To Decay or not to Decay: Modeling Video Memorability Over Time

Anelise Newman*, **Camilo Fosco***, **Vincent Casser***, **Barry McNamara**, **Aude Oliva**

Massachusetts Institute of Technology
{apnewman, camilolu, barryam3, oliva}@mit.edu, vcasser@csail.mit.edu

Abstract

Videos present a unique challenge for understanding visual memorability due to motion and evolution across frames. To enable exploration of this field, we introduce Memento10k, the largest video memorability dataset to date. Based on our analysis of this data, we propose a new formulation for how video memorability decays that differs significantly from image memorability. Finally, we design a prediction network, Temporal MemNet, that achieves state-of-the-art performance by harnessing both visual and temporal content. The model accounts for 90% of human consistency, leaving room for improvement in terms of understanding and imitating how humans process motion in videos.

1 Introduction

Prior work has shown that humans are remarkably consistent in which images they remember and forget [8], with some models able to predict these differences [6, 9]. The more lifelike analog of videos requires us to study how movement and temporal dynamics affect memorability. We introduce Memento10k, a video memorability dataset with $10,000$ short video clips of natural scenes, with high diversity in content and temporal dynamics. Whereas the textbook model for image memorability assumes that memory decays log-linearly at a constant decay rate [1, 9], our data indicates that video memory falls off linearly and that the decay rate depends on the stimulus, with fast decay being characteristic of low-memorability videos. Based on these results, we propose a new mathematical model of memorability decay. We leverage this formula to design Temporal MemNet, a video memorability network that achieves state-of-the-art performance on both Memento10k and VideoMem [4, 3] by combining predictions on visual appearance and optical flow.

2 Memento10k Dataset

Game Setup: Inspired by the techniques from [8, 9], we created *Memento: The Video Memory Game*. Crowdworkers sourced using Amazon's Mechanical Turk (AMT) watched a continuous stream of three-second video clips and pressed the space bar when they saw a repeated video. "Target" repeats, which provide our memorability data, occur at lags of 9-200 videos (up to a 9 minute delay). A "hit" refers to when a crowdworker correctly identified a repeat and a video's "hit rate" is the fraction of times its repeats were correctly detected. We discard data from workers who miss at least 20% of short-term "vigilance" repeats and who false alarm more than 50% of the time.

Dataset Contents: Memento10k is composed of 10,016 videos. All videos are 3-second clips scraped from the Internet[1] and then filtered to include only unprocessed clips shot in a natural context, free of jump cuts, watermarks, or other artificial elements. We provide on average 90+ valid memorability

[1]The Memento videos have partial overlap with the Moments in Time [10] dataset

1st Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM), NeurIPS 2019.

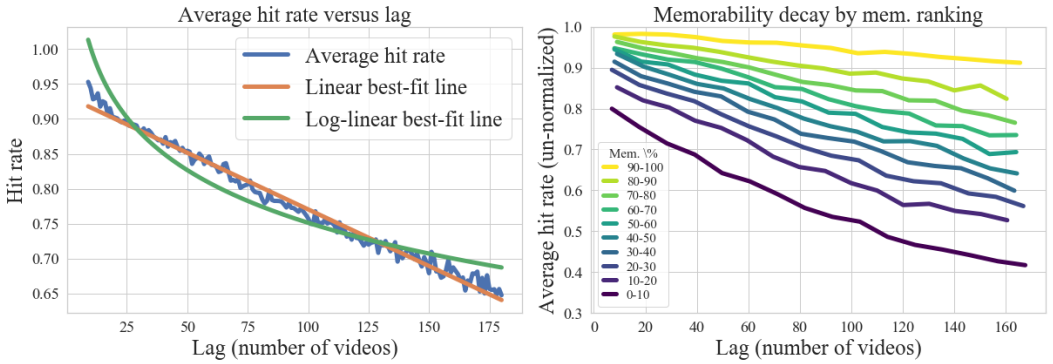

Figure 1: Our data suggests a memory model where each video decays linearly according to an individual decay rate $\alpha^{(v)}$. **Left:** A linear trend is a better approximation for our raw data ($r = -0.993$) than a log-linear trend ($r = -0.963$). **Right:** Differing decay rates among stimuli help explain differences in memorability. We group videos into deciles based on their normalized memorability scores and plot the average hit rate as a function of lag. The slope of these lines varies, with low-memorability videos having a steeper decay rate than high-memorability ones.

annotations per video in addition to category labels for 481 action classes and 251 place classes. Memento10k's large number of annotations per video make it the largest video memorability dataset to date. Its "in-the-wild" content makes it ideal for studying how people process changing visual stimuli in an everyday context. Furthermore, the Memento10k videos are more dynamic than prior video memorability datasets (mean magnitude of optical flow is 15.476 for Memento10k compared to 7.296 for VideoMem), enabling analysis of how motion and dynamics affect memorability.

3 Memory Decay and Calculating Memorability Scores

Since our annotations occur at different lags, we must account for memory decay when calculating memorability scores. We find that video memorability differs from textbook image memorability.

Insight 1: Video memory decays linearly with lag. It is widely accepted that images decay as the log of the lag [1, 11, 9, 7]. By contrast, we find that videos decay linearly for the lags that we studied. Figure 1 (left) shows that a linear trend best fits our raw annotations.

Insight 2: Memorable videos decay more slowly than non-memorable ones. Prior work indicated that all stimuli decay at the same rate [7, 9]; however, we find that some videos decay faster than others, and that this is a main driver of differences in memorability (Figure 1 (right)). Thus, instead of assuming one constant rate of decay, α, for all stimuli, we assume that each video decays at its own individual rate, $\alpha^{(v)}$, which reduces our least squares error by 1.6%.

Calculating memorability scores. We define a video's memorability score, $m_T^{(v)}$, as its average hit rate at a reference lag T (here we set $T = 80$ videos). We can express the memorability of video v as $m_T^{(v)} = \alpha^{(v)}T + c^{(v)}$, where $c^{(v)}$ is the memorability of the video at lag 0. Given $m_T^{(v)}$ and $\alpha^{(v)}$, we can calculate the memorability at a different lag t using the following formula: $m_t^{(v)} = m_T^{(v)} + \alpha^{(v)}(t - T)$. We then use the iterative optimization technique introduced in [9] to alternately solve for the values of $m_T^{(v)}$ and $\alpha^{(v)}$ that best fit our annotations.

4 What makes a video memorable?

Using the technique from section 3, we calculate memorability scores and decay rates for the Memento10k videos. We find that both motion and visual content influence memorability. High-memorability/slow-decay videos are dynamic with interesting or surprising motion patterns, while many low-memorability/fast-decay videos are nearly static. Visually, we observe many similarities with image memorability; the most memorable videos contain people and animate objects and tend to be brighter with higher contrast, while low memorability videos are often dark, cluttered,

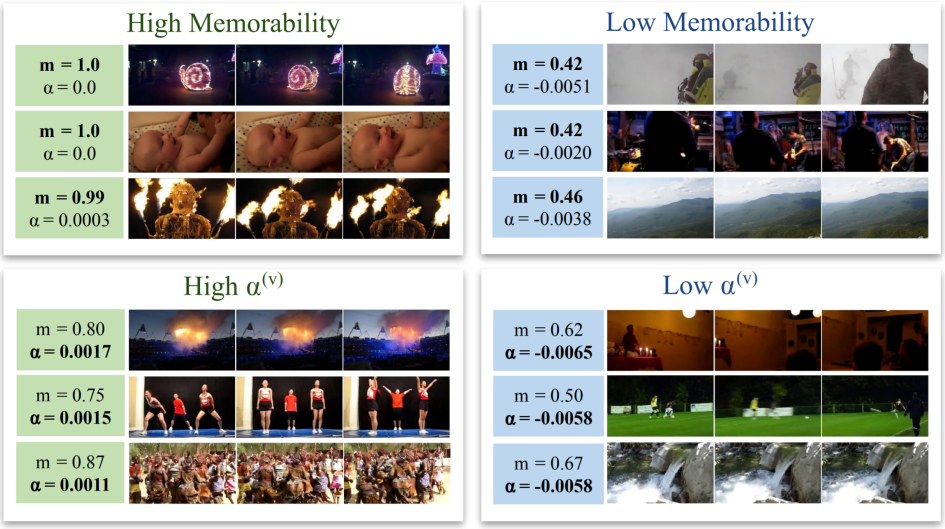

Figure 2: Examples of videos with high and low values of $m_T^{(v)}$ and $\alpha^{(v)}$. Some of the videos in our dataset actually have *positive* values of alpha; this is possibly owing to noise in the data.

and low contrast (see Figure 2). We also examine the relationship between action/scene classes and memorability. The most memorable action classes ("cooking", "cleaning") contain people manipulating objects, thus including both human faces and interesting motion patterns, while the least memorable action classes ("skiing", "playing sports") feature zoomed out landscapes. Similarly, the memorable place categories ("kitchen", "inn") are indoor, human-centric spaces with close-up content, while the least memorable places ("snowfield", "ski slope") are outdoor landscapes.

5 Temporal MemNet: State-of-the-Art in Predicting Video Memorability

Architecture. Temporal MemNet has three streams that extract both visual and motion information. The Visual Appearance (VA) stream passes individual frames through a 2D DenseNet and averages the outputs. The two Temporal Dynamics streams consist of I3D architectures [2] that take as input either the raw video (TD-VIDEO) or the optical flow (TD-FLOW). Outputs of the three streams are combined to make a final prediction. Temporal MemNet predicts $m_T^{(v)}$ and $\alpha^{(v)}$ for each video.

Results. We evaluate and train our model on both Memento10k and on the VideoMem dataset [4]. On VideoMem, we compare to the top available models for video memorability prediction, while on Memento10k, we use the ResNet-101 model used in [4] as a baseline. We observe that all previous models operate only on individual images and thus do not receive information about motion. To explore the importance of temporal dynamics in video memorability, we also perform an ablation study where we evaluate each of Temporal MemNet's three streams individually. The results of these experiments are shown in Table 1. Not only does our model outperform all previous architectures, but also we demonstrate that all three streams (VA, TD-VIDEO, TD-FLOW) contribute to performance. This underscores that human-level video memorability models must be capable of analysing motion. We note, however, that Temporal MemNet still falls short of human consistency (rank correlation of 0.659 versus 0.730), whereas image memorability models have all but closed the gap with human performance [9]. This suggests that understanding motion remains a challenge for current architectures.

6 Conclusion

We expand the field of visual memorability in terms of data variety, understanding of memory, and modeling tools. We find that a video decay model where stimuli decay linearly and at different rates accurately represents human memory data, in contrast to previous work on images. Using our network

VideoMem results

	Approach	Type	RC
	Human Consistency	-	0.616
Ours	**Temporal MemNet**	**Multimodal**	**0.555**
Ours	VA + TD-FLOW only	Multimodal	0.542
Ours	VA only	Image	0.527
Ours	TD-VIDEO only	Video	0.492
Ours	TD-FLOW only	Optical Flow	0.425
Others	HMP+Caption+VA* [5]	Image	0.508
Others	Semantic Embedding* [4]	Image	0.503
Others	ResNet101 [4]	Image	0.498

Memento10k results

Approach	RC
Human Consistency	0.730
Temporal MemNet	**0.659**
VA + TD-FLOW only	0.641
VA only	0.601
TD-VIDEO only	0.618
TD-FLOW only	0.580
ResNet101 [4]	0.578

Table 1: Model comparisons on VideoMem and Memento10k using rank correlation as our metric. Temporal MemNet outperforms the state-of-the-art on VideoMem and Memento10k, but falls short of human consistency.

Temporal MemNet, we demonstrate that modeling motion is critical to achieve top performance. Future work will focus on better understanding and modeling how motion impacts memorability. In the future, predicting video memorability can contribute to applications such as education, video compression, virtual assistants, and diagnosing dementia.

References

[1] Timothy F Brady et al. "Visual Long-Term Memory Has a Massive Storage Capacity for Object Details". In: *Proceedings of the National Academy of Sciences* 105.38 (2008), pp. 14325–14329.

[2] João Carreira and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 4724–4733. DOI: 10.1109/CVPR.2017.502. URL: https://doi.org/10.1109/CVPR.2017.502.

[3] Romain Cohendet et al. "Annotating, Understanding, and Predicting Long-Term Video Memorability". In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM. 2018, pp. 178–186.

[4] Romain Cohendet et al. "VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability". In: *CoRR* abs/1812.01973 (2018). arXiv: 1812.01973. URL: http://arxiv.org/abs/1812.01973.

[5] Rohit Gupta and Kush Motwani. "Linear Models for Video Memorability Prediction Using Visual and Semantic Features". In: *Working Notes Proceedings of the MediaEval Workshop, EURECOM*. 2018.

[6] Phillip Isola et al. "Understanding the Intrinsic Memorability of Images". In: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*. 2011, pp. 2429–2437.

[7] Phillip Isola et al. "What Makes a Photograph Memorable?" In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.7 (2014), pp. 1469–1482.

[8] Phillip Isola et al. "What Makes an Image Memorable?" In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. 2011, pp. 145–152. DOI: 10.1109/CVPR.2011.5995721. URL: https://doi.org/10.1109/CVPR.2011.5995721.

[9] Aditya Khosla et al. "Understanding and Predicting Image Memorability at a Large Scale". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2390–2398.

[10] Mathew Monfort et al. "Moments in time dataset: one million videos for event understanding". In: *IEEE transactions on pattern analysis and machine intelligence* (2019).

[11] Melissa Le-Hoa Võ, Zoya Bylinskii, and Aude Oliva. "Image Memorability In The Eye Of The Beholder: Tracking The Decay Of Visual Scene Representations". In: *bioRxiv* (2017), p. 141044.