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 on from the study of image memorability prediction, which draw continuous attention over the past six years, the computational understanding of video memorability's search field has recently hatched. The growing number of shared videos foster us to find new ways to make their occurrence the most relevant in our everyday lives. There is no available dataset of videos annotated in terms of memorability; such dataset may probably result in a launching of the field, as it had been the case for images. The first goal of the pioneers of this search should be to succeed to propose a protocol to constitute such a dataset. In this article, we propose a protocol of 700 videos annotated with memorability scores, constitute a dataset, study the quality of the collected data, and computationally modeling memorability to predict it.

We finally propose a deep-learning model base to predict video memorability, comparing several methods and class of features.

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

1 INTRODUCTION

Enhancing the relevance of multimedia occurrences in our everyday life requires to imagine new ways to organize – in particular, to retrieve – contents. Like other metrics of video "importance", such as aesthetics or interestingness, memorability can be regarded as a particularly relevant element to help us picking a video among several ones, with the advantage of a possibility to be measured less subjectively.

The computational understanding of video memorability's search field follows on from the study of image memorability prediction which has attracted increasing attention from the seminal work of Isola *et al.* released in 2001 [6]. Recently, models of image memorability prediction achieved very good results with the introduction of deep learning to address the challenge of image memorability prediction [1, 7, 11]. This success resulted on the extension of the challenge to videos. Video memorability prediction is however a new field, and there are only two studies, to our knowledge, that addressed this issue [4, 10].

At least two important problems could explained this scarcity of studies. Firstly, protocols used in the existing studies are not smart enough, but are questionable. Secondly, authors didn't make their data accessible for community, so no database exist. 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 [6], 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 [7]. Nous pouvons parier qu'il en sera de mÃłme pour le champ de la mÃľmorabilitÃľ des vidÃľos.

We think that the existence of such a database could have the importance for video memorability that the ones of [6] and [7] had on image memorability.This article wants to takes his part to develop such a protocol to collect high quality data, and such a dataset.

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

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

ICMR 2018, June 2018, Yokohama, Japan
© 2018 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

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.

1.1 Previous work

The work on image memorability prediction, initiated by [6], produced great results [1, 7, 11]. The availability of two large datasets [6, 7] has been critical in this achievement. However, possibly constrained by the difficulty in conducting long crowdsourcing studies, the authors measured memory performance a few minutes after memorization to obtain their memorability annotations. As shown in [2], this could be a problem if we conceive that memorability reflects a lasting memory performance. Indeed, it has long been shown that long-term memories continue to change long after their memorization through an ongoing process called consolidation [9].

In particular, the early work of the psychologist Ebbinghaus showed that the drop in long-term memory performance is particularly strong in the few days immediately following the memorization.

, with a memory performance in recall that asymptotically tends its long-term performance [3].

Because several factors are susceptible to influence the consolidation process (such as the emotional content, the semantics, the attention...), which is not a linear decrease for all the memories, it follows that the order of memorability of images is susceptible to change from some minutes to a day after memorization, as it has been found in [2] with emotional image from of the International Affective Picture System [8].

A protocol to collect memorability annotations for videos would benefit from the capacity to measure what we will later refer to as "long-term memorability".

To our knowledge, the first attempt at predicting video memorability had been proposed by [4].

The authors partially adapted the protocol proposed by [6] to measure image memorability for videos. In contrast to the "memory game" proposed by Isola *et al.* to collect memorability data, their protocol is however much heavier. They used a classical recognition task to measure memory for videos, which consists in two steps: a free viewing task, followed two days later by a recall task.

We can infer from the sparse information authors gave that the task duration, for each of the 20 participants, was about 24 hours (by taking an average video duration half-way between 15 and 30 sec), spread over 10 sessions (5 free viewing and 5 recall tasks) of about 2 hours each.

Authors used the same proportion of fillers in the free viewing and recall tasks (4/5 of fillers and 1/5 of targets), to "to guarantee that viewers are unaware of targets". If it was totally justified in [6] for which encoding and recall tasks were interlaced, there is here a way to alleviate the task without impacting its quality; indeed, reducing the number of fillers (never reasked videos) in free viewing task would have very week impact on memorability scores (gÃľnÃľralement on se passe mÃłme de fillers dans les tÃćches d'apprentissage du matÃľriel).

Still, as pointed by [10], the long time span of the experiment makes difficult the generalization of this protocol, in particular to build an extensive dataset.

Regarding the prediction part, we also agree with these authors to say that "the method used fMRI measurements for predicting memorability, which would be difficult to generalize".

Another earlier approach was the one of [10]. The participants passed in crowdsourcing a free viewing task followed by a recall task, for which they had to answer question (such as: "Did you see a man juggling?", "Did you see a car on road?").

The major drawback of the study comes from this use of questions instead of a recognition task. The "memorability" measured may be due not to memorability of the videos but (also) to the differences between the questions rather than a difference of memorability of the videos. This bias could be of several kinds: difficulty, comprehension (non anglophone people in AMT), easility of imaging), indexage plus facile... C'est d'autant plu sproblÃľmatique que ces diffÃľnrece could have drives to a high inter-human consistency not due to memorability attribuÃľ Ãă tort Ãă un phÃľnomÃĺne de mÃľmoire. C'est d'autant plus problÃľmatique que le temps de rÃľponse des partcipants entre dans le calcul de la mÃľmorabilitÃľ, ce qui ne peut q'uaccentuer ces biais, et est porbablement responsable de la high human response consistency they observe.

Le fait que les questions doivent Ãłtre crÃľÃľ Ãă la main, et que les choix des questions sont assez limitatifs (par ex avec la question "Did you see a car on road?", il ne peut y avoir de toute la tÃćche qu'une seule fois une voiture) makes difficult the generalization of this protocol to build an extensive dataset. âĂIJIt was manually ensured that no two textual questions nor any two videos in a sequence were similar in content.âĂİ

Les critiques exposÃľs montre Ãă quel point il est difficile de construire la premiÃĺre base de donnÃľes pour mesurer la mÃľmorabilitÃľ des vidÃľos. Isola et al. SâĂŹen sont trÃĺs bien sortis, en se cantonnant Ãă une mesure purement objective de la mÃľmoire âĂŞ tÃćche de reconnaissance classique âĂŞ, ce qui leur a permis de lancer le champ de recherche de la mÃľmorabilitÃľ des images. Il faudrait en faire de mÃłme pour la mÃľmorabilitÃľ des vidÃľos. Cette Ãľtude sâĂŹinscrit dans cette volontÃľ âĂŞ non pas de crÃľer un nouveau datasets, mais dâĂŹessayer un protocole pour voir sâĂŹil est efficient.

1.2 Goals of our study

To sum up, we want a protocol: - which could be extended to create a large dataset in case it shows its quality

- that measure a long-term memorability: as a longitudinal study would be difficult to generalized to create a large dataset, we decide to measure the memoray of participants created previously to the experiment.

D'autre part, all around paper

Finally, we want – thanks to our protocol – to participate in the VM prediction challenge.

The rest of the article is as follows: part 2 concerns the ground truth collection, part 3 the study of the collected data and the calculation of the memorability scores of the sequences, part 4 predicting memorability, and part 5 the conclusion and intended future work.

2 CONSTRUCTION OF THE DATASET

Our first step was to implement a protocol to measure the long-term memorability of 700 video sequences, extracted from 100 mainstream movies. The main characteristic of the protocol, in contrast with previously proposed methods to collect image and video memorability annotations, is the absence of learning/free viewing task, replaced by a questionnaire designed to collect information about the participants' prior memory. The way of proceeding aimed at measuring memory performance after a significant retention period.

2.1 Dataset

We mixed popularity and genres to establish our list of movies.Then, we manually selected seven sequences of 10 seconds from each movie (for a total of 700 sequences).

We made efforts to maintain a high intra-sequence semantic cohesion; in particular, we did not cut through meaningful vectors nor we agglomerated shots that belonged to different scenes. Indeed, we think the images' characteristic by which they constitute obvious semantic units is an important element in the capacity of the prediction models to generalize. By contrast, the meaning a sequence conveyed was susceptible to change depending on where we decided to cut the movie.

Because the semantics is linked to the memorability of images [5], so probably to videos, therefore their memorabilities, we chose this definition susceptible to be extensive to predict other dataset.

We also gave preference to the sequences we called "neutral", by contrast to the "typical" ones. According to our definition, a neutral sequence is a part of a movie which contains no element that would enable someone to easily guess it belongs to a particular movie.

The list of undesirable includes but is not limited to: recognizable famous actors, typical music, style, or any element...

The question we asked ourselves to select the neutral sequence was the following: "Could I guess this sequence belongs to the movie I extracted it from if I didn't know it?"

In most movies, just a few or no 10-sec sequences of this sort exist. For example, in Kill Bill 1, we can

That explains why, at the end, we obtained only 127 neutral sequences for 573 typical ones, that are defined as being non neutral sequences that meet the first criterion.

These two types of sequences – neutral *vs.* typical – should be considered in conjunction with the instruction of the task:

neutral sequences are control.

of the task written above (even if you know that a sequence comes from a movie, just answer if you recognize THE sequence):

They only are used as controls is just a supplementary control (qui joue un rÃťle symÃĺtrique au taux de FA pour la prise de risque).

2.2 Survey design

After they had provided basic demographics, participants filled out a questionnaire composed of 100 items, indicating whether they had seen the films of our list. Each item included a film poster with its name below, followed by the question: "Do you remember watching (fully) this movie?" The experimenter was present in case the participant had doubt about one movie. If the participant answered "yes", he had to answer three additional questions: "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", he had to answer one more question: "Please let us know how confident you are of NOT having seen this movie". (I am not/slightly/50%/considerably/100% confident). The question relative to confidence aimed to ensure the targets and fillers corresponded to seen and non-seen films respectively. The other questions (i.e. when and how many times the films had been seen) were used during result analysis to evaluate the likely effects of visualization repetition and passage on long-term memorability. The questionnaire required about 20 minutes to complete.

Based on the answers to the questionnaire, an algorithm selected 80 targets and 40 fillers among the films associated with the highest degree of certitude, with a maximum of two sequences from the same film. Another criterion of selection for the target was the number of annotations of the sequences, in order to harmonize the nb of annotations per video. The sequences were presented in a random order for presentation

The recognition task lasted 24 minutes. Participants saw 10-seconds sequences separated by an inter-stimuli interval of 2 second. Their task was to press the spacebar when they recognized a sequence. During the instruction, the experimenter insisted on the fact that the only task of the participant was to answer only when he recognized the particular sequence, but not answer if he recognized not even when he could guess that the sequence came from a film he had seen. This, because for number of sequences (i.e. the typical ones), it was possible to guess that the sequence came from a particular film. Even if the participant answers, the video continue to run until its end.

2.2.1 *Participants, facilities and apparatus.*

104 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. The videos were displayed on a ????? inch monitor (TV SONY Bravia ?????) with a display resolution of 1,920 × 1,080. The participants were seated at a distance of

150 centimeters from the screen (three times the screen height) in a room equipped with subdued lights.

3 STUDY OF THE MEMORABILITY ANNOTATIONS

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.

For the results we retain only the 660 sequences that had been seen at least 4 times as targets, and remove sequence selected from films too few participants had seen.

3.1 Overview

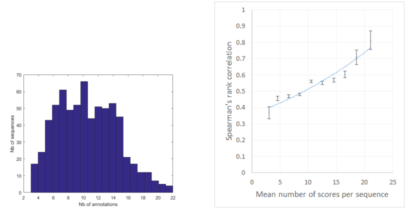
The average memorability of the test was 46.71% (SD of 14.65%). The average false alarm rate was 4.16% (SD of 5.27%).

On average, each sequence was viewed as target by 10.7 participants. On average, each sequence was viewed as filler by 10.5 participants.

3.2 Sequences' memorabilities

3.3 Consistency analysis

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



(a) Number of sequences associated with number of annotations.

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

Figure 1: (a) Number of sequences for each possible nb of annotations per sequence. (b) Human consistency averaged over 25 random splits (right) obtained using the method proposed by [5] for sequences with at least 5 annotations (with standard error).

We implemented the method proposed in [5] 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 1(b) 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?'

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 [7] (1(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 Neutral and typical sequences

3.5 Quality of the movies

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

3.6 Response time

Previous authors have integrated response time in their score of memorability. We pose here the following question: "Is the degree of memorability related to response time?" to answer the question: "Should we exploit response time to correct memorability score?" We hypothesized that the most memorable the sequences, the faster the participant will answer (no instructions to answer faster: just to answer durgin the 10-sec display).

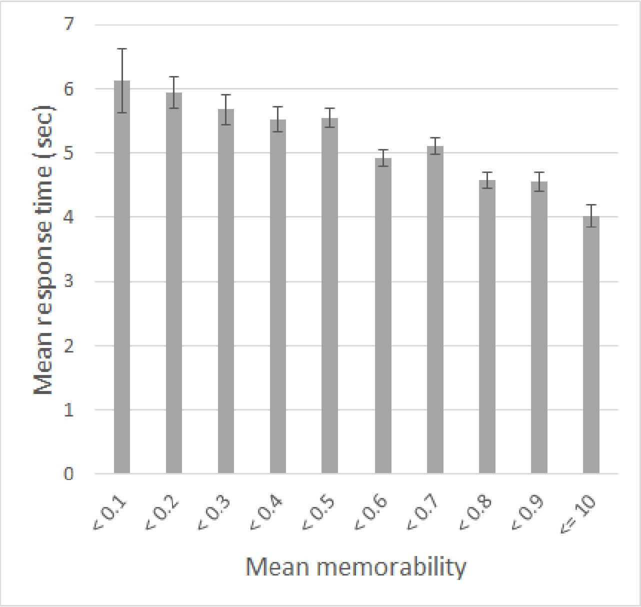


Figure 2: Mean response time for each memorability degree (error bars correspond to SEM).

We observed a Person's correlation of $-0.36 (p < .0001)$ between the response time on the target and their memorability scores.

This is consistent with the figure 2. In other words, the higher the memorability, the lower the response time tends to be, suggesting that a sequence most memorable is also a sequence for which we rapidly detect that it is memorable. This is consistent with what we expected: empirically, we observe during passation that participant tend to answer as soon as they recognized the sequence.

It is interesting to link it to the response time for answer on targets (i.e. correct detections) vs. on fillers (i.e. false alarms). 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). One explanation is that participants hesitated more for fillers they answered on, increasing their response time.

It tells us something on memory, that the memories are not evident but more blurred. This is an important point, taht explained why, in our discussion, we propose to turn to a "totally objective" measure of memory to constitute a dataset for video memorability prediction: if the participants hesitated, so they choice to answer or not was probably impacted by the way they answered the instructions and their own sensibility to the risk.

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 CONCLUSIONS AND FUTURE WORKS

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] Romain Cohendet. 2016. *Prédiction computationnelle de la mémorabilité des images: vers une intégration des informations extrinsèques et émotionnelles.* Ph.D. Dissertation. Nantes.

[3] Hermann Ebbinghaus. 1913. *Memory: A contribution to experimental psychology.* Number 3. University Microfilms.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Peter J Lang. 2005. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical report* (2005).

[9] James L McGaugh. 2000. Memory–a century of consolidation. *Science* 287, 5451 (2000), 248–251.

[10] 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).

[11] Hammad Squalli-Houssaini, Ngoc Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2017. Deep learning for predicting image memorability. (2017).