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

Extended Abstract†

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

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

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

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

ABSTRACT

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

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

KEYWORDS

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

ACM Reference Format:
Romain Cohendet, Karthik Yadati, Ngoc Khanh Duong, and Claire-Helene Demarty. 2018. Annotating, understanding, and predicting long-term video memorability: Extended Abstract. In *Proceedings of ACM Yokohama conference (ICMR 2018).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.475/123_4

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

ICMR 2018, June 2018, Yokohama, Japan

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

1 INTRODUCTION

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

The computational understanding of video memorability's search field follows on from the study of image memorability prediction which has attracted increasing attention from the seminal work of Isola *et al.* released in 2001 [11]. Recently, models of image memorability prediction achieved very good results with the introduction of deep learning to address the challenge of image memorability prediction [1, 13, 20]. This success resulted on the extension of the challenge to videos. Video memorability prediction is however a new field, and there are only two studies, to our knowledge, that addressed this issue [7, 19].

At least two important problems could explained this scarcity of studies.

Firstly, there is no available dataset to train and test models. The fact that there exist no dataset to modeling video memorability is the most serious obstacle which prevent video memorability prediction to take off. This should be the very first goal of pioneers of this new research field: to provide data to community. For images, the pioneers created and made downloadable such a database [11], which enable to made image memorability research flourishing. The second released database, far larger, launched new possibility for deep learning and enable to obtained the aforementioned really good results [13]. We could expect a similar success for video memorability prediction research. We think that the existence of such a database with high quality data could have the importance for video memorability that the ones of [11] and [13] had on image memorability.

This drives us to the second point: the difficulty to build such a dataset. The principle difficulty to define what video memorability and unity should be, and a protocol to measure items according to these definition.

If we consider the studies on image memorability prediction, we could see that, with their database, the cardinal points defined in

the very first study [11] were widely accepted by the community without re-interrogation of what memorability – of what the models – really predicted. First, it is because Isola *et al.* find a proper and objective way to measure memory performance at a large scale; but it remains that how the authors proceeded is just one way to measure memorability.

From what we can imagine that such a dataset for VM will probably have an important influence on future research, one should define very carefully what memorability of video is and what the video as unity means. It requires to define the cardinal points that will guide – and constraint – researchers that will used the dataset.

First, if image in image memorability was obvious, video in video memorability is less obvious. Videos are more complex than images. Contrary to images, videos do not constitute clearly defined units, but have supplementary dimensions – sound and movement – that makes difficult the definition of what a video is. By an extreme example, is it possible for a model to transfer his capacity to predict memorability from movies-like videos to Youtube-like videos?

1.1 Previous work

The work on image memorability prediction, initiated by [11], produced great results [1, 13, 20]. The availability of two large datasets [11, 13] has been critical in this achievement. However, possibly constrained by the difficulty in conducting long crowdsourcing studies, the authors measured memory performance a few minutes after memorization to obtain their memorability annotations. As shown in [4], this could be a problem if we conceive that memorability reflects a lasting memory performance. Indeed, it has long been shown that long-term memories continue to change long after their memorization through an ongoing process called consolidation [16]. In particular, the early work of the psychologist Ebbinghaus showed that the drop in long-term memory performance in recall is particularly strong in the few days immediately following the memorization, after which it asymptotically tends its the almost permanent long-term performance [5]. Because several factors are susceptible to influence the consolidation process (such as the emotional content, the semantics, the attention...), which is not a linear decrease for all the memories, it follows that the order of memorability of images is susceptible to change from some minutes to a day after memorization, as it has been found in [4] with emotional image from of the International Affective Picture System [14]. A protocol to collect memorability annotations for videos would benefit from the capacity to measure what we will later refer to as "long-term memorability".

To our knowledge, the first attempt at predicting video memorability had been proposed by [7]. The authors partially adapted the protocol proposed by [11] to measure image memorability for videos. In contrast to the "memory game" proposed by Isola *et al.* to collect memorability data, their protocol is however much heavier. They used a classical recognition task to measure memory for videos, which consists in two steps: a free viewing task, followed two days later by a recall task. We can infer from the sparse information authors gave that the task duration, for each of the 20 participants, was about 24 hours (by taking an average video duration half-way between 15 and 30 sec), spread over 10 sessions (5 free viewing and 5 recall tasks) of about 2 hours each. Authors used the same proportion of fillers (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, according to them, guarantee that viewers were unaware of targets. If it was mandatory in [11] for which encoding and recall tasks were interlaced, there is here a way to alleviate the task without impacting its quality; indeed, reducing the number of fillers in free viewing task would have very little impact on memorability scores (often authors even use only material interrogated later in learning/free viewing task). Furthermore, as pointed by [19], the long time span of the experiment makes difficult the generalization of this protocol, in particular to build an extensive dataset. Regarding the prediction part, we also agree with these authors to say that the method would be difficult to generalize as it used fMRI measurements for predicting memorability.

Another earlier approach was the one of [19]. The participants performed a crowdsourcing experiment consisted of a free viewing task where they saw a sequences a videos, followed by a recall task for which they had to answer textual question (such as: "Did you see a man juggling?", "Did you see a car on road?"). The major drawback of the study comes from the use of questions instead of a 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, ease of understanding (in particular for the non-anglophone people that works with Amazon Mechanical Turk), of imaging, ease to retrieve a scene by some words more the by others... Especially since the authors took into account the response time of the participants to calculate the memorability scores of the videos. One will note that this potential bias could have also affect the measure of inter-human consistency. Furthermore, the questions are handcrafted, and the choices for types of questions and videos are very limiting (e.g. the question "Did you see a car on road?" implies that in a whole experimental session you can have just one car on a road), and the authors manually ensured that no textual questions nor videos in a sequence were similar in content. This makes difficult the generalization of this protocol to build an extensive dataset.

1.2 Goals of our study

It appears from the review of previous work that video memorability prediction research is in its early stage. We still need to find a proper way to collect quality data in order to build a large dataset to bring video memorability prediction research to a new level. The present study aims to be part of these efforts to reach such a way. As aforementioned, we also want this protocol to measure a "long-term" memorability. This can be achieved with a study that have a long time span, where learning and recall tasks are seperated by a certain amount of time, or by measuring a memory created prior to the experiment. Because crowdsourcing is a priori the only practicable way to experimentally collect great amount of annotations, we focused on the last choice more suitable for crowdsourcing as it requires only one connexion to the measurement application. To assess the quality of the data produced by the protocol, we analyzed the data obtained from different perspectives, in particular in order to shed light on some potential human or task-related factors susceptible to influence memorability as it is measured by the protocol, that are numerous **??**. Thanks to the data collected, we

participate in the VM prediction challenge. The structure of the rest of the article follows these points: part 2 concerns the ground truth collection, part 3 the study of the collected data and the calculation of the memorability scores of the sequences, part 4 predicting memorability, and a conclusive part 5 where we talked about intended future work related to video memorability prediction.

2 CONSTRUCTION OF THE DATASET

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

2.1 Dataset

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

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

Because the semantics is linked to the memorability of images [9], so probably to videos, therefore their memorabilities, we chose this definition susceptible to be extensive to predict other dataset.

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

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

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

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

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

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

neutral sequences are control.

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

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

2.2 Survey design

After they had provided basic demographics, participants filled out a questionnaire composed of 100 items, indicating whether they had seen the films of our list. Each item included a film poster with its name below, followed by the question: "Do you remember watching (fully) this movie?" The experimenter was present in case the participant had doubt about one movie. If the participant answered "yes", he had to answer three additional questions: "Please let us know how confident you are of having seen this movie." (I am not/slightly/50%/considerably/100% confident); "Please let us know when did you last see this movie." (<1 month; <1 year; <5 years; < 10 years; > 10 years); "Please let us know how many times did you watch this movie." (once; 2-4 times; 5-9 times; 10-19 times; >20 times). If he answered "no", he had to answer one more question: "Please let us know how confident you are of NOT having seen this movie". (I am not/slightly/50%/considerably/100% confident). The question relative to confidence aimed to ensure the targets and fillers corresponded to seen and non-seen films respectively. The other questions (i.e. when and how many times the films had been seen) were used during result analysis to evaluate the likely effects of visualization repetition and passage on long-term memorability. The questionnaire required about 20 minutes to complete.

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

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

2.2.1 Participants, facilities and apparatus. 104 participants (22–58 years of age; $mean = 37.1$; $SD = 10.4$; 26% of them female), employees of [removed for double-blind review] participated in he experiment on a volunteer basis. The videos were displayed on a ????? inch monitor (TV SONY Bravia ?????) with a display resolution of 1,920 × 1,080. The participants were seated at a distance of 150 centimeters from the screen (three times the screen height) in a room equipped with subdued lights.

3 STUDY OF THE MEMORABILITY ANNOTATIONS

In this section, we analyze the ground truth data collected thanks to the protocol described in the previous section.

3.1 Memorability scores calculation

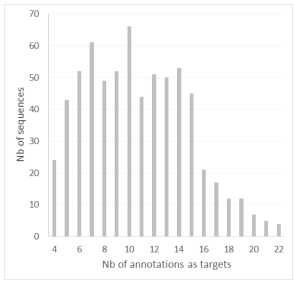
After collecting the data, we measured the average percentage of correct detections for all participants, of 46.71% (SD of 14.65%), and the average false alarm rate, of 4.16% (SD of 5.27%). Then we assigned a memorability score to each sequence, defined as the percentage of correct detections by participants. We retained only the 660 sequences that had been seen at least 4 times as targets, and remove the sequences chosen from films seen by too few participants. On average, each sequence was viewed as target by 10.7 participants, and as filler by 10.5 participants.

The figure 1 provides a sample of sequences sorted according to their memorability (a) and a distribution of the sequences according to their degree of memorability (b).

Figure 1: (a) Sample of the dataset with key frames of sequences sorted from most memorable (left) to less memorable (right). (b) Distribution of the memorability scores.

3.2 Consistency analysis

We implemented the method proposed in [9] to measure the human consistency. We randomly split our 104 participants into two independent halves, and calculate how well sequence memorability scores from the first half of the participants match with sequence memorability scores from the second half of the participants. Averaging over 25 random split half trials, we calculated a Spearman's rank correlation ρ of 0.57 between these two sets of scores.



(a)

Figure 2: (a) Number of sequences for each possible nb of annotations per sequence. (b) Human consistency averaged over 25 random splits (right) obtained using the method proposed by [9] for sequences with at least 5 annotations (with standard error and linear trendline).

We reproduced this calculation to obtain 25 Spearman's correlation coefficients as a function of the mean number of scores per video, presented in figure 2(b). This curve is to be compared with the histogram presented in figure 2(a), which shows that the number of sequences for each number of annotations was unequal. According to the curve, we achieved a mean consistency of .70 around 18 annotations, which is consistent with the previous attempt of [7], were it is the maximum consistency they achieve. It is interesting to compare these results to the one obtained by authors that have collected image memorability annotations [11, 13]. In these studies, the authors reached such a correlation of .70 after 80 annotations per image, which is much later. The experimental protocols and conditions are different between all these studies; in particular, emphasis must be placed on two points. First, in the present studies and [7], we measured a long-term memory performance later than two days after the memorization, that is to say after the biggest losses in memory performance, as was previously stated. By contrast, [11, 13] measured memory performance from dozens of seconds to a few minutes after memorization. Second, videos memorability annotations were collected through in-lab experiments, and images annotations through crowdsourcing experiments. These two confound factors may have contributed to the shortest number of annotations necessary to reach a high human consistency. However, it would be interesting in future work to confirm if an important difference exists between images and videos regarding to the number of annotations necessary to achieve a high human response consistency. Apart from the conclusions we could draw about the universality of the intrinsic memorability of videos compare to images, this would mean that the magnitude of the work to carry out to build an extensive database for video memorability prediction is substantially smaller than one could expect from work on image memorability prediction.

3.3 Neutral and typical sequences

We select two sorts of sequences from the films: neutral and typical ones. For neutral sequences, participant had no element to guess that the sequence came from a particular film; for typical sequence, participants could guess that the sequence came from a particular film. The instructions was made clear, that we wanted that participants answered when they recognized a sequence, not when they guess that they came from a particular film they answered during the previous questionnaire film they had seen. Thus, the response for the neutral sequences is objective, but there is subjectivity in the typical sequence.

A non-parametric Wilcoxon rank sum test showed a significant difference in memorability between the neutral ($\mu = .24$) and typical ($\mu = .53$) sequences: $Z = 10.22, p < .00001$. We expected a result in this sens because neutral sequence contained less contextual elements useful to recognize that a segment comes from a particular film. Because of this compounding factor this results does not mean that participants tended to guess – rather to purely recognize – sequences drawn from films they have seen.

We also observed a difference in human consistency for memorability responses between the neutral ($\mu = .45$) and typical sequences ($\mu = .41$): $Z = 2.75, p < .01$. The human congruency was slightly higher for neutral than for typical sequences. Along with the comments collected from the participants, who have as a majority reported difficulty to know if they guess or just recognized, this suggests that human congruence is higher for the objective measure than for a one with a part of subjectivity. In another context than with colleagues of our company, who highly probably respected the instructions – in particular in crowdsourcing, which as we aforementioned stated that it must be used to built an extensive database – it is probable that this difference would increase. Certainly, this

result has to be confirmed by further studies, but it could be one of the weakness of our protocol to collect extensive data.

We can also note that the false alarm rate was low for neutral sequences ($\mu = .05$) as well as for typical sequences ($\mu = .03$). Specifically, we expect lucky confusions to account for little of correct detections on average for the two sorts of sequences.

3.4 Response time

Figure 3: Mean response time for each memorability degree for targets for which participants answered(error bars correspond to SEM).

The figure 3 shows that the response time to do a correct detection decreases when the memorability of the sequence increase. We also observed a Person's correlation of $-0.36 (p < .0001)$ between the response time on the target and their memorability scores. These result show that the participants tended to answer quicker when the sequences were memorable than non memorable, even though the participants did not receive any instruction to answer quickly. This suggest two things. First, that participants tends to naturally answer rapidly after having recognized the sequence. Second, either that the most memorable sequences are also the sequences the most accessible in memory, or that the most memorable sequences have more early recognizable elements than less memorable sequences. However, the way we select sequences and the linear decrease in response time with memorability increasing showed in figure 3 go in the direction of the first hypothesis.

It is interesting to link it to the response time for answer on targets (i.e. correct detections) vs. on fillers (i.e. false alarms). The global mean response time was $4.88sec$ on targets and $5.90sec$ on fillers. A Wilcoxon rank sum test showed a significant difference ($Z = -5.10, p < .00001$), in the way participants globally answered more rapidly for targets (i.e. correct detections) than for fillers (i.e. false alarm). One explanation is that participants hesitated more for fillers they answered on than to make a correct detection, increasing their response time. This hesitation would reinforce the hypothesis that the most memorable sequences are most easily accessible in memory: if it has just been a question of a difference in the early recognizable elements that had explained that participants answered faster for the most memorable sequences, and not a question of hesitation – i.e. of accessibility in memory –, the response time would have been the same for false alarm than for correct detection.

It also point out the fact that memories of videos are generally not obvious but more blurred. This is an interesting point to connect with the fact that human consistency is lower for typical than for neutral sequences: because of the non obviousness of if they had memories or not, the part of subjectivity for typical sequences could have fostered participants to answer when they hesitated if they were guessing that a sequence came from a film they had seen or remembering it. The higher degree of memorability of typical sequences compare to neutral ones could be partly due to this factor too. Among other, this is also an argument in favour of the most "objective" possible measure to choose to constitute an extensive dataset for video memorability prediction.

To conclude this section, the degree of memorability sounds to be naturally related to response time. Therefore, we could imagine to exploit in the calculation of memorability score. That is what [19] did in their previous attempt to construct their dataset to predict video memorability: they integrated response time in the calculation of the memorability score of their videos as the principal element (even if, in this case, the aforementioned problem of the use of questions could have influenced their results).

3.5 Logistic regression vs. SVM to personalize prediction model

Figure 4: .

4 MEMORABILITY PREDICTION

Until now, we have presented the video collection and the annotation protocol. We then computed memorability scores and analyzed how different users fared in remembering the video sequences from the movies they have already watched. In this section, we move towards building a machine learning model that can learn and then predict the memorability score of a video from its audio-visual features. We pose the problem as a standard regression problem and Figure 5 illustrates different steps in our method. First, we can see that the video sequences are converted into feature vectors and then a regression model is trained that can in-turn predict the memorability scores for new video sequences. In this following sub-sections, we explain our choice of features, models to address the problem in hand.

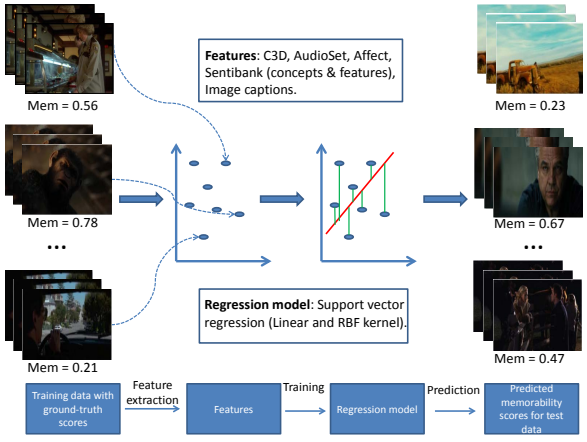


Figure 5: Proposed approach for memorability score prediction.

Before we go into the technical details about the different audio-visual features and regression models we explored, we will talk about how we split our dataset. We split our 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 video sequences, in order to avoid sequences from the same movie being present in the training as well as the evaluation (validation+test) set.

4.1 Feature extraction

We explore a variety of audio-visual features in building a model that can predict memorability scores for new video sequences. We mostly stick to state-of-the-art deep learning features that have been proposed in vision and audio research communities for tasks like video classification [21], event detection [6], image captioning [12], and concept detection [2].

4.1.1 Spatio-temporal visual features (C3D). These features are extracted from the C3D model, a 3-dimensional convolutional network proposed for generic video analysis [21]. 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 [15]. We use the model trained on the Sports-1M dataset [21] that is provided by the authors of the paper. We use the output of the fully connected layer – fc6 of the network with a dimensionality of 4096. Since such dimensionality is high compared to other features, we explore the use of principal component analysis (PCA) for the dimension reduction.

4.1.2 Audio features (AudioSet). Using a recently released large-scale dataset of manually annotated events: AudioSet [6] and the corresponding event detection models, 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.

4.1.3 Emotion related features (SentiBank and Affect). As research in psychology showed that emotion and memory are highly correlated [3], we investigate the use of emotion-related feature in our prediction system. For emotion from visual content, we resort to a visual sentiment concept detector: SentiBank [2]. 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. We sample one frame for every second of the video sequence in our dataset, resulting in 10 frames per video sequence. For each of these 10 frames, we run the SentiBank concept detector and obtain two pieces of information: concepts with probabilities and features. Concepts are adjective–noun pairs and the probability represents how likely this concept is depicted visually in that particular frame. Examples of some concepts in the SentiBank ontology are: young driver, scary face, terrible pain etc. We rank the concepts based on the probability of their occurrence in the frame and take the top-50 concepts. For each of the 50 concepts, we extract 300-dimensional word2vec [17] embeddings, using a model trained on english wikipedia corpus. We then take an average of all the vectors 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 4096-dimensional features 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 two features: 300-dimensional concept vectors and 4096-dimensional feature vectors as features to build a model for memorability prediction.

Emotion will also be embedded in the audio signal and hence we resort to an audio-visual analysis of the video sequence to obtain arousal and valence scores [8]. Arousal is the dimension of emotion that measures the excitement in the video, while valence measures whether the video happy or sad, following a circumplex model of affect [18]. For each frame in the video sequence, we compute the arousal and valence scores using the method proposed in [8]. We concatenate the arousal and valence scores for the first 200 frames in each video sequence resulting in a 400-dimensional feature vector (200 for arousal and 200 for valence) for a video.

4.1.4 Visual semantic features (Image captions). Visual semantics are known to play an important role in memorability [10]. We utilize the state-of-the-art research in image captioning to capture the semantics [12]. We sample one frame for every second of the video sequence in our dataset, resulting in 10 frames per video sequence. 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 word in the caption, we extract 300-dimensional word2vec [17] embeddings and taking an average of 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 feature vector for each video. We use these 300-dimensional vectors for building our model.

4.2 Modelling and evaluation

We used the features discussed in Section 4.1 to train an SVR model for the memorability score prediction. We studied the use of two kernels: linear and RBF for their popularity and wide applicability. We use the grid search strategy to obtain the best hyper-parameters for SVR: C and γ.

Since we are posing the memorability prediction problem as a regression problem, we use the standard regression metric: Mean squared error (*MSE*) for evaluation. Additionally, we also use the spearmann correlation (*SpCorr*) to measure the rank correlation between the predicted memorability scores and the ground-truth. This metric gives us an indication of the correlation between the predicted memorability scores and the ground-truth memorability scores.

4.3 Memorability prediction results

In this section, we will discuss how the models trained on different features perform in predicting memorability scores of new video sequences. In addition to the variety of features explained in Section 4.1, we also explore a combination of all the features by concatenating all the feature vectors into a single feature vector. While concatenating the feature vectors, we use the lower dimensional features for C3D and SentiBank features, obtained after applying PCA to the original set of features. Each of the video sequences in our dataset has multiple annotations and we pick sequences with at least 4 annotations for training the models. In this way, we eliminate movies that were not seen by a majority of the users who participated in our experiment. Additionally, we investigate the

effect of number of annotations, for video sequences in the training set, on the performance of the model.

Since we are using an SVR with linear and rbf kernels, we train the models on the training set (70 movies) and validate the model on the validation set (15 movies), before reporting the final scores on the test set (15 movies). We pick the model parameters: kernel (linear or rbf), C and γ that give the best performance in terms of *MSE*. We retrain the model with these parameters and evaluate on the final test set. In the experiments where we use a dimension reduction method (PCA): C3D and SentiBank features, we retain 95% of variance in the data while reducing the feature dimensions.

Based on our experiments on the validation set, we observe that the image caption feature performs the best while using training video sequences that have atleast 4 annotations. Based on this result, we train different SVR models with varying number of annotations on the training set using image caption features. For example, we train a model, using image caption features, on training sequences with at least 4 annotations and use this model to predict the memorability score for video sequences in the validation set. We repeat this process for different number of annotations: from 4 to 15 annotations increasing the number of annotations by 1 in each iteration. Please note that the validation set in each of the repetitions is fixed and only the training set changes. We observed that the *MSE* does not change significantly in each of the steps and hence we provide a demonstration of how *SpCorr* varies with increasing number of annotations in the training set. Figure 6 demonstrates the result of this experiment, where we observe that beyond 8 annotations the value of *SpCorr* keeps going down. In the wake of this observation, we investigate the performance of all the features when we train the regression model with video sequences that have at least 8 annotations.

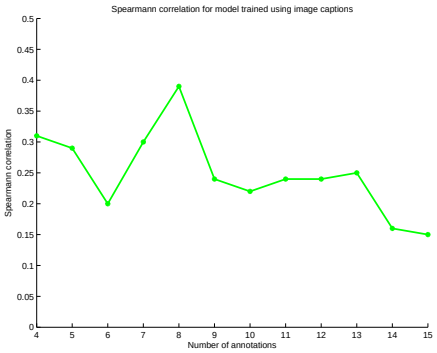


Figure 6: Spearmann correlation on validation set with models trained using image caption features for varying number of annotations.

Now, reverting to the main experiment to measure how each feature performs in predicting the memorability scores for video sequences in the test set. We compute the average number of annotations per sequence in training and test set for both the cases in order to check if there is a mismatch. On an average, each sequence in the test set has around 10 annotations and the test set remains the same across different experiments. Similarly, each sequence in the training set has 9 annotations on an average in the case where we train the models with sequences having at least 4 annotations and 11 annotations in the other case. Looking at these numbers, we can see that the average number of annotations per sequence is quite balanced across the training and test sets. Table 1 reports the memorability prediction results for different features using models trained on sequences with at least 4 (columns 3-5) and 10 (columns 6-8) annotations. Observing the table, we can clearly see that the image captions perform the best, among all the features, in terms of *MSE* and *SpCorr*. These features capture the semantics in the video and are helping in better predicting the memorability scores on new data. Though similar *MSE* scores are obtained for other features like AudioSet, the memorability scores, predicted by the model trained on image caption features, are better correlated with the ground-truth (*SpCorr* = 0.31 and 0.39). A combination of all the features is the second best performing feature combination in terms of *MSE*, but falls behind other individual features (AudioSet and C3D) in terms of *SpCorr*.

One of our initial hypotheses was that emotion would play a important role in memorability of a video sequence, supported by literature from psychology. We used different set of features to encode the emotion in a video: Affect [8] and SentiBank [2]. Observing the scores in Table 1, we can say that the SentiBank features (fourth row from bottom) and affect features (sixth row from bottom) perform reasonably well, when the video sequences in the training set have at least 8 annotations. But the models trained on image caption features out-perform those trained on emotion features. This could be because of the choice of features that encode emotion or the performance of the emotion models not being good enough for such an application. From the analysis on this dataset and the features we used for encoding emotion, we *cannot* conclude that emotion plays an important role in memorability.

Comparing the two sets of results in Table 1, we observe that the model trained on sequences with at least 8 annotations performs better than the model trained on sequences with at least 4 annotations. The only exception being the C3D (PCA) feature, where the model trained on sequences with at least 4 annotations performs marginally better than the model trained on sequences with at least 8 annotations.

5 CONCLUSIONS AND FUTURE WORKS

We presented a machine learning model that can predict memorability score for a given video sequence. After exploring a variety of state-of-the-art deep learning audio-visual features, we can say that a model trained with features from image captions (semantic features) can provide predictions that are most correlated with ground-truth memorability scores. We also investigated the number of annotations required to train a model and found a sweet spot where the performance is optimal. An immediate direction for future research would be to explore other approaches to encode emotion in videos and investigate whether we can improve the prediction performance using these approaches.

Feature	Dimension	SVR Model (4 annotations)			SVR Model (8 annotations)		
		Parameters	*MSE*	*SpCorr*	Parameters	*MSE*	*SpCorr*
C3D	4096	Linear, C=1.0	0.08	0.26	Linear, C=1.0	0.07	0.34
C3D (PCA)	225	Linear, C=1.0	0.08	0.21	RBF, C=1.0, γ=0.01	0.08	0.17
AudioSet	128	RBF, C=1.0, γ=0.01	0.06	0.22	RBF, C=1.0, γ=0.01	0.05	0.24
Affect	400	RBF, C=1.0, γ=0.1	0.07	0.17	RBF, C=1.0, γ=0.1	0.07	0.23
SentiBank concepts	300	RBF, C=1.0, γ=10	0.08	0.13	RBF, C=1.0, γ=0.1	0.08	0.17
SentiBank features	4096	RBF, C=1.0, γ=0.1	0.07	0.21	RBF, C=1.0, γ=0.01	0.07	0.26
SentiBank features (PCA)	225	RBF, C=1.0, γ=0.01	0.07	0.21	RBF, C=1.0, γ=0.01	0.07	0.23
Captions	300	RBF, C=1.0, γ=0.01	0.06	0.31	RBF, C=1.0, γ=0.01	0.05	0.38
Combination	1578	RBF, C=1.0, γ=0.1	0.06	0.23	RBF, C=1.0, γ=0.1	0.05	0.27

Table 1: Prediction results on the test data for different features with models trained on sequences that have at least 4 (columns 3-5) or 10 (columns 6-8) annotations.

APPENDIX

ACKNOWLEDGMENT

REFERENCES

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

[2] 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 *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 459–460. https://doi.org/10.1145/2502081.2502268

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

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

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

[6] 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 ICASSP 2017*. New Orleans, LA.

[7] 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.

[8] A. Hanjalic and Li-Qun Xu. 2005. Affective video content representation and modeling. *IEEE Transactions on Multimedia* 7, 1 (Feb 2005), 143–154. https://doi.org/10.1109/TMM.2004.840618

[9] 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.

[10] 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. https://doi.org/10.1109/TPAMI.2013.200

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

[12] 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. https://doi.org/10.1109/TPAMI.2016.2598339

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

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

[15] S. Liu and W. Deng. 2015. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rdv IAPR Asian Conference on Pattern Recognition (ACPR)*. 730–734. https://doi.org/10.1109/ACPR.2015.7486599

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

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

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

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

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

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, Washington, DC, USA, 4489–4497. https://doi.org/10.1109/ICCV.2015.510