VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability

Romain Cohendet
Technicolor, Rennes, France
romain.cohendet@laposte.net

Claire-Hélène Demarty
InterDigital, Rennes, France
claire-helene.demarty@interdigital.com

Ngoc Q. K. Duong
InterDigital, Rennes, France
quang-khanh-ngoc.duong@interdigital.com

Martin Engilberge
InterDigital, Rennes, France
martin.engilberge@interdigital.com

Abstract

Humans share a strong tendency to memorize/forget some of the visual information they encounter. This paper focuses on understanding the intrinsic memorability of visual content. To address this challenge, we introduce a large scale dataset (VideoMem) composed of 10,000 videos with memorability scores. In contrast to previous work on image memorability – where memorability was measured a few minutes after memorization – memory performance is measured twice: a few minutes and again 24-72 hours after memorization. Hence, the dataset comes with short-term and long-term memorability annotations. After an in-depth analysis of the dataset, we investigate various deep neural network-based models for the prediction of video memorability. Our best model using a ranking loss achieves a Spearman's rank correlation of 0.494 *(respectively* 0.256*) for short-term (resp. long-term) memorability prediction, while our model with attention mechanism provides insights of what makes a content memorable. The VideoMem dataset with pre-extracted features is publicly available*[1].

1. Introduction

While some contents have the power to burn themselves into our memories for a long time, others are quickly forgotten [17]. Evolution made our brain efficient to remember only the information relevant for our survival, reproduction, happiness, *etc*. This explains why, as humans, we share a strong tendency to memorize/forget the same images, which translates into a high human consistency in image memorability (IM) [20], and probably also a high consistency for video memorability (VM). Although, like for any other perceptual concept, we can observe individual differences while memorizing content, in this paper we target the capture and prediction of the part of the memorability that is shared by humans, as it can be assessed by averaging individual memory performances. This *shared-across-observers* part of the memorability, and especially long-term memorability, has a very broad application range in various areas including education and learning, content retrieval, search, filtering and summarizing, storytelling, *etc*.

The study of VM from a computer vision point of view is a new field of research, encouraged by the success of IM since the seminal work of Isola *et al.* [17]. In contrast to other cues of video importance, such as aesthetics, interestingness or emotions, 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 judgement). This certainly participates to the growing interest for its study. IM 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 [17]. This definition has been widely accepted within subsequent work [24, 21, 3, 20, 23]). The introduction of deep learning to address the challenge of IM prediction causes models to achieve results close to human consistency [20, 1, 34, 18, 31, 12]. As a result of this success, researchers have recently extended this challenge to videos [14, 30, 7, 5]. However, this new research field is nascent. As argued in [7], releasing a large-scale dataset for VM would highly contribute to launch this research field, as it was the case for the two important dataset releases in IM [17, 20]. Such a dataset should try to overcome the weaknesses of the previously released datasets. In particular, previous research on IM focused on the measurement of memory performances only a few minutes after memorization. However, passage of time is a factor well-studied in psychology for its influence on memory, while having been

[1] https://www.technicolor.com/dream/research-innovation/video-memorability-dataset

largely ignored by previous work on IM, probably because of the difficulty to collect long-term annotations at a large scale, in comparison with short-term ones. Measuring a memory performance a few minutes after the encoding step is already a measure a long-term memory, since short-term memory usually lasts less than a minute for unrehearsed information [28]. However, memories continue to change over time: going through a consolidation process (*i.e.,* the time-dependent process that creates our lasting memories), some memories are consolidated and others are not [25]. In other words, short-term memory performances might be poor predictors of longer term memory performances. In the following, we refer to measures of long-term memory a few minutes after memorization as measures of *short-term memorability*, and use the term *long-term memorability* for measures of long-term memory performance after one day. Since long-term memorability is more costly and difficult to collect than short-term memorability, it would nevertheless be interesting to know if the former can be inferred from the latter, which would also push forward our understanding of what makes a video *durably* memorable. A way to achieve this consists in measuring memorability for the same videos at two points of time. These two measures would be particularly interesting if spaced by a time interval in which forgetting is quite significant, to maximize the size of the potentially observable differences depending on the different video features. Observing the different forgetting curves in long-term memory (*e.g*. Ebbinghauss seminal work [9]), one can observe that the drop in long-term memory performance in recall follows an exponential decay and is particularly strong in the first hour, and to a lesser extent in the first day, immediately after the memorization. Measuring long-term memory a few minutes after encoding (as done in studies of IM [17, 20]), and again one day or more after (*i.e.,* to obtain a measure close to very long-term memory), sounds therefore a good trade-off.

The main contributions of this work are fivefold:

- We introduce a new protocol to objectively measure human memory of videos at two points of time (a few minutes and then 24-72 hours after memorization) and release VideoMem, the premier large-scale dataset for VM, composed of 10,000 videos with short-term and long-term memorability scores (Sections 3.1 and 3.2).
- Through an analysis of the dataset, we address the problem of understanding VM, by highlighting some factors involved in VM (Section 4).
- We benchmark several video-based DNN models for VM prediction (Section 5.2) against image-based baseline models (Section 5.1).
- We prove that, similarly to IM, semantics is highly relevant for VM prediction, through the study of a state-of-the-art image-captioning model (Section 5.3). This best model reaches a performance of 0.494 for Spearman's rank correlation on VideoMem for short-term memorability and 0.256 for long-term memorability.
- We propose an extension of the best performing model with an attention mechanism to localize what in an image makes it memorable (Section 5.5).

2. Related work

If long-term memory has been studied for over a century in psychology, since the seminal experimental studies of Ebbinghaus [10], its study from a computer vision point of view started quite recently, with [17]. Images and videos had long been used as material to assess memory performances [32, 2, 13], proving that human possessess an extensive long-term visual memory. The knowledge accumulated in psychology helped to measure memory using classical memory tests (see [29] for an extensive overview) such as recognition tests [17, 20, 14, 7] or textual question-based recall surveys [30]. Several factors are highlighted in the psychological literature for their critical influence on long-term memory, including emotion [19], attention [8], semantics [27], several demographic factors [6], memory re-evocation [26], or passage of time [25], also providing computer vision researchers with insights to craft valuable computational features for IM and VM prediction [24, 16, 7].

Focusing on IM in computer vision, most studies made use of one of the two available large datasets, specifically designed for IM prediction, where IM was measured a few minutes after memorization [17, 20], and consequently focused on predicting a so-called short-term IM [24, 21, 3, 20, 1, 23, 31, 12]. The pioneering work of [17] focused primarily on building computational models to predict IM from low-level visual features [17], and showed that IM 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 [24], interestingness and aesthetics [16], or emotions [20]. The best results were finally obtained by using fine-tuned or pre-extracted deep features, which outperformed all other features [20, 1, 31, 12], with models achieving a Spearman's rank correlation near human consistency (*i.e.,* .68) when measured for the ground truth collected in [17, 20].

VM study is more recent. To the best of our knowledge, there exist only three previous attempts at measuring it [14, 30, 7]. Inspired by [17], Han *et al.* built a similar but far much heavier protocol to measure VM: 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. Another approach uses questions instead of a classic visual recognition task to measure VM [30]. As a results, memorability annotations collected for the videos may reflect not only the differences in memory performances but also the differences of complexity between the questions, especially since the authors use

the response time to calculate memorability scores, which might critically depend on the questions' complexity. The most recent attempt at measuring VM, and the only one, to our knowledge, resulting in a publicly available dataset, comes from [7]. The authors introduced a novel protocol to measure memory performance after a significant retention period – *i.e.,* weeks to years after memorization – without needing a longitudinal study. In contrast with previous work, the annotators did not pass through a learning task. It was replaced with a questionnaire designed to collect information about the participants' prior memory of Hollywood-like movies. However, such a protocol implies a limited choice of content: authors needed contents broadly disseminated among the population surveyed, as the participants should have seen some of them before the task (hence the Hollywood-like movies), leading to a number of annotations biased towards most famous content. Furthermore, the absence of control of the memorizing process and the answers of the questionnaire based on subjective judgments make the measure of memory performance not fully objective. To sum up, none of the previous approaches to measure VM is adapted to build a large-scale dataset with a ground truth based on objective measures of memory performance. Results obtained for VM prediction are yet far from those obtained in IM prediction. Han *et al.* proposed a method which combines audio-visual and fMRI-derived features supposedly conveying part of the brain activity when memorizing videos, which in the end enables to predict VM without the use of fMRI scans [14]. However, the method would be difficult to generalize. Shekhar *et al.* investigated several features, including C3D, semantic features obtained from some video captioning process, saliency features, dense trajectories, and color features, before building their memorability predictor [30]. They found that the best feature combination used dense trajectories, captioning, saliency and color features.

3. VideoMem: large-scale video memorability dataset

In Section 3.1, we describe the collection of source videos that compose the VideoMem dataset. We then introduce a new protocol to collect short-term and long-term memorability annotations for videos (Section 3.2), before explaining the computation of VM scores (Section 3.3).

3.1. Video collection

The dataset is composed of 10,000 soundless videos of 7 seconds shared under a license that allows their use and redistribution for research purpose only. In contrast to previous work on VM, where videos came from TRECVID [30, 14] or were extracted from Hollywood-like movies [7], videos in our dataset were extracted from raw footage, mainly from staged settings, dedicated to be further edited by professionals when creating new content, *e.g.* a new motion picture, video clip, television show, advertisements, *etc*. Because such video footage is typically used to save shooting new material, it is usually generic enough to be easily integrated in different sorts of creations. As such, they are context-independent and contain only one semantic scene. By this choice of content, we expect these basic building units to be relevant to train models which generalize on other types of videos. We are also confident that observers never saw the videos before participating in the experiment. Videos are varied and contain different scene types such as animal, food and beverages, nature, people, transportation, *etc*. A few of them contain similarities, *e.g.* same actor, same place but slightly different action, as it is the case in everyday video consumption ($< 1\%$). A small fraction is also slow-motion. Each video comes with its original title, that can often be seen as a list of tags (textual metadata). Example video keyframes are shown in Fig. 1.

The original videos are of high quality (HD or 4k) and of various durations (from seconds to minutes). As it will be described in Section 3.2, our protocol relies on crowdsourcing. For the sake of fluency during the annotation collection and consistency between the videos, we rescaled the videos to HD and re-encoded them in *.webm* format, with a bitrate of 3,000 kbps for 24 fps. To satisfy to the protocol's constraints, *i.e.,* minimal delay before measuring memory performance and maximal duration of the tasks to avoid user fatigue, we also cut the videos to keep only the first 7 seconds. Most videos are short ($<$ a few minutes) and contain one semantic scene. Those 7 seconds should therefore be representative of their content. Videos are soundless, firstly because a large part of the original data came without audio, and secondly, because it is difficult to control the audio modality in crowdsourcing. Accordingly, memorability would be linked only to the visualization of a semantic unit, which sounds a reasonable step forward for VM prediction, without adding a potentially biasing dimension.

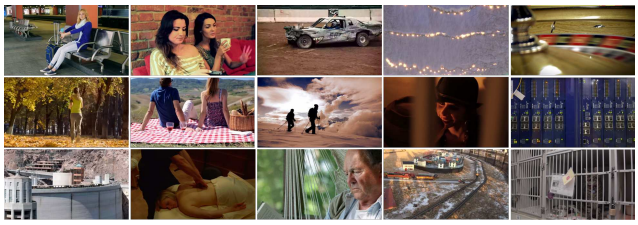
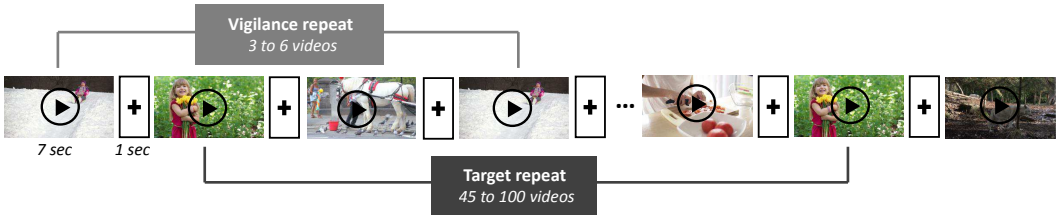


Figure 1: Example keyframes from videos of VideoMem, sorted by decreasing long-term memorability (from left to right, and top to bottom).

3.2. Annotation protocol

To collect VM annotations, we introduced a new protocol which enables to measure both human short-term and



(a) Step #1. Interlaced encoding and recognition tasks.

(b) Step #2. Second recognition task after 24 to 72 hours.

Figure 2: Proposed protocol to collect both short-term and long-term video memorability annotations. The second recognition task measures memory of videos viewed as fillers during step #1, to collect long-term memorability annotations.

long-term memory performances. Inspired by what was proposed in [16, 17] for IM, we also used recognition tests for our memorability scores to reflect objective measures of memory performance. However, our protocol differs in several ways, not mentioning the fact that it is dedicated to videos. Firstly, as videos have an inherent duration, we had to revise 1) the delay between the memorization of a video and its recognition test and 2) the number of videos, for the task not be too easy. Secondly, in contrast to previous work on IM, where memorability was measured only a few minutes after memorization, memory performance is measured twice to collect both short-term and long-term memorability annotations: a few minutes after memorization and again (on different items) 24-72 hours later. The retention interval between memorization and measure is not as important as in [7], where it lasts weeks to years. As previously explained, we hope, however, that this measure reflects very-long term memory performance instead of short-term memory, as forgetting happens to a large extent during the first day following the memorization.

Fig. 2 illustrates our protocol, that works in two steps. Step #1, intended to collect short-term annotations, consists of interlaced viewing and recognition tasks. Participants watch a series of videos, some of them – the *targets* – repeated after a few minutes. Their task is to press the space bar whenever they recognize a video. Once the space bar is pressed, the next video is displayed, otherwise current video goes on up to its end. Each participant watches 180 videos, that contain 40 *targets*, repeated once for memory testing, and 80 *fillers* (*i.e.,* non target videos), 20 of which (so-called *vigilance fillers*) are also repeated quickly after their first occurrence to monitor the participant's attention to the task. The 120 videos (not counting the repetitions) that participate to step #1 are randomly selected among the 1000 videos that received less annotations at the time of the selection. Their order of presentation is randomly generated by following the given rule: the repetition of a *target* (respectively a *vigilance filler*) occurs randomly 45 to 100 (resp. 3 to 6) videos after the *target* (resp. *vigilance filler*) first occurrence. In the second step of the experiment, that takes place 24 to 72 hours after step #1, the same participants are proposed another similar recognition task, intended to collect long-term annotations. They watch a new sequence of 120 videos, composed of 80 *fillers* (randomly chosen totally new videos) and 40 *targets*, randomly selected from the *non-vigilance fillers* of step #1. Apart from the vigilance task (step #1 only), we added several controls, settled upon the results on an in-lab test: a minimum correct recognition rate (15%, step #2 only), a maximum false alarm rate (30%, step #1; 40%, step #2) and a false alarm rate lower than the recognition rate (step #2 only). This allows to obtain quality annotations by validating each user's participation; a participant could participate only once to the study. We recruited participants from diverse countries and origins via the Amazon Mechanical Turk (AMT) crowdsourcing platform.

3.3. Memorability score calculation

After a filtering of the participants to keep only those that passed the vigilance controls, we computed the final memorability scores on 9,402 participants for short-term, and 3,246 participants for long-term memorability. On average, a video was viewed as a repeated target 38 times (and at least 30 times) for the short-term task, and 13 times (at least 9 times) for the long-term task (this difference is inherent to the lower number of participants in step #2, as a large part of participants in step#1 did not come back). We assigned a first raw memorability score to each video, defined as the percentage of correct recognitions by participants, for both short-term and long-term memorability.

The short-term raw scores are further refined by applying

a linear transformation that takes into account the memory retention duration to correct the scores. Indeed, in our protocol, the repetition of a video happens after variable time intervals, *i.e.,* after 45 to 100 videos for a *target*. In [16], using a similar approach for images, it has been shown that memorability scores evolve as a function of the time interval between repeats while memorability ranks are largely conserved. We were able to prove the same relation for videos, *i.e.,* memorability decreases linearly when the retention duration increases (see Fig. 3, left). Thus, as in [20], we use this information to apply a linear correction (shown in Fig. 3) to our raw memorability scores to explicitly account for the difference in interval lengths, with the objective for our short-term memorability scores to be the most representative of the typical memory performance after the maximal interval (*i.e.,* 100 videos). Note that the applied correction has nevertheless little effect on the scores both in terms of absolute and relative values. Note also that we did not apply any correction for long-term memorability scores (Fig. 3, right). Indeed, we observed no specific, strong enough relationship between retention duration and long-term memorability. This was somehow expected from what can be found in the literature : according to our protocol, the second measure was carried out 24 to 72 hours after the first measure. After such a long retention duration, it is expected that the memory performance is no more subjected to substantial decrease due to the retention duration. In the end, the average short-term memorability score is 0.859 (instead of 0.875) and the average long-term memorability score is 0.778, all values showing a bias towards high values.

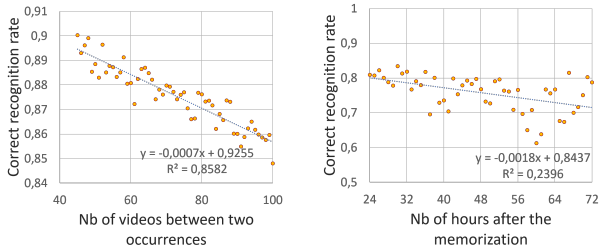
4. Understanding video memorability

4.1. Human consistency *vs.* annotation consistency

Following the method proposed in [16], we measured human consistency when assessing VM. For this purpose, we randomly split our participants into two groups of equal size (4,701 for short-term memorability, 1,623 for long-term memorability), and computed VM scores independently in each group as described in Section 3.3. We then calculated a Spearman's rank correlation between the two groups of scores. Averaging over 25 random half-split trials, an average Spearman's rank correlation, *i.e.,* a global human consistency, of 0.481 is observed for short-term memorability and of 0.192 for long-term memorability.

Such a method divides the number of annotations that is taken into account for the score computation at least by a factor of 2. Moreover, it may end with groups with unbalanced number of annotations per video as the split is randomly applied on the participants, not taking into account which videos they watched. For this reason, we proposed a new metric named *annotation consistency*, more representative of the performance consistency of the users. We reproduced the previous process of human consistency computation but on successive subparts of the dataset by considering for each sub-part only videos which received at least N annotations. Each subpart is then split in two groups of participants while ensuring a balance number of participants per video. By doing so, we obtain the annotation consistency as a function of the number of annotations per video, as presented in Fig. 4. This allows us to interpolate the following values: Annotation consistency reaches 0.616 (respectively 0.364) for the short-term (resp. long-term) task, for a number of annotations of 38 (resp. 13). Both values represent strong (resp. moderate) correlations according to the usual Spearman scale of interpretation. Hence, choosing larger mean number of annotations provides more stable annotations, *i.e.,* 0.616 (resp. 0.364) rather than 0.481 (resp. 0.192) for the short-term (resp. long-term) task.

The value of 0.616 for short-term memorability is to be compared to 0.68 for images as found in [20]. Slightly lower, VideoMem consistency was nevertheless obtained with less annotations than in [20], which is consistent with [7]. The maximum consistency is also slightly higher for VM than for IM (0.81 against 0.75 in [17] and 0.68 in [20]). An explanation is that videos contain more information than images and thus are more easily remembered. However, one should keep in mind that the protocols to collect annotations differ in several ways, making these results not fully comparable. Fig. 4 also shows that long-term and short-term consistencies follow the same evolution.



(a) Step #1. Recognition rate decreases linearly over time.

(b) No significant change in memory performance between 24 and 72 hours after memorization.

Figure 3: Mean correct recognition rate *vs.* the retention interval between the memorization and the measure of memory performance. Blue lines represent linear fitting.

4.2. Memorability consistency over time

In this study, we are interested in assessing how well memorability scores remain consistent over time, *i.e.,* if a video highly memorable after a few minutes of retention remains also highly memorable after 24 to 72 hours. The Spearman's rank correlation coefficient between the long-term and short-term memorability scores for

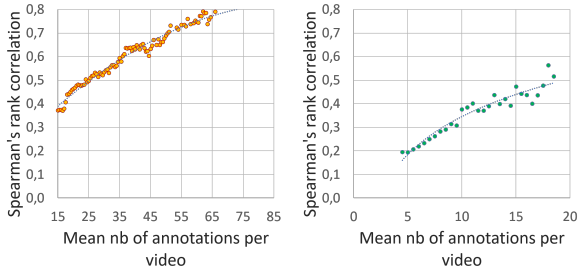


Figure 4: Annotation consistency *vs.* mean number of annotations per video (left: short-term, right: long-term).

the 10,000 videos exhibits a moderate positive correlation ($\rho = 0.305, p < .0001$) between the two variables, as also shown in Fig. 5. To discard a potential bias that would come from the highest number of annotations in step #1 compared to step #2, we computed the correlation for the 500 most annotated videos in the long-term task (that have at least 21 annotations) and then again for the 100 most annotated (at least 28 annotations), observing similar Spearman values of $\rho = 0.333$, $p < .0001$ and $\rho = 0.303$, $p < .0001$, respectively. This result suggests that memory evolves with time and in a non-homogeneous manner depending on the videos: a video highly memorable a few minutes after visualization might not remain highly memorable in long-term memory. This finding is consistent with the hypothesis we proposed in the introductory section, that the information important for a content to be memorized might not be the same for short-term and long-term memorization.

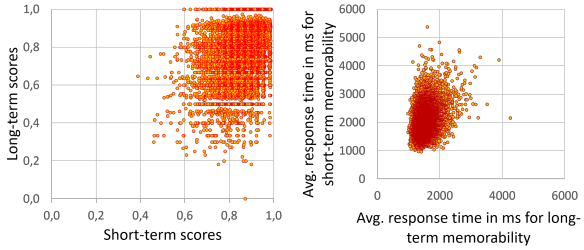


Figure 5: Short-term *vs.* long-term scores (left) and average response times (correct detections only) (right).

4.3. Memorability and response time

We observed negative Pearson correlations between the mean response time to correctly recognize targets and their memorability scores, both for short-term ($r = 0.307$, $p < .0001$) and long-term (0.176, $p < .0001$) memorability, as also illustrated in Fig. 6. This tends to prove that, globally, participants tended to answer more quickly for the most memorable videos than for the less memorable ones. This is consistent with [7], where the authors propose two explanations to this result: either the most memorable videos are also the most accessible in memory, and/or the most memorable videos contain more early recognizable elements than the less memorable ones. As videos in VideoMem consist of semantic units with often one unique shot – with most of the information already present from the beginning – the first explanation sounds more suitable here. This also suggests that participants tend to quickly answer after recognizing a repeated video (even though they did not receive any instruction to do so), maybe afraid of missing the time to answer, or to alleviate their mental charge. This correlation highlights that the average response time might be a useful feature to further infer VM in computational models.

The correlation is, however, lower for long-term memorability. One explanation might be that, after one day, remembering is more difficult. In connection with this explanation, we observed a significant difference between the mean response time to correctly recognize a video during step #1 and during step #2 ($1.43sec.$ *vs.* $3.37sec.$), as showed by a Student's t-test ($t(9999) = -122.59, p < 0001$). Note that the Pearson correlation (0.291) between average response time per video for short-term and long-term memorability is close to the Pearson correlation (0.329) observed between short-term and long-term memorability scores (see Fig. 5, right). Note that the mean response time for a false alarm was $3.17sec.$ for step #1 and $3.53sec.$ for step #2.

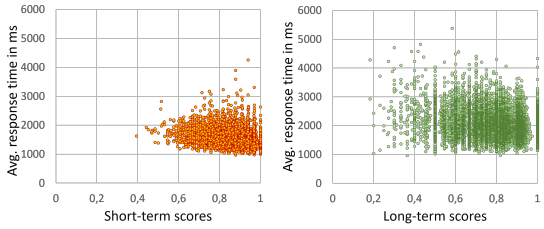


Figure 6: Average response time (correct recognitions only) as a function of memorability scores, for short-term (left) and long-term memorability (right).

5. Predicting video memorability

In this section we focus on predicting VM using various machine learning approaches. We pose the VM score prediction as a standard regression problem. We first benchmark several state-of-art video-based models on our data (Section 5.2), against performances of IM models (Section 5.1). We then focus on assessing how a very recent state-of-the-art image captioning based model, fine-tuned on our data, performs for VM prediction. The aim is here to see if the finding in [31, 7] that semantics highly intervenes in IM prediction still stands for VM prediction. In Section 5.4, we analyze the prediction results of all models and give insights to understand the correlation between IM and VM.

Last, in Section 5.5, we modify the advanced IC model by adding an attention mechanism that helps us better understand what makes a content memorable. Note that, for training (when applied) and evaluating the considered models, we split VideoMem dataset into training (6500 videos), validation (1500), and test (2000) sets, where the test set contains 500 videos with a greater number of annotations. Similarly to previous work in IM and VM, the prediction performance is evaluated in term of the Spearman's rank correlation between the ground truth and the predicted scores.

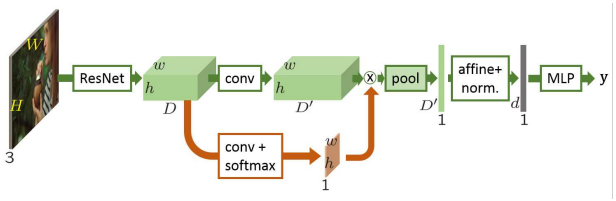


Figure 7: Semantic embedding model without (green pipeline) and with an attention mechanism (full workflow).

5.1. Image memorability-based baselines

In order to investigate the correlation between IM and VM and to build some first baselines on the dataset, we directly used two state-of-the-art models for IM prediction to compute a memorability score for 7 successive frames in the video (one per second): MemNet proposed in [20] and Squalli *et al.* in [31]. The final VM score for one video is obtained by averaging the 7 corresponding frame scores.

5.2. Video-based models

In a first attempt to capture the inherent temporal information of the videos, we investigated the performances of two classic, yet temporal, features: C3D [33] and HMP [4] as input features to some MLP layers. We tested them alone and concatenated, using some grid search for hyperparameter optimisation. Best results were obtained for the features alone, with the parameters: two hidden layers with 10 neurons for HMP and one hidden layer with 100 neurons for C3D, optimizer=IBLGS, activation=tanh, learning rate (lr)=1e-3. Second, instead of using a fix feature extractor, we directly fine-tuned the state-of-the-art ResNet3D model (based on ResNet34) [15]. For this, we replaced the last fully connected layer of ResNet3D by a new one dedicated to our considered regression task. This last layer was first trained alone for 5 epochs (Adam optimizer, batchsize=32, lr=1e-3), then the whole network was re-trained for more epochs (same parameters, but lr=1e-5).

5.3. Semantic embedding-based model

As scene semantic features derived from an image captioning system (IC) [22] have been shown to well characterize the memorability of images [31] and videos [7], we also investigated the use some IC system. Also, following the idea of model fine-tuning, we fine-tuned a state-of-art visual semantic embedding pipeline used for image captioning [11], on top of which a 2-layer MLP is added, to regress the feature space to a single memorability score. The overall architecture is shown in Fig. 7, in the green pipeline. As the model in [11] remains at the image-level, we first predict scores for the same 7 frames as in Section 5.1, then compute the final prediction at video level by averaging those 7 values. It is fine tuned on both VideoMem and LaMem [20] datasets, for short-term memorability only, because LaMem only provides short-term annotations. The training is done using the Adam optimizer and is divided in two steps: in the first 10 epochs only the weights of the MLP are updated while those of the IC feature extractor remain frozen. Later the whole model is fine-tuned. The learning rate is initialized to 0.001 and divided in half every three epochs. It is important to note that the original IC model was trained with a new ranking loss (*i.e.,* Spearman surrogate) proposed in [11]. This new loss has proved to be highly efficient for ranking tasks as claimed in [11]. For the fine-tuning however, the training starts with a ℓ_1 loss as initialization step, before coming back to the ranking loss. The reason is that the original model was indeed trained for scores in [-1;1], while our memorability scores are in [0;1]. Thus the ℓ_1 loss forces the model to adapt to this new range.

5.4. Prediction results

From the results in Table 1, we may draw several conclusions. Additional results are presented in the supplementary material. First, it is possible to achieve already quite good results in VM prediction using models designed for IM prediction. This means that the memorability of a video is correlated to some extent with the memorability of its constituent frames. For both C3D and HMP-based models, it seems that the simple MLP layers put on top of those features did not successfully capture the memorability. This might be explained by the fact that most of the videos contain no or little motion (62%), whereas 11% only contain high motion. However, the comparison between short-term and long-term performances exhibits some interesting information: HMP performs better than C3D for short-term and the inverse is true for long-term, as if direct motion information was more relevant for short-term than for long-term memorability. This is a first finding on what distinguishes the two notions. Also, the two fine-tuned models, dedicated to the task, show significantly higher performances. The fine-tuned ResNet3D, although purely video-based, is exceeded by the fine-tuned semantic embedding-based model. However, for the latter, data augmentation was performed using the LaMem dataset [20], which was not possible for the former as LaMem only contains image memorability information. This indeed biases the comparison between

Models	short-term memorability			long-term memorability		
	validation	test	test (500)	validation	test	test (500)
MemNet (Sec. 5.1)	0.397	0.385	0.426	0.195	0.168	0.213
Squalli *et al.* (Sec. 5.1)	0.401	0.398	0.424	0.201	0.182	0.232
C3D (Sec. 5.2)	0.319	0.322	0.331	0.175	0.154	0.158
HMP (Sec. 5.2)	0.469	0.314	0.398	0.222	0.129	0.134
ResNet3D (Sec. 5.2)	0.508	0.462	0.535	0.23	0.191	0.202
Semantic embedding model (Sec. 5.3)	**0.503**	**0.494**	**0.565**	**0.26**	**0.256**	**0.275**

Table 1: Results in terms of Spearman's rank correlation between predicted and ground truth memorability scores, on the validation and test sets, and on the 500 most annotated videos of the dataset (test (500)) that were placed in the test set.

the two models, but current results still show that, as expected, leveraging both a dedicated fine-tuning and the use of high level semantic information from some image captioning system, gives an already quite high prediction performance. For all models, we note that performances were lower for long-term memorability. One interpretation might be that the memorability scores for long-term are based on a smaller number of annotations than for short-term, so they probably capture a smaller part of the intrinsic memorability. However, it may also highlight the difference between short-term and long-term memorability, the latter being more difficult to predict as it is more subjective, while both being still – though not perfectly – correlated. The performances of our models on the 500 most annotated videos are better. This reveals that our dataset might benefit from a larger number of annotations. Last, compared to annotation consistency values, performances remain lower, showing that there is still room for improvement.

5.5. Intra-memorability visualization

To better understand what makes a video frame memorable, we added an attention mechanism to our best model. It will then learn what regions in each frame contribute more to the prediction. For this purpose, a convolutional layer is added in parallel with the last convolutional layer of the feature extractor part. It outputs a 2D attention map which goes through a softmax layer and is multiplied with the last convolution map of the visual pipeline as shown in Fig. 7 (orange branch). An empirical study of the resulting attention maps tends to separate them in two categories. In the first one, when image frames contain roughly one main object and no or rare information apart from this main object (this might be because the background is dark or uniform), it seems that the model focuses, as expected intuitively, on the main object and even, in the case of large enough faces, on details of the faces, as if trying to remember the specific features of faces. Example results for images in the first category can be found in Fig. 8, first row. In the second category that groups all other frames, with several main and secondary objects, cluttered background, *etc.,* it seems on the contrary that the model focuses on all but the main objects/subjects of the images, as if trying to remember little details that will help it differentiate the image from another similar one. Or said differently, the second category shows results that might be interpreted as a second memorization process, once the first one – focusing on the main object – is already achieved. Examples for the second category can be found in the second row of Fig. 8. More results and insights are given in the supplementary material.

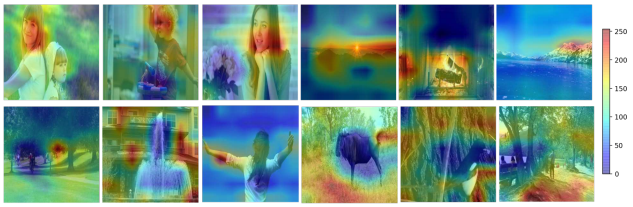


Figure 8: Visualization of the attention mechanism's output. The model focuses either on close enough faces or main objects when the background is dark or uniform (row #1), or it focuses on details outside the main objects (row #2).

6. Conclusions

In this work, we presented a novel memory game based protocol to build VideoMem, a premier large-scale VM dataset. Through an in-depth analysis of the dataset, we highlighted several important factors concerning the understanding of VM: human *vs.* annotation consistency, memorability over time, and memorability *vs.* response time. We then investigated various models for VM prediction. Our proposed model with *spatial* attention mechanism allows to visualize, and thus better understand what type of visual content is more memorable. Future work would be devoted to further study the differences between short-term and long-term memorability, and improve prediction results with a particular focus on temporal aspects of the video, *e.g.* by adding *temporal* attention model and recurrent neural network blocks to the workflow.

References

[1] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. Deep learning for image memorability prediction: the emotional bias. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 491–495, 2016.

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

[3] Bora Celikkale, Aykut Erdem, and Erkut Erdem. Visual attention-driven spatial pooling for image memorability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 976–983, 2013.

[4] Arridhana Ciptadi, Matthew S Goodwin, and James M Rehg. Movement pattern histogram for action recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 695–710. Springer, 2014.

[5] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. Mediaeval 2018: Predicting media memorability task. In *Proceedings of the MediaEval Workshop*, 2018.

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

[7] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 11–14, 2018.

[8] Nelson Cowan. *Attention and memory: An integrated framework*. Oxford University Press, 1998.

[9] Hermann Ebbinghaus. *Memory; a contribution to experimental psychology*. New York city, Teachers college, Columbia university, 1913.

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

[11] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.

[12] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6363–6372, 2018.

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

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

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? *arXiv preprint*, arXiv:1711.09577, 2017.

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

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

[18] Peiguang Jing, Yuting Su, Liqiang Nie, and Huimin Gu. Predicting image memorability through adaptive transfer learning from external sources. *IEEE Transactions on Multimedia*, 19(5):1050–1062, 2017.

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

[20] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2390–2398, 2015.

[21] Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. Relative spatial features for image memorability. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 761–764, 2013.

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

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

[24] Matei Mancas and Olivier Le Meur. Memorability of natural scenes: The role of attention. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 196–200, 2013.

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

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

[27] M Ross Quillan. Semantic memory. Technical report, Bolt Beranek and Newman Inc Cambridge MA, 1966.

[28] Russell Revlin. *Cognition: Theory and Practice*. Palgrave Macmillan, July 2012.

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

[30] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2017.

[31] Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Gwenaëlle Marquant, and Claire-Hélène Demarty. Deep learning for predicting image memorability. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375, 2018.

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

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the International Conference on Computer Vision (CVPR)*, pages 4489–4497, 2015.

[34] Soodabeh Zarezadeh, Mehdi Rezaeian, and Mohammad Taghi Sadeghi. Image memorability prediction using deep features. In *Proceedings of the IEEE International Conference on Electrical Engineering (ICEE)*, pages 2176–2181, 2017.