This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

What makes a photograph memorable?

Phillip Isola, Jianxiong Xiao, *Member, IEEE,* Devi Parikh, *Member, IEEE,* Antonio Torralba, *Member, IEEE,* and Aude Oliva

Abstract—When glancing at a magazine, or browsing the Internet, we are continuously exposed to photographs. Despite this overflow of visual information, humans are extremely good at remembering thousands of pictures along with some of their visual details. But not all images are equal in memory. Some stick in our minds while others are quickly forgotten. In this paper we focus on the problem of predicting how memorable an image will be. We show that memorability is an intrinsic and stable property of an image that is shared across different viewers, and remains stable across delays. We introduce a database for which we have measured the probability that each picture will be recognized after a single view. We analyze a collection of image features, labels, and attributes that contribute to making an image memorable, and we train a predictor based on global image descriptors. We find that predicting image memorability is a task that can be addressed with current computer vision techniques. While making memorable images is a challenging task in visualization, photography, and education, this work is a first attempt to quantify this useful property of images.

Index Terms—Scene understanding, image memorability, global image features, attributes

1 Introduction

People have the remarkable ability to remember thousands of pictures they saw only once [1], [2], even when they were exposed to many other images that look alike [3], [4]. We do not just remember the gist of a picture, but we are able to recognize which precise image we saw along with some of its details [5], [3], [4], [6]. However, not all images are remembered equally well. Some pictures stick in our minds whereas others fade away. The reasons why images are remembered may be highly varied; some pictures might contain friends, a fun event involving family members, or a particular moment during a trip. Other images might not contain any recognizable monuments or people and yet also be highly memorable [5], [3], [2]. In this paper we are interested in this latter group of pictures: what makes a generic photograph memorable?

Whereas most studies on human visual memory have been devoted to evaluating how good average picture memory can be, no work has systematically studied differences between individual images and if those differences are consistent across different viewers. Can a specific photograph be memorable to all of us, and can we estimate what makes it distinctive?

Similar to other subjective image properties, memorability is likely to be influenced by the user context and also be subject to some degree of inter-subject variability [7]. However, despite this expected variability when evaluating subjective properties of images, there is often also a sufficiently large degree of consistency between different users' judgments, suggesting it is possible to devise automatic systems to estimate these properties directly from images, ignoring user differences. As opposed to other image properties, there are no previous studies that try to quantify individual, everyday photos in terms of how memorable they are, and there are no computer vision systems that try to predict image memorability. This is contrary to many other photographic properties that have been addressed in the literature such as photo quality [8], aesthetics [9], [10], interestingness [11], saliency [12], attractiveness [13], composition [14], [15], color harmony [16], and importance [17], [18]. Also, there are no databases of photographs calibrated in terms of the degree of memorability of each image.

In this paper, we characterize an image's memorability as the probability that an observer will detect a repetition of a photograph at various delays after exposition, when presented amidst a stream of images. This setting allows us to measure long-term memory performance for a large collection of images [1]. We mine this data to identify which features of the images correlate with memorability, and we train memorability predictors on these features. Whereas further studies will be needed to validate these predictions on other datasets, the present work constitutes an initial benchmark for quantifying image memorability. A previous version of this work appeared partly in [20] and [21].

Just like aesthetics, interestingness, and other metrics of image importance, memorability quantifies something about the utility of a photograph toward our everyday lives. For many practical tasks, memorability is an especially desirable property to maximize. For example, this may be the case when creating educational materials, logos, advertisements, book covers, websites, and much more. Understanding memorability, and being able to automatically predict it, lends itself to a wide variety of applications in each of these areas. By analyzing memorability, educators could create textbook diagrams that stick in students' minds, or mnemonic cartoons

- *Massachusetts Institute of Technology*
  E-mail: {phillipi, jxiao, torralba, oliva}@mit.edu
- *Virginia Polytechnic Institute and State University*
  E-mail: parikh@vt.edu

1. Short-term memory typically can only hold 3 or 4 items at once [19] and is generally tested over durations of just a few seconds; since participants in our experiment had to hold many more images in memory and were tested minutes to nearly one hour after the first presentation, the experiments tackle long-term memory.

Digital Object Indentifier 10.1109/TPAMI.2013.200

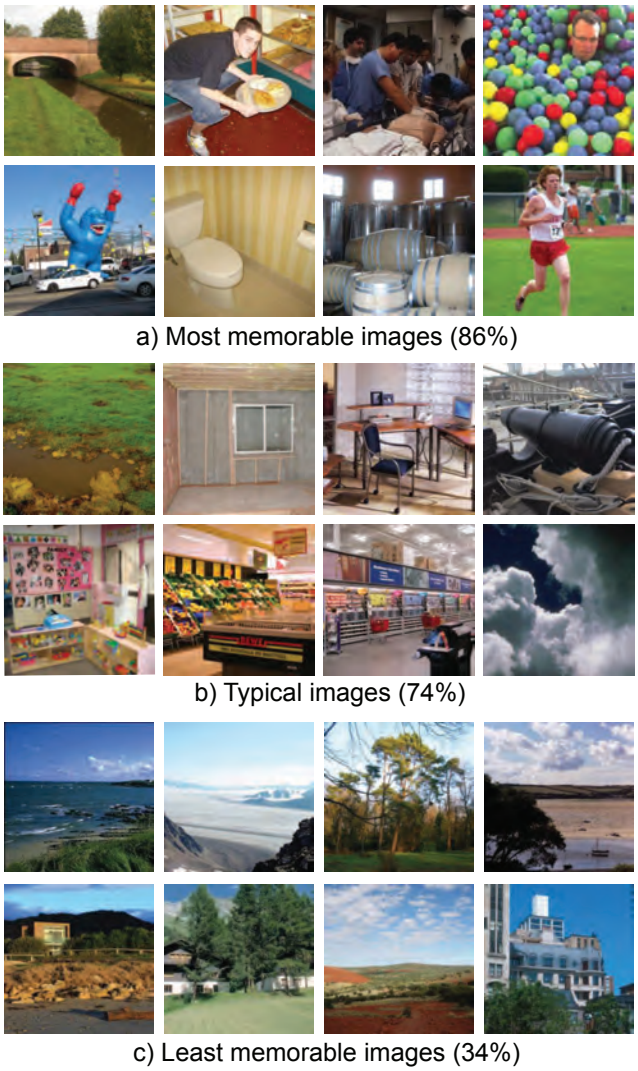


Fig. 1: Each set of 8 images was selected according to one half of the participants in our study as being (a) the 8 most memorable images, (b) 8 average memorability images, and (c) the 8 least memorable images. The number in parentheses gives the percent of times that these images were remembered by an independent set of participants.

that help students learn a foreign language. Memorability might also find application in user interface design. Memorable icons could clarify a messy desktop, and memorable labels could be stuck to pill jars and entryways in retirement homes. Memorability could also be used as a metric to pick out the most meaningful images from a photo collection or video. For example, a video could be summarized with just its most memorable frames, omitting the intervening images that would have been forgotten anyway. Farther in the future, we hope understanding memorability could lead to more fundamental advances in computer vision and artificial intelligence. If we can figure out what we humans remember, then we may be able to design intelligent systems that acquire knowledge that is similarly ecologically meaningful.

2 Measuring Image Memorability

Although we all have the intuition that some images will capture our attention and will be easier to remember than others, quantifying this intuition has only been addressed in limited settings in previous experiments. Previous research has looked at the effects of emotional images on memory [22] [23], face photo memorability [24] [25], and the memorability of facial caricatures [26] [27]. However, a comprehensive study of the memorability of individual, natural photos has been lacking. Are the photos remembered by one person more likely to be remembered also by somebody else? In this section, we characterize the consistency of image memory across different observers and time delays. In order to do so, we built a database of images (Figure 2), and we measured the probability of observers remembering each image (Figure 1 shows example images that span a wide range of memorabilities).

2.1 How to measure image memorability?

Cognitive psychologists have been studying the mechanisms and representations of human memory for nearly half a century. Studies have examined memory at multiple scales (e.g., perceptual, short-term, and long-term storage) and with a variety of tasks. Classical paradigms include asking observers if a given image has been seen before (repeat detection method) and two alternative forced choice paradigms (i.e. two images are presented at test, one novel and one old). Here we are interested in modeling an ecological and explicit measure of image memorability – namely, which images will tend to be best recognized when re-encountered – and so we choose a repeat detection task. The repeat detection paradigm also allows us to test familiarity of a given image at different delays after encoding the image (by showing the repeat image after a few seconds, minutes, or hours). Thus, for present usage, we simply define the 'memorability' of each image as how often the participants will tend to correctly detected a repetition of the image. Since motivation, attention, and participant ability are all known to modulate raw memory performance, we do not expect raw detection rates to be constant across all participants and contexts. Therefore, we chose to analyze memorability using rank scores, which we expect should be more stable across changes in user focus and ability.

2.2 The Visual Memory Game

In order to measure image memorability, we presented workers on Amazon Mechanical Turk with a Visual Memory Game. In the game, participants viewed a sequence of images, each of which was displayed for 1 second, with a 1.4 second gap in between image presentations (Figure 3). Their task was to press the space bar whenever they saw an identical repeat of an image at any time in the sequence [5] [3]. Participants received feedback whenever they pressed a key (a green symbol shown at the center of the screen for correct detection, and a gray X for an error).

Image sequences were broken up into levels that consisted of 120 images each. Each level took 4.8 minutes to perform. At the end of each level, the participant saw his or her correct response average score for that level, and was allowed to take a short break. Participants could complete at most 30 levels, and were able to exit the game at any time. A total of 665 workers

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

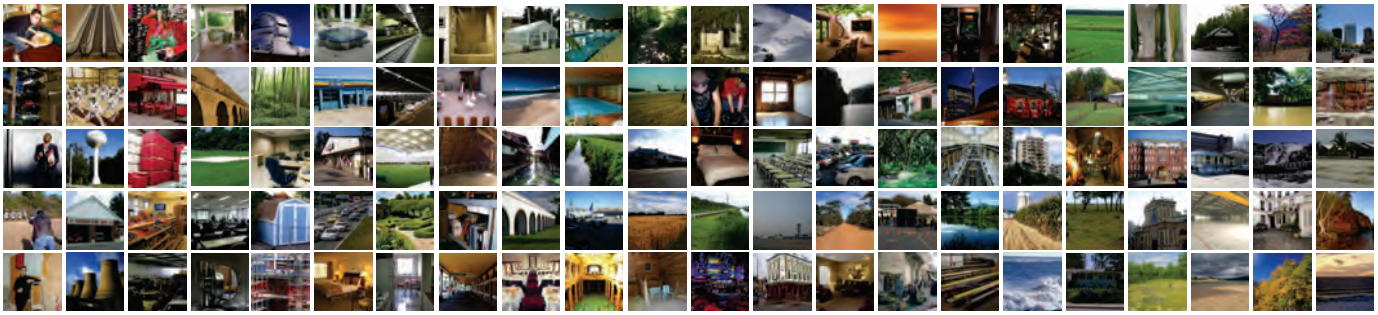


Fig. 2: Sample of the database used for the memory study. The images are sorted from more memorable (left) to less memorable (right).

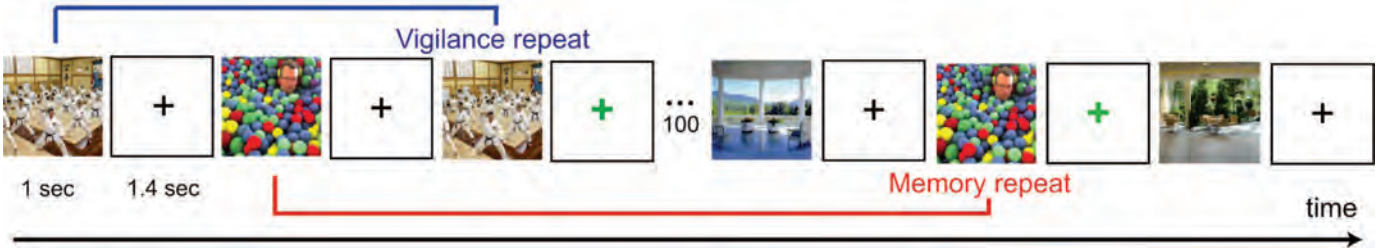


Fig. 3: Mechanical Turk workers played a "Memory Game" in which they watched for repeats in a long stream of images.

from Mechanical Turk ($> 95\%$ approval rate in Amazon's system) performed the game. Over 90% of our data came from 347 of these workers. We payed workers \$0.30 per level in proportion to the amount of the level completed, plus a \$0.10 bonus per fully completed level. This adds up to about \$5 per hour. The average worker stayed in the game for over 13 levels.

Unbeknownst to the participants, the sequence of images was composed of 'targets' (2222 images) and 'fillers' (8220 images). Target and filler images represented a random sampling of the scene categories from the SUN dataset [28] [1] All images were scaled and cropped about their centers to be 256x256 pixels. The role of the fillers was two-fold: first, they provided spacing between the first and second repetition of a target; second, responses on repeated fillers constituted a 'vigilance task' that allowed us to continuously check that participants were attentive to the task [5], [3]. Repeats occurred on the fillers with a spacing of 1-7 images, and on the targets with a spacing of 91-109 images. Each target was sequenced to repeat exactly once, and each filler was presented at most once, unless it was a vigilance task filler, in which case it was sequenced to repeat exactly once.

Stringent criteria were used to continuously screen worker performance once they entered the game. First, the game automatically ended whenever a participant fell below a 50% success rate on the last 10 vigilance task repeats or above a 50% error rate on the last 30 non-repeat images. When this happened, all data collected on the current level was discarded. Rejection criterion reset after each level. If a participant failed any of the vigilance criteria, they were flagged. After receiving three such flags they were blocked from further participation in the experiment. Otherwise, participants were able to restart the game as many times as they wished until completing the max 30 levels. Upon each restart, the sequence was reset so that the participant would never see an image they had seen in a previous session. Finally, a qualification and training 'demo' preceded the actual memory game levels.

After collecting the data, we assigned a 'memorability score' to each target image, defined as the percentage of correct detections by participants. On average, each target was scored by 78 participants. The average memorability score was 67.5% (SD of 13.6%). Average false alarm rate was 10.7% (SD of 7.6%).

Given this low false alarm rate, and the fact that false alarm rates do not correlate with hit rates ($\rho = 0.01$), we expect that false memories do not play a large role in our memorability scores, and thus our scores are a good measure of correct memories.

Throughout this paper, we refer to our the memorability scores collected through our memory game as "ground truth" memorability scores.

2.3 Is memorability consistent across observers?

Are the images that are more memorable (or forgettable) for a group of observers also more likely to be remembered (or forgotten) by a different group of observers?

To evaluate human consistency, we split our participant pool into two independent halves, and quantified how well image scores measured on the first half of the participants matched image scores measured on the second half of the participants. Averaging over 25 random split half trials, we calculated a Spearman's rank correlation (ρ) of 0.75 between these two sets of scores. We sorted photos by their scores given by the first half of the participants and plotted this against memorability according to the second half of the participants (Figure 4). This shows that, for example, if a repeat is correctly detected 80% of the time by one half of the participants, we can expect the other half of the participants to correctly detect this repeat

1. In addition, 717 of the 8220 filler images were textural images; 178 of these were actually sequenced as targets but since we did not include them from our subsequent memorability analysis (which focused on generic photos of natural scenes), we refer to them for present purposes as fillers.

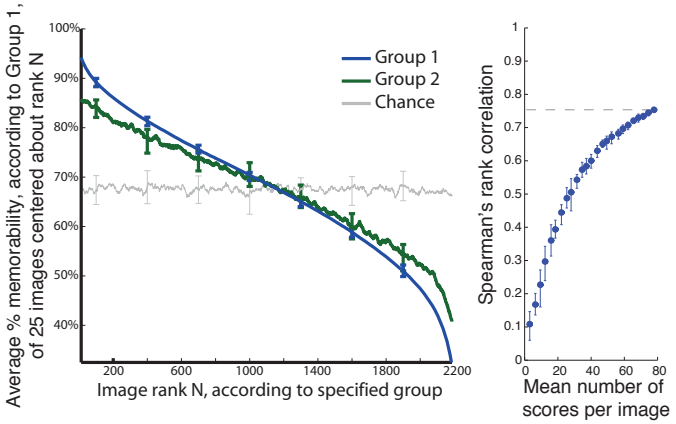


Fig. 4: Measures of human consistency. Participants were split into two independent sets, Group 1 and Group 2. Left: Images were ranked by memorability scores from participants in one or the other group and plotted against the average memorability scores given by participants in Group 1. For clarity, we convolved the resulting plots with a length-25 box filter along the x-axis. The gray chance line was simulated by assigning the images random ranks (i.e. randomly permuting the x-axis). Right: Spearman's rank correlation between subject Groups 1 and 2 as a function of the mean number of scores per image. Both left and right analyses were repeated for 25 random splits and mean results are plotted. Error bars show 80% confidence intervals over the 25 trials.

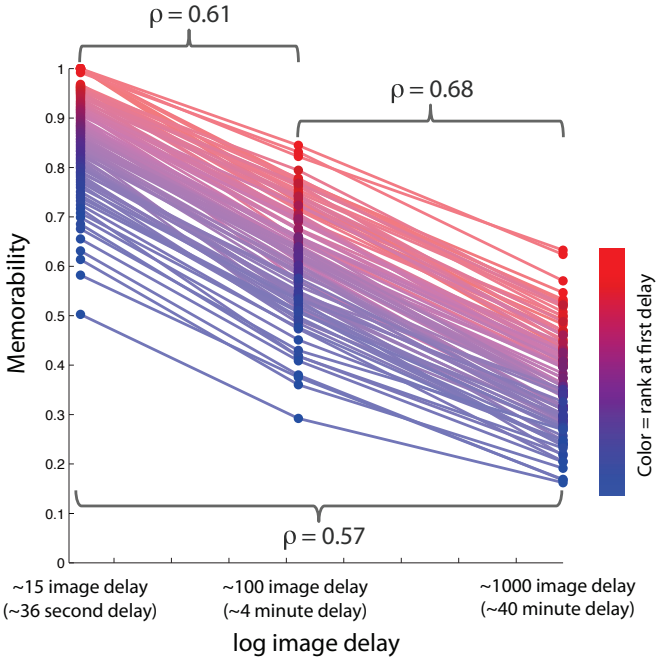


Fig. 5: Image memorability versus delay between repeat and initial presentation. Color depicts memorability rank at shortest delay. Lines interpolate between the measurements at each of the three delays. Spearman's rank correlations between memorabilities measured at each pair of delays are given above plot. For clarity of visualization, each plotted point and line is the mean memorability of 22 images binned in the order of memorability at the shortest delay.

around 78% of the time, corroborating that this photo is truly memorable. At the other end of the spectrum, if a repeat is only detected 50% of the time by one half of the participants, the other half will tend to detect it only 54% of the time – this photo is consistently forgotten. It thus appears that there really is sizable variation in photo memorability. (Figure 4). Thus, our data has enough consistency that it should be possible to predict image memorability. Individual differences and random variability in the context each participant saw add noise to the estimation; nonetheless, this level of consistency suggests that information intrinsic to the images might be used by different people to remember them. In section 3, we search for this image information.

2.4 Is memorability consistent over time?

In the previous sections, we showed that memorability tested after a few minutes is a stable property of images independent of randomized user and image sequence. But is memorability also stable over various time delays? We ran a variant of our Memory Game to test the effect of delay on image memorability. The procedure was the same as reported above (including vigilance and target repeats) except that target repeats were sequenced to appear at one of three possible delays, tapping into long term visual representations: ~15 images back (with jitter this condition corresponded to 11-19 images back) , ~100 back (96-104 images back) and ~1000 back (996-1004 images back). So that the longest delay repeats would appear with equal frequency to the shorter delay repeats, we did not start any target repeats until after an initial 1080 images had been presented (about 40 minutes of playing the memory game; note that we presented vigilance repeats as usual during this phase).

We measured the memorability of each image at each delay as the proportion of times a repeat of the image at that delay was correctly detected, and collected about 30 scores per delays. Figure 5 shows the memorability scores (percent of correct responses) for the three delays: for clarity, each plotted line is the mean memorability of 22 images binned in the order of memorability at the shortest delay. Strikingly, even after the shortest delay (11-19 images back; i.e. 24-48 seconds back), there were already large memorability differences between the images, and these differences were remarkably similar to those at both longer delays: rank memorabilities at one delay correlated strongly with those at the other delays: ρ = 0.61, 0.68 and 0.57 for the three pairwise comparisons (Figure 5). Thus, it appears that rank memorability is stable over time.

For practical applications, this degree of stability is quite fortunate. What is relatively memorable after ~15 intervening images is also relatively memorable after ~1000 intervening images. Thus, in order to predict memorability, we do not need to model a complex time-dependent function; instead, for our present purposes, we will treat rank memorability as time-independent, and investigate its properties at the ~100 image delay.

2.5 Role of context

A large body of research on human memory suggests that we remember things in proportion to how well they stand out from their local context (e.g., [2], [29], [30], [31], [32]). Our present quest is to uncover factors that are *intrinsic* to an image and make it memorable, independent of extrinsic variables such as observer, time delay, and local visual context. By

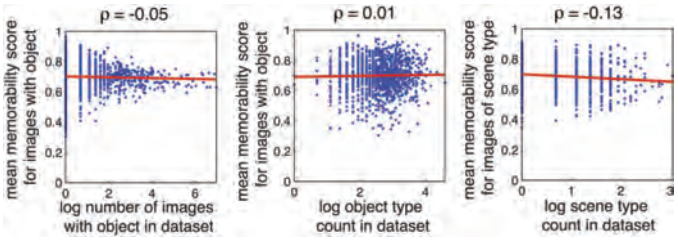


Fig. 6: Semantic frequencies in our dataset do not explain much of the variance in memorability. Red line is linear least squares fit.

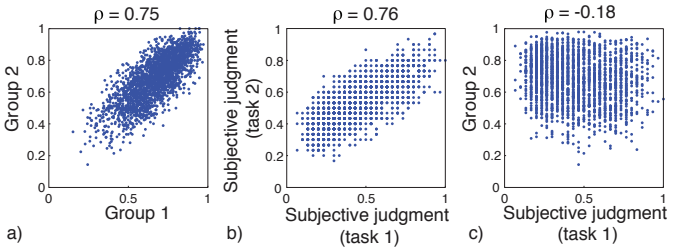


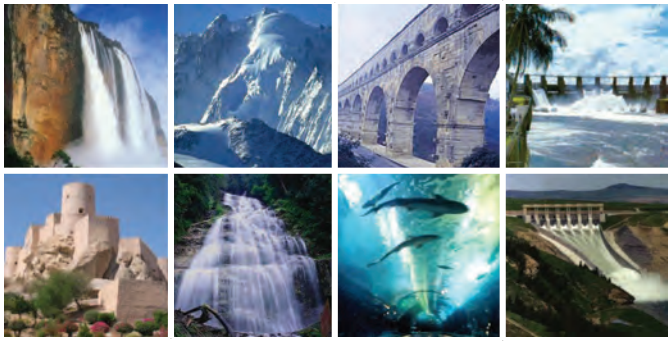
Fig. 7: In each scatter plot, each dot corresponds to one image. a) Comparison of memorability scores measured on Participant group 1 versus those measured on Participant group 2 in the memory game. The plot shows that there is a strong correlation between two different sets of subjects. b) Memorability scores from task 1 (memory) vs. task 2 (repeat). c) Scores from task 1 vs. memorability measured during the memory game (group 2).

randomizing the sequence each participant in our experiment sees, we ensure that our measurements do not depend on the precise order in which the photos were presented. However, it remains unclear to what degree overall dataset statistics could have affected the memorability scores. To test for simple interactions with dataset context, we measured the correlation between image content frequency in our dataset and mean memorability over images with this content. For frequencies of images containing a particular object, frequency of objects, and frequencies of scene category we found no strong correlation ($\rho = -0.05$, 0.01, and -0.13 respectively; Figure 6). This suggests that these simple forms of dataset bias cannot explain our results. Ultimately, to test more subtle possible interactions with context, it will be important to measure memorability on additional datasets and measure how well our present results generalize.

2.6 Subjective judgments do not predict memorability

In the previous section, we have shown that there is consistency in image memorability between separate groups of observers and over a wide range of time delays from image presentation. In this section we want to explore a different aspect of our measurements. When working with collections of images, users are generally forced to make subjective decisions such as choosing which images are most pleasing, or of highest quality. Here we want to know how successful one user would be if he or she were to guess which images are the most memorable in a collection. To test this, we ran two experiments on Mechanical Turk.

- Task 1 (Memory Judgment): we asked 30 participants to indicate if they believe that an image is memorable or not. In each HIT, we showed 36 images to each participant and they had to provide for each image a binary answer to the question "Is this a memorable image?".
- Task 2 (Repeat Judgment): we also ran a separate task on the same set of images asking 30 participants to perform the next task: "For each of the images shown below, please indicate if you would remember seeing it or not i.e. If you were to come across this image in the morning, and then happen to see it again at the end of the day, do you think you would realize that you have seen this image earlier in the day?"



a) Predicted by participants as being most memorable images

b) Predicted by participants as being least memorable images

Fig. 8: This figure is similar to figure 1 but using the judgments of participants to select which images they believe are memorable and which ones are not. (a) shows the 8 images participants thought would be most memorable and (b) shows the 8 image participants thought would be least memorable. In fact, however, set (a) has an average memorability of 70% and set (b) has an average memorability of 74%, as measured in our memory game. This shows that people's intuitions about which images are memorable can be wrong.

For these two tasks we used the same set of 2222 target images as in the previous experiment. For each image we computed a score by averaging the 30 participant responses. Both tasks provided similar results with a rank correlation between the two of $\rho = 0.76$ (this value is similar to the correlation between the two groups of participants obtained in the memory experiment from section 2.3). This is illustrated in Figure 7. Figure 7.a shows the scatter plot of the experiment of section 2.3) and Figure 7.b shows the scatter plot comparing the two binary Mechanical Turk tasks.

However, Figure 7.c shows that the subjective judgments on which images are memorable do not predict the actual memory results obtained during the memory game (rank correlation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

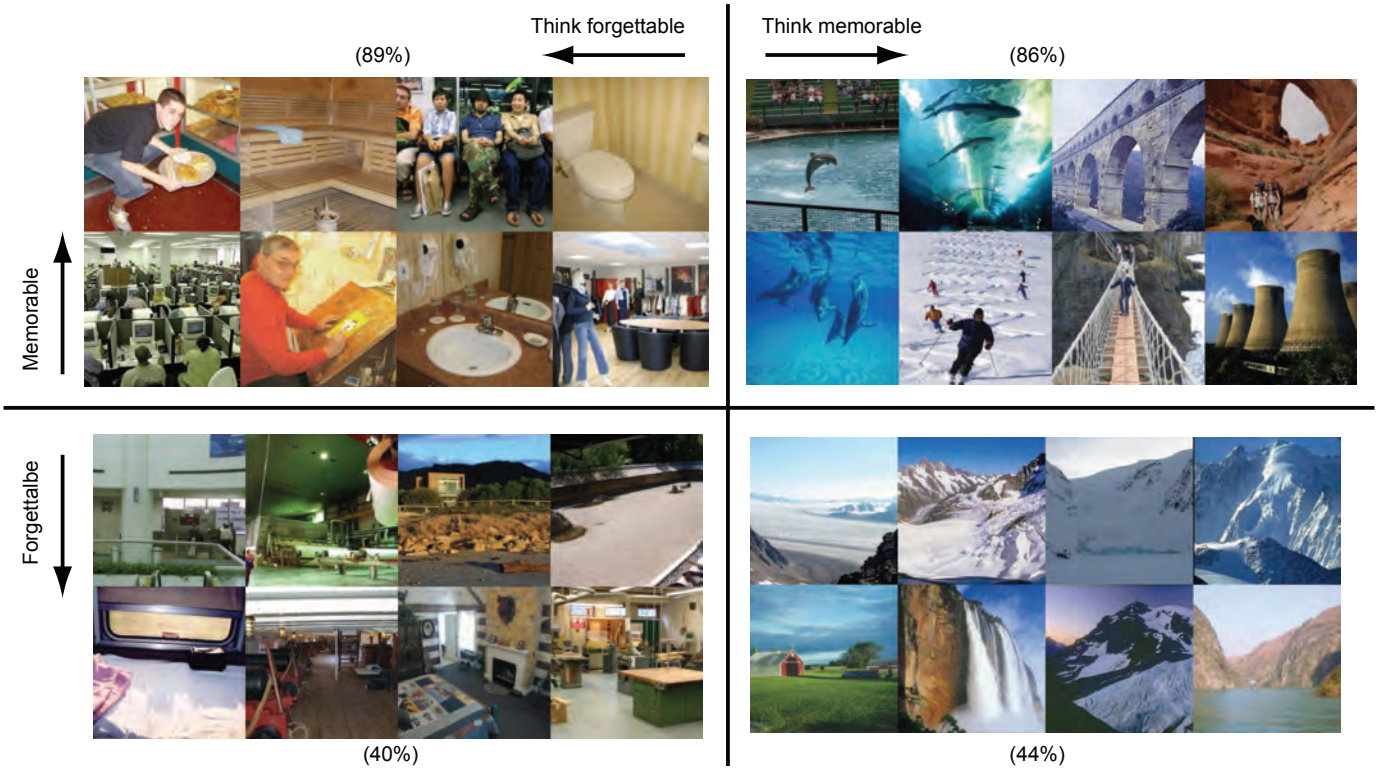


Fig. 9: This figure shows 4 sets of images illustrating the images in the four corners of the scatter plot from Figure 7.c. The number beside each set of images corresponds to the average memorability measured by the memory game on each set of 8 images.

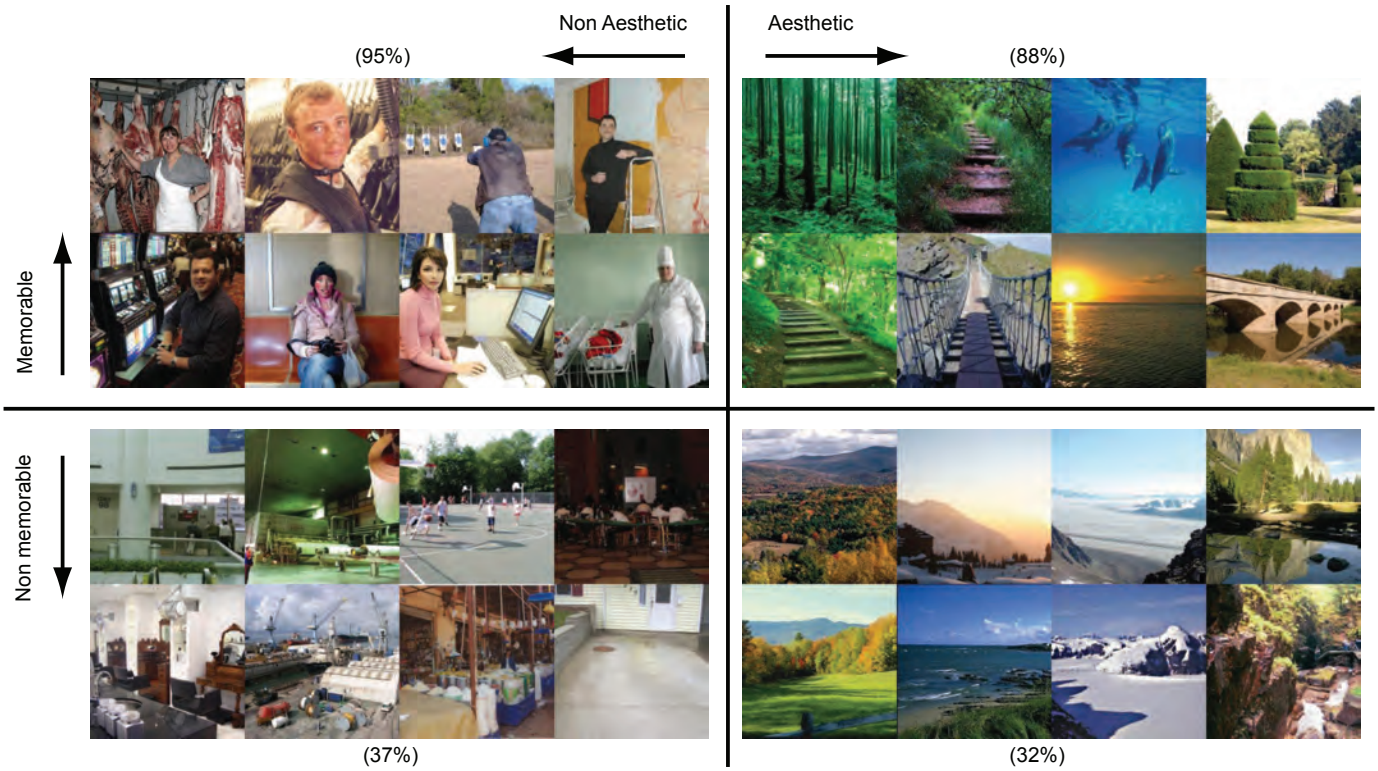


Fig. 10: In our dataset, image memorability is distinct from image aesthetics. The vertical axis separates images that are considered aesthetic (right) vs. images that are consider not aesthetic (left). The horizontal axis separates the images that are memorable (on top) vs images that are not memorable (bottom). The number beside each set of images corresponds to the average memorability measured by the memory game on each set of 8 images.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

a) most aesthetic

b) least aesthetic

Fig. 11: Most and least aesthetic images from our dataset as chosen by 30 participants. The top eight most aesthetic images have an average memorability of 57%, while the least aesthetic images have an average memorability of 84%.

a) most interesting

b) least interesting

Fig. 12: Most and least interesting images from our dataset as chosen by 30 participants.The top eight most interesting images have an average memorability of 70%, while the least interesting images have an average memorability of 78%.

between task 1 and the memory game is $\rho = -0.19$ and between task 2 and the memory game is $\rho = -0.02$). Although the memory game provides just one way of measuring memorability, our results suggest that users sometimes have the wrong intuition about memorability. Figure 8 shows the images that observers believed would be most (a) and least (b) memorable. These images are very different from the ones shown in Figure 1.

Figure 9 further shows how subjective intuitions about which images are memorable can be very wrong. This figure shows 4 sets of images illustrating the images in the four corners of the scatter plot from Figure 7.c. The top-left corner shows 8 images that participants rated as being among the least memorable images while doing task 1. However, those images were among the most memorable images during the memory game. Analogously, images in the bottom-right corner were rated as among the most memorable images in task 1, but they were among the least memorable images during the memory game.

Interestingly, despite that memorability is highly consistent across observers, people do not have a good intuition about which images are memorable and which ones are not. In contrast with these subjective intuitions, our ground truth memorability scores provide an objective measure of how an image will affect an observer's memory.

3 What makes an image memorable?

Among the many reasons why an image might be remembered by a viewer, we investigate first the role of various image-based and semantic properties of the images: color, simple image features, object statistics, object semantics, scene semantics, and high-level attributes. First, we will show that some of the aspects that observers believe contribute to make an image more memorable do not predict which images are memorable.

3.1 Memorability, aesthetics, and interestingness

One important question to explore is the relationship between image memorability and other subjective image properties such as aesthetic judgments or image interestingness. To measure image aesthetic value and interestingness we ran two separate Mechanical Turk experiments on the 2222 target images. Participants were asked the questions "Is this an aesthetic image?" and "Is this an interesting image?" and had to answer this "Yes" or "No" for 36 images per HIT. For each image we computed an aesthetic and an interestingness score by averaging the answers given by 30 participants.

Figure 11 shows the most and least aesthetic images and Figure 12 shows the most and least interesting images out of the 2222 images from our dataset. We found that interestingness and aesthetics subjective judgments are strongly correlated ($\rho = 0.85$, see Figure 13.a).

FIgure 13.b and c show the scatter plot of memorability (measured in the memory game) as a function of the image aesthetic score and image interestingness score. Each dot in the plot corresponds to one image. These two image properties correlate weakly with image memorability. $\rho = -0.36$ between aesthetics and memorability and $\rho = -0.23$ between

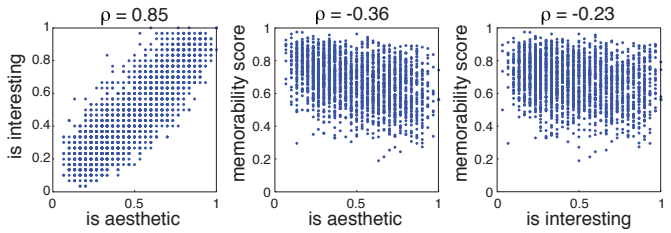


Fig. 13: In each scatter plot, each dot corresponds to one image. Judgments of aesthetic and interestingness are highly correlated in our dataset (a). However, aesthetic (b) and interestingness (c) have very weak correlation with memorability.

interestingness and memorability. The negative values indicate that, in our database, images that were less aesthetic and less interesting turned out to be more memorable than beautiful and interesting images.

Interestingly, image aesthetics and interestingness strongly correlate with the subjective judgments of image memorability ($\rho = 0.83$ and $\rho = 0.86$ respectively for task 1). This illustrates that participants had the wrong intuition that beautiful and interesting images will produce a lasting memory.

Figure 10 shows 4 sets of 8 images each showing the images on the four corners of the scatter plot from Figure 13.c. This figure shows how many of the most aesthetic images are also among the least memorable ones (e.g., the 8 images from the bottom-right corner of Figure 10).

Together, these results show that image memorability is an image property that is distinct from two other commonly used subjective image properties.

3.2 Color and simple image features

Are simple image features enough to determine whether or not an image will be memorable? We looked at the correlation between memorability and basic pixel statistics. Mean hue was weakly predictive of memory: as mean hue transitions from red to green to blue to purple, memorability tends to go down (ρ = -0.16). This correlation may be due to blue and green outdoor landscapes being remembered less frequently than more warmly colored human faces and indoor scenes. Mean saturation and value, on the other hand, as well as the first three moments of the pixel intensity histogram, exhibited weaker correlations with memorability (Figure 14). These findings concord with other work that has shown that perceptual features are not retained in long term visual memory [2], [6]. In order to make useful predictions, more descriptive features are likely necessary.

3.3 Object statistics

Object understanding is necessary to human picture memory [33], [2]. Using LabelMe [34], each image in our target set was segmented into object regions and each of these segments was given an object class label by a human user (e.g., “person”, “mountain”, “stethoscope”) (see [35] for details). In this section, we quantify the degree to which our data can be explained by non-semantic object statistics.

Do such statistics predict memorability? For example do the number of objects one can attach to an image determine its

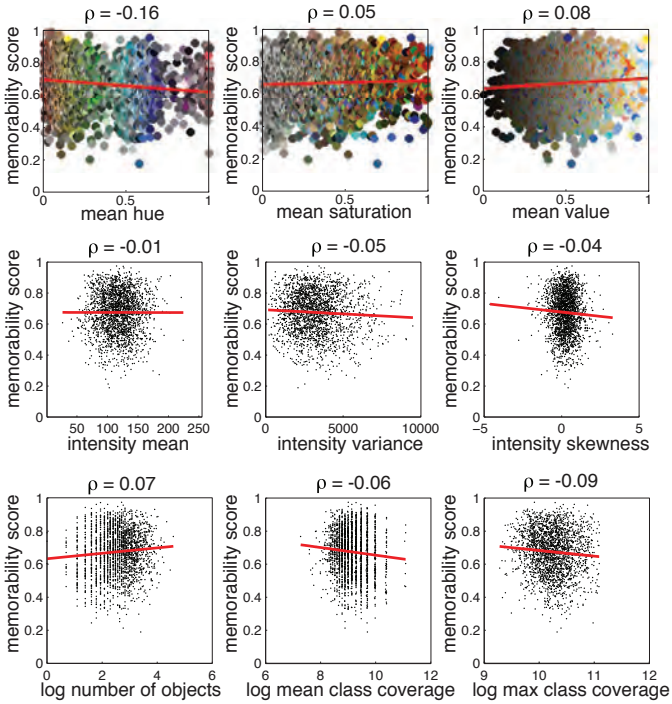


Fig. 14: Simple image features, as well as non-semantic object statistics, do not correlate strongly with memorability score. Red line is linear least squares fit.

memorability, or is it critical that an object class takes up a large portion of an image in order for the image to stick in memory? We find the answer to be no: none of these statistics make good predictions on their own. Simple object statistics (log number of objects, log mean pixel coverage over present object classes, and log max pixel coverage over object classes) did not correlate strongly with memorability (ρ = 0.07, -0.06, and -0.09 respectively) (Figure 14).

To investigate the role of more subtle interactions between these statistics, we trained a support vector regression (ϵ-SVR [36]) to map object statistics to memorability scores. For each image, we measured several object statistics: the number of objects in the image per class, and the number of pixels covered by objects of each class in the entire image as well as in each quadrant of the image. For each of these statistics, we thereby obtained joint distribution on (object class, statistic). We then marginalized across class to generate histograms that only measure statistics of the image segmentation, and contain no semantic information: ‘Object Counts’, ‘Object Areas’, and, concatenating pixel coverage on the entire image with pixel coverage per quadrant, ‘Multiscale Object Areas’. We used these histograms as features for our regression and applied histogram intersection kernels.

For each of 25 regression trials, we split both our image set and our participant set into two independent, random halves. We trained on one half of the images, which were scored by one half of the participants, and tested on the left out images, which were scored by the left out participants. During training, we performed grid search to choose cost and ϵ hyperparameters for each SVR.

We quantified the performance of our predictions similarly

	Object Counts	Object Areas	Multiscale Object Areas	Object Label Presences	Labeled Object Counts	Labeled Object Areas	Labeled Multiscale Object Areas	Scene Category	Attributes	Objects, Attributes, and Scenes	Other Humans
Top 20	68%	68%	73%	83%	81%	84%	84%	81%	87%	88%	86%
Top 100	68%	68%	72%	79%	79%	82%	82%	78%	83%	83%	84%
Bottom 100	67%	64%	64%	57%	57%	56%	56%	57%	54%	55%	47%
Bottom 20	67%	64%	65%	54%	54%	52%	53%	56%	53%	51%	40%
ρ	0.04	0.04	0.20	0.43	0.44	0.47	0.47	0.37	0.51	0.54	0.75

TABLE 1: Comparison of predicted versus measured memorabilities. Images are sorted into sets according to predictions made on the basis of a variety of object and scene features (denoted by column headings). Average ground truth memorabilities are reported for each set. e.g., The "Top 20" row reports average ground truth memorability over the images with the top 20 highest predicted memorabilities. ρ is the Spearman rank correlation between predictions and measurements.

to how we analyzed human consistency above. First, we calculated average ρ between predicted memorabilities and ground truth memorabilities. Second, we sorted images by predicted score and selected various ranges of images in this order, examining average ground truth memorability on these ranges (Table 1). As an upper-bound, we compared to a measure of the available consistency in our data, in which we predicted that each test set image would have the same memorability according to our test set participants as was measured by our training set participants ('Other Humans').

Quantified in this way, our regressions on object statistics appear ineffective at predicting memorability (Table 1). However, predictions made on the basis of the Multiscale Object Areas did begin to show substantial correlation with measured memorability scores (ρ = 0.20). Unlike the Object Counts and Object Areas, the Multiscale Object Areas are sensitive to changes across the image. As a result, these features may have been able to identify cues such as "this image has a sky," while, according to the other statistics, a sky would have been indistinguishable from a similarly large segment, such as a closeup of a face.

3.4 Object and scene semantics

As demonstrated above, objects without semantics are not effective at predicting memorability. This is not surprising given the large role that semantics play in picture memory [33], [2]. To investigate the role of object semantics, we performed the same regression as above, except this time using the entire joint (object class, statistic) distributions as features. This gave us histograms of 'Labeled Object Counts', 'Labeled Object Areas', 'Labeled Multiscale Object Areas', and, thresholding the labeled object counts about zero, 'Object Label Presences'. Each image was also assigned a scene category label as described in [28] ('Scene Category'). We applied histogram intersection kernels to each of these features, and also tested a combination of Labeled Multiscale Object Areas and Scene Category features using a kernel sum ('Objects and Scenes').

Semantics boosted performance (Table 1). Even the Object Label Presences alone, which simply convey a set of semantic labels and otherwise do not describe anything about the pixels in an image, performed well above our best unlabeled object statistic, Multiscale Object Areas (ρ = 0.43 and 0.20 respectively). Moreover, Scene Category, which just gives a single label per image, appears to summarize much of what makes an image memorable (ρ = 0.37). These performances support the idea that object and scene semantics are a primary substrate of memorability [2], [3], [33].

3.5 Semantic attributes

Scene semantics go beyond just object content and scene category. Hence, we investigate 127 semantic attributes that capture the spatial layout of the scene (e.g., open, enclosed, cluttered, etc.), aesthetics (e.g., postcard-like, unusual, etc.), dynamics (e.g., static, dynamic, moving objects, etc), location (e.g., famous place), emotions (e.g., frightening, funny, etc.), actions (e.g., people walking, standing, sitting, etc.) as well as demographics and appearance of people (e.g., clothing, accessories, race, gender, etc.). Please see [18] for details.

As above, we train SVRs to map attributes to memorability scores. Here, we use an RBF kernel, and achieve a performance of $\rho = 0.51$. This performance is striking because these attributes outperform all our above feature sets while also being more concise (i.e. lower entropy [21]). This suggests that high-level semantic attributes are an especially efficient way of characterizing the memorability of a photo.

When we combine all our semantic features together with a kernel sum (Labeled Multiscale Object Areas + Scene Category + Attributes), we achieve a maximum performance of $\rho = 0.54$.

3.6 Visualizing what makes an image memorable

Since object content appears to be important in determining whether or not an image will be remembered, we further investigated the contribution of objects by visualizing object-based "memory maps" for each image. These maps shade each object according to how much the object adds to, or subtracts from, the image's predicted memorability. More precisely, to quantify the contribution of an object i to an image, we take a prediction function, f, that maps object features to memorability scores and calculate how its prediction m changes when we zero features associated with object i from the current image's feature vector, $(a_1, \cdots, a_n)$. This gives us a score s_i for each object in a given image:

$$m_1 = f(a_1, \cdots, a_i, \cdots, a_n) \quad (1)$$

$$m_2 = f(a_1, \cdots, 0, \cdots, a_n) \quad (2)$$

$$s_i = m_1 - m_2 \quad (3)$$

For the prediction function f, we use our SVR on Labeled Multiscale Object Areas, trained as above, and we plot memory maps on test set images (Figure 16). Thus, these maps show predictions as to what will make a novel image either remembered or not remembered. The validity of these maps is supported by the fact that the SVR we used to generate them

Fig. 15: Objects sorted by their predicted impact on memorability. Next to each object name we report how much an image's predicted memorability will change, on average, when the object is included in the image's feature vector versus when it is not. For each object name, we also display two test set images that contain the object: on the left is the example image with the highest memorability score among all test set images that contain (over 4000 pixels of) the object. On the right is the example with the lowest score. Only objects that appear (cover over 4000 pixels) in at least 20 images in our training set are considered.

(the Labeled Multiscale Object Areas regression) makes predictions that correlate relatively well with measured memory scores (ρ = 0.47, see Table 1).

This visualization gives a sense of how objects contribute to the memorability of particular images. We are additionally interested in which objects are important across all images. We estimated an object's overall contribution as its contribution per image, calculated as above, averaged across all test set images in which it appears with substantial size (covers over 4000 pixels). This method sorts objects into an intuitive ordering: people, interiors, foregrounds, and human-scale objects tend to contribute positively to memorability; exteriors, wide angle vistas, backgrounds, and natural scenes tend to contribute negatively to memorability (Figure 15). While we require human annotations to create these visualizations, Khosla et al. have recently shown that they can generate similar memorability maps automatically from unlabeled images [37].

	Pixels	GIST	SIFT	SSIM	HOG 2x2	All Global Features
Top 20	73%	82%	82%	83%	84%	83%
Top 100	73%	79%	79%	80%	80%	80%
Bottom 100	61%	58%	57%	58%	58%	56%
Bottom 20	59%	57%	55%	55%	56%	54%
ρ	0.22	0.38	0.41	0.43	0.43	0.46

TABLE 2: Comparison of global feature predictions versus ground truth memory scores. Uses same measures as described in Table 1.

4 Predicting Image Memorability

4.1 Predicting memorability of generic images

As we have seen in the previous sections there is a significant degree of consistency between different sets of viewers on how memorable are individual images. In addition, we have seen that some of the consistency can be explained in terms of the objects, scenes, and attributes present in the picture. In this section, we describe an automatic predictor of memorability, which uses only features algorithmically extracted from an image. Here, we followed a similar approach to works studying other subjective image properties [8], [15], [28].

As with the object regressions, we trained an SVR to map from image features to memorability scores. We tested a suite of global image descriptors that have been previously found to be effective at scene recognition tasks [28] as well as being able to predict the presence/absence of objects in images [38], [39], [40]. The facility of these features at predicting image semantics suggests that they may be able to predict, to some degree, those aspects of memorability that derive from image semantics.

These global features are GIST [41], and spatial pyramid histograms of SIFT [42], HOG2x2 [38], [39], [28], and SSIM [40] features. We additionally looked at pixel histograms, which capture color distributions in an image: for each image, we built the 'pixel histogram' as the concatenation of three 21-bin histograms of intensity values, one for each color channel of the RGB image. We used an RBF kernel for GIST and histogram intersection kernels for the other features. Lastly, we also combined all these features with a kernel product ('All Global Features').

We evaluated performance in the same way as we evaluated the object regressions, and we found that the combination of global features performs best, achieving a rank correlation of 0.46. This correlation is less than human predictions, but close to our best predictions from labeled annotations. Figure 17 shows sample images from predicted sets. Figure 19 shows sample images on which our global features regression performed poorly.

To set a high watermark, and to get a sense of the redundancy between our image features and our annotations, we additionally trained an SVR on a kernel sum of all our global features plus Labeled Multiscale Object Areas, Scene Categories, and Attributes ('Global Features and Annotations'). This combination achieved a rank correlation of $\rho = 0.57$. See Table 2 and Figure 18 for detailed results.

The memorability variation we have predicted may appear to be dominated by coarse categorical differences between images: e.g., photos of people are more memorable than photos of landscapes. Can we also predict memorability differences within categories? To investigate this, we selected subsets of our dataset and analyzed and predicted variation within those subsets.

4.2 Memorable photos of people

Photos of people are among the most memorable in our dataset (average memorability score of 82%). Such photos are also especially prevalent in everyday contexts – we share photos of each other on Facebook, remember the faces of the thousands of friends and celebrities [25], and may be swayed

Fig. 16: Visualization of how each object contributes to the memorability of sample images spanning a range of memorability predictions. We estimate contribution as the difference between predicted memorability when the object is included in the image versus when it is removed from the image. In red we show objects that contribute to higher predicted memorability and in blue are objects that contribute to lower predicted memorability. Brightness is proportional to the magnitude of the contribution. Average measured memorability of each sample set is given in parentheses.

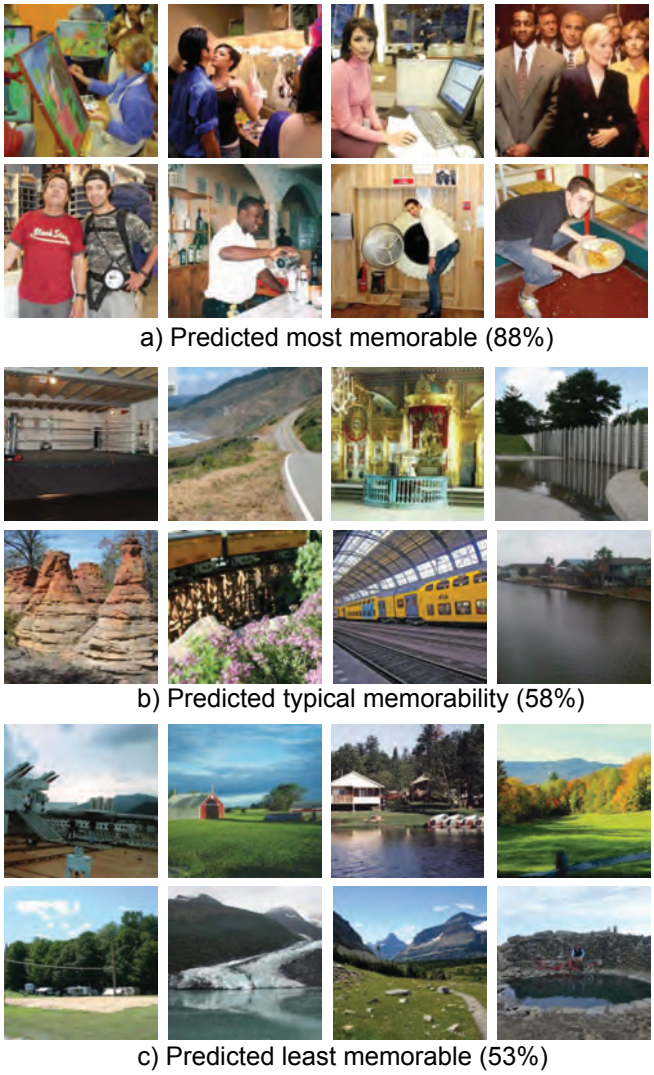


Fig. 17: The 8 images predicted, on the basis of global image features, as being the most memorable out of all test set images (a), 8 images with average memorability predictions (b), and the 8 images predicted as being the least memorable of all test set images (c) . The number in parentheses gives the mean ground truth memorability score for images in each set. The predictions produce clear visual distinctions, but may fail to notice more subtle cues that make certain images more memorable than others.

by advertisements delivered by beautiful spokespeople. Consequently, it may be especially useful to be able to predict the memorability of photos of people.

We took a first step in this direction by testing our algorithm just on the photos of people in our dataset (defined as photos with at least 5,000 pixels labeled as *person*, or a synonym, and with the attribute *face visible*). Within this subset, which consisted of 209 photos, split half consistency between our participants was $\rho = 0.53$ – robust variation in memorability exists even within this constrained subset. Using our best automatic predictor ('All Global Features'), we achieved a rank correlation between predictions and measurements of $\rho = 0.16$. A summary of our predictions, and the ground truth variability, is given in Figure 20. We additionally tried training SVRs on just photos of people, in order to perhaps better fit to the specific variation in this class of photos. This training scheme did not substantially improve performance ($\rho = 0.17$).

4.3 Memorable photos of nature

Photos of nature tend to be less memorable than artificial scenes (average memorability score of 61%), but are all photos of the natural world forgettable? We analyzed the subset of photos in our dataset categorized as *outdoor-natural* in the SUN dataset [28], and with less than 1,000 pixels labeled as *person* (this gave us 373 photos in total). We analyzed this subset in the same way as we analyzed the people subset: split half consistency among experiment participants was $\rho = 0.74$ and our best predictor, trained on all photos and tested on nature photos, reached $\rho = 0.32$ (training just on nature photos

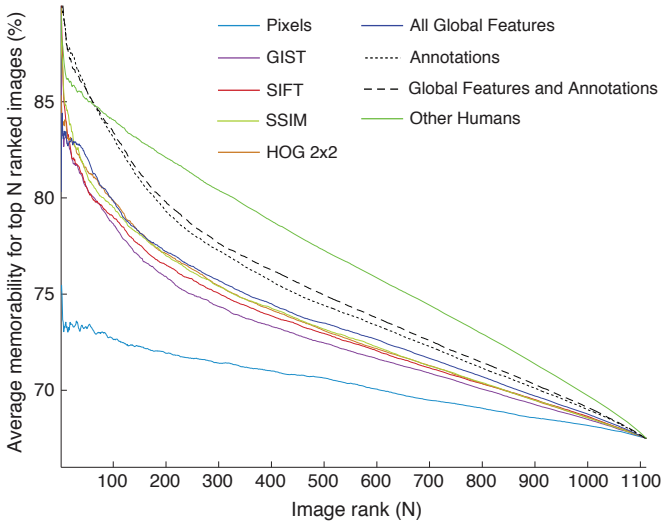


Fig. 18: Comparison of regressions results averaged across 25 random split half trials. Images are ranked by predicted memorability and plotted against the cumulative average of ground truth memorability scores. Error bars omitted for clarity.

Fig. 19: The 8 images whose predicted memorability rank, on the basis of global features, most overshot ground truth memorability rank (a) and most undershot ground truth memorability rank (b). The mean rank error between predicted and measured ranks across each set of images is given in parentheses.

gives $\rho = 0.29$). Some photos of nature are consistently more memorable than others (Figure 20).

4.4 Memorability of aesthetic images

Another image subset of particular interest is those images marked as being aesthetic. We envision memorability scores as being a useful and novel way of quantifying image utility. However, for many applications, we may want images that are not just memorable, but are also good in other ways. For example, a photographer may want to identify images that are both memorable and beautiful – photos of office chairs and toilets, despite the fact that they may be memorable, probably will not do. We are thus interested in combining multiple photo quality metrics at once. Given ground truth aesthetics ratings, can we automatically pick out the images that are both beautiful and memorable?

Here we selected the top 250 photos with the highest value of the *Is this an aesthetic image?* attribute defined in section 3.1. Split half consistency among experiment participants was $\rho = 0.76$ and our predictions, trained on all photos and tested on the selected aesthetic photos, reached $\rho = 0.31$ (training on just the selected aesthetic photos gives $\rho = 0.28$).

5 Conclusion

Making memorable images is a challenging task in visualization and photography, and is generally presented as a vague concept hard to quantify. Surprisingly, there has been no previous attempt to systematically measure this property on image collections, and to apply computer vision techniques to extract memorability automatically. Measuring subjective properties of photographs is an active domain of research with numerous applications. Our present work could be used to extract, from a collection of images, the ones that are most likely to be remembered by viewers. This could be applied to selecting images for illustrations, covers, user interfaces, educational materials, memory clinical rehabilitation, and more.

In this paper we have shown that predicting image memorability is a task that can be addressed with current computer vision techniques. We have measured memorability using a restricted experimental setting in order to obtain a meaningful quantity: we defined an image's memorability score as the probability that a viewer will detect a repeat of the image within a stream of pictures. We have shown that there is a large degree of consistency among different viewers, even at different time delays, and that some images are more memorable than others even when there are no familiar elements (such as relatives or famous monuments). This work is a first attempt to quantify this important property of individual images. Future work will investigate the relationship between image memorability and other measures such as object importance [17], [18], saliency [12], and photo quality [8].

Acknowledgments

We would like to thank Timothy Brady and Talia Konkle for helpful discussions and advice. This work is supported by the National Science Foundation under Grant No. 1016862 to A.O, NSF CAREER Award No. 0747120, Intelligence Advanced Research Projects Activity via Department of the Interior contract D10PC20023, and ONR MURI N000141010933 to A.T, as well as Google and Xerox awards to A.O and A.T. P.I. is supported by a National Science Foundation Graduate Research Fellowship. J.X. is supported by a Google U.S./Canada Ph.D. Fellowship.

References

[1] L. Standing, "Learning 10,000 pictures," *Quarterly Journal of Experimental Psychology*, 1973.

Fig. 20: Memorability predictions within particular image type subsets. Rows labeled "predicted" give the images our system predicts as most memorable (left) and most forgettable (right) within each subset. Rows labeled "ground truth" give the images found as most memorable (left) and most forgettable (right) using our memory game measurements. For the "predicted" rows, ρ values measure rank correlation between predictions and ground truth. For the "ground truth" rows, ρ values measure rank correlation between independent sets of empirical measurements.

[2] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva, "Conceptual distinctiveness supports detailed visual long-term memory for real-world objects," *JEP:G*, 2010.

[3] ——, "Scene memory is more detailed than you think: the role of categories in visual long-term memory," *Psych Science*, 2010.

[4] S. Vogt and S. Magnussen, "Long-term memory for 400 pictures on a common theme," *Experimental Psychology*, 2007.

[5] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, "Visual long-term memory has a massive storage capacity for object details," *Proc Natl Acad Sci, USA*, 2008.

[6] T. Brady, T. Konkle, G. Alvarez, and A. Oliva, "Are real-world objects represented as bound units? Independent forgetting of different object details from visual memory." *Journal of Experimental Psychology: General*, 2012.

[7] R. Hunt and J. Worthen, *Distinctiveness and Memory*. New York, NY: Oxford University Press, 2006.

[8] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *ECCV*, 2008.

[9] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *ICCV*, 2011.

[10] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," *CVPR*, 2012.

[11] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *CVPR*, 2011.

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, 1998.

[13] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Data-driven enhancement of facial attractiveness," *ToG*, 2008.

[14] B. Gooch, E. Reinhard, C. Moulding, and P. Shirley, "Artistic composition for image creation," in *Proc. of the 12th Eurographics workshop on Rendering Techniques*, 2001.

[15] L. Renjie, C. L. Wolf, and D. Cohen-or, "Optimizing photo composition," Tel-Aviv University, Tech. Rep., 2010.

[16] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y. Xu, "Color harmonization," *ToG*, 2006.

[17] M. Spain and P. Perona, "Some objects are more equal than others: Measuring and predicting importance," in *ECCV*, 2008.

[18] A. Berg, T. Berg, H. Daume III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi, "Understanding and Predicting Importance in Images," *CVPR*, 2012.

[19] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behavioral and Brain Sciences*, 2001.

[20] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable," in *CVPR*, 2011.

[21] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the instrinsic memorability of images," in *NIPS*, 2011.

[22] S. T. Charles and M. Mather, "Aging and emotional memory: the forgettable nature of negative images for older adults." *Journal of Experimental Psychology: General*, 2003.

[23] A. D'Argembeau, M. Van der Linden, C. Comblain, and A.-M. Etienne, "The effects of happy and angry expressions on identity and expression memory for unfamiliar faces," *Cognition & Emotion*, vol. 17, no. 4, pp. 609–622, Jan. 2003.

[24] J. F. Cross and J. Cross, "Sex, race, age, and beauty as factors in recognition of faces," *Attention*, 1971.

[25] W. Bainbridge, P. Isola, I. Blank, and A. Oliva, "Establishing a database for studying human face photograph memory," *Proceedings of the Cognitive Science Society*, 2012.

[26] B. Tversky, "Memory for faces: Are caricatures better than photographs?" *Memory & cognition*, 1985.
[27] B. Gooch and E. Reinhard, "Human facial illustrations: Creation and psychophysical evaluation," *Tog*, 2004.
[28] J. Xiao, J. Hayes, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010.
[29] M. W. Eysenck, *Depth, elaboration, and distinctiveness*. Levels of processing in human memory, 1979.
[30] J. S. Nairne, "Modeling distinctiveness: Implications for general memory theory," *Distinctiveness and memory*, 2006.
[31] K. A. Rawson, "How does knowledge promote memory? The distinctiveness theory of skilled memory," *Journal of Memory and Language*, 2008.
[32] S. R. Schmidt, "Encoding and retrieval processes in the memory for conceptually distinctive events." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1985.
[33] W. Koutstaal and D. Schacter, "Gist-based false recognition of pictures in older and younger adults," *Journal of Memory and Language*, 1997.
[34] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *IJCV*, 2008.
[35] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," *CVPR*, 2010.
[36] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.
[37] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," in *NIPS*, 2012.
[38] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, 2010.
[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
[40] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *CVPR*, 2007.
[41] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *IJCV*, 2001.
[42] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

Phillip Isola is a Ph.D. candidate in Brain and Cognitive Sciences at the Massachusetts Institute of Technology. He received a B.S. degree in Computer Science from Yale University in 2008. His research interests include human vision and computer vision, and the interplay between the two. He received a National Science Foundation Graduate Research Fellowship in 2010.

Jianxiong Xiao is a Ph.D. candidate in the Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT). Before that, he received a B.Eng. and a M.Phil. from the Hong Kong University of Science and Technology. Starting in September 2013, he will be an assistant professor in the Department of Computer Science in Princeton University. His research interests are in computer vision, with a focus on scene understanding. His work has received the Best Student Paper Award at the European Conference on Computer Vision (ECCV) in 2012, and has appeared in the popular press. Jianxiong was awarded the Google U.S./Canada Ph.D. Fellowship in Computer Vision in 2012 and MIT CSW Best Research Award in 2011. He is a student member of IEEE.

Devi Parikh is an Assistant Professor in the Bardley Department of Electrical and Computer Engineering at Virginia Tech. Prior to this she was a Research Assistant Professor at TTI-Chicago, an academic computer science institute affiliated with University of Chicago. She received her M.S. and Ph.D. degrees from the Electrical and Computer Engineering department at Carnegie Mellon University in 2007 and 2009 respectively. She received her B.S. in Electrical and Computer Engineering from Rowan University in 2005. Her research interests include computer vision, pattern recognition and AI in general and visual recognition problems in particular. She was a recipient of the Carnegie Mellon Deans Fellowship, National Science Foundation Graduate Research Fellowship, Outstanding Reviewer Award at CVPR 2012, Google Faculty Research Award in 2012 and the 2011 Marr Best Paper Prize awarded at ICCV.

Antonio Torralba received the degree in telecommunications engineering from Telecom BCN, Spain, in 1994 and the Ph.D. degree in signal, image, and speech processing from the Institut National Polytechnique de Grenoble, Grenoble, France, in 2000. He is an Associate Professor of Electrical Engineering and Computer Science at the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge. From 2000 to 2005, he spent postdoctoral training at the Brain and Cognitive Science Department and the Computer Science and Artificial Intelligence Laboratory, MIT. Dr. Torralba is an Associate Editor of the IEEE Trans. on Pattern Analysis and Machine Intelligence, and of the International Journal in Computer Vision. He received the 2008 National Science Foundation (NSF) Career award, the best student paper award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2009, and the 2010 J. K. Aggarwal Prize from the International Association for Pattern Recognition (IAPR).

Aude Oliva After a French baccalaureate in Physics and Mathematics and a B.Sc. in Psychology, Aude Oliva received two M.Sc. degrees in Experimental Psychology, and in Cognitive Science and a Ph.D from the Institut National Polytechnique of Grenoble, France. She joined the MIT faculty in the Department of Brain and Cognitive Sciences in 2004 and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) as a Principal Research Scientist in 2012. Her cross-disciplinary work in Computational Perception and Cognition builds on the synergy between human and machine vision, and how it applies to solving high-level recognition problems like understanding scenes, perceiving space, recognizing objects, modeling eye movements and visual memory, as well as predicting subjective properties of images like image memorability. Her work has been featured in the scientific and popular press and has made its way in textbooks of Perception, Cognition, Computer Vision, Design, as well as in museums of Art and Science. She is the recipient of a National Science Foundation CAREER Award.