Computational Understanding of Image Memorability

by

Zoya Bylinskii

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Author . . . .
Department of Electrical Engineering and Computer Science
January 30, 2015

Certified by. . . . .
Antonio Torralba
Associate Professor
Thesis Supervisor

Certified by. . . . .
Aude Oliva
Principal Research Scientist
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

Dedicated to those who have taught me to be a scientist (you know who you are).

Computational Understanding of Image Memorability

by

Zoya Bylinskii

Submitted to the Department of Electrical Engineering and Computer Science
on January 30, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Previous studies have identified that images carry the attribute of memorability, a predictive value of whether a novel image will be later remembered or forgotten. In this thesis we investigate the interplay between intrinsic and extrinsic factors that affect image memorability.

First, we find that intrinsic differences in memorability exist at a finer-grained scale than previously documented. Moreover, we demonstrate high consistency across participant populations and experiments. We show how these findings generalize to an applied visual modality - information visualizations. We separately find that intrinsic differences are already present shortly after encoding and remain apparent over time. Second, we consider two extrinsic factors: image context and observer behavior.

We measure the effects of image context (the set of images from which the experimental sequence is sampled) on memorability. Building on prior findings that images that are distinct with respect to their context are better remembered, we propose an information-theoretic model of image distinctiveness. Our model can predict how changes in context change the memorability of natural images using automatically-computed image features. Our results are presented on a large dataset of indoor and outdoor scene categories.

We also measure the effects of observer behavior on memorability, on a trial-by-trial basis. Specifically, our proposed computational model can use an observer's eye movements on an image to predict whether or not the image will be later remembered. Apart from eye movements, we also show how 2 additional physiological measurements - pupil dilations and blink rates - can be predictive of image memorability, without the need for overt responses. Together, by considering both intrinsic and extrinsic effects on memorability, we arrive at a more complete model of image memorability than previously available.

Thesis Supervisor: Antonio Torralba
Title: Associate Professor

Thesis Supervisor: Aude Oliva
Title: Principal Research Scientist

Acknowledgments

Many people have earned my heartfelt thanks, through my continuing journey in science. I offer my thanks to:

- Aude Oliva, who has been an incredible source of support, mental and physical energy, invaluable brainstorming, and amazing foresightedness - a superwoman.
- Antonio Torralba, who has never ceased to dazzle me with his insightfulness, inventiveness, and most importantly - optimism.
- Sven Dickinson, my first academic advisor, who has had a tremendous impact on where I am and where I'm going - a perpetual source of fuel.
- John Tsotsos, Stan Sclaroff, and Hanspeter Pfister, who have shown me that a good scientist is first and foremost a good person, and then a good researcher.
- My amazing colleagues and collaborators, most notably Melissa Le-Hoa Võ and Michelle Borkin, who never stop smiling.
- Phillip Isola, who has been a mentor, a friend, and a truly inspiring researcher with a never-ending supply of ideas.
- Adrian Dalca, who has been a vital source of academic and personal support, and incredible kindness.
- My family and friends in Toronto, Vancouver, and Boston, who keep me ticking.

And a very special thank you to the people, who by nature and nurture, have instilled in me the desire to know and understand as much as possible, Marina and Vasili Gavrilov. And to my biggest source of inspiration, Alexei Bylinskii.

Funding for this work has been provided by the National Sciences and Engineering Research Council of Canada (NSERC), the National Science Foundation (NSF), Xerox, Google, and the MIT Big Data Initiative at CSAIL.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis lies at the intersection of the computational and psychological sciences, containing many novel findings about image memorability, from human experiments conducted both online and in the lab, and from computational models. **Image memorability**, pioneered by the Oliva Lab [30, 31, 29, 38, 3, 9, 37, 36, 16, 61], is an objective and quantifiable measure of an image that is independent of the observer and can be computationally predicted. Thus, image memorability is driven by intrinsic features in an image that dictate how memorable or forgettable an image will be across a population. However, extrinsic effects like the experimental context in which an image appears, or how attentive an individual is while looking at an image, can modulate the memorability of an image to finally determine whether an image will be remembered or forgotten on a particular trial. These extrinsic effects are quantified and modeled in this thesis.

Here we build on the initial work in image memorability conducted by [30, 31, 29, 3] who first showed that memorability is an intrinsic property of images, independent of observer. The consistency across people as to which images are remembered and which forgotten allows memorability to be a property that can be computationally estimated from images, and thus opens up the doors to computational applications, from educational material to promotional material, visual design, and human computer interfaces. In order to make these applications possible, a good understanding of what drives and impacts image memorability is thus crucial.

The work that is described in this thesis is motivated by a number of questions that have been prompted by previous studies of memorability, including:

1. How generalizable are the findings about the consistency of human visual memory?
2. How do extrinsic effects such as context and observer differences affect image memorability?
3. How quickly do differences in memorability become apparent, and how is memorability modulated by time?
4. Can physiological measurements be used to predict memorability, without the need for explicit (overt) responses from human participants?

This thesis is an attempt to answer these questions. In Chapter 2, we show that human consistency at remembering and forgetting images holds at a within-category level - a finer grained level than previously found. We show that intrinsic memorability generalizes also to information visualizations - an entirely different form of visual imagery (Chapter 6).

In Chapters 3 and 4, we show how the extrinsic effects of context and observer differences, correspondingly, modulate the intrinsic memorability of images. We present an information-theoretic framework to quantify context differences and image distinctiveness using state-of-the-art computer vision features, and we show correlations with image memorability scores. We discuss where context has the greatest effect on memorability (Chap. 3), and show how the images that are most distinct relative to their image context are also the most memorable. In Chapter 6 we also hint at distinctiveness being a factor influencing the memorability of information visualizations.

Moreover, we demonstrate how physiological measurements, including eye movements (Chap. 4), pupils (Chap. 5), and blinks (Chap. 5) can be predictive of memorability. Specifically, in Chapter 4 we present a computational model that, given an individual's eye movements on an image, can predict with reasonable accuracy whether the individual will remember the image at a later point in time. In other words, without explicit responses from a user, it is already possible to determine if an image has been successfully encoded into memory using eye movements as an indicator of how much attention was paid to the image.

In Chapter 5, we show that pupils dilate more, and blinks decrease, when retrieving less memorable images than when retrieving more memorable images (due, likely, to differences in cognitive effort). We demonstrate how these effects change over time by varying the lags at which images and their repeats occur. Additionally, we show that differences in memorability are already apparent at the shortest lag (only 20 seconds after image presentation), and become even more pronounced at later lags (Chap. 5).

Some of the computational applications made possible by this research are outlined in Chapters 6 and 7. The robustness and predictability of memorability, as well as the ability to use automatic, physiological measurements (such as eye movements, pupils, and blinks) to measure a higher-level cognitive process such as memory puts power at the hands of the application designer.

Chapter 2

Understanding Intrinsic Effects on Memorability

Does human consistency in image memorability generalize to finer-grained scene categories? Do people remember and forget the same images in different categories?[1]

2.1 Related work

Recent work in image memorability [30, 31, 29, 3] has reported high consistency rates among participants in terms of which images are remembered and which forgotten, indicating that memorability is a property that is intrinsic to the image, despite individual differences between observers. The high consistency was first demonstrated for a database of images from hundreds of scene categories [31], and later shown to extend to narrower classes of images - faces [3]. To show that this consistency is not a special property of face stimuli but holds more generally, here we replicate this result across 21 separate natural scene experiments, each consisting of hundreds of instances of a single scene category (both indoor and outdoor). This is the first image memorability study to consider fine-grained scene categories. We further extend these findings and show they continue to hold for an entirely different visual domain - information visualizations - in Chapter 6.

[1] This chapter is closely related to publication [16]

2.2 FIGRIM (Fine-Grained Image Memorability) dataset

For our studies, we constructed the *FIne-GRained Image Memorability (**FIGRIM**) dataset*. We created this novel dataset by sampling high-resolution (at least $700 \times 700px$) images from 21 different indoor and outdoor scene categories from the SUN Database [67]. We chose all SUN scene categories with at least 300 images of the required dimensions. Image duplicates and near-duplicates were manually removed[2]. The images were downsampled (to avoid introducing resolution discrepancies), and cropped to $700 \times 700px$[3]. From each scene category, 25% of the images were randomly chosen to be *targets* and the rest of the images became *fillers* (table 2.1 lists the number of targets and fillers per scene category). The targets are the images for which we obtained memorability scores. Sample dataset images are provided in Fig. 2-2.

We are publicly releasing the full FIGRIM dataset[4] with a range of popular image features (including Gist and convolutional neural net features) precomputed for all $9K$ images of the dataset, as well as memorability scores for each of the 1754 target images. For the target images, we provide separate memorability scores for the image presented in the context of its own scene category (discussed in Chap. 2) and different scene categories (discussed in Chap. 3). Additionally, for the collection of 630 target images used for our eyetracking experiments (discussed in Chap. 4), we are providing the eyetracking data and responses of a total of 42 participants (16.2 ± 1.6 participants per image).

2.3 Crowdsourcing (within-category) experiment AMT 1

We ran Amazon Mechanical Turk (AMT) studies following the protocol of Isola et al. [31] to collect **memorability scores** (i.e. performance on a recognition memory task) for each of the scene categories, separately. We set up memory games on AMT

[2] We calculated the Gist descriptor [47] of each image, displayed its 5 nearest neighbors, and removed identical copies and near-duplicates. Some remaining duplicates were removed after post-processing the experimental data.

[3] Images were later resized to $512 \times 512px$ to fit comfortably in browser windows for the online AMT experiments (Chap. 2-3), and to $1000 \times 1000px$ for the eyetracking experiments (Chap. 4).

[4] Available at `http://figrim.mit.edu`

category	targets	fillers	datapts per target	$\overline{\text{HR}}$ (%)	$\overline{\text{FAR}}$ (%)	HR cons. (ρ)	FAR cons. (ρ)
amusement park	68	296	64.2 (SD: 15.5)	10.2 (SD: 9.7)	0.85 (SD: 0.3)	0.80 (SD: 0.3)	84.8 (SD: 3.2)
playground	74	330	63.3 (SD: 14.4)	14.7 (SD: 12.7)	0.78 (SD: 0.4)	0.84 (SD: 0.3)	86.4 (SD: 2.7)
bridge	60	260	61.2 (SD: 13.2)	13.2 (SD: 12.0)	0.77 (SD: 0.4)	0.84 (SD: 0.2)	90.2 (SD: 4.4)
pasture	60	264	59.2 (SD: 17.5)	11.5 (SD: 9.5)	0.86 (SD: 0.3)	0.83 (SD: 0.4)	86.2 (SD: 3.7)
bedroom	157	652	58.9 (SD: 14.7)	13.5 (SD: 10.9)	0.77 (SD: 0.2)	0.81 (SD: 0.2)	84.5 (SD: 3.9)
house	101	426	58.0 (SD: 13.3)	14.4 (SD: 10.3)	0.73 (SD: 0.3)	0.80 (SD: 0.3)	82.7 (SD: 3.7)
dining room	97	410	57.8 (SD: 13.6)	14.1 (SD: 10.8)	0.77 (SD: 0.4)	0.79 (SD: 0.3)	83.8 (SD: 2.8)
conference room	68	348	57.1 (SD: 13.7)	12.5 (SD: 8.8)	0.77 (SD: 0.4)	0.80 (SD: 0.3)	85.2 (SD: 3.3)
bathroom	94	398	57.1 (SD: 12.8)	16.3 (SD: 13.9)	0.73 (SD: 0.4)	0.82 (SD: 0.3)	86.6 (SD: 3.4)
living room	138	573	56.9 (SD: 14.1)	14.4 (SD: 9.6)	0.77 (SD: 0.3)	0.73 (SD: 0.3)	81.2 (SD: 2.7)
castle	83	389	56.4 (SD: 17.2)	12.8 (SD: 8.9)	0.87 (SD: 0.2)	0.77 (SD: 0.4)	91.5 (SD: 3.3)
kitchen	120	509	56.2 (SD: 14.0)	16.8 (SD: 10.7)	0.74 (SD: 0.3)	0.80 (SD: 0.2)	80.5 (SD: 3.5)
airport terminal	75	323	55.6 (SD: 13.6)	14.9 (SD: 10.8)	0.76 (SD: 0.3)	0.86 (SD: 0.2)	95.9 (SD: 3.7)
badlands	59	257	52.9 (SD: 20.3)	15.6 (SD: 15.1)	0.82 (SD: 0.3)	0.90 (SD: 0.2)	80.1 (SD: 7.0)
golf course	88	375	52.9 (SD: 17.6)	15.2 (SD: 9.9)	0.84 (SD: 0.3)	0.77 (SD: 0.2)	80.2 (SD: 3.9)
skyscraper	62	271	52.8 (SD: 17.0)	13.5 (SD: 10.6)	0.85 (SD: 0.3)	0.76 (SD: 0.3)	84.4 (SD: 4.3)
tower	86	376	52.7 (SD: 14.3)	18.9 (SD: 13.0)	0.75 (SD: 0.4)	0.83 (SD: 0.3)	82.2 (SD: 3.0)
lighthouse	56	247	52.1 (SD: 15.2)	15.2 (SD: 12.4)	0.78 (SD: 0.4)	0.88 (SD: 0.2)	90.3 (SD: 4.3)
mountain	69	302	50.2 (SD: 21.7)	14.9 (SD: 11.7)	0.87 (SD: 0.2)	0.83 (SD: 0.2)	79.3 (SD: 2.9)
highway	71	348	50.0 (SD: 12.9)	15.0 (SD: 10.4)	0.69 (SD: 0.5)	0.85 (SD: 0.3)	85.9 (SD: 4.6)
cockpit	68	320	49.5 (SD: 17.2)	18.2 (SD: 14.7)	0.70 (SD: 0.5)	0.88 (SD: 0.2)	80.6 (SD: 3.5)

Table 2.1: *FIGRIM* dataset statistics for AMT 1 (within-category), with a total of 1754 target and 7674 filler images. The $\overline{\text{HR}}$ and $\overline{\text{FAR}}$ scores are computed over the targets, for which we have an average of 85 experimental datapoints per image. The average $\overline{\text{HR}}$ across all the scene categories is 56.0% (SD : 4.2%), and the average $\overline{\text{FAR}}$ is 14.6% (SD : 2.0%).

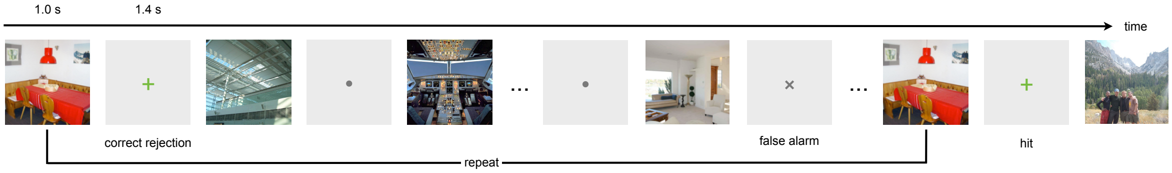


Figure 2-1: An example AMT experimental sequence. During image presentation, the participant presses a key if the image has already appeared in the sequence, and receives feedback at the end of the image presentation. A false alarm occurs when on first presentation, the participant indicates that the image has repeated. No key press during first presentation is recorded as a correct rejection. A hit occurs when a repeated image is correctly remembered, and otherwise, the response is recorded as a miss.

where sequences of 120 images (a mix of target and filler images sampled from a *single* scene category) were presented for 1 second each, with a distance of 91-109 images between an image and its repeat, and consecutive images separated by a fixation cross lasting 1.4 seconds. Images and repeats occurred on average 4.5 minutes apart, thus allowing us to capture memory processes well beyond short-term and working memory. Some filler images repeated at much shorter intervals of 1-7 images and were used as vigilance tests to recognize when a participant wasn't paying attention to the game[5]. Participants were instructed to press a key when they detected an image repeat, at which point they received feedback. No image repeated more than once. Participants could complete multiple memory games, since we ensured that a different set of images was presented each time. Figure 2-1 depicts an example experimental sequence.

2.4 Memorability scores and comparison to other experiments

On average, 80 workers saw each target image and its repeat (see table 2.1 for a complete breakdown), providing us with enough data points per image to collect reliable statistics about the memorability of each image. We define a **hit** to be a correct response to an image presented for the second time. A **miss** is when an image was repeated, but not recognized. **False alarms** and **correct rejections** are

[5] Vigilance repeats also occurred at the very beginning of the game, within the first 100 images (at which point target repeats did not yet appear) to maintain an even rate of image repeats.

incorrect and correct responses (respectively) to target images shown for the first time. We define *hit rate* (**HR**) and *false alarm rate* (**FAR**):

$$\text{HR(I)} = \frac{\text{hits(I)}}{\text{hits(I)} + \text{misses(I)}} \times 100\% \tag{2.1}$$

$$\text{FAR(I)} = \frac{\text{false alarms(I)}}{\text{false alarms(I)} + \text{correct rejections(I)}} \times 100\% \tag{2.2}$$

We also define $\overline{\textbf{HR}}$ and $\overline{\textbf{FAR}}$ to be category averages - computed over all images belonging to a single category. The $\overline{\text{HR}}$ scores vary from 49.5% to 64.2% ($M = 56.0\%, SD = 4.2\%$)[6]. $\overline{\text{FAR}}$ scores vary between 10.2% and 18.9% ($M = 14.6\%, SD = 2.0\%$), following a partial *mirror effect* [20, 64], where high HR are often accompanied by low FAR. The Spearman rank correlation between the $\overline{\text{HR}}$ and $\overline{\text{FAR}}$ scores is -0.66 ($p < 0.01$). Individual scores for all the categories can be found in table 2.1, and for comparison, memorability scores from other experiments are included in table 7.1. For instance, scene categories with lower memorability scores have similar performances to experiments with face stimuli [3] and data visualizations (see Chap. 6 or [9]). Lower variability across the stimuli in these categories could lead to lower scores (see Chap. 3, Sec. 3.5 for the possible reason). By the same token, scene categories with higher memorability scores have similar memory performances as the experiments with hundreds of scenes [31], likely due to larger variability across stimuli.

Figure 2-2 includes a sample of some of the most memorable and forgettable images in a few *FIGRIM* categories. The most memorable categories are *amusement parks* and *playgrounds*, scenes consisting of a large variety of objects in different configurations, and often containing people. Interestingly, 8/9 of the indoor categories are in the top 13 most memorable scene categories (the last indoor category, *cockpits* is the least memorable category overall). Qualitatively, the most memorable instances across categories tend to contain people, animals, text, and objects like cars and flags. Overall, memorable images tend to be distinct from the other images in their category – they may have unusual objects, layouts, or perspectives. This latter point will be quantified in Chapter 3.

[6] Throughout the rest of the paper, M will refer to 'mean' and SD to 'standard deviation'.

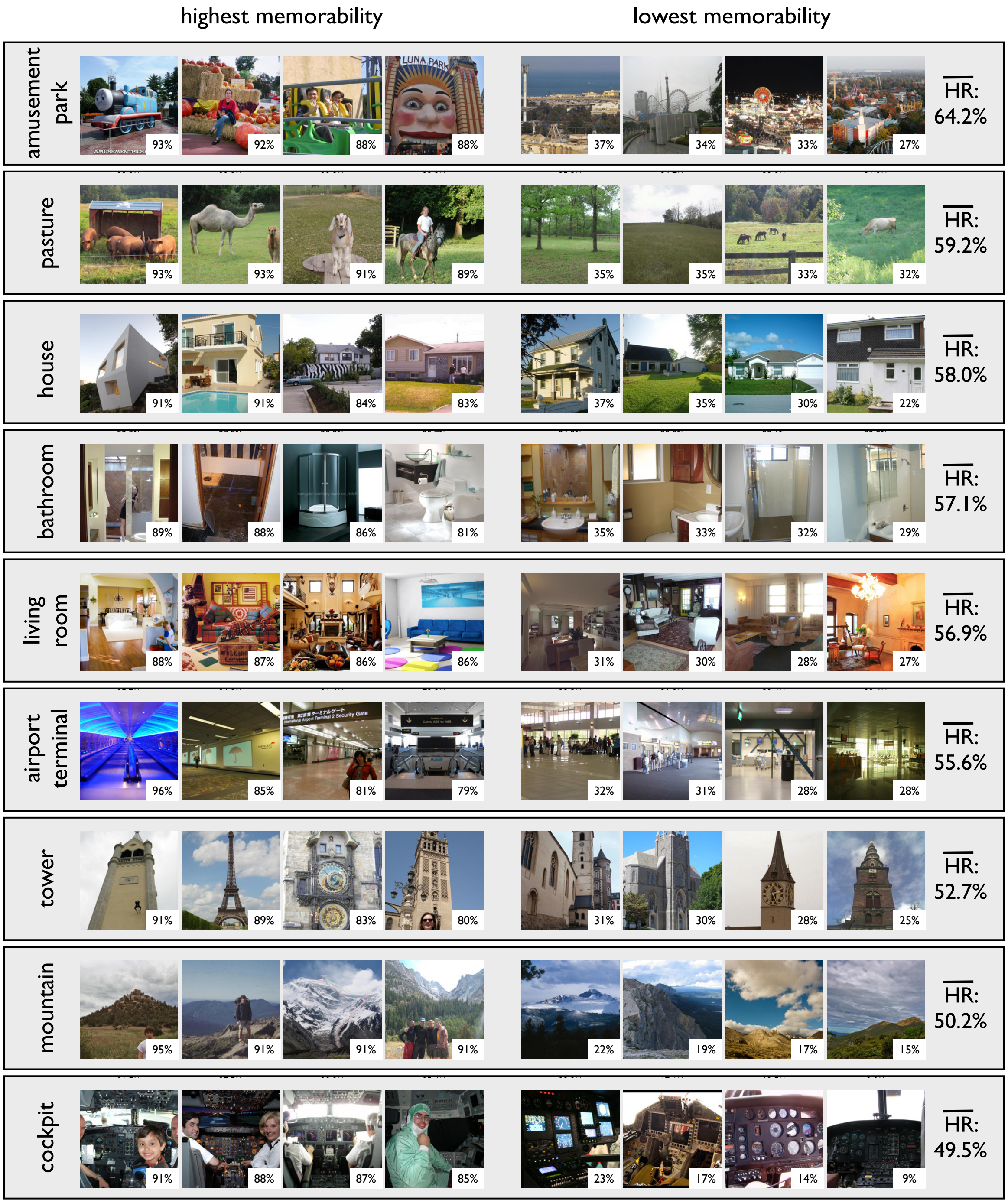


Figure 2-2: A sample of the most memorable and forgettable images from 9 of the 21 categories in the *FIGRIM* dataset, sorted from most to least memorable category, with the $\overline{\text{HR}}$ per category reported. Inset are the HR scores of the individual images.

2.5 Some scene categories are intrinsically more memorable

How consistent is the relative ranking (the ordering in table 2.1) of the scene categories? For instance, if we select a different subset of images, is the average memorability of the amusement park images still going to be at the top? We took half the images from each category, and computed the $\overline{\text{HR}}$ scores for all the categories. We also computed the $\overline{\text{HR}}$ scores for the other half of the images in all the categories. Over 25 such half-splits, the rank correlation between these 2 sets of $\overline{\text{HR}}$ scores was 0.68 (with significant p-values). Thus, the relative memorability of the scene categories is stable, and some scene categories are intrinsically more memorable than others.

2.6 Within categories, some images are intrinsically more memorable

Previous studies have demonstrated that memorability is consistent across participant populations for a general set of scene images (HR: $\rho = 0.75$, FAR: $\rho = 0.66$) [31] and for the specific classes of faces (HR: $\rho = 0.68$, FAR: $\rho = 0.69$) [3]. Here we show that this also holds at a fine-grained level across very different categories of scenes, thereby both replicating and extending previous results.

The consistencies of the image memorability scores were measured separately for each of the scene categories (see table 2.1). This was done by splitting the participants of AMT 1 into two independent groups, computing the memorability scores of images based on the participants in each group separately, ranking the images according to the memorability scores, and computing the Spearman rank correlation between the two possible rankings. Results are averaged over 25 such half-splits of the participant data. For all of the scene categories, consistency of HR scores ranges from 0.69 to 0.86 and from 0.79 to 0.90 for FAR scores. These high values demonstrate that memorability is a consistent measure across participant populations, indicating real differences in memorability across images.

Chapter 3

Computationally Modeling Context Effects on Memorability

How can image distinctiveness be quantified for natural images? How does context affect image memorability? When does context matter most?[1]

3.1 Related work

Previous studies have suggested that items that stand out from (and thus do not compete with) their context are better remembered [39, 50, 63, 27, 53, 17, 66, 56, 2, 65]. For instance, Standing observed a large long-term memory capacity for images that depict oddities [56]. Konkle et al. demonstrated that object categories with conceptually distinctive exemplars showed less interference in memory as the number of exemplars increased [39]. Additionally, for the specific categories of face images, studies have reported that a distinctive or atypical face (i.e., a face distant from the average) is more likely to be remembered [4, 13, 59]. Nevertheless, recent work on predicting image memorability [31, 29, 38] has largely ignored the effects of image context (the set of images from which the experimental sequence is sampled) on memory performance, instead focusing on the modeling of intrinsic image features.

Here, we are able to rigorously quantify, using a large-scale natural scene database,

[1] This chapter is closely related to publication [16]

category	targets	fillers	datapts per target	$\overline{\text{HR}}$ (%)	$\overline{\text{FAR}}$ (%)	HR cons. (ρ)	FAR cons. (ρ)
21 scenes	1754	7296	74.3 (SD: 7.5)	66.0 (SD: 13.9)	11.1 (SD: 9.5)	0.74 (SD: 0.2)	0.72 (SD: 0.1)

Table 3.1: *FIGRIM* dataset statistics for AMT 2 (across-category). The targets are the same for AMT 1 and AMT 2. The difference in the number of fillers between AMT 1 and AMT 2 is accounted for by demo images that were presented to participants at the beginning of each experiment, and are included with the fillers. Each category in AMT 1 had 20 demo images, while AMT 2 had a total of 42 demo images, sampled from all the categories.

the observation that images that are unique or distinct with respect to their image context are better remembered. We steer away from subjective human ratings, and instead compute statistics over automatically-extracted computer vision features. By systematically varying the image context across experiments (AMT 1 and AMT 2), we are able to computationally model the change in context at the feature level, and predict corresponding changes in image memorability.

3.2 Crowdsourcing (across-category) experiment AMT 2

We ran another AMT study on the combined target and filler images across all the scene categories, and collected a new set of memorability scores, following the same protocol as before (see Chap. 2, Sec. 2.3). The dataset statistics are provided in table 3.1. The average memorability scores for this experiment are: HR: $M = 66.0\%, SD = 13.9\%$, FAR: $M = 11.1\%, SD = 9.5\%$. Per-image memorability scores correlate strongly with those measured in the within-category experiment AMT 1 (Spearman $\rho = 0.60$ for HR and $\rho = 0.75$ for FAR), demonstrating that the intrinsic memorability of images holds across different image contexts.

3.3 In-lab control experiment

We selected a subset of the target images from the AMT experiments in order to verify replicatability of the online data using in-lab experiments. From each scene category from AMT 1 we obtained the 15 target images with the highest and 15 with the lowest memorability scores. This was done to capture the range of memorabilities of images in each of the scene categories. These 630 images became the targets for our in-lab experiments. We recruited 20 participants for our experiments.

An experimental sequence was composed of about 1000 images, of which 210 were targets that repeated exactly once in the sequence, spaced 91-109 images apart. Images in the test sequence were presented for 2 sec, separated by a fixation cross lasting 0.5 sec. Participants were instructed to respond (by pressing the spacebar) anytime they noticed an image repeat in the sequence, at which point they would receive feedback. In a single experimental session, the targets consisted of 30 images taken from each of 7 randomly selected scene categories, making up a total of 210 targets. The filler images were chosen in equal proportions from the same set of scene categories as the targets. Participants could choose to complete up to 3 sessions (each with a disjoint set of 7 categories) on separate days. The memorability scores for the in-lab experiment are HR: $M = 64.9\%, SD = 21.3\%$, FAR: $M = 6.0\%, SD = 8.9\%$.

Note that by changing the number of scene categories in an experiment (from 1 in AMT 1, to 7 in this in-lab experiment, to 21 in AMT 2), we are also increasing the variability of the experimental image context. To demonstrate the effect of number of scene categories on memorability, we sorted the HR scores of the overlapping targets in all 3 experiments by the scores of AMT 2 and binned them into *high*, *middle*, and *low* memorability. In figure 3-1, as the number of scene categories increases, the overall memorability scores of all the images in the experiment also increase (even for the least memorable images). At the same time, the difference between the (high, middle, low) memorability bins remains statistically significant, indicating that some images are intrinsically more memorable and others forgettable.

The rank correlation of the HR scores for the 630 target images used in the in-lab experiment with the scores for the same images in AMT 1 is 0.75, and with AMT 2 is 0.77. Thus, across all 3 of the experiments (two online, one in-lab), the relative ranking of these target images are highly consistent, providing further evidence that image memorability is to a large extent an intrinsic property of images that holds across different populations of human participants, different image contexts, and different experimental settings.

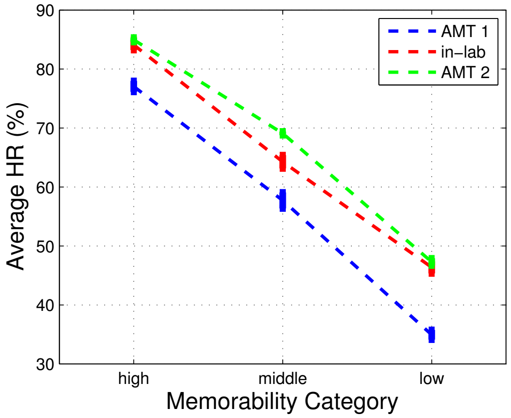


Figure 3-1: Memorability scores for images in the context of 21 scenes (AMT 2) are higher than in the context of 7 scenes (in-lab), and higher still than in the context of 1 scene (AMT 1). At the same time, the most memorable images remained the most memorable, and the most forgettable remained the most forgettable. Standard error bars have been plotted.

3.4 Contextually distinct images are more memorable

We call images **contextually distinct** if they are distinct with respect to their image context (the set of images from which the experimental sequence is sampled). To model context effects, we first estimate the probability distribution over features in an image's context. Then we define the distinctiveness of the image as the probability of its features under this distribution. We consider two different contexts: (a) within-category context composed of images from a single category (AMT 1), and (b) across-category context composed of images from all categories (AMT 2). To estimate the probability distribution of features in the context, we use kernel density estimation [28].

For each image I, we compute a feature vector $f_i = F(I)$, where F can be any feature mapping. We model the probability of features f_i appearing in image context C as:

$$P_c(f_i) = \frac{1}{\|C\|} \sum_{j \in C} K(f_i - f_j) \tag{3.1}$$

where K can be any kernel function, and $\|C\|$ indicates the size of the context, measured in number of images. We use an Epanechnikov kernel and leave-one-out-cross-validation to select the kernel bandwidth. We tried a number of features[2] (see Tables 3.2 and 3.3), but found that scene-based CNN features worked best, and that is the feature space we use for the rest of this chapter. The features come from a convolutional neural network (CNN), a popular feature space recently shown to outperform other features in computer vision [41, 51]. Specifically, we used the **Places-CNN** from [69] trained to classify scene categories. We took the 4096-dimensional features from the response of the Fully Connected Layer 7 ($fc7$) of the CNNs, which is the final fully-connected layer before producing class predictions. We then reduced this feature vector to 10 dimensions using PCA. This was found to prevent overfitting and increase efficiency in estimating the kernel densities. Note that in contrast to simple visual descriptors like Gist, the deep features are trained to predict image semantics, and this may account for some of the performance boost.

In figure 3-2a, we correlate the memorability score of an image, $\mathrm{HR}(I)$, with its distinctiveness with respect to the image context, $D(I;C)$. Mathematically, we define[3]:

$$D(I;C) = -\log P_c(f_i) \tag{3.2}$$

Furthermore, we denote C_2 as the across-category context of AMT 2, and C_1 as the within-category context of AMT 1. We find that $D(I;C_2)$ is positively correlated with $\mathrm{HR}(I)$ (Pearson $r = 0.24$, $p < 0.01$), as plotted in 3-2a. The correlation also holds when images are compared to images within the same category (correlation between $D(I;C_1)$ and $\mathrm{HR}(I)$ is $r = 0.26$, $p < 0.01$). Thus, more *contextually distinct* images are more likely to be memorable. We present the same correlations according to some alternative measurements of memorability (definitions provided in Appendix A) in Tables 3.2 and 3.3.

[2] The *object-based CNN features* come from a pre-trained model from Caffe available at `http://caffe.berkeleyvision.org` tuned to perform object classification [33]. The Gist features, as defined in [47], are calculated using the *LabelMe Toolbox* available at `http://labelme.csail.mit.edu/Release3.0/browserTools/php/matlab_toolbox.php` [52].

[3] In information theory, this is alternatively termed *self-information* and *surprisal.*

feature space	HR	FAR	ACC	MI	DPRIME
GIST	0.12 (HS)	-0.19 (HS)	0.21 (HS)	0.21 (HS)	0.19 (HS)
object-based CNN	0.19 (HS)	-0.14 (HS)	0.22 (HS)	0.22 (HS)	0.20 (HS)
scene-based CNN	0.26 (HS)	-0.26 (HS)	0.35 (HS)	0.36 (HS)	0.34 (HS)

Table 3.2: $D(I; C)$ of images (C = AMT 1) correlated with different memorability measurements. Here, HS = significant at the $p = 0.01$ level.

feature space	HR	FAR	ACC	MI	DPRIME
GIST	0.12 (HS)	-0.01 (NS)	0.12 (HS)	0.11 (HS)	0.11 (HS)
object-based CNN	0.13 (HS)	-0.05 (S)	0.15 (HS)	0.12 (HS)	0.12 (HS)
scene-based CNN	0.24 (HS)	-0.17 (HS)	0.32 (HS)	0.33 (HS)	0.32 (HS)

Table 3.3: $D(I; C)$ of images (C = AMT 2) correlated with different memorability measurements. Here, HS = significant at the $p = 0.01$ level, S = significant at the $p = 0.05$ level, NS = not significant.

3.5 More varied image contexts are more memorable overall

We also measure the **context entropy** by averaging $D(I; C)$ over all the images in a given image context. This is just the information-theoretic entropy:

$$\begin{aligned} H(C) &= \mathbb{E}_c[D(I; C)] \\ &= \mathbb{E}_c[-\log P_c(f_i)] \end{aligned} \tag{3.3}$$

Here, $\mathbb{E}_c$ is just expectation over the image context specified by C. As in figure 3-2b, the Pearson correlation between $H(C)$ and $\overline{HR} = \mathbb{E}_c[\text{HR(I)}]$, is $r = 0.52$ ($p = 0.01$) - and more results provided in Table 3.4. Thus, categories that contain many contextually distinct images are more memorable overall. For instance, the *mountain* category contains a relatively stable collection and configuration of scene elements: mountains and sky. The *amusement park* category, however, consists of a much larger variability of images: images of roller-coasters, concession stands, or other rides, consisting of different elements. Thus entropy in feature space can explain some of the differences in average HR we observe across categories in AMT 1.

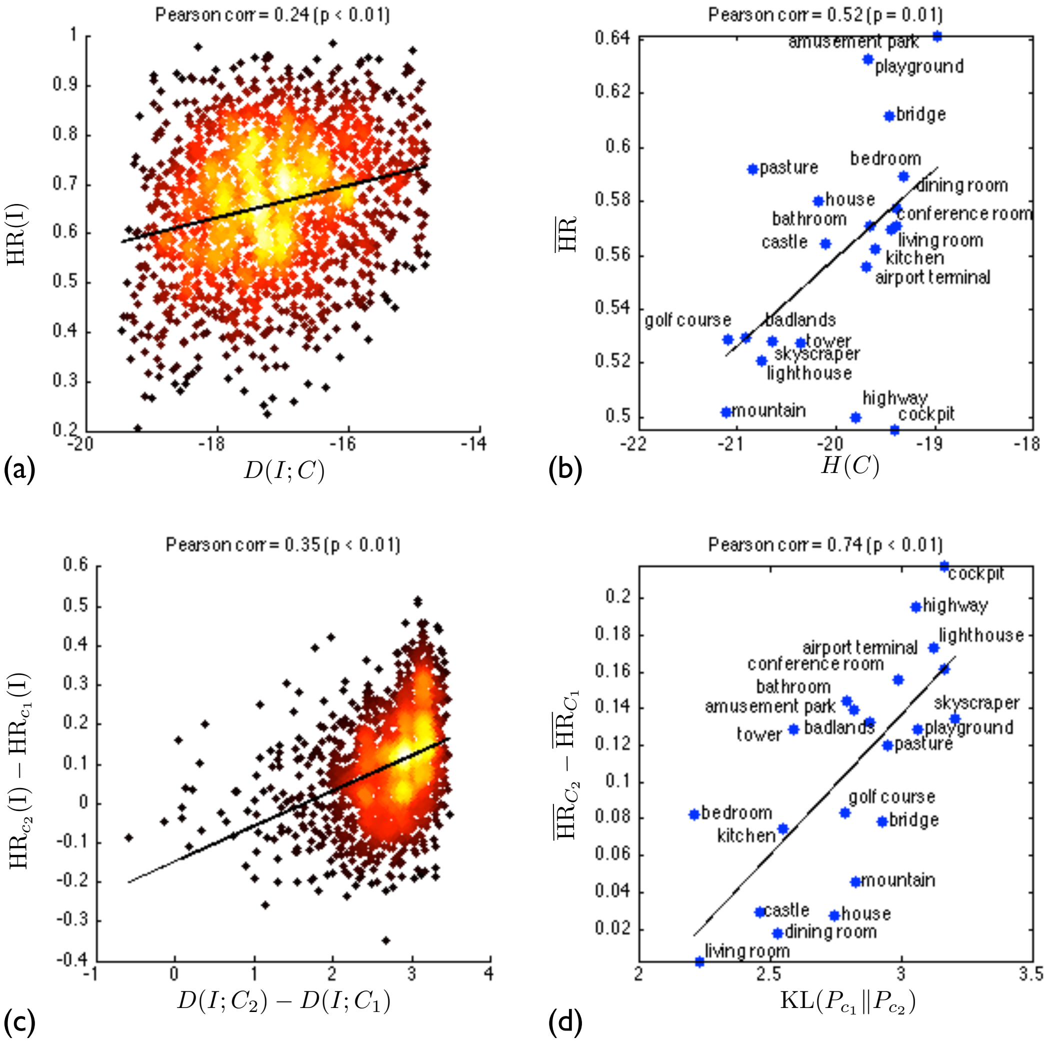


Figure 3-2: The effects of context on memorability. In figures (a) and (c), each dot is a single target image from the *FIGRIM* dataset, for a total of 1754 images. Brighter coloring represents a greater density of points. In figures (b) and (d), all images in a given category are collapsed into a single summary number. The trends we see are: (a) Images are more memorable if they are less likely (more contextually distinct) relative to the other images in the same image context; (b) Image contexts that are more varied (have larger entropy) lead to higher memorability rates overall; (c) Images that become more distinct relative to a new context become more memorable; (d) Scene categories that are more distinct relative to other categories become more memorable in the context of those other categories.

feature space	HR	FAR	ACC	MI	DPRIME
GIST	0.48 (S)	-0.23 (NS)	0.34 (NS)	0.35 (NS)	0.36 (NS)
object-based CNN	0.51 (S)	-0.29 (NS)	0.34 (NS)	0.38 (NS)	0.32 (NS)
scene-based CNN	0.52 (S)	-0.16 (NS)	0.28 (NS)	0.37 (NS)	0.34 (NS)

Table 3.4: $H(C)$ of images in AMT 1 (within-category) correlated with different memorability measurements.

feature space	HR	FAR	ACC	MI	DPRIME	RANK
GIST	0.24 (HS)	0.10 (HS)	0.15 (HS)	0.10 (HS)	0.12 (HS)	0.10 (HS)
object-based CNN	0.32 (HS)	0.00 (NS)	0.25 (HS)	0.21 (HS)	0.24 (HS)	0.15 (HS)
scene-based CNN	0.35 (HS)	0.00 (NS)	0.29 (HS)	0.25 (HS)	0.28 (HS)	0.14 (HS)

Table 3.5: Change in *contextual distinctiveness* between AMT 1 and AMT 2 correlated with different memorability measurements.

3.6 Modeling image memorability as a function of image context

AMT experiments 1 and 2 systematically vary the context for images, while keeping the images constant. This allows us to isolate the effects of context from other possible confounds[4]. To model the change in context, we compute the difference in the distinctiveness of an image relative to its own scene category versus all scene categories. In figure 3-2c we see that changing the context of an image to make it more distinct relative to the other images in its context, increases its memorability. The Pearson correlation between $D(I; C_2) - D(I; C_1)$ and $\text{HR}_{C_2}(I)$ - $\text{HR}_{C_1}(I)$ is 0.35 ($p < 0.01$) with more results provided in Table 3.5.

We can also consider change in memorability at the category level. In figure 3-3 we see that across all categories, $\overline{\text{HR}}$ for each category goes up in the context of images from other categories. However, how much change there is in image memorability when we switch contexts depends on the scene category.

How does a scene category's memorability change when the category is combined with other categories? We measure this change in context as the *Kullback—Leibler*

[4] Spurious correlations are possible when both contextual distinctiveness and memorability correlate with a third causal factor, but when we systematically change the context while keeping everything else fixed (particularly, the experimental images), we can isolate the effects of context alone.

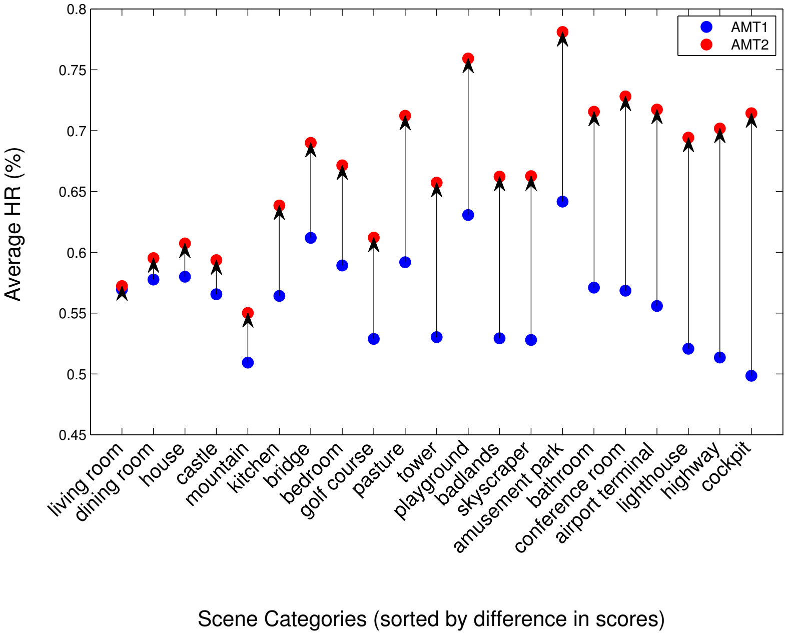


Figure 3-3: The average memorability of the images in each scene category went up when images were presented in the context of images from other scene categories (AMT 2) compared to when they were presented only in the context of images from the same category (AMT 1).

divergence between the density functions computed over contexts C_1 and C_2 as:

$$\mathrm{KL}(P_{c_1}||P_{c_2}) = \mathbb{E}_{c_1}[-\log P_{c_2}(f)] - \mathbb{E}_{c_1}[-\log P_{c_1}(f)] \tag{3.4}$$

The first term is the probability of the images in a category under the context of AMT 2, and the second term is the probability of the images under its own category in AMT 1. Intuitively, this measures how much more (or less) likely a category's images are under the context of AMT 2 compared to that of AMT 1. In figure 3-2d, the Pearson correlation between the change in context entropy and change in memorability is $r = 0.74$ ($p < 0.01$) with more results provided in Table 3.6.

Consider the *cockpit* category: many of the cockpit images look the same when viewed only with other cockpits; however, when mixed with images from other scenes, they become very distinct: there is no other scene category with similar images. Compare this with *dining rooms* that can also look like *living rooms* and *kitchens*,

feature space	HR	FAR	ACC	MI	DPRIME
GIST	0.64 (HS)	0.41 (NS)	0.47 (S)	0.39 (NS)	0.41 (NS)
object-based CNN	0.55 (S)	0.24 (NS)	0.35 (NS)	0.46 (S)	0.47 (S)
scene-based CNN	0.74 (HS)	0.44 (S)	0.51 (S)	0.50 (S)	0.60 (HS)

Table 3.6: Change in *context entropy* between AMT 1 and AMT 2 correlated with different memorability measurements.

and thus are not as visually distinct when combined with images from these other scene categories.

3.7 When context matters most

To better understand when context matters most, consider the images that were memorable with respect to their own category, but became more forgettable when combined with other categories. In figure 3-4, we can see that these images tend to look different from other images in their category, and are more similar to images of other categories. These images may have been more memorable in the first place because they stood out from other images in their category. They no longer stand out when combined with other categories.

To quantify this intuition, we mapped the Places-CNN deep features to category labels by training a linear multi-class SVM on the filler images of the FIGRIM dataset with labels of 21 scene categories. We then evaluated our classifier on the target images of the FIGRIM dataset to automatically predict the most likely scene category for each image (the overall scene classification accuracy was 91.56%). These predicted category labels are included with each image in fig. 3-4. Notice that for the images that decreased in memorability, more of the predicted labels come from other categories. Compare this to the images that increased in memorability when combined with other categories - they are more likely to be correctly classified.

We can also consider the probability, under the scene classifier, of the correct category label. These probabilities are included with each image in fig. 3-4. Images with a higher probability value are more typical examples of their scene category. Across all 1754 target images, the Pearson correlation between the probability of the correct category label and the change in memorability due to context (from AMT 1 to AMT 2) is $r = 0.30$ ($p < 0.01$). In other words, the images least likely to belong to their own category experience the greatest drop in memorability when they are combined with images of other categories.

Which images remain memorable within and across categories? It is the images that will look distinct enough from the other images in their category, but will not look like images from other categories either. Consider the images in the top right quadrant in figure 3-5. These are images that are memorable across contexts. Take for example the bridge in front of the red sky. It looks like no other scene category other than a bridge, but it also looks like no other bridge (the red sky is unique). Compare this to the bridge in the bottom right, which looks more like a pasture. Among bridges, it is memorable, but among pastures it is not. Thus, the memorability of the images in the top right quadrant is least affected by context, but the memorability of images in the bottom right quadrant is most affected by context.

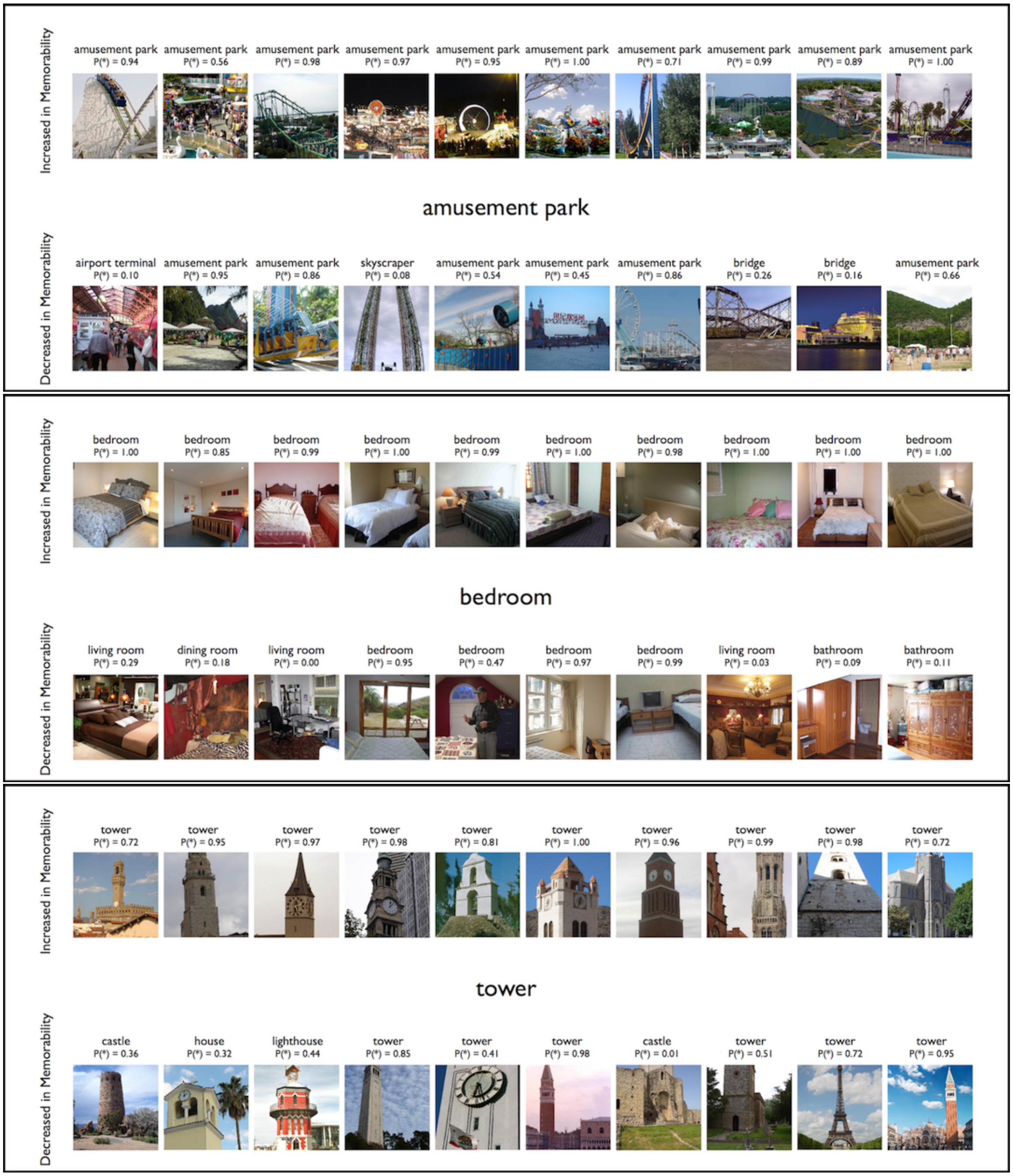


Figure 3-4: We evaluated a scene classifier on the images in that increased most and those that dropped most in memorability when combined with other categories. Here we show 3 categories. For each image, we provide the classifier's predicted category label, as well as the probability of the correct category label (where * is replaced with the correct category). Images that drop in memorability are more likely to be confused with other categories.

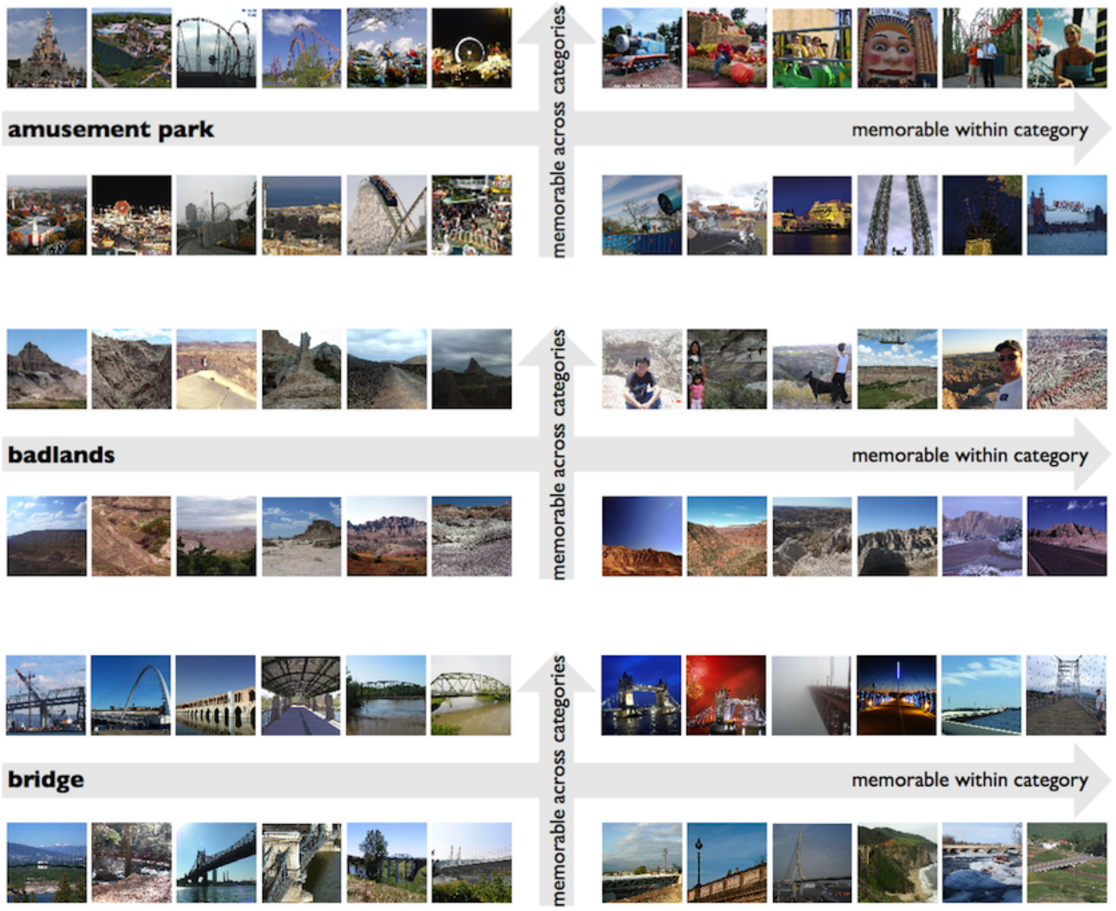

amusement park
memorable within category
memorable across categories
badlands
memorable within category
memorable across categories
bridge
memorable within category
memorable across categories

Figure 3-5: Memorability scores of images in the top right quadrant of each plot are least affected by context whereas the scores of images in the bottom right quadrant are most affected by context. Images in the top right are distinct with respect to both contexts, while images in the bottom right are distinct only with respect to their own category.

Chapter 4

Eye movements to predict individual image memories

To what extent can one population of human participants reliably predict another? Can an individual's eye-movements on an image be used to predict if the image will be later remembered?[1]

4.1 Related work

Little work has considered the intersection between image memorability and visual attention [42, 46, 19, 14]. Mancas et al. [42] use saliency features to show a slight improvement over the automatic image memorability predictions in [31]. They use a form of visual attention (i.e. saliency, eye movements) to improve on image memorability. We refer to image memorability as a **population predictor** because it ignores trial-by-trial variability, effectively averaging over a population of participants or experiments. We, instead, use visual attention to improve the trial-by-trial predictions of memory (an **individual trial predictor**). Bulling and Roggen [14] use eye movement features to predict image familiarity, classifying whether images have been seen before or not. They assume that all images seen again are remembered, particularly due to the long exposure times (10 seconds) used per image, and

[1] This chapter is closely related to publication [16]

by testing on a small dataset of 20 faces. They also use eye movement analysis as a *population predictor* to decide whether an image was *previously seen*, while we use eye movement analysis as an *individual trial predictor*, taking into account individual differences in making predictions of whether an image will be *later remembered.*

Our work is also related to recent studies on the use of eye movements for decoding an observer's task [8, 23]. These studies consider features extracted from the eye movements of individual participants to determine the task they are performing (e.g., what question they are answering about an image), modeled on the original Yarbus experiment [68]. These studies utilize a very small set of images (ranging from 15-64) with a very constrained theme (grayscale photographs taken between 1930-1979 with at least two people [23]; paintings depicting "an unexpected visitor" [8]). In our study, we measure the eye movements of participants on 630 target images sampled from 21 different indoor and outdoor scene categories. We extract features from eye movements to determine whether or not an image is correctly encoded (measured by whether it is correctly recognized on a successive exposure). We are able to solve our decoding task using only 2 seconds of viewing time per image, whereas the previous studies worked with durations of 10 sec [14, 23], 30 sec [8], 50 sec [57], and 60 sec [8]. For this purpose, we learn image-specific classifiers to distinguish fixations on one image versus fixations on other images.

4.2 Eyetracking experiments

We used a similar set-up to the in-lab experiment from Chap. 3, Sec. 3.3, but with important differences to collect eye-movements in an un-biased manner (outlined in Fig. 4-1). Images were presented to participants at 1000 × 1000px. We used the same set of 630 targets as in the in-lab experiment, but split the images over 4 separate experimental sessions (of 157-158 target images, randomly sampled from all categories). Target images were repeated 3 times in the sequence, spaced 50-60 images apart. Images remained on the screen for 2 seconds, and participants gave a

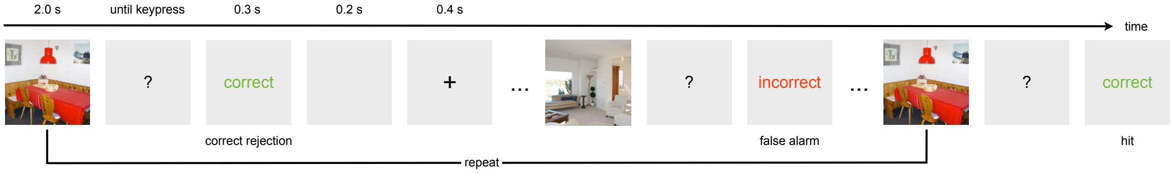


Figure 4-1: An example eyetracking experimental sequence. The main differences from the AMT experiment in Fig. 2-1 are the slightly longer image presentation times, the collection of key presses *after* image presentation at the prompt, and the forced-choice response.

forced-choice response *at the end* of each image presentation to indicate whether the image appeared previously or not. After a keypress response and feedback, a fixation cross came on the screen for 0.4 seconds, followed by the next image.

Eyetracking was performed on an SR Research EyeLink1000 desktop system at a sampling rate of 500Hz, on a 19 inch CRT monitor with a resolution of 1280 × 1024 pixels, 22 inches from the chinrest mount. The experiments started with a randomized 9-point calibration and validation procedure, and at regular intervals throughout the experiment drift checks were performed, and if necessary, recalibration. Each experiment lasted 75-90 minutes, and participants could take regular breaks throughout. All participant eye-fixations and keypresses were recorded. We recruited a total of 42 participants for our study (16.2±1.6 participants per image). The memorability scores for this experiment were: HR: $M = 75.8\%, SD = 14.4\%$, FAR: $M = 5.2\%, SD = 7.4\%$.

4.3 Classification model

Given a set of fixations on an image, we want to know: will the viewer remember this image at a later point in time? The key idea is that if a viewer's fixations differ from the fixations expected on an image, the viewer may not have encoded the image correctly. Thus, when evaluating a novel set of fixations, we want the probability that these fixations came from this image - as opposed to some other image. If the probability is high, we label the fixations as successful encoding fixations, since we believe they will lead to a correct recognition of the image later. Otherwise, we assume the image was not properly encoded, and will be forgotten. To provide some further intuition, a few examples are provided in figure 4-2. We construct a computational model by training a separate classifier for each image, differentiating fixations that belong to this image from fixations on all other images.

After preprocessing[2], we convert an observer's fixations on an image into a **fixation map** by binning the fixations into a 20 × 20 grid, normalizing the binned map, and smoothing it by convolution with a Gaussian with $\sigma = 2$ grid cells. Coarse sampling and smoothing is necessary to regularize the data.

For each image, we train an ensemble classifier $G_i = g(I)$ to differentiate fixation maps on this image (positive examples) from fixation maps on all other images (negative examples). For training, we only consider **successful encoding fixations** - the fixations made on an image the first time it appeared in the image sequence, and led to a correct recognition later in the sequence.

We use a RUSBoost classifier [54], which handles the class imbalance problem[3], and **balanced accuracy** as a metric of performance because it avoids inflated performance estimates on datasets with unequal numbers of positives and negatives [12]. It is calculated as:

$$\begin{aligned} \text{balanced accuracy} = & \frac{0.5 \times \text{true positives}}{\text{true positives} + \text{false negatives}} \\ & + \frac{0.5 \times \text{true negatives}}{\text{true negatives} + \text{false positives}} \end{aligned} \tag{4.1}$$

Over 5 train-test splits, the balanced accuracy of our classifier on determining whether a set of fixations comes from a specific image vs some other image is 79.7% (SD: 13.9%), where chance is at 50%. This high performance indicates that we are able to successfully learn diagnostic fixation patterns for an image to distinguish it from all other images. However, not all images produce diagnostic fixation patterns, and thus predictive power varies by image (section 4.6).

[2] We processed the raw eye movement data using the EyeLink Data Viewer, removed all fixations shorter than 100 ms or longer than 1500 ms, and kept all others that occurred within the 2000ms recording segment (from image onset to image offset).

[3] N being the total number of images, we have order $N - 1$ negatives, since those come from all other images while the positives come from a single image.

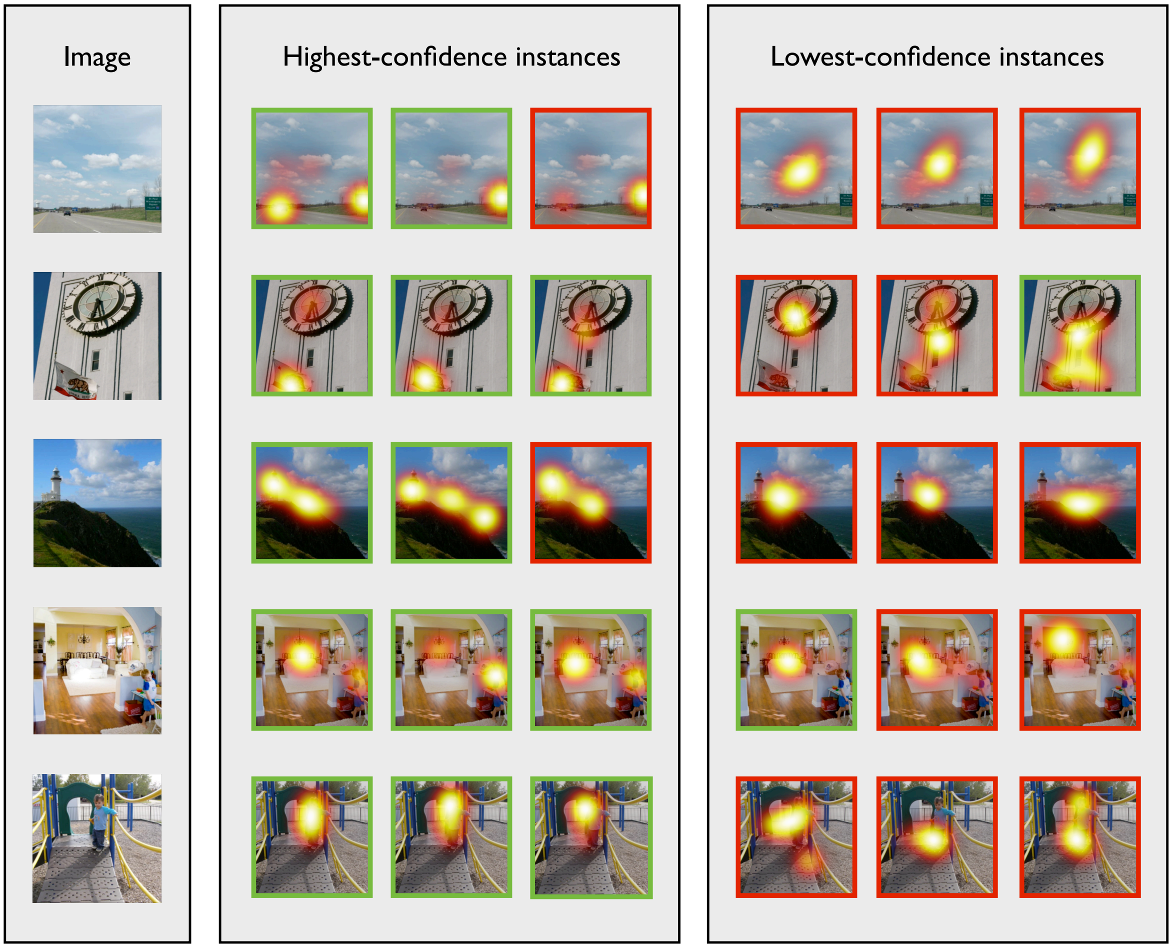


Figure 4-2: Examples of individual viewers' fixation maps (at encoding) overlaid on top of the images viewed. For each of these 5 example images, we include the 3 highest-confidence and 3 lowest-confidence instances under the image's classifier (trained to differentiate fixations on this image from fixations on other images). Fixations that later led to a correct recognition of the image are outlined in green, and those where the image was unsuccessfully remembered are in red. This depicts some of the successes and failure modes of our model.

4.4 Eye movements are predictive of whether an image will be remembered

Here we use the model developed in the previous section to make image memorability predictions on a trial-by-trial basis. As demonstrated in previous chapters, people are highly consistent in which images they remember and forget. Thus as a baseline we use an image's memorability score (HR from AMT 2, Chap. 3) to make trial-by-trial predictions. We refer to this as a population predictor because these memorability scores are obtained by averaging over observers.

We compare this population predictor with an individual trial predictor which uses a viewer's eye movements on a particular trial to predict whether an image will be remembered. Our individual trial marker involves measuring the confidence of a viewer's fixation map under the classifier G_i. Our split of the original eyetracking data is threefold: we have a set of participants on each image for training the classifier G_i to differentiate fixations on image I from fixations on other images; another set of participants on each image is used for estimating the threshold required to differentiate successful from unsuccessful encoding fixations (where we use the ground truth data on whether the participants successfully recognized the image); finally, we evaluate the fixation maps of the last set of participants using the learned threshold to classify fixations as successful or unsuccessful. For picking the threshold, we perform a grid search over 200 values, and optimize for balanced accuracy.

Over 15 different threefold splits of data, we obtain a balanced accuracy of 66.02% (SD : 0.83) at determining whether a set of encoding fixations is successful and will lead to a correct recognition of the image at a later point in time. Compare this to 60.09% (SD : 1.55%) when using the memorability score of an image - our population predictor which does not take into account the trial-to-trial variability. Additional baselines that we considered were the similarity of the fixation map to a center prior, achieving an accuracy of 56.35% (SD: 0.60%), and the coverage of the fixation map (proportion of image fixated), achieving an accuracy of 55.89% (SD: 0.58%). Thus, neither of the baselines could explain away the predictive power of our model[4].

4.5 When individual differences matter most

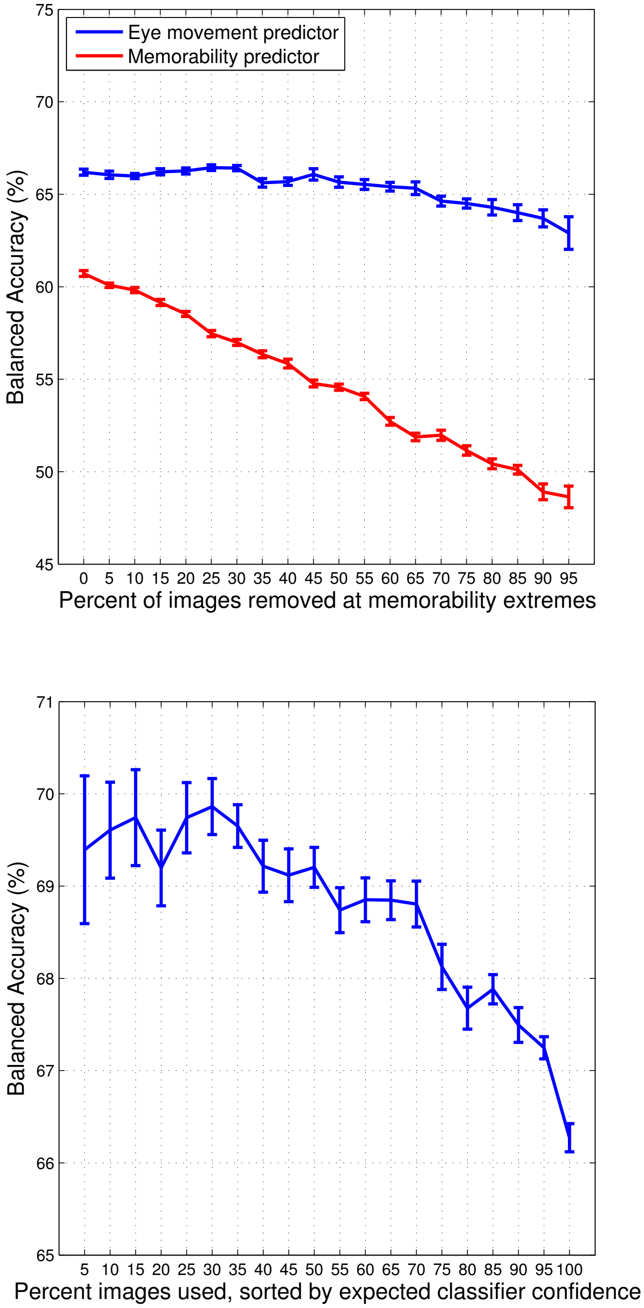
Consider the cases where images are not consistently memorable or forgettable across individuals. We sort images by their AMT scores (which we obtain from AMT 2, Chap. 3), and progressively remove images at the memorability extremes. The resulting prediction performance is plotted in figure 4-3a. Memorability scores fall to chance at predicting individual trials precisely because the images at the memorability extremes were most predictive. Meanwhile, our eye movement features retain predictive power, indicating that individual differences become most relevant for the middle memorability images. These are the images that may not be memorable at-a-glance, and may require the viewer to be more "attentive".

4.6 Not all images are equally predictable

Our classifier is more likely to make a successful prediction on a given trial when it is confident. A classifier is confident when it can easily discriminate successful from unsuccessful fixations on an image. For instance, an image with all of the important content in the center might not require the viewers to move their eyes very much and this makes prediction particularly difficult because successful and unsuccessful fixations may not be that different.

Our model construction allows us to easily estimate the expected confidence of our classifier on an image. For a given image I, we compute the expected confidence of classifier G_i as the average confidence value over its positive training examples - the successful fixation maps on image I. When we sort images by this measure (see fig. 4-4), we obtain the results in fig. 4-3b. We can achieve a balanced accuracy above

[4] *Successful fixations tend to be alike; every unsuccessful set of fixations is unsuccessful in its own way*: the fixations may be center-biased (the viewer does not move his eyes to look around), they may be off-center or even off-the-image (the viewer is distracted), the fixations might be randomly distributed over the image and have high coverage (the viewer is not paying attention to the task), etc. Thus baseline models that try to separate successful from unsuccessful fixations using simple principles, like coverage or center bias, will not have full predictive power.



(a)

(b)

Figure 4-3: (a) When we prune images at the memorability extremes, memorability scores fall to chance as a predictor of per-trial memory performance, while eye movements remain important for making trial-by-trial predictions. (b) Our classifier makes more accurate predictions when it has higher expected confidence. Standard error bars are included for both plots.

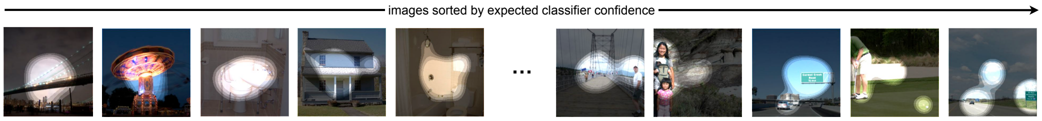


Figure 4-4: Images sorted by expected classifier confidence (from least to most). A classifier with high confidence on its positive training examples will do better at differentiating successful from unsuccessful fixations on an image. Overlaid on top of each image is the average fixation map computed over all successful encodings of the image.

70% for the images where our classifier has the highest confidence. Thus, we can automatically select images that our classifier will likely do well on. This becomes an important feature for applications where we have a choice over the input images that can be used, and need to have a system to robustly predict from eye fixations, whether an image will be later remembered.

Chapter 5

Contributions to Cognitive Science: Pupils as Indicators of Memory Processes

How quickly can an image fade from memory? Can physiological measurements like pupillary responses and blink rates be used to track image memorability without the need for overt responses? [1]

5.1 Related work

Memory for objects and scenes is massive and image details can be stored for hours or even days [11, 25, 39, 40]. Isola et al. have demonstrated that images in a collection that are the most memorable at shorter intervals are still the most memorable even after 40 minutes [30].

We confirm that memorability is an intrinsic property of images that is stable over time, using a logarithmic set of time intervals (lags of 8, 16, 32, 64 images). Differences in image memorability are already present briefly after encoding, at the shortest lag (only 20 seconds). Moreover, differences in memorability produce different rates of forgetting. The most memorable images degrade least in memorability across time.

[1] This chapter is closely related to publication [61]

The graded effect of memorability also shows up in the pupil and blink rates, as differences in physiological responses during retrieval.

While it has been known since the early 60s that the overt pupillary response can be linked to covert cognitive constituents like information processing and mental load [10, 24, 35, 22], recent studies have started using pupil dilations more deliberately to investigate recognition memory [21, 34, 44, 48, 49, 62]. For instance, Võ and colleagues [62] have demonstrated that pupils dilate more to “old”, studied items than to “new” ones, and termed this effect the **Pupil Old/New Effect** or **PONE**.

Usually, studies interested in pupil measurements disregard blinks as missing data. However, in addition to pupil dilations, blinks provide mutually exclusive, but complementary indices of information processing [55]. Blink rates have been shown to decrease under conditions of high visual and/or cognitive load, since a reduced blink rate supports a continuous input of visual information especially necessary when cognitive demands are high [45].

With regards to image memorability, these finding suggest that since a correct retrieval of less memorable images requires more cognitive effort, we should see an increase in pupil dilations as well as a decrease in blink rate. Furthermore, memories encoded longer ago should also require more cognitive effort to retrieve and should be accompanied by similar physiological markers.

5.2 In-lab experiments

From the Isola et al. [31] database of 2222 images with accompanying memorability scores, we chose 80 images each with the highest, lowest, and medium memorability scores, while balancing the occurrence of indoor/outdoor scenes, people, and animals in the 3 memorability categories. There was no difference in mean luminance between low and high memorability categories for neither LAB, HSV, nor RGB values, all $t(158) < 1$. The images were resized to be 512×512 pixels. A sample from the 240 target images used in our present study can be found in Fig. 5-1.

The experimental design included two main manipulations: image memorability

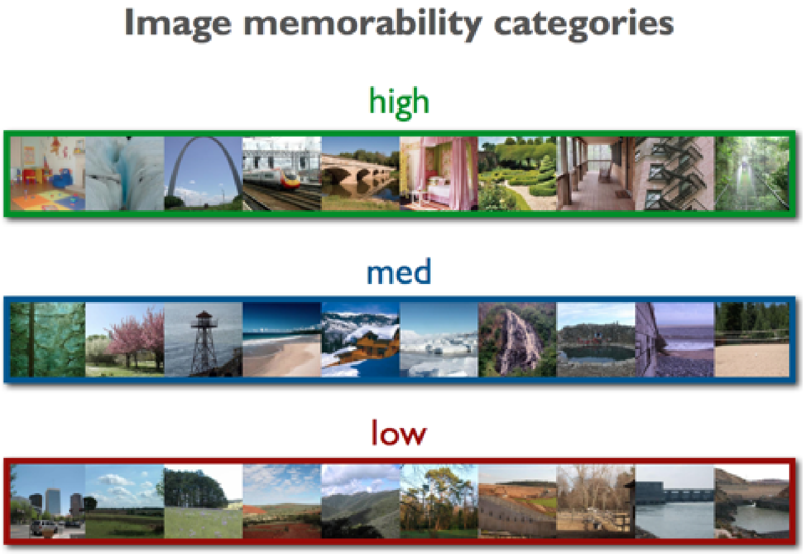


Figure 5-1: Sample of top, medium, and low memorability images used from the memorability images of [31] as the targets for our experiments.

(low, medium, high) and lag (8, 16, 64, 256). Each participant was presented with a sequence of 1216 images, 240 of which were target images presented twice. The other 736 images were fillers, which were sampled randomly from the remaining images in the database. We randomly assigned target images to prespecified positions within the image sequence repeating at one of 4 lags categories (8, 16, 64, 256). Pupil dilations and blink rates were recorded with the EyeLink1000 set-up described in Sec. 4.2. 15 subjects participated.

Stimulus-locked recording segments of 2000 ms were baseline-corrected by subtracting the average pupil dilation over the 100 ms preceding stimulus onset. Peak horizontal dilations in the 2000 ms after stimulus onset were submitted separately to ANOVAs with memorability (low, medium, high) and lag (8, 16, 64, 256) as within-subject factors. Blinks were recorded as events where the pupil size was very small, missing in the camera image, and/or severely distorted by eyelid occlusion. We computed the blink rate as the average percentage of blinks in each of the conditions.

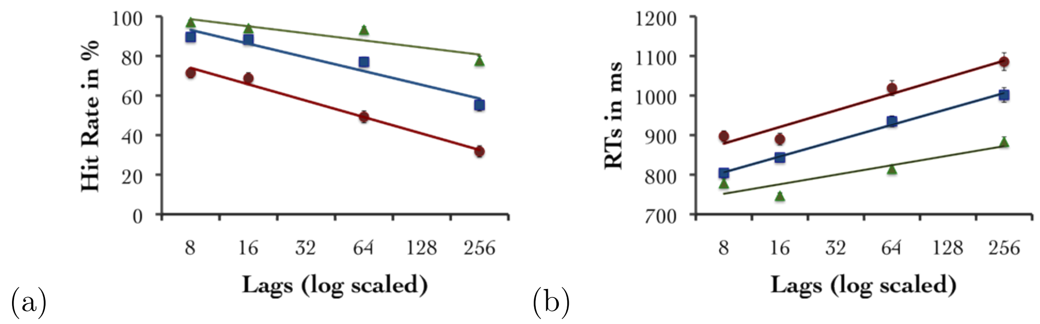


Figure 5-2: (a) The most memorable images remain the most memorable for different lag lengths, and decay at a slower rate even for longer lags. (b) Response times follow the opposite pattern.

5.3 Differences in memorability show up early and change over time

HR decreased with lower image memorability[2], $F(13, 2) = 63.33$, $p < 0.01$, $\rho\eta^2 = 0.70$ and with increasing lag, $F(13, 3) = 97.76$, $p < 0.01$, $\rho\eta^2 = 0.63$. We also found an interaction of memorability and lag, $F(13, 6) = 10.22$, $p < 0.05$, $\rho\eta^2 = 0.18$, as the rate of forgetting increased from high-mem images (-3.58), over medium-mem images (-6.94) to low-mem images (-8.30). Hit rates already differed significantly at the shortest lag: high (97%) vs. low (71%), $t(12) = 3.79$, $p < 0.05$. See Fig. 5-3a.

RTs (response times) were measured from scene onset to button press and are only considered for successful recognition trials. Similar to HR, we see a main effect of memorability, $F(12, 2) = 105.72$, $p < 0.01$, $\rho\eta^2 = 0.59$, a main effect of lag, $F(12, 3) = 53.18$, $p < 0.01$, $\rho\eta^2 = 0.61$, and an interaction of memorability and lag, $F(12, 6) = 2.39$, $p < 0.05$, $\rho\eta^2 = 0.09$, with a difference in slopes across lags as a function of memorability: high = 24.07, medium = 40.08, and low = 41.91. RTs already differed significantly at the shortest lag: high (779ms) vs. low (897ms), $t(12) = 9.48$, $p < 0.01$. See Fig. 5-3b.

[2] Note that memorability scores were fixed from the previous study of [31], whereas the HR referred to here is the performance by the eye tracked participants in the current study.

5.4 Pupil size and blink rates are predictive of differences in memorability

The Pupil Old/New Effect (PONE) is measured as the difference in pupillary responses to old images correctly identified as "old" (Hits) versus new images correctly identified as "new" (Correct Rejections, CRs). There was a significant PONE during retrieval with pupils dilating more to Hits vs. CRs ($M = 241$, $t(13) = 5.98$, $p < 0.01$). Differences in image memorability produced a graded PONE response, $F(13, 2) = 3.33$, $p < 0.05$, $\rho\eta^2 = 0.20$, with low-mem images eliciting a greater PONE than high-mem images, $t(13) = 2.23$, $p < 0.05$. The PONE was also modulated by the lag between image encoding and its successful retrieval, $F(13, 3) = 3.98$, $p < 0.05$, $\rho\eta^2 = 0.23$, with an increased PONE at the longest lag compared to lag-16, $t(13) = 2.57$, $p < 0.05$.

The Blink Old/New Effect (BONE) was calculated as the difference in mean blinking rate for Hits versus CRs. There was a significant BONE, with reduced blinking rates for Hits vs. CRs during retrieval ($M = -16\%$, $t(13) = 3.47$, $p < 0.01$). The BONE was further modulated by image memorability, $F(13, 2) = 6.04$, $p < 0.01$, $\rho\eta^2 = 0.32$, with blink rate significantly reduced for low-mem compared to high-mem images, $t(13) = 2.83$, $p < 0.05$. The BONE was not significantly modulated by the lag between image encoding and its successful retrieval, $F(13, 3) = 1.98$, $p = 0.13$, $\rho\eta^2 = 0.13$, but there was a marginally significant decrease of blink rate at lag-256 compared to lag-16, $t(13) = 2.13$, $p = 0.05$. We therefore propose that this **Blink Old/New Effect** could be used as a complementary measure of memory processes.

5.5 Discussion

Image memorability has strong and robust effects on both recognition memory performance and eye activity measures. This study not only corroborated earlier findings that memorability is an intrinsic property of an image that is shared across different viewers and remains stable over time, but also clearly showed that low memorable

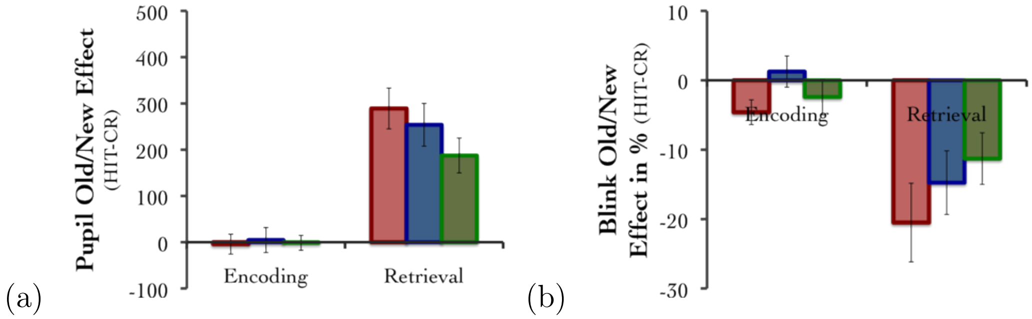


Figure 5-3: (a) Pupils are larger at retrieval than encoding, with the greatest effects elicited for the low-mem images (hardest to retrieve). (b) There are fewer blinks at retrieval than encoding, with the fewest blinks for low-men images.

images compared to highly memorable images: 1) show a decline in recognition performance of over 25% only 20 seconds after initial scene viewing, 2) produce steeper rates of forgetting than more memorable images, and 3) are accompanied by increased pupillary responses and decreased blink rates.

While it is not surprising that some images are less memorable than others, it does seem striking how quickly differences in image memorability become evident. For the advertising business it would be devastating to use images that will be forgotten before a customer could even navigate to the webpage to make a purchase.

In addition to the almost instantaneous decline in memory performance, mental representations of less memorable images also faded away at a faster rate. After only 10 minutes, recognition memory for these images dropped dramatically from 71% to only 32%, while recognition memory for highly memorable images merely declined from 97% to 78% in the same time interval. Moreover, the successful retrieval of scene representations from low memorability images involved more effort, as seen through prolonged RTs, greater pupil dilations, and decreased blink rates.

In sum, we have shown that the intrinsic memorability of an image has both immediate and long-lasting effects on recognition performance and can be tracked using two easily accessible and complementary physiological measures: the pupillary response and endogenous blink rate. Image memorability is therefore indeed mirrored in the eye of the beholder.

Chapter 6

Application to Information Visualizations

Does consistency of image memorability generalize to other domains like information visualizations? What makes a visualization memorable?[1]

6.1 Motivation

In earlier parts of this thesis we have seen how natural scene images tend to be consistently memorable or forgettable across individuals, demonstrating that there are intrinsic factors that contribute to image memorability. Here we extend these findings to the more applied area of visual imagery: information visualization. In the age of Big Data, visualizing all this data becomes a key challenge. Information (or data) visualizations become crucial for communicating ideas, analyses, and findings to company employees at industry meetings, to scientists via academic publications, to students in education settings, and to the general public via news and other media.

We aim to understand if memorability for visualizations is also consistent across a population, and what key factors may contribute to making some visualizations *intrinsically* more memorable than others. We set out to answer the basic question: "What makes a visualization memorable?" Clearly, a more memorable visualization

[1] This chapter is closely related to publication [9]

is not necessarily a more comprehensible one. However, knowing what makes a visualization memorable is a step towards answering higher level questions like "What makes a visualization engaging?" or "What makes a visualization effective?".

We studied the memorability of visualizations as images to better understand their intrinsic memorability. While we did not specifically study the memorability or comprehensibility of the underlying data presented in the visualization in the current work, identifying which type of visual information is *memorable* or *forgettable* provides a basis for understanding a number of cognitive aspects of visualizations. This is because given limited cognitive resources and time to process novel information, capitalizing on memorable displays is an effective strategy. Research in cognitive psychology has shown that conceptual knowledge is an organizing principle for the storage and retrieval of information in memory. For instance, details of a story or a picture that are consistent within an existing schema are more likely to be remembered than those that are not [1, 39]. Recent large-scale visual memory work has shown that existing categorical knowledge supports memorability for item-specific details [39]. In other words, many additional visual details of the image come for free when retrieving memorable items. Understanding the memorability of visualizations provides a baseline for leveraging these cognitive capabilities.

6.2 Related work

Recently, there have been a number of studies aiming to evaluate the impact of embellishments on visualization memorability and comprehension [5, 6, 7, 18, 26, 43, 60]. Bateman et al. conducted a study to test the comprehension and recall of graphs using an embellished version and a plain version of each graph [5]. They showed that the embellished graphs outperformed the plain graphs with respect to recall, and the embellished versions were no less effective for comprehension than the plain versions. There has been some support for the comprehension results from a neurobiological standpoint, as it has been hypothesized that adding "visual difficulties" may enhance comprehension by a viewer [7, 26]. Other studies have shown that the effects of stylistic choices and visual metaphors may not have such a significant effect on perception and comprehension [6, 60]. While there have been studies evaluating memorability and perception of graphical layouts for specific types of graphs, such as the work by Marriott et al. for network diagrams [43], there has not yet been a memorability study to target a wide variety of visualizations.

Moreover, a number of these studies were conducted with a limited number of participants and target visualizations. In some studies the visualization targets were designed by the experimenters, introducing inherent biases and over-simplifications [5, 7, 60]. We reduced our biases by compiling a large database of thousands of real-world visualizations and enrolling a large and diverse set of participants on Amazon's Mechanical Turk. And while previous studies confound perception, recall, and comprehension, we focus purely on memorability of the visualizations as images to remove any obfuscation by other variables.

In our study we apply the same methods of measuring memorability (as described earlier in this thesis, and first developed in [31]) to visualizations. In contrast to the prior work that focused on natural images and real-world objects, visualizations are artificial representations of data. Our study contributes not only to the field of visualization but also adds memorability results for artificial images to the cognitive psychology literature.

6.3 MASSVIS (Massive Visualization) dataset

In order to have a large number of real world examples for our memorability experiment we started by scraping the web to collect 5,693 data visualizations. To ensure a breadth of visualization types, design aesthetics, and visualization domains, we focused on the visualization sources listed in Table 6.1. Of the 5,693 visualizations, only 2,070 single visualizations (i.e., stand-alone visualizations with one panel) were kept for further analysis. Our dataset is called the *MASSive VISualization (**MASSVIS**)* dataset, and will be made publicly available. A thorough discussion of the dataset, as well as a novel taxonomy for visualizations, are provided in [9].

Source	Total (single)	Website(s)	Per website (single)
Government or World Organizations	607 (528)	US Treasury Dept. World Health Organization	141 (117) 464 (411)
News Media	1187 (704)	Wall Street Journal Economist National Post	609 (309) 519 (378) 55 (17)
Infographics	1721 (490)	Visual.ly	1721 (490)
Scientific Publications	2,178 (348)	Nature	2,178 (348)
TOTAL	5,693 (2,070)		

Table 6.1: List of visualization sources, their websites, and the respective number of visualizations in the MASSVIS dataset.

Attribute	**Measure**
Black & White	[yes, no]
Number of Distinct Colors	[1, 2-6, $\geq$7]
Data-Ink Ratio	[good, medium, bad]
Visual Density	[low, medium, high]
Human Recognizable Objects	[yes, no]
Human Depiction	[yes, no]

Table 6.2: Attributes used to label visualizations.

In order to determine which visualization elements affect memorability, we further defined a series of visual attributes (Table 6.2). The first two attributes, “black & white” and “number of distinct colors” are meant to give a general sense of the amount of color in a visualization. A measure of chart junk and minimalism is encapsulated in Edward Tufte’s “data-ink ratio” metric [58], which approximates the ratio of data to non-data elements. The “visual density” rates the overall density of visual elements in the image without distinguishing between data and non-data elements. Finally, we have two binary attributes to identify pictograms, photos, or logos: “human recognizable objects” and “human depiction”. We explicitly chose to have a separate category for human depictions due to prior research indicating that human representations have an effect on memorability [31].

For use in our memorability experiment (Sec. 6.4), we selected a subset of 410 images ($\sim$20% of the single images in our database) to be “target” visualizations, for which we collected memorability scores[2]. The target visualizations were also chosen

[2] Of these, 17 were subsequently filtered out because their aspect ratios were deemed too skewed for the comparison to other visualizations to be fair. Visualizations with aspect ratio greater than 3:1 made the text hard to read, and pictographic elements hard to decipher.

to match the distribution of original visualization sources as well as the distribution of visualization categories of the total 2,070 single visualization population. Thus the target population is representative of the observed diversity of real-world visualization types. Fillers were sampled from the rest of the single visualizations. All images were resized to lie within a maximum dimension of 512×512 pixels (while preserving aspect ratios), so as to fit comfortably into a webpage containing the memorability game.

6.4 Online crowdsourcing experiments

The methodology was the same as for the AMT experiments described in Chap. 2, Sec 2.3. On average, we collected 87 responses (SD: 4.3) per target image. Given the responses collected, for performing a relative sorting of our data instances we used the *d-prime* metric[3], defined in Appendix A. This is a common metric used in signal detection theory, which takes into account both signal (HR) and noise (FAR). We use this as a *memorability score* for our visualizations. A high score will require HR to be high and FAR to be low. This will ensure that visualizations that are easily confused for others (high FAR) will have a lower memorability score.

6.5 Some visualizations are intrinsically more memorable

The scores obtained were HR: M = 55.36%, SD = 16.51% and FAR: M = 13.17%, SD = 10.73%. We also measured the consistency of our memorability scores (using the procedure discussed in 2.6). Averaging over 25 such random half-splits, we obtain Spearman's rank correlations of 0.83 for HR, 0.78 for FAR, and 0.81 for d-prime. This high correlation demonstrates that the memorability of a visualization is a consistent measure across participants, and indicates real differences in memorability between visualizations. Thus, despite possible differences in knowledge and experience levels across participants, memorability is intrinsic to the visualizations.

[3] Also called the *sensitivity index.*

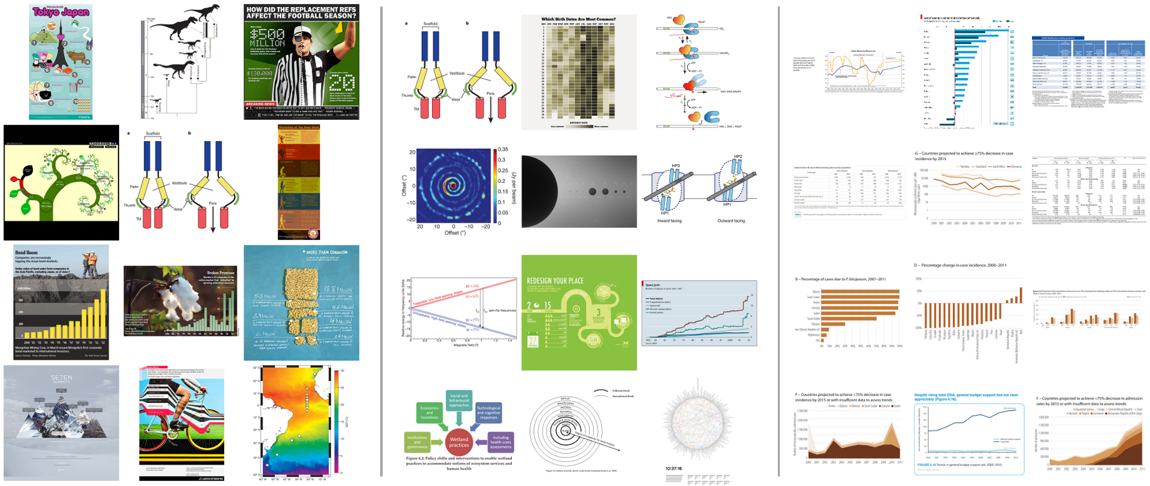

Tokyo Japan
HOW DID THE REPLACEMENT REFS AFFECT THE FOOTBALL SEASON?
$500 MILLION
Which Birth Dates Are Most Common?
REDESIGN YOUR PLACE

Figure 6-1: **Left:** The top 12 overall most memorable visualizations (most to least memorable from top left to bottom right). **Middle:** The top 12 most memorable visualizations without pictograms. **Right:** The bottom 12 least memorable visualizations.

6.6 Visualization attributes are predictive of memorability

Of our 410 target visualizations, 145 contained either photographs, cartoons, or other pictograms of human recognizable objects (here referred to as "pictograms"). Visualizations containing pictograms have on average a higher memorability score (M=1.93) than visualizations without pictograms ($M = 1.14, t(297) = 13.67, p < 0.001$). Thus, just as with scene images, a visualization containing a human recognizable object will more likely be remembered.

Due to this strong main effect of pictograms, we include separate results for visualizations with and without pictograms. As shown in the left-most panel of Fig. 6-1, all but one of the most memorable images (ranked by their d-prime scores) contain human recognizable pictograms. The one visualization without a human recognizable image, the molecular diagram in the middle of the second row, is the most memorable image of our non-pictogram visualizations (see Fig. 6-1, middle panel). The least memorable visualizations are presented in the right-most panel of Fig. 6-1.

As shown in Fig. 6-2a visualizations with 7 or more colors have a higher memorability score ($M = 1.71$) than visualizations with 2-6 colors ($M = 1.48, t(285) = 3.97, p < 0.001$), and even more than visualizations with 1 color or black-and-white gradient ($M = 1.18, t(220) = 6.38, p < 0.001$). Across visualizations without pictograms, the

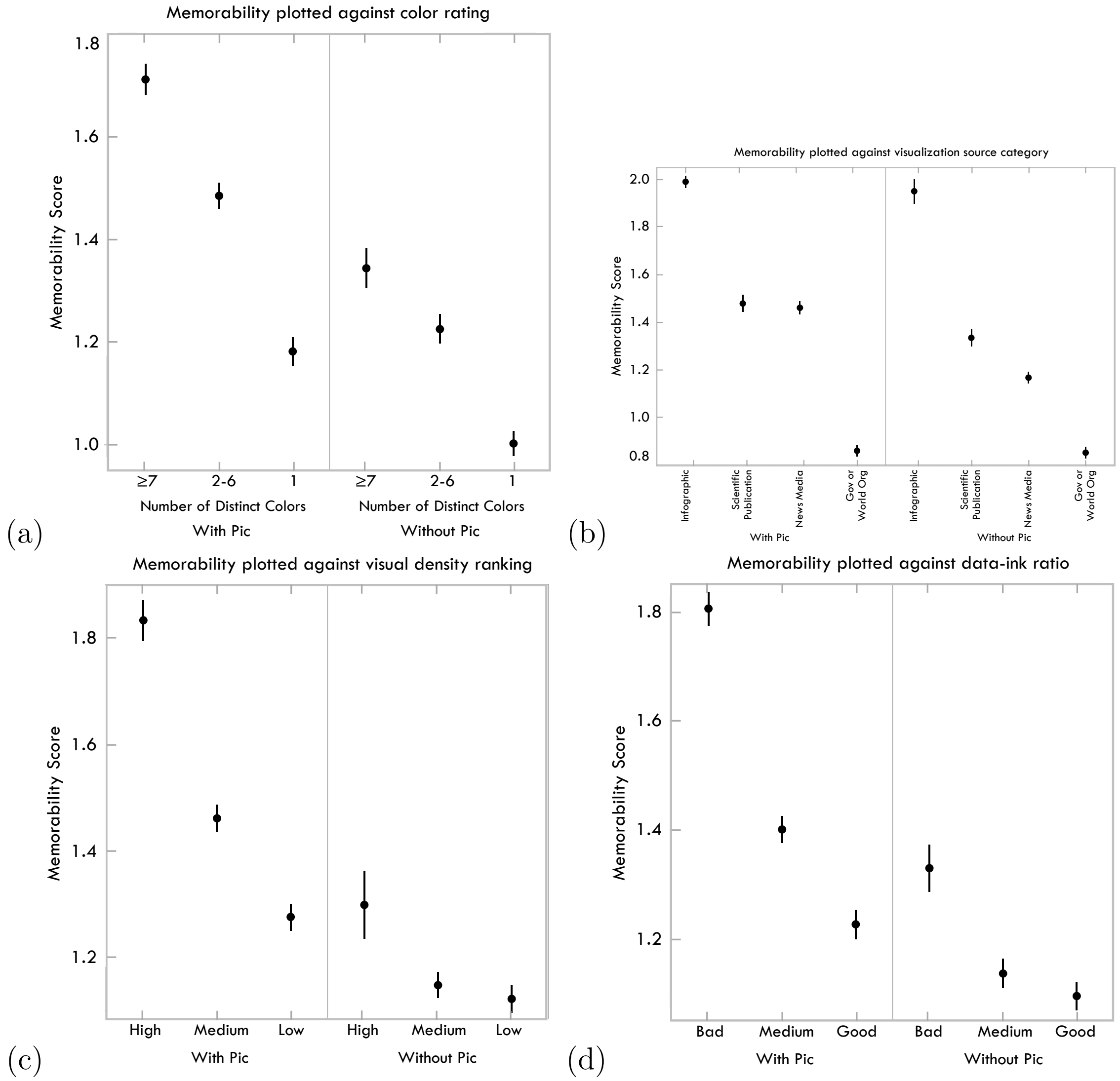


Figure 6-2: (a) Memorability scores for visualizations based on the number of colors contained. (b) Memorability scores for visualizations based on original source category. (c) Memorability scores for visualizations based on visual density. (d) Memorability scores for visualizations based on the data-to-ink attribute ratings. Across all 4 plots: the left side corresponds to all visualizations, and the right to visualizations without pictograms.

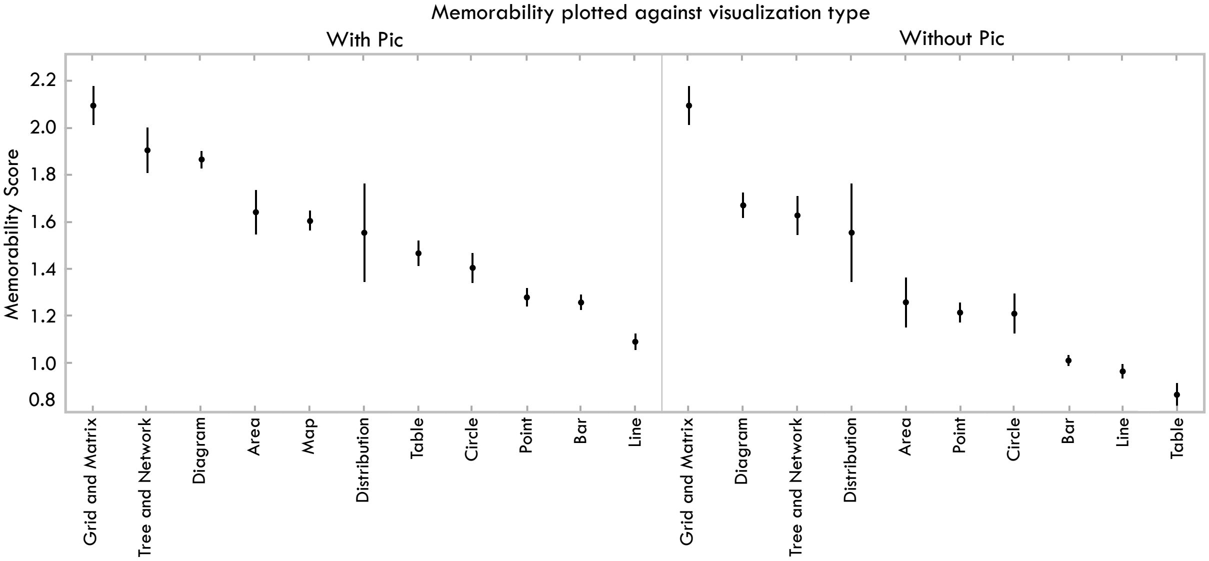


Figure 6-3: Memorability scores for visualizations based on visualization type. The left side corresponds to all visualizations, and the right to visualizations without pictograms.

difference between visualizations with 7 or more colors ($M = 1.34$) and those with 1 color ($M = 1.00$) remains statistically significant ($t(71) = 3.61, p < 0.001$). Across all visualizations, a high visual density rating of "3" has higher memorability ($M = 1.83$) than a low visual density rating of "1" ($M = 1.28, t(115) = 6.08, p < 0.001$) - see Fig. 6-2c. We also observed a significant effect of data-to-ink ratio on memorability scores with a "bad" ($M = 1.81$), i.e., low data-to-ink ratio, being higher than a "good" rating ($M = 1.23, t(208) = 6.92, p < 0.001$) - see Fig. 6-2d. The 3 levels of data-ink ratio are pairwise significantly different from each-other (according to corrected t-tests). As shown in Fig. 6-3, diagrams were statistically more memorable than points, bars, lines, and tables. These trends remain even across visualizations without pictograms, other than some minor ranking differences (e.g. tables without pictograms becomes least memorable).

The middle panel of Fig. 6-1 displays the most memorable visualizations that do not contain pictograms. Why are these visualizations more memorable than the ones in the right-most panel? Qualitatively they are higher contrast, have more color, and are easier to see and discriminate as images. Another possible explanation is that "distinct" types of visualizations, such as diagrams, are more memorable than

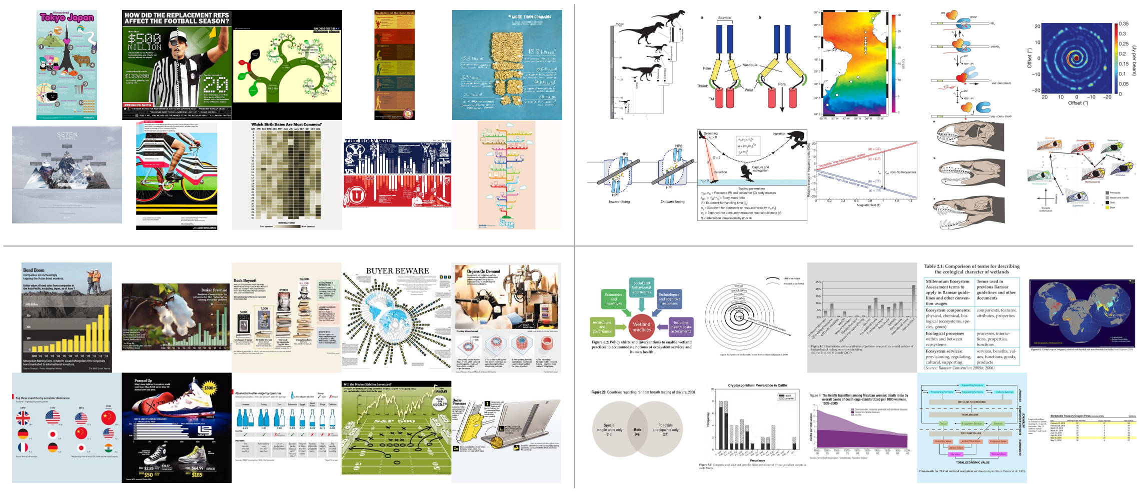

Tokyo Japan
HOW DID THE REPLACEMENT REFS AFFECT THE FOOTBALL SEASON?
$500 MILLION
Which Birth Dates Are Most Common?
THE HOLY WAR
BUYER BEWARE
Organs On Demand

Figure 6-4: The top ten most memorable visualizations for each of the four visualization source categories: infographic (top left), scientific publications (top right), news media (bottom left), and government / world organization (bottom right). In each quadrant, the visualizations are ordered most to least memorable from top left to bottom right.

"common" types of visualizations, such as bar charts. This trend is also evident in Fig. 6-3 in which grid/matrix, trees and networks, and diagrams have the highest memorability scores and tend to all look different from one another, whereas bar charts and line graphs have the lowest memorability and are uniform with limited visual variability (e.g., all the bar charts look alike). Another contributing factor is that target visualizations represented a distribution of types found "in the wild." Thus, of the 410 target visualizations, trees and networks totaled 11 targets and grid/matrix totaled 6 targets. Their low frequency may have contributed to their distinctiveness.

6.7 How visualization memorability differs by publication source

As shown in Fig. 6-2b, regardless of whether the visualizations did or did not include pictograms, infographic visualizations were the most memorable ($M = 1.99, t(147) = 5.96, p < 0.001$ when compared to the next highest category, scientific publications with $M = 1.48$), while the least memorable were the government and world organizations visualizations ($M = 0.86, t(220) = 8.46, p < 0.001$ when compared to the next lowest category, news media with $M = 1.46$). These results were significant according to corrected t-tests, for visualizations with and without pictograms. In fact, with pictograms removed, scientific publications ($M = 1.95$) become significantly more memorable than news media ($M = 1.17, t(23) = 6.92, p < 0.001$). The top ten most memorable visualizations from each source category are shown in Fig. 6-4.

Note that the infographic visualizations from Visual.ly come from a more design-focused venue, and are intentionally created to be flashy and to include stylized elements. These visualizations are pre-judged by people before being published, and thus compete for the viewer's attention. These visualizations are more likely to be bright, bold, and contain pictorial visual elements to grab a reader's attention. Thus this type of publication venue's motivational bias may translate into design features that lead to higher memorability.

Another possible influence of visualization source is venue-specific aesthetics. Many visualizations, particularly those from the news media and government sources, tend to publish with the same visual aesthetic style. This may be due to either the venue maintaining a consistent look so viewers will automatically recognize that a visualization was published by them, or because they have editorial standards to create visualizations that appear similar. This may have a negative impact on memorability scores because visualizations of similar aesthetics lack distinctiveness.

6.8 Discussion

The results of our memorability experiment show that like scenes and faces, visualizations are consistently memorable across people, which may hint at generic, abstract, features of human memory. In particular, the inclusion of human-recognizable objects enhances memorability. And similar to previous studies we found that visualizations with low data-to-ink ratios and high visual densities (i.e., more chart junk) were more memorable than minimal, "clean" visualizations. We found that distinct visualization types (pictoral, grid/matrix, trees and networks, and diagrams) had significantly higher memorability scores than common graphs (circles, area, points, bars, and lines). Overall, novel and distinct visualizations can be better remembered, and this finding is consistent with results for natural scenes and objects.

Understanding what makes a visualization memorable is only the first step to understanding how to create effective data presentations. Making a visualization more memorable means making some part of the visualization “stick” in the viewers mind. We do not want just any part of the visualization to stick, but rather we want the most important relevant aspects of the data or trend the author is trying to convey to stick. If we can accomplish this, then we will have a method for making data more memorable. This will have diverse applications in education, business, and more generally, in how data is presented to wide audiences.

Chapter 7

Conclusion

7.1 Contributions and discussion

In this thesis, the intrinsic and extrinsic effects on image memorability have been thoroughly investigated and quantified. With regards to intrinsic effects, we have been able to show high consistency in memorability scores. Specifically that:

- consistency exists at the within-category level, demonstrated for each of 21 different indoor and outdoor scene categories (Chap. 2)
- even whole scene categories (or image collections) can be consistently more memorable that others (Chap. 2)
- consistency holds across experimental settings and different participant populations (Chap. 3)
- images most memorable after shorter time intervals are also most memorable after longer time intervals (Chap. 5)
- there is also consistency for non-natural images - i.e. information visualizations (Chap. 6)

All of these findings suggest that there is a component of image memory intrinsic to the images themselves, making automatic prediction a very real possibility. Intrinsic effects are not, however, sufficient for predicting image memorability at a per-trial level. Extrinsic effects like the context in which an image appears or the particular observer, will influence whether or not an image will be remembered or forgotten on a given trial. Thus, intrinsic image memorability is modulated by extrinsic effects.

In Chap. 3, we presented an information-theoretic framework for modeling the context of an image collection using automatically-computed visual features. We have applied this framework to a large collection of natural scenes (the 9K images in the FIGRIM dataset, presented in Chap. 2). By systematically varying image context between AMT 1 (Chap. 2) and AMT 2 (Chap. 3), we have been able to quantify how and when context affects memorability. Although previous memory studies have indicated that items that are distinct with respect to their context are better remembered, we have been able to quantify this intuition in a fully-automatic manner using our large scene dataset. Moreover, we have shown that more variable contexts are more memorable overall. Thus as one increases the variety or distinctiveness of the images in a collection, one can increase the number of images that can be remembered. Does this mean that performance on image recognition tasks can increase indefinitely as long as the images being presented together (in the same context) are sufficiently different? This is probably not the case due to a possible saturation effect - see figure 7-1.

To further consolidate these points, consider the comparison across a number of image memorability experiments presented in table 7.1. Note that there is consistency in the distribution of memorability scores (the average HR and FAR scores) across experiments. For experiments composed of a single stimuli category - faces [3], visualizations (MASSVIS), within-scene experiment (FIGRIM), the HR and FAR scores are very similar. Likewise, the scores are also similar for experiments composed of multiple different stimuli (scene) types - the many-scene experiments of [31] and the across-scene experiment (FIGRIM). Thus, there seem to be some natural bounds to the number of images that can be remembered for a given diversity of image context. With further experiments, it would be interesting to determine the exact function of memorability with changes in context variability.

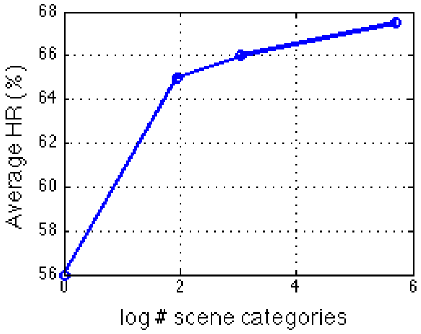


Figure 7-1: Average memorability scores for scene contexts composed of different numbers of scene categories: 1 (AMT 1), 7 (in-lab), 21 (AMT 2), over 300 [31]. Thus, as the variability of images in a given image context increases, the memorability scores go up (more images can be remembered). However, memory performance is not likely to increase indefinitely, eventually reaching a plateau.

Dataset	targets	fillers	datapts per target	mean HR (%)	mean FAR (%)	HR cons. (ρ)	FAR cons. (ρ)
FIGRIM (Chap. 2)	1754	7296	74	66.0 (SD: 13.9)	11.1 (SD: 9.5)	0.74	0.72
Isola [31]	2222	8220	78	67.5 (SD: 13.6)	10.7 (SD: 7.6)	0.75	0.66
Faces [3]	2222	6468	82	51.6 (SD: 12.6)	14.4 (SD: 8.7)	0.68	0.69
MASSVIS (Chap. 6)	410	1660	87	55.4 (SD: 16.5)	13.2 (SD: 10.7)	0.83	0.78

Table 7.1: A comparison of the memorability scores across different datasets, showing consistency in results and stability of memory performance. Additionally note that for the *FIGRIM* dataset, when each category was separately tested, the average memorability scores over 21 categories were: 56.0% (SD : 4.2%) for HR and 14.6% (SD : 2.0%) for FAR, showing consistency with the instance-based databases of faces and visualizations.

Another set of questions addressed in this thesis is how physiological markers such as eye movements, pupil dilations, and blinks, can serve as indicators of memorability. In Chapter 4, we developed a computational model to predict, given an individual's fixations on an image, whether the individual will remember the image at a later time point. Thus, how people look at an image can be informative of how (and whether) they encode it, and whether they can later successfully retrieve it. In Chapter 5 we have shown that pupils dilate more, and blink rates decrease, during the retrieval of a lower memorability image than during the retrieval of a more memorable image. Both of these physiological markers have been found to be indicative of cognitive effort, and this fits our observations that lower-memorability images take more effort to retrieve.

Taken together, all of the findings presented in this thesis can contribute to a single model of memorability, with both intrinsic and extrinsic effects taken into account. Importantly, since high consistency can be found across experiments and participant populations, automatic prediction becomes a possibility. A fully computational model of memorability is then only a few steps away.

7.2 Future applications

Previous studies have shown that image memorability can be computationally predicted from image features [31], opening up applications such as automatically generating memorability maps for images [38], modifying image memorability [36, 37], and designing better data visualizations [9].

Taking into account the extrinsic effects discussed in this paper will lead to more complete models that are better able to approximate human performance on specific memory tasks. Apart from the extrinsic effects we have discussed in this paper, other ones can affect the memorability of individual images, including the observer's expertise, time spent studying each image, attention biases, etc. How memorable something is may additionally be affected by its *familiarity* and *utility*. Note that familiarity, which involves multiple repetitions of an item, has not been considered in our studies but is an important factor in natural environments. The effect of familiarity on memory has a long history in psychology [64, 15, 32, 4]. Utility would correspond to how important a given item is to the observer. For instance, faces have high utility, and images with faces have been found to be more memorable. It remains to be understood and computationally modeled how exactly all these factors combine to make an image more or less memorable.

Building extrinsic effects into memorability models will open up new application areas for the customization of visual material, including user interfaces and educational tools. Imagine an automatic system that monitors the eye movements of a student on a set of lecture slides or data presentations and uses this information to determine whether or not the student is properly encoding the content. If not, the system may either alert the student to increase attentiveness at this point in time, or else the system may continue to re-present the material again until it has acquired some confidence that the student has finally mastered the content.

In the case of all of the physiological markers presented in this thesis, no overt response from a human is required, and prediction can be made automatically. A finer-grained understanding of how these physiological markers vary with memorability, as well as the consideration of additional physiological markers (pulse, sweat, etc.), can open doors to even more applications.

Appendix A

Memorability Measurements

Here we include the definitions for different memorability measurements:

$$\begin{aligned}
\mathrm{HR(I)} &= \frac{\mathrm{hits(I)}}{\mathrm{hits(I)} + \mathrm{misses(I)}} \times 100\% \\
\mathrm{FAR(I)} &= \frac{\text{false alarms(I)}}{\text{false alarms(I)} + \text{correct rejections(I)}} \times 100\% \\
\mathrm{ACC(I)} &= \frac{\mathrm{hits(I)} + \text{correct rejections(I)}}{\mathrm{total(I)}} \times 100\% \\
\mathrm{DPRIME(I)} &= Z(\mathrm{HR}) - Z(\mathrm{FAR})
\end{aligned}$$

where Z is the inverse of the cumulative Gaussian distribution and:

$$\mathrm{total} = \mathrm{hits(I)} + \mathrm{misses(I)} + \text{false alarms(I)} + \text{correct rejections(I)}$$

Additionally, given the following 2×2 matrix:

$\frac{\mathrm{hits(I)}}{\mathrm{total(I)}}$	$\frac{\mathrm{misses(I)}}{\mathrm{total(I)}}$
$\frac{\text{false alarms(I)}}{\mathrm{total(I)}}$	$\frac{\text{correct rejections(I)}}{\mathrm{total(I)}}$

Mutual information (between a response and whether an image was a repeat) is calculated as:

$$\mathrm{MI(I)} = \sum_i \sum_j p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \text{ (where } i \text{ and } j \text{ index into the matrix above)}$$

Bibliography

[1] R. C. Anderson and J. W. Pichert. Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, 17(1):1–12, 1978.

[2] F. Attneave. *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results.* Holt, Rinehart & Winston; 1st Edition, 1959.

[3] W. A. Bainbridge, P. Isola, and A. Oliva. The Intrinsic Memorability of Face Images. *Journal of Experimental Psychology: General*, 142(4):1323–1334, 2013.

[4] J. C. Bartlett, S. Hurry, and W. Thorley. Typicality and familiarity of faces. *Memory & Cognition*, 12(3):219–228, 1984.

[5] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI '10*, pages 2573–2582. ACM, 2010.

[6] A. F. Blackwell and T.R.G. Green. Does metaphor increase visual language usability? In *Visual Languages, 1999. Proceedings. 1999 IEEE Symposium on*, pages 246–253. IEEE, 1999.

[7] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768, 2012.

[8] A. Borji and L. Itti. Defending Yarbus: eye movements reveal observers' task. *Journal of Vision*, 14(3):1–22, 2014.

[9] M. Borkin, A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What Makes a Visualization Memorable? In *IEEE Transactions on Visualization and Computer Graphics (Infovis)*, 2013.

[10] J. Bradshaw. Pupil size as a measure of arousal during information processing. *Nature*, 216(5114):515–516, 1967.

[11] T. F. Brady and A. Oliva. Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. *Psychological science*, 19:678–685, 2008.

[12] K.H. Brodersen, C.S. Ong, K.E. Stephan, and J.M. Buhmann. The balanced accuracy and its posterior distribution. In *International Conference on Pattern Recognition (ICPR)*, 2010.

[13] V. Bruce, A.M. Burton, and N. Dench. What's distinctive about a distinctive face? *The Quarterly Journal of Experimental Psychology*, 47(1):119–141, 1994.

[14] A. Bulling and D. Roggen. Recognition of visual memory recall processes using eye movement analysis. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2011.

[15] T. Busey. Formal models of familiarity and memorability in face recognition. In M.J. Wenger & J.T. Townsend, editor, *Computation, geometric and process perspectives on facial cognition: Contexts and challenges*. Lawrence Erlbaum Associates, Inc., 2001.

[16] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. Under review.

[17] M. W. Eysenck. Depth, elaboration, and distinctiveness. In *Levels of processing in human memory*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1979.

[18] S. Few. The chartjunk debate: A close examination of recent findings. Visual Business Intelligence Newsletter, 2011.

[19] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2):1–17, 2008.

[20] M. Glanzer and J.K. Adams. The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16:5–16, 2010.

[21] S. D. Goldinger and M. H. Papesh. Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, 21(2):90–95, 2012.

[22] E. Granholm and S. R. Steinhauer. Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, 52(1):1–6, 2004.

[23] M.R. Greene, T. Liu, and J.M. Wolfe. Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, 62:1–8, 2012.

[24] E. H. Hess and J. M. Polt. Pupil size as related to interest value of visual stimuli. *Science*, 132(3423):349–350, 1960.

[25] A. Hollingworth. Scene and position specificity in visual memory for objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32:58–69, 2006.

[26] J. Hullman, E. Adar, and P. Shah. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2213–2222, 2011.

[27] R. R. Hunt and J. B. Worthen. *Distinctiveness and Memory.* Oxford University Press, New York, 2006.

[28] A. Ihler and M. Mandel. Kernel Density Estimation Toolbox for MATLAB (R13). `http://www.ics.uci.edu/~ihler/code/kde.html`. Accessed: 2014-07-21.

[29] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the Intrinsic Memorability of Images. In *Conference on Neural Information Processing Systems (NIPS)*, 2011.

[30] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(7):1469–1482, 2014.

[31] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[32] L. L. Jacoby. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30(5):513–541, 1991.

[33] Y. Jia. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. `http://caffe.berkeleyvision.org/`, 2013.

[34] A. Kafkas and D. Montaldi. Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. *The Quarterly Journal of Experimental Psychology*, 64(10):1971–1989, 2011.

[35] D. Kahneman and J. Beatty. Pupil diameter and load on memory. *Science*, 154(3756):1583–1585, 1966.

[36] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva. Modifying the Memorability of Face Photographs. In *International Conference on Computer Vision (ICCV)*, 2013.

[37] A. Khosla*, J. Xiao*, P. Isola, A. Torralba, and A. Oliva. Image Memorability and Visual Inception. In *SIGGRAPH Asia Technical Briefs*, 2012. * indicates equal contribution.

[38] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of Image Regions. In *Conference on Neural Information Processing Systems (NIPS)*, 2012.

[39] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Conceptual Distinctiveness Supports Detailed Visual Long-Term Memory for Real-World Objects. *Journal of Experimental Psychology: General*, 139:558–578, 2010.

[40] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science*, pages 1–7, 2010.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Conference on Neural Information Processing Systems (NIPS)*, 2012.

[42] M. Mancas and O. Le Meur. Memorability of natural scene: the role of attention. In *IEEE International Conference on Image Processing (ICIP)*, 2013.

[43] K. Marriott, H. Purchase, M. Wybrow, and C. Goncu. Memorability of visual features in network diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2477–2485, 2012.

[44] M. Naber, S. Frassle, U. Rutishauser, and W. Einhauser. Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *Journal Of Vision*, 13(2):11–11, 2013.

[45] D. L. Neumann and O. V. Lipp. Spontaneous and reflexive eye activity measures of mental workload. *Australian Journal of Psychology*, 54(3):174–179, 2002.

[46] D. Noton and L. Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, 11(9):929–942, 1971.

[47] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision (IJCV)*, 42(3):145–175, 2001.

[48] S. C. Otero, B. S. Weekes, and S. B. Hutton. Pupil size changes during recognition memory. *Psychophysiology*, 48(10):1346–1353, 2011.

[49] G. Porter, T. Troscianko, and I. D. Gilchrist. Effort during visual search and counting: Insights from pupillometry. *Quarterly journal of experimental psychology*, 60(2):211–229, 2007.

[50] K. A. Rawsona and J. P. Van Overscheldeb. How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, 58(3):646–668, 2008.

[51] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014.

[52] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision (IJCV)*, 77(1–3):157–173, 2008.

[53] S.R. Schmidt. Encoding and retrieval processes in the memory for conceptually distinctive events. *Journal of Experimental Psychology: learning, memory, cognition*, 11(3):565–578, 1985.

[54] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(1):185–197, 2010.

[55] G. J. Siegle, N. Ichikawa, and S. Steinhauer. Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, 45(5):679–687, 2008.

[56] L. Standing. Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology*, 25(2):207–222, 1973.

[57] B. W. Tatler, N. J. Wade, H. Kwan, J. M. Findlay, and B. M. Velichkovsky. Yarbus, eye movements, and vision. *i-Perception*, 1(1):7–27, 2010.

[58] E. Tufte. *Envisioning Information*. Cheshire (Conn.), 1990.

[59] T. Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, 43(2):161–204, 1991.

[60] A. Vande Moere, M. Tomitsch, C. Wimmer, B. Christoph, and T. Grechenig. Evaluating the effect of style in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2739–2748, 2012.

[61] M. L.-H. Võ, Z. Bylinskii, and A. Oliva. Image memorability in the eye of the beholder: Tracking the decay of visual scene representations. Under review.

[62] M. L.-H. Võ, A. M. Jacobs, L. Kuchinke, M. Hofmann, M. Conrad, A. Schacht, and F. Hutzler. The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, 45(1):130–140, 2008.

[63] S. Vogt and S. Magnussen. Long-term memory for 400 pictures on a common theme. *Experimental Psychology*, 54(4):298–303, 2007.

[64] J.R. Vokey and J.D. Read. Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory Cognition*, 20(3):291–302, 1992.

[65] H. von Restorff. Bereichsbildungen im Spurenfeld (The effects of field formation in the trace field). *Psychologische Forschung*, 18:299–342, 1933.

[66] S. Wiseman and U. Neisser. Perceptual organization as a determinant of visual recognition memory. *The American Journal of Psychology*, 87(4):675–681, 1974.

[67] J. Xiao, J. Hayes, K. Ehinger, A. Oliva, and A. Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[68] A.L. Yarbus. *Eye movements and vision.* Plenum Press, New York, 1967.

[69] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In *Conference on Neural Information Processing Systems (NIPS)*, 2014.