Vision Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Intrinsic and extrinsic effects on image memorability

Zoya Bylinskii [a,b,*], Phillip Isola [b,c], Constance Bainbridge [b], Antonio Torralba [a,b], Aude Oliva [b]

[a] Department of Electrical Engineering and Computer Science, MIT, Cambridge 02141, USA
[b] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge 02141, USA
[c] Department of Brain and Cognitive Sciences, MIT, Cambridge 02141, USA

ARTICLE INFO

Article history:
Received 4 August 2014
Received in revised form 3 March 2015
Available online xxxx

Keywords:
Image memorability
Eye movements
Scene dataset
Fine-grained categories
Visual distinctiveness
Context

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. Here 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. Second, we test two extrinsic factors: image context and observer behavior. 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 automatically predict how changes in context change the memorability of natural images. In addition to context, we study a second extrinsic factor: where an observer looks while memorizing an image. It turns out that eye movements provide additional information that can predict whether or not an image will be remembered, on a trial-by-trial basis. Together, by considering both intrinsic and extrinsic effects on memorability, we arrive at a more complete and fine-grained model of image memorability than previously available.

1. Introduction

Recent work on image memorability has shown that independent of observer, certain images are consistently remembered and others forgotten (Bainbridge, Isola, & Oliva, 2013; Borkin et al., 2013; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011; Isola et al., 2014), indicating that memorability is an intrinsic property of images that can be estimated with computer vision features (Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011; Isola et al., 2014; Khosla, Xiao, Torralba, et al., 2012, Khosla et al., 2013). These previous image memorability studies raise a number of questions, including: does the consistency of human memory generalize? How might extrinsic effects such as context and observer differences affect image memorability?

In this paper, we report that: (1) human consistency at remembering and forgetting images holds at a within-category level, and (2) extrinsic effects predictably affect whether an image will be later remembered or forgotten. Here we consider the effects of the context in which images are seen, as well as the observer's eye movement patterns on a trial-by-trial basis.

Previous work on image memorability has not computationally addressed either image context or trial-by-trial observer behavior. Moreover, although many decades of prior research on memory have considered context and the effects of item/image distinctiveness of memorability (Hunt & Worthen, 2006; Konkle et al., 2010; Nairne, 2006; Standing, 1973), these effects have not been rigorously quantified on large datasets of natural scenes. Prior work has relied on subjective human judgments of distinctiveness (Bainbridge et al., 2013; Konkle et al., 2010). In contrast, we provide an objective, automatic measure: we model distinctiveness as an information-theoretic property computable from raw visual data.

For our studies, we collected the *FIne-GRained Image Memorability (***FIGRIM***) dataset*[1] composed of over 9K images, which we used to test human memory performance on 21 different scene categories, each containing hundreds of images. We used this dataset to collect memorability scores for 1754 target images, whereby we systematically varied the image context. In this paper we refer to the set of images from which the experimental sequence is sampled as **image context**. 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

* Corresponding author at: 32-D542, 32 Vassar St., Cambridge, MA 02141, USA.
E-mail address: zoya@mit.edu (Z. Bylinskii).

[1] We are publicly releasing the full *FIGRIM* dataset with popular image 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 and different scene categories. Available at: <http://figrim.mit.edu/>.

http://dx.doi.org/10.1016/j.visres.2015.03.005
0042-6989/

with image memorability scores. We discuss which images are most affected by context to gain a better understanding of the interplay between intrinsic and extrinsic factors on image memorability.

To account for additional extrinsic effects caused by the variation in observer behavior from trial to trial, we collected eyetracking data for over 2.7 K of the *FIGRIM* images. For 630 target images and using eye movements alone we can predict, on a trial-to-trial basis, which images will be remembered and which forgotten with 66% accuracy. Thus, we demonstrate how eye movements have predictive power on a trial-by-trial basis for image memorability.

2. Background

2.1. Image memorability

Recent work in image memorability (Bainbridge et al., 2013; Borkin et al., 2013; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011; Isola et al., 2014) 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 (Isola, Xiao, et al., 2011), and later extended to narrower classes of images – faces (Bainbridge et al., 2013) and visualizations (Borkin et al., 2013). In this paper, we show that this consistency is not a special property of the stimuli considered, and that it can not be explained away by a simple distinction between images (e.g. indoor scenes tend to be memorable, outdoor scenes forgettable). We demonstrate that the high consistencies hold within 21 different indoor and outdoor scene categories, each consisting of hundreds of instances. This is the first image memorability study to consider fine-grained scene categories. Previous studies have shown that image memorability can be computationally predicted from image features (Isola, Xiao, et al., 2011) which opens up applications such as automatically generating memorability maps for images (Khosla, Xiao, Torralba, et al., 2012), modifying image memorability (Khosla et al., 2013; Khosla, Xiao, Isola, 2012), and designing better data visualizations (Borkin et al., 2013). In this paper, we additionally model extrinsic effects on memorability, which have not yet been explored in the image memorability literature, and can open up new application areas.

2.2. Distinctiveness in visual long-term memory

Previous studies have suggested that items that stand out from (and thus do not compete with) their context are better remembered (Attneave, 1959; Eysenck, 1979; Hunt & Worthen, 2006; Konkle et al., 2010; Rawsona & Overscheldeb, 2008; Schmidt, 1985; Standing, 1973; Wiseman & Neisser, 1974; Vogt & Magnussen, 2007; von Restorff, 1933). For instance, Standing observed a large long-term memory capacity for images that depict oddities (Standing, 1973). Konkle et al. demonstrated that object categories with conceptually distinctive exemplars showed less interference in memory as the number of exemplars increased (Konkle et al., 2010). 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 (Bartlett, Hurry, & Thorley, 1984; Bruce, Burton, & Dench, 1994; Valentine, 1991). In the domain of data visualizations, Borkin et al. noticed that unique visualization types had significantly higher memorability scores than common graphs and that novel and unexpected visualizations were better remembered (Borkin et al., 2013). In this paper, we quantify the intuitions that distinctive images are more memorable using an information theoretic framework, and we compute the distinctiveness of images with reference to their image context (the set of images from which the experimental sequence is sampled). We steer away from subjective human ratings, and instead compute statistics over automatically-extracted image features. By systematically varying the image context across experiments, we are able to computationally model the change in context at the feature level, and predict corresponding changes in image memorability.

2.3. Memorability and visual attention

Little work has considered the intersection between image memorability and visual attention (Bulling & Roggen, 2011; Foulsham & Underwood, 2008; Mancas & Le Meur, 2013; Noton & Stark, 1971). Mancas and Le Meur (2013) used saliency features to show a slight improvement over the automatic image memorability predictions in Isola, Xiao, et al. (2011). 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. Thus, Mancas et al. used saliency to improve a population predictor. We, instead, use eye-movements to improve the trial-by-trial predictions of memory for specific individuals (an **individual trial predictor**). Bulling and Roggen (2011) used eye movement features to predict image familiarity, classifying whether images have been seen before or not. They assumed that all images seen again are remembered, particularly due to the long exposure times (10 s) used per image, and by testing on a small dataset of 20 faces. They also used 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*.

2.4. Decoding task using eye movements

Our work is also related to recent studies on the use of eye movements for decoding an observer's task (Borji & Itti, 2014; Greene, Liu, & Wolfe, 2012). These studies considered 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 (Yarbus, 1967). These studies utilized a very small set of images (ranging from 15 to 64) with a very constrained theme (grayscale photographs taken between 1930 and 1979 with at least two people (Greene et al., 2012); paintings depicting "an unexpected visitor" (Borji & Itti, 2014)). In our study, we measured the eye movements of participants on 630 target images sampled from 21 different indoor and outdoor scene categories. We extracted 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 were able to solve our decoding task using only 2 s of viewing time per image, whereas the previous studies worked with durations of 10 s (Bulling & Roggen, 2011; Greene et al., 2012), 30 s (Borji & Itti, 2014), 50 s (Tatler et al., 2010), and 60 s (Borji & Itti, 2014). For this purpose, we learned image-specific classifiers to distinguish fixations on one image versus fixations on other images.

3. Memorability experiments

3.1. FIGRIM dataset

We created a novel dataset by sampling high-resolution (at least 700 × 700 px) images from 21 different indoor and outdoor

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005

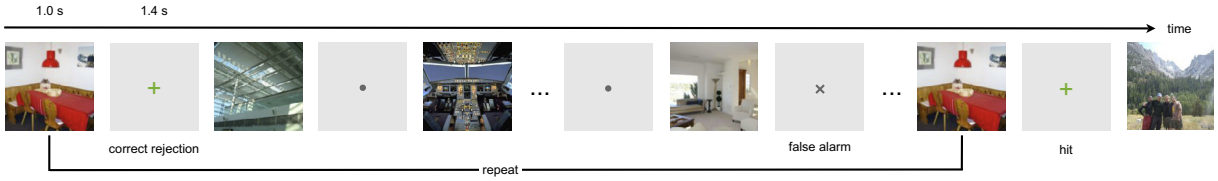


Fig. 1. An example AMT experimental sequence. During image presentation, the participant presses a key if the image has already appeared in the sequence, and receives feedback at the end of the image presentation. A false alarm occurs when on first presentation, the participant indicates that the image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

scene categories[2] from the SUN Database (Xiao et al., 2010). Image duplicates and near-duplicates were manually removed.[3] The images were downsampled and cropped to 700×700 px.[4] From each scene category, 25% of the images were randomly chosen to be *targets* and the rest of the images became *fillers* (Table 1 in the appendix lists the number of targets and fillers per scene category). The targets are the images for which we obtained memorability scores.

3.2. AMT 1: within-category experiment

We ran Amazon Mechanical Turk (AMT) studies following the protocol of Isola, Xiao, et al. (2011) 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[5] where sequences of 120 images (a mix of target and filler images sampled from a *single* scene category) were presented for 1 s 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 s.[6] Some filler images repeated at much shorter intervals of 1–7 images and were used as vigilance tests to recognize when a participant was not paying attention to the game. Participants were instructed to press a key when they detected an image repeat, at which point they received feedback (a red or green cross). 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. Fig. 1 depicts an example experimental sequence.

On average, 80 participants saw each target image and its repeat, providing us with enough data points per image to collect reliable statistics about the memorability of each image.[7] 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 incorrect and correct responses (respectively) to target images shown for the first time. We define *hit rate* (**HR**) and *false alarm rate* (**FAR**):

$$\mathrm{HR(I)} = \frac{\mathrm{hits(I)}}{\mathrm{hits(I)} + \mathrm{misses(I)}} \times 100\% \quad (1)$$

$$\mathrm{FAR(I)} = \frac{\mathrm{false\ alarms(I)}}{\mathrm{false\ alarms(I)} + \mathrm{correct\ rejections(I)}} \times 100\% \quad (2)$$

We also define $\overline{\mathbf{HR}}$ and $\overline{\mathbf{FAR}}$ to be category averages – computed over all images belonging to a single category. The $\overline{\mathrm{HR}}$ scores vary from 49.5% to 64.2% ($M = 56.0\%,\ SD = 4.2\%$).[8] $\overline{\mathrm{FAR}}$ scores vary between 10.2% and 18.9% ($M = 14.6\%,\ SD = 2.0\%$), following a partial *mirror effect* (Glanzer & Adams, 2010; Vokey & Read, 1992), where high HR are often accompanied by low FAR. The Spearman rank correlation between the $\overline{\mathrm{HR}}$ and $\overline{\mathrm{FAR}}$ scores is -0.66 ($p < 0.01$). Note that this is to be expected by signal detection theory as sensitivity increases: targets and distractors become more discriminable, leading simultaneously to high HR and low FAR. Memorability scores for all the categories can be found in Table 1, and for comparison, memorability scores from other experiments are included in Table 3 (in the appendix). For instance, a previous experiment that combined images from hundreds of scene categories (Isola, Xiao, et al., 2011) reported average HR and FAR scores of 67.5% and 10.7%, respectively.

Fig. 2 includes a sample of some of the most memorable and forgettable images from 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 Section 5.

3.3. AMT 2: across-category experiment

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 dataset statistics in Table 2, appendix). The average memorability scores for this experiment are: HR: $M = 66.0\%,\ SD = 13.9\%$, FAR: $M = 11.1\%,\ SD = 9.5\%$.

3.4. In-lab control experiment

We replicated the AMT experiments in the lab using a subset of 630 target images. 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. The exact experimental set-up can be found in the appendix. 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

[2] We chose all SUN scene categories with at least 300 images of the required dimensions.

[3] We calculated the Gist descriptor (Oliva & Torralba, 2001) 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.

[4] Images were later resized to 512×512 px for the online AMT experiments (to fit comfortably in browser windows), and to 1000×1000 px for the eyetracking experiments.

[5] Compliance with the Declaration of Helsinki is acknowledged in Section 8.

[6] Images and repeats occurred on average 4.5 min apart, thus allowing us to capture memory processes well beyond short-term and working memory.

[7] AMT participant demographics are discussed in Ross et al. (2010).

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

Fig. 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.

21 in AMT 2), we increase 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 Fig. 3, 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.

4. Intrinsic effects on memorability

4.1. Some images are intrinsically more memorable, even at the category level

Previous studies have demonstrated that memorability is consistent across participant populations for a general set of scene images (HR: $\rho = 0.75$, FAR: $\rho = 0.66$) (Isola, Xiao, et al., 2011) and for the specific classes of faces (HR: $\rho = 0.68$, FAR: $\rho = 0.69$) (Bainbridge et al., 2013) and data visualizations (HR: $\rho = 0.83$, FAR: $\rho = 0.78$) (Borkin et al., 2013). Here these results are

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005

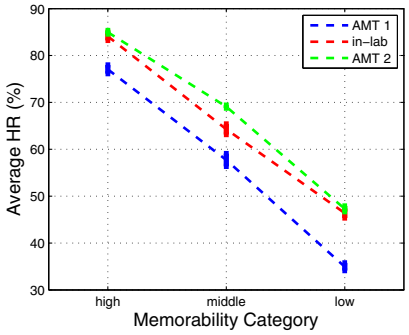


Fig. 3. Memorability scores for images in the context of 21 scenes (AMT 2) are higher than in the context of 7 scenes (in-lab), and higher still than in the context of 1 scene (AMT 1). At the same time, the most memorable images remained the most memorable, and the most forgettable remained the most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

extended to the fine-grained category level for a variety of scene categories.

The consistencies of the image memorability scores were measured separately for each of the scene categories (see Table 1 in the appendix for all the values). 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 were averaged over 25 such half-splits of the participant data. For all of the scene categories, consistency of HR scores ranged from 0.69 to 0.86 and from 0.79 to 0.90 for FAR scores. These high correlations demonstrate that memorability is a consistent measure across participant populations, indicating real differences in memorability across images.

4.2. Some scene categories are intrinsically more memorable

How consistent is the relative ranking (the ordering in Table 1) of the scene categories? For instance, if we selected a different subset of images, would the average memorability of the amusement park images still 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 of images, 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.

4.3. Image memorability is consistent across experiments

Per-image memorability scores measured in AMT 2 also correlated 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.

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.

5. Context effects on memorability

A large body of literature suggests that items that stand out from their context are better remembered (Attneave, 1959; Eysenck, 1979; Hunt & Worthen, 2006; Konkle et al., 2010; Rawsona & Overscheldeb, 2008; Schmidt, 1985; Standing, 1973; Vogt & Magnussen, 2007; von Restorff, 1933; Wiseman & Neisser, 1974). However, recent work on predicting image memorability (Isola, Xiao, et al., 2011; Isola, Parikh, et al., 2011; Khosla, Xiao, Torralba, et al., 2012) has largely ignored the effects of image context on memory performance.

By systematically varying the context for our target images between AMT 1 and AMT 2, we directly measure context effects on image memorability. We use state-of-the-art computer vision features within an information-theoretic framework to quantify context differences and image distinctiveness. We are able to rigorously quantify, using our large-scale natural scene database, the observation that images that are unique or distinct with respect to their image context are better remembered.

5.1. 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 estimated the probability distribution over the images in an image's context (in some feature space). The distinctiveness of an image is its negative log likelihood under this distribution. We considered 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 a given context, we used kernel density estimation (Ihler & Mandel, 2014).

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

$$P_c(f_i) = \frac{1}{\|C\|}\sum_{j\in C} K(f_i - f_j) \tag{3}$$

where K can be any kernel function, and $\|C\|$ indicates the size of the context, measured in number of images. We used an Epanechnikov kernel and leave-one-out-cross-validation to select the kernel bandwidth.[9] The features come from a convolutional neural network (CNN), a popular feature space recently shown to outperform other features in computer vision (Krizhevsky, Sutskever, & Hinton, 2012; Razavian et al., 2014). Specifically, we used the **Places-CNN** from Zhou et al. (2014) trained to classify scene categories. We used the 4096-dimensional features from the response of the Fully-Connected Layer 7 ($fc7$) of the CNN, which is the final fully-connected layer before producing class predictions. We reduced this feature vector to 10 dimensions using PCA to prevent overfitting and increase efficiency in estimating the kernel densities.

Our results are not restricted to this feature space, and hold more generally. In particular, we obtained similar (though weaker) trends when using the simpler Gist descriptor (Oliva & Torralba, 2001), for which we provide results in the Supplemental Material. In contrast to simple visual descriptors like Gist, the deep features are trained to predict image semantics.[10] This is supported by research showing that conceptual (semantic) similarity is more

[9] Bandwidth selection was performed just once on all the images across all the scene categories, and this same bandwidth was used for estimating the distributions for each category.

[10] Some visualizations of these features can be found in Zhou et al. (2014, 2015).

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

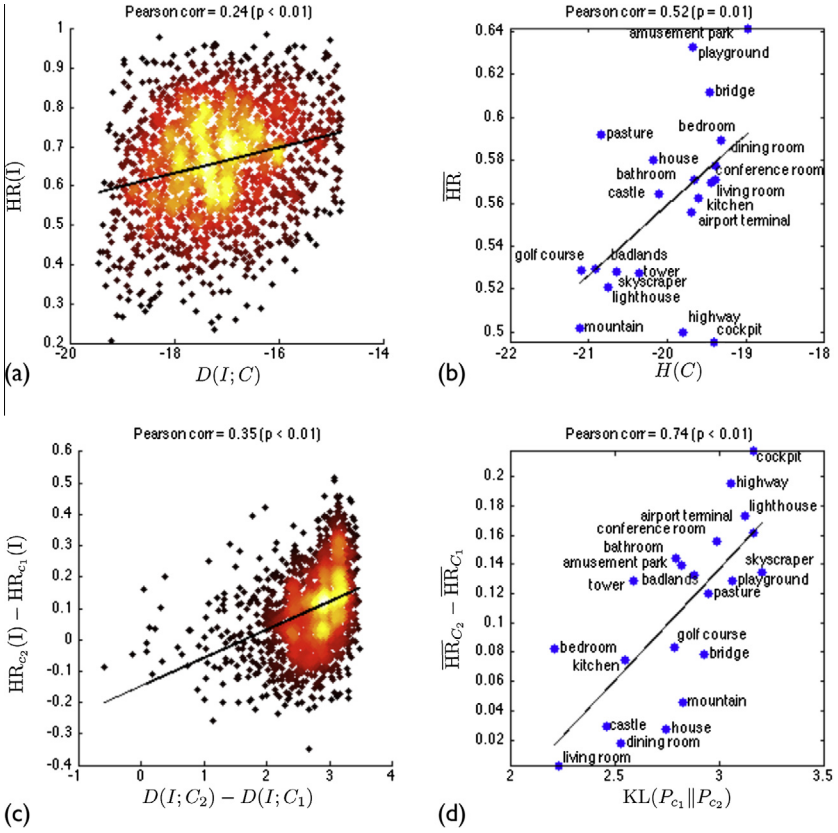


Fig. 4. The effects of context on memorability. In figures (a) and (c), each dot is a single target image from the *FIGRIM* dataset, for a total of 1754 images. Brighter coloring represents a greater density of points. In figures (b) and (d), all images in a given category are collapsed into a single 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.

predictive of long term visual memory performance than perceptual similarity (Brady, Konkle, & Alvarez, 2011; Konkle et al., 2010).

In Fig. 4a, the memorability score of an image, HR(I), is correlated with its distinctiveness with respect to the image context, $D(I;C)$. Mathematically, we define[11]:

$$D(I;C) = -\log P_c(f_i) \tag{4}$$

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 found that $D(I;C_2)$ is positively correlated with HR(I) (Pearson $r = 0.24$, $p < 0.01$), as plotted in 4a. The correlation also holds when images are compared to images within the same category (correlation between $D(I;C_1)$ and HR(I) is $r = 0.26$, $p < 0.01$). Thus, more *contextually distinct* images are more likely to be memorable.

The Supplemental Material contains the same analyses on alternative measurements of memorability: *d-prime*, *mutual information*, and *accuracy*. $D(I;C)$, the distinctiveness of an image I with respect to its context C, remains positively correlated with these alternative measurements of memorability.

5.2. More varied image contexts are more memorable overall

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

$$\begin{aligned} H(C) &= \mathbb{E}_c[D(I;C)] \\ &= \mathbb{E}_c[-\log P_c(f_i)] \end{aligned} \tag{5}$$

Here, $\mathbb{E}_c$ is expectation over image context C. As in Fig. 4b, the Pearson correlation between $H(C)$ and $\overline{\mathrm{HR}} = \mathbb{E}_c[\mathrm{HR(I)}]$, is $r = 0.52$ ($p = 0.01$). Thus, categories that contain many contextually distinct images are more memorable overall. For instance, the *mountain* category, which has one of the lowest $H(C)$ values, contains a relatively stable collection and configuration of scene elements: mountains and sky. On the other hand, the *amusement park* category, which has the highest $H(C)$ value, consists of a much larger variability of images: images of roller-coasters, concession stands, and rides. Thus entropy in feature space can explain some of the differences in average HR we observe across categories in AMT 1.

5.3. Changing image context can change image memorability

AMT experiments 1 and 2 systematically varied the context for images, while keeping the images constant. This allowed us to isolate the effects of context from other possible confounds.[12] To model the change in context, we computed the difference in the distinctiveness of an image relative to its own scene category versus all scene categories. In Fig. 4c we see that changing the context of an

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

[12] 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.

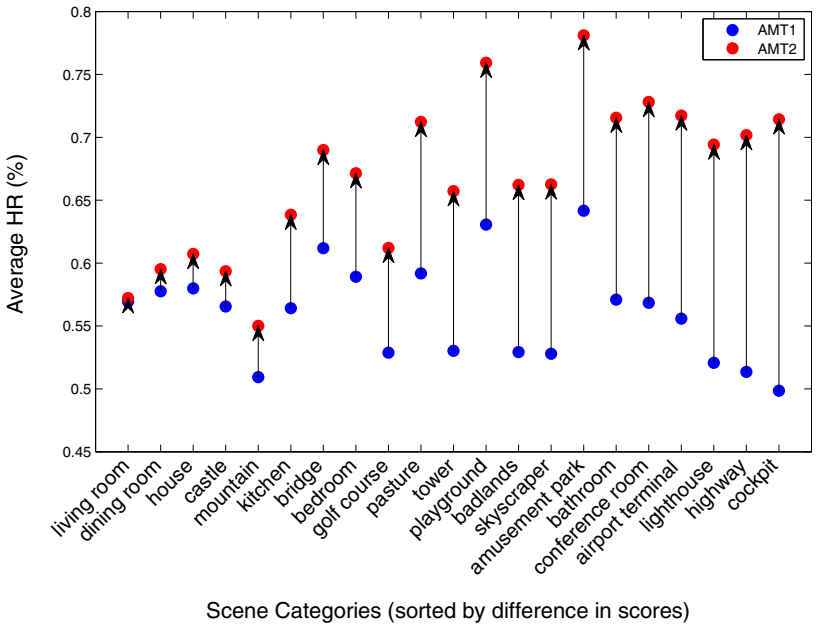


Fig. 5. The average memorability of the images in each scene category went up when images were presented in the context of images from other scene categories (AMT 2) compared to when they were presented only in the context of images from the same category (AMT 1).

image to make it more distinct relative to the context increases its memorability. The Pearson correlation between $D(I;C_2) - D(I;C_1)$ and $\mathrm{HR}_{C_2}(I) - \mathrm{HR}_{C_1}(I)$ is 0.35 ($p < 0.01$).

We also considered differences in memorability at the category level. In Fig. 5 we see that across all categories, $\overline{\mathrm{HR}}$ 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 measured this change in context as the *Kullback–Leibler divergence* between the density functions computed over contexts C_1 and C_2 as:

$$\mathrm{KL}(P_{c_1}||P_{c_2}) = \mathbb{E}_{c_1}[-\log P_{c_2}(f)] - \mathbb{E}_{c_1}[-\log P_{c_1}(f)] \quad (6)$$

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 AMT 1. In Fig. 4d, the Pearson correlation between the change in context entropy and change in memorability is $r = 0.74$ ($p < 0.01$). Consider the *cockpit* category, with the greatest $KL(P_{c_1}||P_{c_2})$ value: many of the cockpit images look alike. 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*, with one of the lowest $KL(P_{c_1}||P_{c_2})$ values, that often look like *living rooms* and *kitchens*, and thus are not as visually distinct when combined with images from these other scene categories.

5.4. Atypical category exemplars are most affected by context

Another way of looking at the distinctiveness story is through a discriminative lens (as an alternative to the generative information theoretic framework presented in the previous sections). Consider the images that were memorable with respect to their own category, but became more forgettable when combined with other categories. In Fig. 6, we can see that these images tend to look more like other categories than their own category.

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. 6. Notice that for the images that decreased in memorability when combined with other categories, the predicted labels are more likely to be incorrect. Compare this to the images that increased in memorability when combined with other categories – they are more likely to be correctly classified.

We also consider the probability, under the scene classifier, of the correct category label. These probabilities are included with each image in Fig. 6. 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 have memorability scores that are least affected by context? Images that are distinct with respect to many contexts – in this case, those that are distinct from their own category, but do not look like images from other categories either. For example, the images in the top right quadrant in Fig. 7 are memorable across contexts. Take for example the bridge in front of the red sky. It is clearly an image of 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, for applications where one intends an image to be robustly memorable, one must consider the different contexts in which this image is likely to occur and ensure the image will stand out from all these contexts.

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005

Fig. 6. We evaluated a scene classifier on the images whose memorability changed when combined with other categories. We show 3 categories (the rest are in the Supplemental Material). For each image, we include the classifier's predicted category label and the probability of the correct category label (where * is replaced with the correct category). Images more likely to be confused with other categories were the ones that dropped most in memorability.

6. Eyetracking experiments

We used a similar set-up to the in-lab experiment from Section 3.4, but with important differences to collect eye-movements in an un-biased manner.[5] Images were presented to participants at 1000 × 1000 px. 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.[13] Images remained on the

[13] Although we do not explicitly use the data from the third repetition, we note here that 78% of the time participants forgot the image on the second repetition, they remembered it on the third repetition. Thus in an application setting, if we can automatically predict when a participant will forget an image, we can show the image again to improve memorability performance.

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005

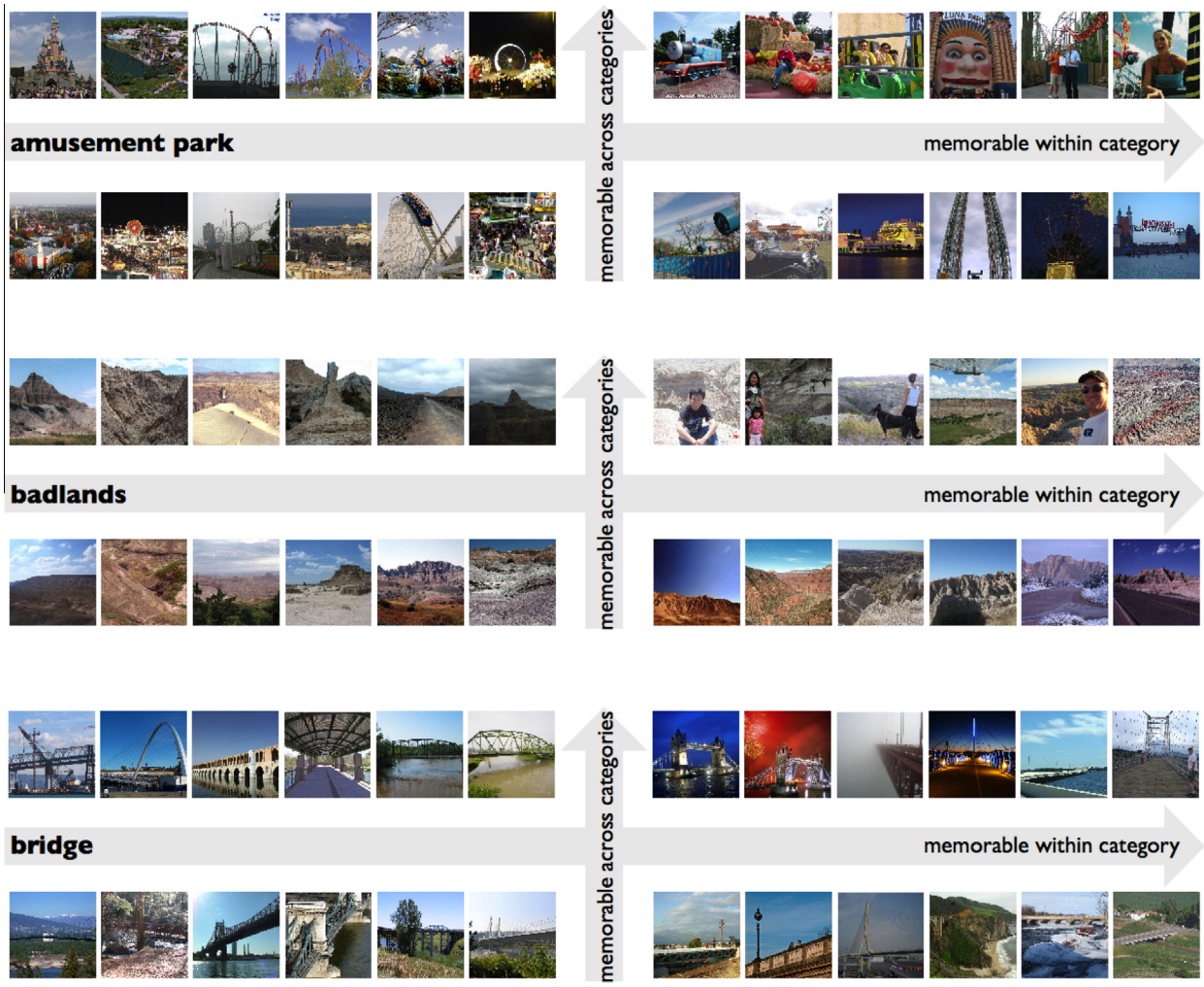


Fig. 7. Memorability scores of images in the top right quadrant of each plot are least affected by context whereas the scores of images in the bottom right quadrant are most affected by context. Images in the top right are distinct with respect to both contexts, while images in the bottom right are distinct only with respect to their own category.

screen for 2 s, and participants gave a forced-choice response *at the end* of each image presentation to indicate whether the image appeared previously or not. After a keypress response and verbal feedback, a fixation cross came on the screen for 0.4 s, followed by the next image. See Fig. 8 for an example experimental sequence.

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 image stimuli subtended 30 degrees of visual angle. 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 min, and participants could take regular breaks throughout. All participant eye-fixations and keypresses were recorded. We recruited a total of 40 participants for our study ($M = 14.1,\ SD = 1.2$ participants per image), 24 of which were female, with overall mean age 21.2 years ($SD = 3.3$). The memorability scores for this experiment were: HR: $M = 75.8\%,\ SD = 14.4\%$, FAR: $M = 5.2\%,\ SD = 7.4\%$.

7. Observer effects on memorability

Can the experience, behavior, or other characteristics of a specific individual on a specific trial be used to make more accurate predictions about memory performance than by using population estimates? Here our goal is to make predictions on a trial-by-trial basis, using an individual's eye-movements to determine if an image will be later remembered.

7.1. 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 Fig. 9. We constructed 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,[14] we converted an observer's fixations on an image into a **fixation map** by binning the fixations into a

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

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005

Fig. 8. An example eyetracking experimental sequence. Differences from the AMT experiment in Fig. 1 include the slightly longer image presentation times, the collection of key presses *after* image presentation at the prompt, and the forced-choice response.

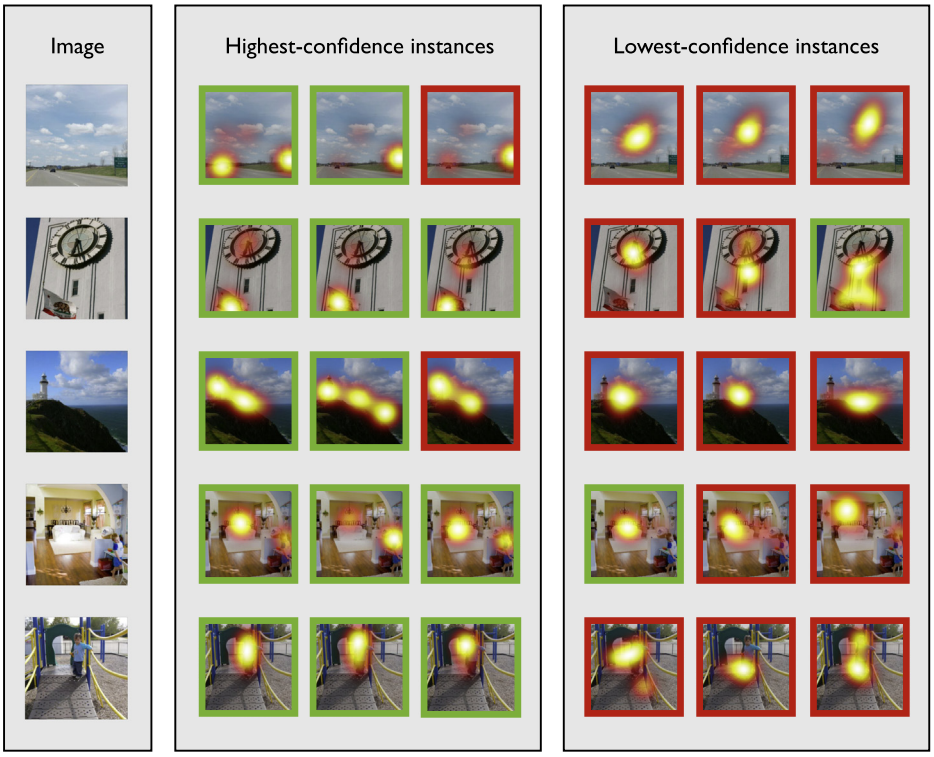


Fig. 9. Examples of individual viewers' fixation maps (at encoding) overlaid on top of the images viewed. For each of these 5 example images, we include the 3 highest-confidence and 3 lowest-confidence instances under the image's classifier (trained to differentiate fixations on this image from fixations 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

20×20 grid, normalizing the binned map, and smoothing it by convolution with a Gaussian with $\sigma = 2$ grid cells. Coarse sampling and smoothing was necessary to regularize the data.

For each image, we trained 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 considered **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 used a RUSBoost classifier (Seiffert et al., 2010), which handles the class imbalance problem,[15] and **balanced accuracy** as a metric of performance because it avoids inflated performance estimates on datasets with unequal numbers of positives and negatives (Brodersen et al., 2010). It is calculated as:

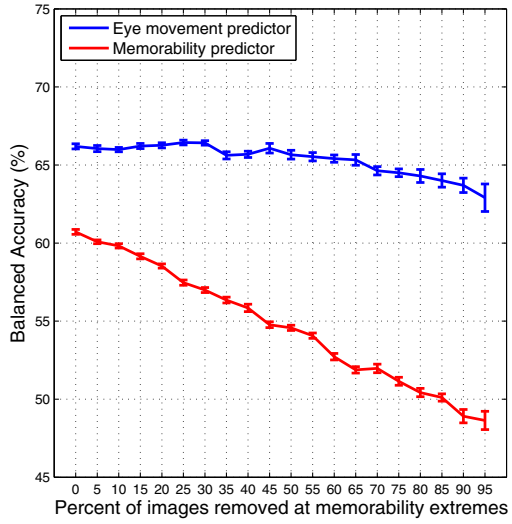
$$\text{balanced accuracy} = \frac{0.5 \times \text{true positives}}{\text{true positives} + \text{false negatives}} + \frac{0.5 \times \text{true negatives}}{\text{true negatives} + \text{false positives}} \quad (7)$$

[15] 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.

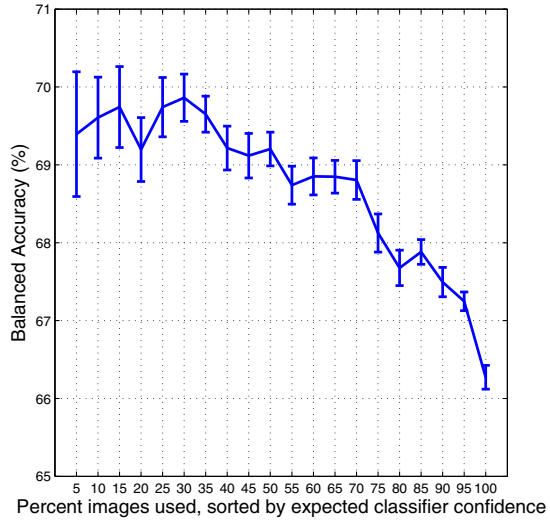
Over 5 train-test splits, the balanced accuracy of our classifier on determining whether a set of fixations comes from a specific image vs some other image is 79.7% (*SD*: 13.9%), where chance is at 50%. This high performance indicates that we are able to successfully learn diagnostic fixation patterns 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 (see Section 7.4).

7.2. Eye movements are predictive of whether an image will be remembered

As demonstrated in the AMT memorability studies, people are highly consistent in which images they remember and forget. Thus as a baseline we use an image's memorability score (i.e. HR from AMT 2) to make trial-by-trial predictions for whether a particular individual will remember a particular image. We refer to this as a population predictor because these memorability scores are obtained by averaging over participants. This predictor achieves an accuracy of 60.09% (*SD*: 1.55%) at making trial-by-trial predictions, significantly above chance (50%). However, this predictor will not be robust across all images (see Section 7.3), and



(a)



(b)

Fig. 10. (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.

thus we use the model developed in the previous section to construct a better predictor that takes into account the individual trial.

Our individual trial predictor uses a viewer's eye movements to predict whether an image will be remembered. During the training phase, we first learn a classifier G_i to differentiate fixations on image I from fixations on other images (as discussed in Section 7.1). Next, we evaluate both successful and unsuccessful fixations on image I under the classifier G_i to obtain confidence values for each set of fixations. We perform a grid search over 200 values to pick a threshold on the confidence values that maximizes balanced accuracy on differentiating successful from unsuccessful fixations.[16] At test time, for a held-out set of participants, we evaluate a participant's encoding fixations under the classifier G_i to obtain a confidence value. We threshold this confidence value with the threshold chosen during training to produce the final prediction: whether the participant's fixations are successful or unsuccessful.

Over 15 different splits of participant 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 the 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.[17]

7.3. Individual differences are key when population predictors fall to chance

Consider the cases where images are not consistently memorable or forgettable across individuals. We sorted images by their AMT 2 scores, and progressively removed images at the memorability extremes. The resulting prediction performance is plotted in Fig. 10a. Memorability scores fell to chance at predicting individual trials precisely because the images at the memorability extremes were most predictive. Meanwhile, our eye movement features retained 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".

7.4. Not all images are equally predictable

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. Thus, we may want to separate images into those on which confident predictions can be made from those on which prediction will be difficult. 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).

Sorting the images by this expected confidence measure (see Fig. 11), we obtain the results in Fig. 10b. Our classifier makes the best predictions on the images for which the training data was easily separable (corresponding to high expected confidence), achieving a balanced accuracy of almost 70% on the test data – i.e. new participants.

Thus, it is possible to automatically select images that our classifier is expected to do well on. This becomes an important feature for applications where we have a choice over the images that can be used, and need to have a system to robustly predict from eye fixations, whether an image will be later remembered.

[16] Note that this two-step learning process was chosen to alleviate the problems of overfitting to insufficient data. Separating successful fixations on an image from fixations on all other images produces a much more robust decision boundary than when directly separating successful from unsuccessful fixations. However, since the final task is to separate successful from unsuccessful fixations on a single image, we add an additional step to adjust only a single scalar parameter to make the final prediction.

[17] *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 look around), they may be off-center or even off-the-image (the viewer is distracted), or they may be randomly distributed over the image (the viewer is not concentrated on 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.

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005

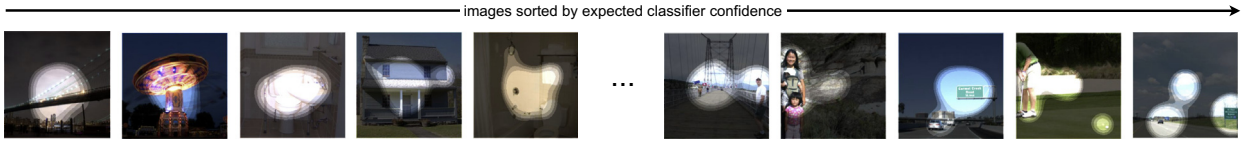


Fig. 11. Images sorted by expected classifier confidence (from least to most). A classifier with high confidence on its positive training examples will do better at differentiating successful from unsuccessful fixations on an image. Overlaid on top of each image is the average fixation map computed over all successful encodings of the image.

7.5. Future applications

The types of investigations presented in this paper have potential applications for education as they provide us with (1) tools to understand the settings in which images can be best remembered (and how memory can change across contexts), and (2) tools to predict an individual's memory for images, thus opening up avenues for customization and intervention. Memorability experiments on information visualizations (Borkin et al., 2013) demonstrate generalization of findings to non-scene stimuli – specifically, data presentation. 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 "paying attention". 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 properly encoded the content.

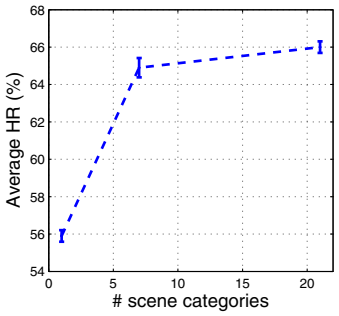


Fig. 12. Average memorability scores for scene contexts composed of different numbers of scene categories: 1 (AMT 1), 7 (in-lab), 21 (AMT 2). It is interesting to note as an additional data point that in the memorability experiment by Isola, Xiao et al. (2011) with hundreds of scene categories, the average HR is 67.5%. 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.

8. Conclusion

In this paper we have replicated and extended previous findings that memorability scores are highly consistent across participants, which suggests there is a component of image memory intrinsic to the images themselves. We have shown that this consistency holds at the *within*-category level, for a total of 21 different indoor and outdoor scene categories. Additionally, high consistency exists *across* experiments, with varying contexts, experimental set-ups, and participant populations (online and in-lab).

Nevertheless, we have also demonstrated how (and in which cases) extrinsic effects can alter the memorability of images. We have presented an information-theoretic framework for predicting the effect of context on memorability. Images that are most distinct from their image context are the most memorable, and image contexts with the highest entropy have the highest overall memorability scores. 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 Fig. 12.

We have also considered cases where memorability scores may not be sufficient for predicting trial-by-trial memory performance. Specifically for images that are not at the memorability extremes, we can provide better predictions by taking into account the extrinsic effects of the observer. We introduced a model for using the eye movements of a viewer when first presented with an image to predict whether the image will be later remembered. Thus we offer an application of memorability decoding from eye movements. 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 (Curby, Glazek, & Gauthier, 2009; Herzmann & Curran, 2011; Vicente & Wang, 1998), temporal effects (Isola et al., 2014; McGaugh, 1966), attentional and task biases (Chun & Turk-Browne, 2007; Tuckey & Brewer, 2003; Walker, Vogl, & Thompson, 1997), etc. How memorable something is may additionally depend on 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 (Bartlett et al., 1984; Busey, 2001; Jacoby, 1991; Mandler, 2008; Vokey & Read, 1992). Utility would correspond to how important a given item is to the observer, and is related to expertise. For instance, faces have high utility, and images with faces have been found to be more memorable (Isola, Xiao, et al., 2011; Isola et al., 2014). It remains to be understood how exactly all these factors combine to make an image more or less memorable.

We can thus consider intrinsic image memorability as a starting point (base level), and all the extrinsic effects as modifiers that finally determine whether or not a particular image will be remembered on a particular trial. Considered from another perspective, intrinsic image memorability is an average across contexts, observers, and settings, where these extrinsic effects are effectively marginalized out.

Understanding and modeling all the effects on memory are necessary to build computational models that will more accurately predict image memorability for specific settings or users. With high prediction accuracies, many interesting applications become possible, including customizable user interfaces and educational tools.

Compliance with the declaration of Helsinki

All studies have been conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). MIT IRB was attained. Participants read about the experiment before giving written consent to participate, and could opt out at any time. They were compensated for their time. AMT anonymized all participant data, including names.

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005

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

Category	Targets	Fillers	$\overline{HR}$ (%)	$\overline{FAR}$ (%)	HR cons. (ρ)	FAR cons. (ρ)	Datapoints/Target
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
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.

Category	Targets	Fillers	Datapoints/target	$\overline{HR}$ (%)	$\overline{FAR}$ (%)	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
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: 55.9% (*SD*: 4.2%) for HR and 14.6% (*SD*: 2.0%) for FAR, showing consistency with the instance-based databases of faces (Bainbridge et al., 2013) and visualizations (Borkin et al., 2013).

Dataset	Targets	Fillers	Datapoints/target	Mean HR (%)	Mean FAR (%)	HR cons. (ρ)	FAR cons. (ρ)
FIGRIM	1754	7296	74	66.0 (*SD*: 13.9)	11.1 (*SD*: 9.5)	0.74	0.72
Isola (Isola, Xiao, et al., 2011)	2222	8220	78	67.5 (*SD*: 13.6)	10.7 (*SD*: 7.6)	0.75	0.66
Faces (Bainbridge et al., 2013)	2222	6468	82	51.6 (*SD*: 12.6)	14.4 (*SD*: 8.7)	0.68	0.69
Visualizations (Borkin et al., 2013)	410	1660	87	55.4 (*SD*: 16.5)	13.2 (*SD*: 10.7)	0.83	0.78

Acknowledgments

We thank Bolei Zhou for his help generating Places-CNN features, Wilma Bainbridge for helpful discussion, and anonymous reviewers for many helpful suggestions. This work is supported by the National Science Foundation under Grant No. 1016862 to A.O, as well as Google, Xerox, and MIT Big Data Initiative at CSAIL awards to A.O and A.T. Z.B. is supported by a Postgraduate Doctoral Scholarship from the Natural Sciences and Engineering Research Council of Canada. P.I. was supported by a National Science Foundation Graduate Research Fellowship.

Appendix

Procedure for in-lab experiment

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.[5] We recruited a total of 29 participants for our study ($M = 14.8$, $SD = 2.4$ participants per image), 16 of which were female, with overall mean age 24.9 years ($SD = 3.8$). In a single session, a participant would see a sequence of about 1000 images, of which 210 were targets that repeated exactly once in the sequence, spaced apart by 91–109 images. Images in the test sequence were presented for 2 s each, separated by a fixation cross lasting 0.5 s. 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. Images were presented on a 19 inch CRT monitor with a resolution of 1280×1024 pixels, 22 inches from the chinrest mount. Images subtended 30° of visual angle.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2015.03.005.

References

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results* (1st ed.). Rinehart & Winston.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face images. *Journal of Experimental Psychology: General, 142*(4), 1323–1334.

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

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision, 14*(3), 1–22.

Borkin, M., Vo, A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A. et al. (2013). What Makes a Visualization Memorable? In: *IEEE Transactions on Visualization and Computer Graphics (Infovis)*.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision, 11*(5), 1–34.

K. Brodersen, C. Ong, K. Stephan, J. Buhmann, The balanced accuracy and its posterior distribution, in: International Conference on Pattern Recognition (ICPR), 2010.

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

Bulling, A., & Roggen, D. (2011). Recognition of visual memory recall processes using eye movement analysis. In *International joint conference on pervasive and ubiquitous computing (UbiComp)*.

Busey, T. (2001). Formal models of familiarity and memorability in face recognition. In M. W. J. Townsend (Ed.), *Computation, geometric and process perspectives on facial cognition: Contexts and challenges*. Lawrence Erlbaum Associates Inc.

Chun, M. C., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology, 17*, 177–184.

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 94–107.

Eysenck, M. W. (1979). Depth, elaboration, and distinctiveness. In *Levels of processing in human memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Fixations during encoding and recognition. *Journal of Vision, 8*(2), 1–17.

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

Greene, M., Liu, T., & Wolfe, J. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research, 62*, 1–8.

Herzmann, G., & Curran, T. (2011). Experts memory: An ERP study of perceptual expertise effects on encoding and recognition. *Memory & Cognition, 39*(3), 412–432.

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory*. New York: Oxford University Press.

Ihler, A., & Mandel, M. (2014). Kernel density estimation toolbox for MATLAB (R13). <http://www.ics.uci.edu/ihler/code/kde.html> (accessed: 2014-07-21).

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

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

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 36*(7), 1469–1482.

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

Khosla, A., Xiao, J., Isola, P., Torralba, A., & Oliva, A. (2012). Image memorability and visual inception. In *SIGGRAPH Asia Technical Briefs*.

Khosla, A., Xiao, J., Torralba, A., & Oliva, A. (2012) Memorability of image regions. In *Conference on Neural Information Processing Systems (NIPS)*.

A. Khosla, A., Bainbridge, W. A., Torralba, A. & Oliva, A. (2013). Modifying the memorability of face photographs. In *International Conference on Computer Vision (ICCV)*.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139*, 558–578.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Conference on neural information processing systems (NIPS)*.

Mancas, M., & Le Meur, O. (2013). Memorability of natural scene: The role of attention. In *IEEE international conference on image processing (ICIP)*.

Mandler, G. (2008). Familiarity breeds attempts: A critical review of dual-process theories of recognition. *Perspectives on Psychological Science, 3*(5), 390–399.

McGaugh, J. L. (1966). Time-dependent processes in memory storage. *Science, 153*(3742), 1351–1358.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory*. Oxford University Press.

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

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV), 42*(3), 145–175.

Rawsona, K. A., & Overscheldeb, J. P. V. (2008). How does knowledge promote memory? Skilled memory. *Journal of Memory and Language, 58*(3), 646–668.

Razavian, A. S., Azizpour, H., Sullivan, J. & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE conference on computer vision and pattern recognition (CVPR) workshops*.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: Shifting demographics in mechanical turk. *CHI EA '10*. ACM, pp. 2863–2872.

Schmidt, S. (1985). Encoding and retrieval processes in the memory for conceptually distinctive events. *Journal of Experimental Psychology: Learning, Memory, Cognition, 11*(3), 565–578.

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics, 40*(1), 185–197.

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

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

Tuckey, M. R., & Brewer, N. (2003). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied, 9*(2), 101–118.

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

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review, 105*(1), 33–57.

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

Vokey, J., & Read, J. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory Cognition, 20*(3), 291–302.

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

Walker, W. R., Vogl, R. J., & Thompson, C. P. (1997). Autobiographical memory: Unpleasantness fades faster than pleasantness over time. *Applied Cognitive Psychology, 11*(5), 399–413.

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

Xiao, J., Hayes, J., Ehinger, K., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from Abbey to Zoo. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press.

Zhou, B., Lapedriza, A. Xiao, J., Torralba, A.& Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Conference on Neural Information Processing Systems (NIPS)*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object detectors emerge in deep scene CNNs. arXiv:1412.6856.

Please cite this article in press as: Bylinskii, Z., et al. Intrinsic and extrinsic effects on image memorability. *Vision Research* (2015), http://dx.doi.org/10.1016/j.visres.2015.03.005