doubt about one movie.

Based on the answers to the questionnaires, the algorithm selected the movies associated with the maximum certitude level (i.e. 5) to constitute targets and fillers sets. // At the end, the algorithm selected 80 targets and 40 fillers for the participants, and place them in a random order for presentation // The criteria of selection was the number after the response to the questionnaire the nb of annotations of the sequences, in order to harmonize the nb of annotations per video. The sequences used as targets were selected among the sequences corresponded to the movies the participants were sure (i.e. rate of 9 in the scale) they have already seen. The algorithm selected the less annotated sequences to try maintaining a number of annotation equivalent between sequences. This was to maximize the probability that targets really are targets, and fillers really are fillers. according to they number of annotations.

Then the participants was provided with the instructions about how to complete the memory task. In particular, the experimenter insisted on the fact that we did not want the participant guess that the sequence was il the film they have already seen, but that we want they recognize.

// During about 24 minutes, you will see a series of 10-seconds video scenes. Press the SPACEBAR during the 10-seconds video scene if you RECOGNIZE/REMEMBER seeing it before. Be careful!! You only press the spacebar when you RECOGNIZE/REMEMBER the video scene, not the movie from which the video scene had been extracted. We do not want you guess that the video scene was in a movie you have seen: just press the spacebar if you remember that you saw a particular video scene.

Then the experimental phase was then launched, corresponding to the memory task.

The total time of the experiment was about 50 minutes.

3 STUDY OF THE MEMORABILITY SCORES

In this section, we describe in detail the ground truth data collected thanks to the protocol described in the previous section. We conclude by calculating the memorability scores finally used to feed our prediction models.

3.1

3.1.1 Sequences' memorabilities. A sequence's memorability is defined as the percentage of correct detections by participants. On average, each video was viewed as target by 9.3 participants. Average sequence memorability was 48.25% (SD of 28.21%). On average, each video was viewed as filler by 9.2 participants. Average false alarm rate was 5.95% (SD of 14.47%).

3.2 Number of annotations per sequence

The histogram presented in the figure 1 corresponds to the number of sequences for each possible number of annotations.

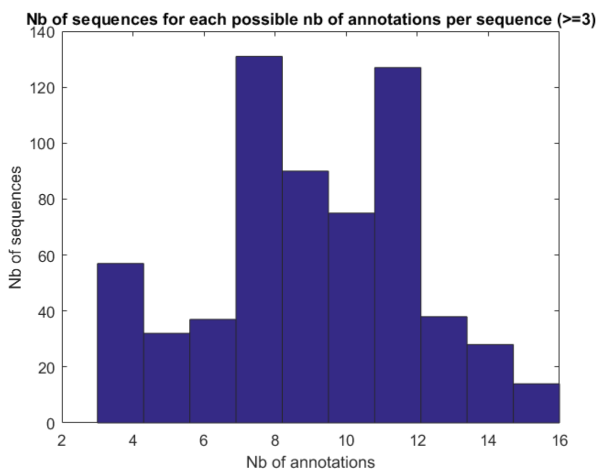
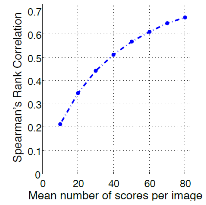


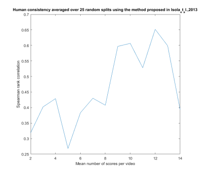
Figure 1: Nb of sequences for each possible nb of annotations per sequence (>=3).

3.3 Consistency analysis

We implemented the method proposed in [3] to measure the human consistency. It answers the question: "Are the videos that are more memorable (or forgettable) for a group of observers also more likely to be remembered (or forgotten) by a different group of observers? The figure 2 presents the curve of the human consistency for each video sequence. It answers the question: "How many annotations for each do we need for the human consistency stop varying?'



(a) (a) Human consistency for images in [5]



(b) (b) Human consistency for videos for us

Figure 2: Human consistency averaged over 25 random splits (right) obtained using the method proposed by [3].

This curve show us that from 14 annotations (using our method) video memorability data has enough consistency

Individual and contextual differences, besides mandatory random variability, explain the $1 - .65$ part of the memorability that is not universally derivable from the intrinsic informations of the videos.

If we compare the curve obtained by [5] (2(a))and the curve we obtained (b), we can note than we attain human consistency far more quickly than Khosla *et al.*. At least three arguments go in the way of this results. First, we work with videos and they work with images ; maybe video memorability is more universal than image memorability because of the higher length of this kind of stimulus (but We can also note that the maximum human consistency is close for images and videos!!). Second, we obtain a measure of real long term memorability, and not a memorability measured some minutes after encoding step ; maybe this measure more representative of what is really memorable. Finally, we can advance the fact that in-lab experiment enable to obtain better memorability measure than crowdsourcing one, resulting to attain maximum human consistency earlier.

3.4 Generic vs. Typical sequences

3.5 Quality of the movies

dvdrip vs. HD 720-1080 => Difference of memorability? => Interesting

3.6 Response time

The question here is: "Could we exploit response time to correct memorability score?"

We hypothesized that the most memorable the sequences, the faster the participant will answer.

We observed a Person's correlation of $-0.35 (p < .0001)$ between the response time on the target and their memorability scores. This means that the videos with the higher memorability score tended to be answered fastly when correct detection by the participants, suggesting that a sequence most memorable is also a sequence for which we rapidly detect that it is memorable.

The global mean response time was $4.87 sec$ on targets and $5.96 sec$ on fillers. A Student t-test for different sample size show a significant difference ($t(2836) = -5.34, p < .0001$). This means that the participants globally answered more rapidly for targets (i.e. correct detections) than for fillers (i.e. false alarm), probably because of their hesitation for fillers.

3.7 Evolution of the memorability along time

3.8 New manner to compute memorability scores: take into account time and FA

3.8.1 Participants' performance. The average memory performance was the following: the average percentage of correct detection was 48.2% (SD of 14.1%) and the average percentage of false alarms was 4.78% (SD of 5.63%).

3.9 Logistic regression vs. SVM to personalize prediction model

3.10 Film genre and IMDB ratings

3.11 Context

3.12 Features linked to memorability

3.13 Indoor *.vs* outdoor scenes

3.14 Memorability score calculation

According to what was presented before in tis section, this is how we finally decided to compute our memorability scores...

4 MEMORABILITY PREDICTION

5 DISCUSSION

6 CONCLUSIONS

APPENDIX

ACKNOWLEDGMENT

REFERENCES

[1] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. 2016. Deep Learning for Image Memorability Prediction: the Emotional Bias. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 491–495.

[2] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. 2015. Learning computational models of video memorability from fMRI brain imaging. *IEEE transactions on cybernetics* 45, 8 (2015), 1692–1703.

[3] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1469–1482.

[4] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 145–152.

[5] 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.

[6] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. *arXiv preprint arXiv:1707.05357* (2017).