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


## ABSTRACT

Memorability can be regarded as a useful metric of video importance to help make a choice between competing videos. Research on computational understanding of video memorability is however in its early stages. There is no available dataset for modelling purposes, and the few previous attempts provided protocols to collect video memorability data that would be difficult to generalize. Furthermore, the computational features needed to build a robust memorability predictor remain largely undiscovered. In this article, we propose a new protocol to collect long-term video memorability annotations. We measure the memory performances of 104 participants from weeks to years after memorization to build a dataset of 660 videos for video memorability prediction. This dataset is made available for the research community. We then analyze the collected data in order to better understand video memorability, in particular the effects of response time, duration of memory retention and repetition of visualization on video memorability. We finally investigate the use of various types of audio and visual features and build a computational model for video memorability prediction. We conclude that high level visual semantics help better predict the memorability of videos.



## KEYWORDS

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

**ACM Reference Format:**

. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of ACM Yokohama conference (ICMR 2018).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.475/123_4


## 1 INTRODUCTION

Enhancing the relevance of multimedia occurrences in our everyday life requires new ways to organize – in particular, to retrieve – digital content. Like other metrics of video "importance", such as aesthetics [13] or interestingness [12], memorability can be regarded as useful to help make a choice between competing videos. In addition, memorability has the advantage of being clearly definable and objectively measurable (i.e., using a measure that is not influenced by the observer's personal judgment). Image memorability has initially been defined as the probability for an image to be recognized **a few minutes** after a single view, when presented amidst a stream of images [21]. This definition has been widely accepted within subsequent work (e.g., [9, 24, 25, 27, 29]).




*ICMR 2018, June 2018, Yokohama, Japan*





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


The computational understanding of video memorability (VM) follows on from the study of image memorability (IM) prediction which has attracted increasing attention since the seminal work of Isola *et al.* [21]. With the recent introduction of deep learning to address the challenge of IM prediction, models also achieved very good results [2, 24, 37]. As a result of this success, researchers have recently extended this challenge to videos. However, to the best of our knowledge, only two available studies focus on VM prediction [17, 36].

Several problems could explain this scarcity of studies on VM. Firstly, there is no publicly available dataset to train and test models. This is probably the most serious obstacle to the rapid expansion of research in VM prediction. Following the foot steps of researchers in IM [21, 24], providing data and ground truth for VM should be one of our first objectives. The second point, closely related to the previous one, is the lack of a common definition for VM. The previous attempts to predict VM [17, 36] were based on different measures of memorability. Furthermore, in comparison to images, videos have supplementary dimensions – sound and visual movement – that critically contribute to the semantic and emotional information conveyed and this makes it difficult to come up with a common definition of VM. But the videos used and the way memorability is measured have a critical impact on what we understand by VM. Similarly, the definition of IM [21] inevitably limited researchers. In particular, previous research on IM focused on the measurement of memory performances only a few minutes after memorization. But these short-term memory performances might be poor predictors of longer term memory performances. For example, the memorability of emotional images would decrease in a non-harmonious way between two measures of memory performance (a few minutes and one day after memorization), inducing in a dataset a change in the ranking of memorability values over time [10]. To this end, VM data would be expected to benefit from a protocol that would measure lasting long-term memory performance.

Regarding modelling, the previous attempts at predicting VM [17, 36] highlighted several features which contribute to the prediction of VM, such as semantic, saliency and color features. But the work is far from complete and our capacity to propose efficient computational models will participate to answer the challenge of VM prediction.

While participating in the expansion of research on VM, the contributions of this paper are threefold:

- We propose a new protocol that measures very long-term memory performances (from weeks to years after memorization) to collect ground truth data of good quality and make the corresponding dataset available for the research community (section 3).
- We assess and analyze the collected data and provide useful insights on the understanding of VM (section 4).

- We build a computational model based on machine learning techniques which allows to predict VM score of a given video. For this purpose, we investigate the use of various types of audio and visual features, ranging from low-level characteristics to emotional and (highly) semantic features (section 5).

## 2 PREVIOUS WORK

In this section, we review previous work on image and video memorability prediction, focusing on annotation protocols and modelling aspects.

### 2.1 Measurement of memorability

Long-term memory has been studied for over a century in psychology, from multiple perspectives, since the seminal experimental studies of Ebbinghaus [14] until the more recent neuro-imaging studies [3]. This work provided researchers, interested in computational understanding of IM and VM, with several memory tests (see [34] for an extensive overview), such as the recognition test [17, 21, 24] or the textual questions-based recall survey [36]. It also demonstrated that humans have an extensive long-term visual memory, which enables them to recall a great amount of images [38] and image details [6], as well as of videos [15]. This is one of the observations at the origin of the work on memorability in computer science [21]. Several factors have also been highlighted for their critical influence on long-term memory, such as emotion [23], semantics [33], several demographic factors [11], memory re-evocation [32], or passage of time [30]. These different factors are all important to better understand memorability, and are found valuable computational features for image and video memorability prediction.

Focusing on the work in computer vision, most studies on IM prediction made use of one of the two available large datasets designed specifically to meet this challenge [21, 24]. To address more specific problems, several other datasets have also been publicly released concerning memorability of face photographs [1], visualization pictures [4], emotional images [11] and scene categories [7]. To build the most used datasets presented in [21, 24], the authors, possibly constrained by the difficulty in conducting long crowdsourcing studies, measured memory performance a few minutes after memorization to obtain their memorability annotations. This could be a problem if we conceive that memorability reflects a lasting memory performance [10]. Indeed, it has been shown that long-term memories continue to change long after their memorization through an ongoing process called consolidation, which lasts weeks to years [30]. Because several factors influence the consolidation process (e.g., emotions, sleep, re-evocation), which does not equally affect all memories, the order of memorability ratings measured for content, and especially videos, is susceptible to change over time [10]. A protocol to collect memorability annotations would benefit from the capacity to capture long-lasting memory performances, averaged to obtain what we will later refer to as "long-term memorability".

To our knowledge, the first attempt at measuring VM [17] partially adapted the protocol proposed to measure IM [21] to videos. The resulting protocol is however much heavier than the memory game protocol [21]. They followed a classical recognition process, which consists of two steps: a free viewing task, followed two days later by a recall task. The task duration, for each of the 20 participants, was about 24 hours, spread over 10 sessions (five free viewing tasks and five recall tasks) of about two hours each. The authors used the same proportion of fillers (i.e., non repeated videos) in the free viewing and recall tasks (i.e., 4/5 of fillers and 1/5 of repeated videos named targets) to guarantee that viewers were unaware of targets. If it was mandatory in [21] for which encoding and recall tasks were interlaced, there is a way here to alleviate the task without impacting its quality; indeed, reducing the number of fillers in the free viewing task would have very little impact on the memorability scores (often authors even use only material interrogated later in the learning/free viewing task). Furthermore, the long time span of the experiment makes the generalization of this protocol difficult, in particular if one targets the construction of an extensive dataset. Moreover, authors measured memory after two days, but, as aforementioned, it is known that the consolidation process lasts weeks to years: it would be interesting to collect even longer term memory performances.

In another earlier approach for VM measurement, the participants performed a crowdsourcing experiment that consisted of a free viewing task during which they saw a series of videos, followed by a recall task in which they had to answer textual questions (such as: "Did you see a man juggling?", "Did you see a car on road?") [36]. The major drawback of the study comes from the use of questions instead of a classic visual recognition task. Indeed, the memorability scores computed for the videos may reflect not only the differences in memory performances but also the differences between the questions in terms of difficulty. The authors tested the complexity of the chosen textual questions using the Flesch-Kincaid Grade Level readability metric, which is designed to quantify how difficult to understand an English text is [26]. However, this measure might not be sufficient to represent the questions in all their complexities, such as the ease for a person to draw images from the words, to connect the words to the associated concepts in memory, to associate the words to the corresponding scene, etc. Moreover, there is an unequal distribution of the reading comprehension of English among Amazon Mechanical Turk workers, who are not necessary anglophone people. This problem of a potential unequal complexity of the question becomes even more important in view of the use by authors of the response time to calculate memorability scores, which might also critically depend on the complexity of the questions. One will note that this potential bias could have also affected the measure of inter-human consistency. Furthermore, the questions were handcrafted, and the choices for types of questions and videos were very limited (e.g., the question "Did you see a car on road?" implies that in a whole experimental session only one car on a road should appear to connect the memory performance to one particular video). The authors also manually ensured that no textual questions nor videos in a session were similar in content. This makes it difficult to generalize this protocol to the construction of an extensive dataset.

## 2.2 Memorability modelling

Previous attempts at predicting image and video memorability highlighted quite a few features that correlate with memorability.

The pioneering work of Isola *et al.* focused primarily on building computational models to predict IM from low-level visual features [21]. From their work, it appeared that the degree of an image's memorability can be predicted to a certain extent. Several characteristics have also been found to be relevant for predicting memorability in subsequent work, for example saliency [29], interestingness and aesthetics [19], or emotions [24]. The best results were finally obtained by using fine-tuned deep features, which outperformed all other features in [24], reaching a rank correlation of .64 which is near human consistency (.68) when measured for the ground truth collected in the study. This result was later confirmed by other researchers [2, 37].

Regarding VM prediction, Han *et al.* proposed a method which combines the power of audio-visual and fMRI-derived features [17]. They preliminary built a computational model learned with fMRI features, which supposedly conveys the brain activity of memorizing videos. This enabled them to finally predict VM without the use of fMRI scans in a second step. However, the method would be difficult to generalize. Shekhar *et al.* conducted a performance analysis of several computationally extracted features before building their memorability predictor [36]. The analysis encompassed C3D deep learning features, semantic features obtained from some video captioning process, saliency features, dense trajectories, and color features. They found that the most predictive feature combination used captioning features, dense trajectories, saliency and color features. The features that performed the best when used alone were image captioning features, i.e., those conveying more semantics. Due to the particularity of the dataset of Shekhar *et al.* – that is, their aforementioned use of questions to measure memorability –, it would be interesting to confirm if this combination works equally well on another dataset, and in particular if image captioning features also perform the best.

# 3 MEMORABILITY DATASET CONSTRUCTION

## 3.1 Video collection

We want our protocol to measure memory performance after a significant retention period. This can be achieved either with a longitudinal study, or by measuring a memory created prior to the experiment. We chose the latter because it enabled us to immediately measure very long-term memory. Thus, the main characteristic of the proposed protocol, in contrast with previous work, is the absence of learning (often, free viewing) task. It is replaced with a questionnaire designed to collect information about the participants' prior memory. In order to repeat measures of memory performances for different persons on the same material, to obtain average performance, we worked with movies famous enough to have been seen by several of our participants.

We first established a list of 100 occidental movies, taking care of mixing popularity and genres. We then manually selected seven videos of 10 seconds from each movie. To maintain a high intra-video semantic cohesion, we did not make cuts that would impair the understanding of the scene, nor did we aggregate shots that belong to different scenes. Indeed, since the semantics is linked to the memorability of images [19], we can expect it is linked to the memorability of videos too.

We also gave preference to the videos we called "neutral", by contrast to the "typical" ones. According to our definition, a neutral video is a part of a movie which contains no element that would enable someone to easily guess this video belongs to a particular movie. The list of undesirable elements includes but is not limited to: recognizable famous actors, typical music, style, etc. Typical videos are simply defined as non-neutral videos. In most movies, just a few or no 10-second neutral videos exist. That explains why we obtained only 127 neutral videos for 573 typical ones (while we expected two neutral and five typical videos per movie, i.e., 200 neutral and 500 typical videos in total).

## 3.2 Annotation protocol

The protocol is composed of two tasks. Firstly, participants had to fill in a questionnaire intended to collect data about whether they have seen the 100 selected movies. Secondly, participants performed a recognition task on videos selected based on their responses to the questionnaire.

104 participants (22 – 58 years of age; age average = 37.1; stdev = 10.4; 26% females; mostly educated persons – engineers or researchers mainly), participated in the experiment on a voluntary basis. The experiment was taking place in a well-controlled environment: a room insulated from noise and equipped with subdued lights. The videos, of HD or DVD quality, were displayed on a 60 inch monitor. The participants were seated at a distance of about 220 centimeters from the screen (three times the screen height). Having provided basic demographics, participants answered the questionnaire. For each of the 100 movies, they were asked whether they remembered watching fully the movie. In case of a positive answer, three additional questions followed: 1/ their confidence of watching the movie (*not confident / slightly confident / 50% confident / considerably confident / 100% confident*), 2/ the last time they saw the movie (*less than month / 1 year / 5 years / 10 years / more than 10 years*), and 3/ the number of times they saw the movie (*once / 2-4 times / 5-9 times / 10-19 times / more than 20 times*). In case of a negative answer, only one question was asked, related to their confidence of not having seen the movie. The questionnaire required about 20 minutes to complete.

Based on the answers to the questionnaire, an algorithm automatically selected 80 targets (i.e., videos from seen movies) and 40 fillers (i.e., videos from never seen movies) among the movies associated with the highest degree of certitude, with a maximum of two videos from the same movie. The fillers enable to quantify how much lucky confusions account for the correct recognitions. During the video selection, the current number of annotations was also taken into account to equally balance the annotations among all the videos. Given such 120 videos selected, participants performed a recognition task where they saw the videos separated by an inter-stimuli interval of 2 seconds. They had to press the space bar when they recognized a video in particular, and not when they were guessing that a particular video came from a movie they had seen (which was possible only for the typical videos).

### 3.3 Memorability score calculation

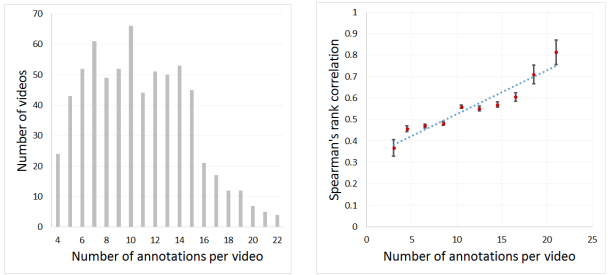
After collecting the data, we kept only the 660 videos that had been seen at least 4 times as targets (from the initial set of 700 videos). On average, each video of our dataset has been viewed as a target by 10.7 participants; which corresponds to the mean number of observations that enters in the calculation of a memorability score. We then assigned a memorability score to each video, defined as the correct recognition rate of the video when viewed as target. The average percentage of correct detections for all participants was 46.71% (stdev = 14.65%), and the average false alarm rate (i.e., the percentage of answers on fillers) was 4.16% (stdev = 5.27%). Figure 2(a) provides a distribution of the videos according to their degree of memorability. The dataset is publicly released here[1].

## 4 STUDY OF THE MEMORABILITY ANNOTATIONS

In this section, we conduct an analysis of the ground truth data collected through the protocol described above. We firstly perform a human consistency analysis. Then we compare neutral videos with typical videos. We finally study which factors affect the memorability among response time, duration of memory retention and repetition of visualization. In what follows, error bars in the graphs correspond to standard error of the mean, $\mu$ to the mean, $Mdn$ to the median and $N$ to the number of observations in the statistics.

### 4.1 Consistency analysis

We follow the method proposed in [19] to measure human consistency when assessing memorability of videos. We randomly split our 104 participants into two independent groups of equal size, and calculate how well video memorability scores from the first group of participants match with video memorability scores from the second group. Averaging over 25 random half-split trials, an average Spearman's rank correlation (i.e., a global human consistency) of 0.57 is observed between these pairs of scores.




(a) (b)

**Figure 1: (a) Distribution of the number of videos according to the number of annotations per video. (b) Human consistency (with linear trendline).**

We reproduced this calculation to obtain human consistency as a function of the number of annotations per video, presented in Figure 1(b). This graph is to be compared with the histogram presented in Figure 1(a), which shows that the number of videos for each number of annotations was unequal. According to the graph, we achieved a consistency of .70 from about 18 annotations, which is consistent with the previous finding [17]. This number also corresponds to the maximal consistency obtained when collecting IM scores [21, 24], but for a much bigger number of annotations (80) per image. It must be noted that the protocols are different between the IM experiments conducted in [21, 24] and ours or the work in [17]. We conducted a measure of long-term memory performance after at least two days of memorization, whereas in [21, 24] it is measured after a dozen of seconds to a few minutes. In addition, VM annotations were collected through in-lab experiments, and IM annotations through crowdsourcing experiments. However, it would be interesting in the future to confirm if an important difference exists between images and videos regarding the number of annotations necessary to achieve a high human consistency. Apart from the conclusions we could draw about the universality of the intrinsic memorability of videos compared to images, this would mean that the magnitude of the work to carry out to build an extensive database for VM prediction is substantially smaller than one could expect from work on IM prediction.

### 4.2 Neutral and typical videos

In our experiment, participants were given clear instructions that they had to really recognize any video they were presented as already seen, and not only guess that a video was coming from a movie whose title was proposed in the questionnaire. In this section, we perform an analysis to compare neutral videos, which contain no element that would enable participants to guess that a video belongs to a particular movie, and typical videos. Indeed, if neutral videos received objective answers from participants, it might be more subjective for typical videos, that could be more or less easily related to the movie they belong to.

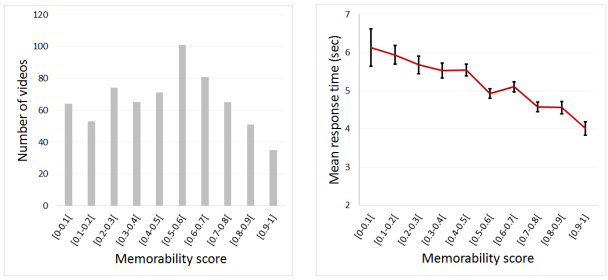
A Wilcoxon rank-sum test indicated that the memorability was greater for neutral ($Mdn = .20, \mu = .24$) than for typical ($Mdn = .53, \mu = .53$) videos, with $Z = 10.22, p < .00001$. Apart from the subjectivity aspect, we expected such a result because neutral videos contain less contextual elements, useful for recognition. Thus, this result does not necessarily mean that participants tended to guess – rather than simply recognize – videos selected from movies they have seen.

As for the human consistency on memorability, a Wilcoxon rank-sum test indicated that it was slightly greater for neutral ($Mdn = .44, \mu = .45$) than for typical ($Mdn = .40, \mu = .41$) videos, with $Z = 2.75, p < .01$. Along with the comments collected from the participants, who have as a majority reported difficulty to know if they were guessing or really recognizing the videos, this result suggests that human congruency is higher for more 'objectively' recognized segments than for ones with subjectivity as part of the equation. One could note there is probably not just pure subjectivity here. Context might have biased the results as it might have helped some participants in their recognition task. If confirmed, this would constitute a weakness of our protocol to collect extensive data, that one should in that case counteract by adapted measures of control.

[1] *www.remove-for-blind-review*. The original videos are not released because of copyright issues, but instead we provide the process we used to extract them from the movies and the features used in section 5.

We can also note that the average false alarm rate (that is, the percentage of wrongly recognized filler videos) was low for neutral videos (.05%) as well as for typical videos (.03%). Specifically, we expect lucky confusions to account for little of correct detections on average for the two sorts of videos.

## 4.3 Response time




(a) (b)

**Figure 2: (a) Distribution of the number of videos depending on memorability score ranges. (b) Mean response time for correct recognitions against memorability scores.**

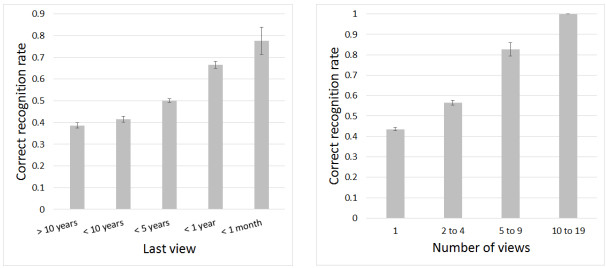
Figure 2 (b) shows that the response time to do a correct detection decreases when the memorability of the video increases. We also observed a Pearson's correlation of $-0.36$ ($p < .0001$) between the response time on the targets and their memorability scores. These two results indicate that participants tended to answer quicker when the videos were more memorable, even though the participants did not receive any instruction to do so. This suggests that people tend to naturally answer rapidly after having recognized the video. This also suggests either that the most memorable videos are also the most accessible in memory, and/or that the most memorable videos contain more early recognizable elements than the less memorable ones. In [36], the response time of the participants was taken to be the measure of video memorability. The authors chose this measure to avoid a long gap between viewing and recall stage. Our results validate – to some extent – their modus operandi: the fact that the response time decreases linearly when the memorability increases suggest that the response time is a good indicator of the memorability of the videos (at least, in a recognition task).

## 4.4 User context and memorability

To provide us with insights on which context-related factors collected through our questionnaire were linked to memorability, we processed to a logistic regression, using demographics and answers to the questionnaire as regressors, and the detection of a target video (with two possible discrete outcomes, detected or not) as observations to fit. Regarding the participants' nationality, we grouped them into occidental (69 pers) and non-occidental (35 pers) categories, motivated by our use of occidental movies, which could have more meaning for occidental than for non-occidental people. We also tested age and gender to reveal a potential bias in our movies' choice, that may be more memorable for people of a certain age and gender. The model, in case of a single observation $n$, can be written as:

$$y_n = \begin{cases} 1 & if\, \beta_0 + \sum_{k=1}^{K} x_{n,k}\beta_k + \epsilon_n > 0 \\ 0 & else \end{cases} \quad (1)$$

where $y_n$ denotes the dependent variable which can take two values, 1 for recognition or 0 for omission of the $n$-th target observation, $x_{n,k}$ our k-th feature value (last view, number of views, nationality, age, gender), $\beta_k$ are the coefficients to be estimated, and $\epsilon_n$ indicates the error term.




(a) (b)

**Figure 3: Correct recognition rate depending on (a) when occurred the last viewing, and (b) the number of views.**

According to the results of the logistic regression, the retention duration (last view) is highly negatively correlated with the probability to recognize a video ($\beta = -.37, p < .0001$). Figure 3 (a) shows that this decrease in memory for videos over time is continuous. This result indicates that long-term memory of videos continues to be altered over time for years. In the context or experiment, it implies that, to provide an accurate representation of an average long-term memory performance, a memorability score should correspond to a memory measure carried out as late as possible after the memorization.

The results of the logistic regression also show that the number of views is highly correlated with the probability to recognize a video ($\beta = .44, p < .0001$). As expected, the more a movie was seen, the better the videos were memorized. Figure 3 (b) shows that this continues to be true even with more than 9 viewings (however, the number of observations – 12 – was very low for videos which belong to movies with 10 or more views). One should note that the repetition of a viewing could not be the (only) factor involved in the above phenomenon; in particular, viewing again a movie may be the sign of a special interest which would explained a better memorization (e.g., via a greater attentional and emotional investment). The fact remains that repetition is an important factor to ask people when measuring their prior memory. Furthermore, a protocol used to build an extensive database for VM prediction should, in case of multiple measures of memory (e.g., after the memorization and then after a longer delay), avoid measuring twice the same items, because this repetition could artificially increase the performance measured for the last items.

We observed no significant effects of the demographic factors (nationality, age and gender). This suggests that the videos were

equally susceptible to be recognized by the different participants (or, that the relations between these factors and the observations are too complex to have been captured by the model).

## 5 MEMORABILITY PREDICTION

Until now, we have presented the video collection and the annotation protocol together with some insights on human VM. In this section, we move towards building a machine learning model that can learn and then predict the VM score of a video from its audio-visual features. The main goal of modelling is to understand if VM is predictable, and if yes identify which features: generic, perceptual, or semantic, are suitable for such prediction. We pose the problem as a standard regression problem and Figure 4 illustrates different steps in our method. In the following sub-sections, we explain our choice of features and models to address the problem in hand.

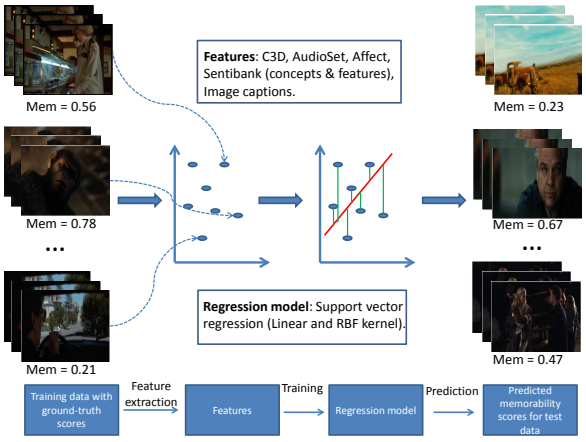



**Figure 4: Proposed approach for memorability score prediction.**

To build our predictive model, we split the dataset, at the level of movies, into training (70%), validation (15%) and test (15%) data, which translates into 70 movies in the training set and 15 movies each in the validation and test sets. We chose to split our dataset at the level of movies, instead of the videos, in order to avoid videos from the same movie being present in the training as well as the evaluation (validation+test) set. To ensure that such random split did not lead to any mismatch, we computed the average number of annotations per video in each of them. We observe that the numbers of annotations are balanced: each video in the test set has around 10 annotations while there are 9 annotations for each one in the train set on an average.

### 5.1 Feature extraction

The task of remembering a specific video has a high cognitive complexity in general, suggesting that it requires a semantic understanding of the content and/or some other perceptual factors such as the emotion conveyed by the video. Many users who participated in our experiments indicated that it is a difficult task. While trying to build a machine learning model for such a task, we explore different kinds of features that can be extracted from the audio-visual signal. We investigate a variety of generic state-of-the-art features ([39], [16]) and compare them with other semantic ([22]) and perceptual (emotion) features ([5]).

*5.1.1 Spatio-temporal visual features (C3D).* These features are extracted from the C3D model, a 3-dimensional convolutional network proposed for generic video analysis [39]. The main motivation to use C3D is that it encodes both the spatial and temporal information in the video. The model has been proposed for video analysis and is not an extension of a model for image analysis, unlike other state-of-the-art models like VGG16 [28]. We use the publicly available model trained on the Sports-1M dataset [39] and extract the output of the fully connected layer – fc6 of the network with a dimensionality of 4096. We additionally explore the use of principal component analysis (PCA) (named C3D (PCA) in Table 1) for the dimension reduction, as the original dimensionality is very high when compared to other features.

*5.1.2 Audio features (AudioSet).* Using a recently released AudioSet [16] model, which was trained on a large dataset for event detection, we extract 128-dimensional embeddings for each audio track associated with a video in our dataset. We use these embeddings for training the regression models. The motivation to use these features is that they are state-of-the-art in the audio event detection research and events could play a major role in how people remember sequences in movies. Additionally, we wanted to investigate how the audio channel contributes to building a model for VM prediction.

*5.1.3 Emotion related features (SentiBank and Affect).* As research in psychology showed that emotion and memory are correlated [8], we investigate the use of emotion-related feature in our prediction system. For modelling emotion from the visual content, we resort to a visual sentiment concept detector: SentiBank [5]. SentiBank is a set of 1200 trained visual concept detectors providing a mid-level representation of sentiment from visual content. We use the binary code for concept detection, from images, provided by the authors. The SentiBank concept detector provides two pieces of information: concepts with probabilities and features. Concepts are adjective–noun pairs and the probability represents how likely each concept is depicted visually in an image. Examples of some concepts in the SentiBank ontology are: *young driver*, *scary face*, *terrible pain*, etc.

We sample one frame for every second of the video in our dataset, resulting in 10 frames per video. We run the SentiBank concept detector on each of these 10 frames and rank the concepts based on the probability of their occurrence in the frame and take the top-50 concepts. We extract 300-dimensional word embeddings (Word2Vec [31]), for each of the 50 concepts and take an average to obtain a single vector per frame. We repeat this process for all the 10 frames and take the average of all the vectors to obtain a single feature vector for each video. Sentibank detectors also provide a 4096-dimensional feature for each frame and we take the average across all the frames to obtain one 4096-dimensional feature vector for each video. In the end, we use a 300-dimensional concept vector and a 4096-dimensional feature vector.

In addition to SentiBank concepts, we investigate other ways to capture emotional content in a video. Following a circumplex model of affect (the experience of emotion) [35], we define arousal as the

dimension of affect that measures the excitement in the video, while valence measures whether the video invokes positive or negative emotion. We resort to an audio-visual analysis of the video to obtain its arousal and valence scores using the method described in [18]. For each frame in the video, we compute the arousal and valence scores using the method proposed in [18]. In order to keep a fixed dimensionality of the feature vector, we take the first 200 frames in the video because of the varying frame rates across the videos. We concatenate the arousal and valence scores for the first 200 frames in each video resulting in a 400-dimensional feature vector (200 for arousal and 200 for valence) for a video.

*5.1.4 Visual semantic features (Image captions).* Visual semantics are known to play an important role in image memorability ([20], [37]). We utilize the state-of-the-art research in image captioning to capture such high-level semantics of the video [22]. We sample one frame for every second of the video in our dataset, resulting in 10 frames per video. For each of these 10 frames, we run the caption detector (code provided by the authors) and obtain a caption for the frame. For each non-functional word in the caption, we extract 300-dimensional word embeddings (Word2Vec [31]) and take an average across all the words to obtain a single vector per frame. We repeat this process for all the 10 frames and take the average of all the vectors to obtain a single 300-dimensional feature vector for each video.

## 5.2 Modelling and evaluation metric

We use the features discussed in Section 5.1 to train a Support Vector Regression (SVR) model for the VM score prediction. The choice of SVR is guided by the nature of the problem as well as by the small size of the dataset. We have chosen to go with the same regressor for all the features because our focus is mainly on identifying which features are more important for VM prediction. This way we ensure that the difference in performances is because of the features themselves. Note that, in addition to the variety of features explained in Section 5.1, we also explore a combination of all the features by concatenating them into a single feature vector. While performing such a concatenation, we use the low-dimension version of the features for C3D and SentiBank features, obtained after applying a dimension reduction method (PCA) to the original set. In the experiments where we use PCA for C3D and SentiBank features, we retain 95% of variance in the data while reducing the feature dimensions.

We use the grid search strategy to obtain the best hyper-parameters for SVR in term of the Mean Squared Error (*MSE*) between the predicted scores and the ground-truth on the validation set. The choice of hyper-parameters in the grid search are: kernel = {linear, RBF}, $C = \{0.1, 1, 10, 100, 1000\}$ and $\gamma = \{0.01, 0.1, 1, 10, 100\}$.

The prediction performance is evaluated by the Spearman correlation (*SpCorr*), which measures the rank correlation between the predicted memorability scores and the ground-truth scores. This metric is chosen as (1) we focus more on the relative memorability between videos rather than on their absolute memorability scores, and (2) it gives an indication of how close the predicted memorability scores are to the human consistency when annotating memorability scores.

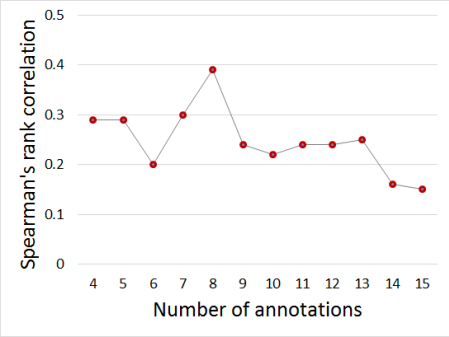



**Figure 5: Spearman correlation on the validation set with models using image captioning features trained on videos with varying number of annotations.**

## 5.3 Memorability prediction results

In this section, we will discuss how the models trained on different features perform while predicting memorability scores of new videos. Table 1 reports the prediction results obtained by the SVR model when trained on different features, for the validation and test sets. Two sets of results are reported: *SpCorr* ($\geq$ 4 annotations) and *SpCorr* ($\geq$ 8 annotations). The former corresponds to the prediction capability of the model when trained on videos with at least 4 annotations, and the latter corresponds to the results when trained on videos with at least 8 annotations where the ground-truth is better annotated.

As it can be seen, visual semantic features derived from image captioning clearly offer better prediction results compared to all other considered features. This is not surprising as they capture visual attributes along the scene, which are known to play an important role in human memory. This is also inline with a previous study in IM [37] where image captioning-based features helped better predict IM than the conventional CNN features. When predicting VM on the test set, C3D features showed to be quite effective to the task as they encode the visual spatio-temporal information of the video. On the contrary, audio information captured by AudioSet features does not seem to be enough for VM prediction. One of our initial hypotheses was that emotion would play an important role in VM, supported by literature from psychology [8]. However, when observing the results in Table 1, both Affect [18] and SentiBank [5] performed quite poorly compared to the image captioning features. Another observation from Table 1 is that the combination of all the features (last row in the table) does not appear in the top-3 best performing features. One of the reasons for this could be that there is a lot of redundance when combining all the features into a single feature vector, or it could be that the size of the combined feature vector is too big when compared to the size of the dataset. Thus, in future work we could look at selectively combining the features to investigate if that improves the performance.

As our dataset contains videos with different numbers of annotations, we further investigate the effect of the number of annotations on the prediction performance. For this purpose, we train a first SVR model, using image captioning features, on videos with at least 4 annotations and use this model to predict the memorability

| Feature | Feature type | Dimension | *SpCorr*(≥ 4 annotations) | | *SpCorr* (≥ 8 annotations) | |
|---|---|---|---|---|---|---|
| | | | *validation set* | *test set* | *validation set* | *test set* |
| C3D | visual spatio-temporal | 4096 | 0.20 | 0.26 | 0.31 | 0.34 |
| C3D (PCA) | visual spatio-temporal | 225 | 0.24 | 0.21 | 0.18 | 0.17 |
| AudioSet | audio related | 128 | 0.23 | 0.22 | 0.21 | 0.24 |
| Affect | affect related | 400 | 0.19 | 0.17 | 0.26 | 0.23 |
| SentiBank concepts | emotion related | 300 | 0.16 | 0.13 | 0.15 | 0.17 |
| SentiBank features | emotion related | 4096 | 0.25 | 0.21 | 0.27 | 0.26 |
| SentiBank features (PCA) | emotion related | 225 | 0.22 | 0.21 | 0.22 | 0.23 |
| Captions | visual semantics | 300 | **0.29** | **0.31** | **0.39** | **0.38** |
| Combination (PCA) | combine all features | 1578 | 0.24 | 0.23 | 0.29 | 0.27 |

**Table 1: Prediction results in terms of Spearman correlation scores on validation and test data for different features with models trained on videos that have at least 4 (columns 4-5) or 8 (columns 6-7) annotations.**

score for videos in the validation set. We repeat this process for different numbers of annotations per video (from 4 to 15) in the training set. Please note that the validation set in each of the repetitions is fixed and only the training set changes. We provide a demonstration of how *SpCorr* varies with an increasing number of annotations in the training set in Figure 5. As can be seen, *SpCorr* first increases up to 5 annotations and then remains more or less constant before decreasing (beyond 10 annotations). In the wake of this observation, we also investigated the performance of all the features when we train the regression model with videos that have at least 8 annotations as reported in the second part of Table 1: *SpCorr* ( ≥ 8 annotations).

Comparing the two sets of results in Table 1, we observe that globally the models trained on videos with at least 8 annotations perform better than the models trained on videos with at least 4 annotations. These results are comparable to the human consistency analysis shown in Section 4.1. We finally performed an additional 10-fold cross-validation on videos with at least 8 annotations (train+validation set) using the image captioning features. From this study, we re-confirmed that there is no overfitting issue in our model and we observed the average value of *SpCorr* across the 10 folds to be 0.33, which is close to the performance on the test set.

## 6 CONCLUSIONS

In this paper, we have presented a novel protocol to collect long-term memorability annotations for videos, which enabled us to build an important dataset to support research in this subject. We then performed a range of statistical studies on this dataset to understand important factors in the annotation process as well as how they can affect the video memorability. One of our key observations is that high human consistency in video memorability can be obtained by only about 18 annotations, which is significantly lower than for images according to the previous studies. We finally proposed computational models for video memorability prediction where we investigated the use of various audio-visual features for the task. For this, we observed that the visual semantic features offer the best prediction result, which re-confirms the correlation between the visual attributes and memorability. Our current work focuses on building a large scale video memorability dataset using crowdsourcing annotations.

## REFERENCES


[1] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. 2013. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* 142, 4 (2013), 1323.

[2] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. 2016. Deep Learning for Image Memorability Prediction: the Emotional Bias. In *Proc. ACM Int. Conf. on Multimedia (ACMM)*. 491–495.

[3] Robert S Blumenfeld and Charan Ranganath. 2007. Prefrontal cortex and long-term memory encoding: an integrative review of findings from neuropsychology and neuroimaging. *The Neuroscientist* 13, 3 (2007), 280–291.

[4] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.

[5] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. SentiBank: Large-scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content. In *Proc. ACM International Conference on Multimedia (ACMM)*. 459–460.

[6] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* 105, 38 (2008), 14325–14329.

[7] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. 2015. Intrinsic and extrinsic effects on image memorability. *Vision research* 116 (2015), 165–178.

[8] Larry Cahill and James Mcgaugh. 1996. A Novel Demonstration of Enhanced Memory Associated with Emotional Arousal. 4 (01 1996), 410–21.

[9] Bora Celikkale, Aykut Erdem, and Erkut Erdem. 2013. Visual attention-driven spatial pooling for image memorability. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshops*. 976–983.

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

[11] Romain Cohendet, Anne-Laure Gilet, Matthieu Perreira Da Silva, and Patrick Le Callet. 2016. Using individual data to characterize emotional user experience and its memorability: Focus on gender factor. In *Proc. Int. Conf. on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[12] Claire-Hélène Demarty, Mats Viktor Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Hanli Wang, Ngoc Q. K. Duong, and Frédéric Lefebvre. 2016. Mediaeval 2016 predicting media interestingness task. In *Proc. MediaEval Workshop*.

[13] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1657–1664.

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

[15] Orit Furman, Nimrod Dorfman, Uri Hasson, Lila Davachi, and Yadin Dudai. 2007. They saw a movie: long-term memory for an extended audiovisual narrative. *Learning & memory* 14, 6 (2007), 457–467.

[16] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE Int. Conf. on Audio, Speech and Language Processing (ICASSP)*.

[17] 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.

[18] A. Hanjalic and Li-Qun Xu. 2005. Affective video content representation and modeling. *IEEE Transactions on Multimedia* 7, 1 (Feb 2005), 143–154.

[19] 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.

[20] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What Makes a Photograph Memorable? *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 7 (July 2014), 1469–1482.

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

[22] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (April 2017), 664–676.

[23] Elizabeth A Kensinger and Daniel L Schacter. 2008. Memory and emotion. *Handbook of emotions* 3 (2008), 601–617.

[24] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. 2390–2398.

[25] Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. 2013. Relative spatial features for image memorability. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 761–764.

[26] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Technical Report. Naval Technical Training Command Millington TN Research Branch.

[27] Souad Lahrache, Rajae El Ouazzani, and Abderrahim El Qadi. 2016. Bag-of-features for image memorability evaluation. *IET Computer Vision* 10, 6 (2016), 577–584.

[28] S. Liu and W. Deng. 2015. Very deep convolutional neural network based image classification using small training sample size. In *Proc. Asian Conference on Pattern Recognition (ACPR)*. 730–734.

[29] Matei Mancas and Olivier Le Meur. 2013. Memorability of natural scenes: The role of attention. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*. 196–200.

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

[31] Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. 2013 (01 2013).

[32] Lynn Nadel and Morris Moscovitch. 1997. Memory consolidation, retrograde amnesia and the hippocampal complex. *Current opinion in neurobiology* 7, 2 (1997), 217–227.

[33] M Ross Quillan. 1966. *Semantic memory.* Technical Report. Bolt Beranek and Newman Inc Cambridge MA.

[34] Alan Richardson-Klavehn and Robert A Bjork. 1988. Measures of memory. *Annual review of psychology* 39, 1 (1988), 475–543.

[35] James Russell. 1980. A Circumplex Model of Affect. 39 (12 1980), 1161–1178.

[36] 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 Conference on Computer Vision and Pattern Recognition*. 2730–2739.

[37] Hammad Squalli-Houssaini, Ngoc Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2018. Deep learning for predicting image memorability. In *Proc. IEEE Int. Conf. on Audio, Speech and Language Processing (ICASSP)*.

[38] Lionel Standing. 1973. Learning 10000 pictures. *Quarterly Journal of Experimental Psychology* 25, 2 (1973), 207–222.

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. 4489–4497.