MemCat: A new category-based image set quantified on memorability

Lore Goetschalckx Corresp., [1] , **Johan Wagemans** [1]

[1] Brain & Cognition, KU Leuven, Leuven, Belgium

Corresponding Author: Lore Goetschalckx
Email address: lore.goetschalckx@kuleuven.be

Images differ in their memorability in consistent ways across observers. What makes an image memorable is not fully understood to date. Most of the current insight is in terms of high-level semantic aspects, related to the content. However, research still shows consistent differences within semantic categories, suggesting a role for factors at other levels of processing in the visual hierarchy. To aid investigations into this role as well as contributions to the understanding of image memorability more generally, we present MemCat. MemCat is a category-based image set, consisting of 10K images representing five broader, memorability-relevant categories (animal, food, landscape, sports, and vehicle) and further divided into subcategories (e.g., bear). They were sampled from existing source image sets that offer bounding box annotations or more detailed segmentation masks. We collected memorability scores for all 10K images, each score based on the responses of on average 99 participants in a repeat-detection memory task. Replicating previous research, the collected memorability scores show high levels of consistency across observers. Currently, MemCat is the second largest memorability image set and the largest offering a category-based structure. MemCat can be used to study the factors underlying the variability in image memorability, including the variability within semantic categories. In addition, it offers a new benchmark dataset for the automatic prediction of memorability scores (e.g., with convolutional neural networks). Finally, MemCat allows the study of neural and behavioral correlates of memorability while controlling for semantic category.

MemCat: A new category-based image set quantified on memorability

Lore Goetschalckx[1], Johan Wagemans[1]

[1]Brain & Cognition, KU Leuven, Leuven, Vlaams-Brabant, Belgium

Corresponding author:
Lore Goetschalckx[1]
Tiensestraat 102, Leuven, Vlaams Brabant, 3000, Belgium
Email address: lore.goetschalckx@kuleuven.be

Abstract

Images differ in their memorability in consistent ways across observers. What makes an image memorable is not fully understood to date. Most of the current insight is in terms of high-level semantic aspects, related to the content. However, research still shows consistent differences within semantic categories, suggesting a role for factors at other levels of processing in the visual hierarchy. To aid investigations into this role as well as contributions to the understanding of image memorability more generally, we present MemCat. MemCat is a category-based image set, consisting of 10,000 images representing five broader, memorability-relevant categories (animal, food, landscape, sports, and vehicle) and further divided into subcategories (e.g., bear). They were sampled from existing source image sets that offer bounding box annotations or more detailed segmentation masks. We collected memorability scores for all 10,000 images, each score based on the responses of on average 99 participants in a repeat-detection memory task. Replicating previous research, the collected memorability scores show high levels of consistency across observers. Currently, MemCat is the second largest memorability image set and the largest offering a category-based structure. MemCat can be used to study the factors underlying the variability in image memorability, including the variability within semantic categories. In addition, it offers a new benchmark dataset for the automatic prediction of memorability scores (e.g., with convolutional neural networks). Finally, MemCat allows the study of neural and behavioral correlates of memorability while controlling for semantic category.

Introduction

A large body of work within the visual memory field has been devoted to questions about its capacity and fidelity (for a review, see Brady, Konkle, & Alvarez, 2011). Often, these studies make abstraction of the to-be-remembered stimuli and potential differences between them. Yet, work by Isola, Xiao, Parikh, Torralba, and Oliva (2014), using everyday images, showed that they do not all share the same baseline likelihood of being remembered and recognized later. Instead, images differ in "memorability" in ways that are consistent across participants and this can be measured reliably (Isola et al., 2014).

To assess memorability, Isola et al. (2014) used a repeat-detection memory task, in which participants watch a sequence of images and respond whenever they see a repeat of a previously shown image. The researchers assigned a memorability score to 2222 scene images based on the proportion of participants recognizing the image upon its repeat. They found that memorability rank scores were highly consistent across participants. In other words, there was a lot of agreement as to which images were remembered and recognized, and which ones were easily forgotten. This suggests that memorability can indeed be considered an intrinsic image property and that whether you will remember a certain image does not only depend on you as an individual, but also on the image itself. The result has furthermore been replicated with a more traditional long-term visual memory task, with a separate study and test phase (Goetschalckx, Moors, & Wagemans, 2018). Moreover, image memorability rankings have been shown to be stable across time (Goetschalckx, Moors, & Wagemans, 2018; Isola et al., 2014), across contexts (Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015), and across encoding types (intentional versus incidental; Goetschalckx, Moors, & Wagemans, 2019). Finally, while they might be related to some extent, image memorability does not simply boil down to popularity (Khosla, Raju, Torralba, & Oliva, 2015), aesthetics (Isola et al., 2014; Khosla et al., 2015), interestingness (Gygli, Grabner, Riemenschneider, Nater, & Van Gool, 2013; Isola et al., 2014), or the ability of an image to capture attention (Bainbridge, 2017).

The findings spurred new research aimed at understanding and predicting memorability. When it comes to merely predicting the memorability score of an image, the best results so far are achieved using convolutional neural networks (CNNs; e.g., Khosla et al., 2015). When it comes to truly understanding, on the other hand, CNNs have often been critiqued to be black boxes (however, see Benitez, Castro, and Requena, 1997 for counterarguments). It is not always clear to us humans why a CNN predicts a certain score for one image and not another. Nonetheless, Khosla et al.'s (2015) analyses provided some further insight, mostly pointing at differences between broader image categories and content types. For example, units in the network displaying the highest correlation with memorability seemed to respond mostly to humans, faces, and objects, while those with the lowest correlation seemed to respond to landscapes and open scenes. Furthermore, the most memorable regions of an image, according to the CNN, often capture people, animals or text. These findings are in line with earlier work, which also predominantly revealed high-level semantic attributes. Isola et al. (2014), for example, showed that the predictive performance of a model trained on mere object statistics (e.g., number of objects) was boosted considerably when the object labels were taken into account. In addition, a model trained on the overall scene labels alone, already predicted memorability scores with a Spearman's rank correlation of .37 to the ground truth. Memorable images often had labels referring to people, interiors, foregrounds, and human-scaled objects, while labels referring to exteriors, wide-angle vistas, backgrounds, and natural scenes were associated with low image memorability scores.

Together, these findings suggest that a fair share of the variability in memorability resides in differences between semantic categories. Perhaps this is not surprising considering the central position occupied by categories in the broader cognitive system. It has been said that carving up the world around us into meaningful categories of stimuli that can be considered equivalent is a core function of all organisms (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). It helps us understand novel events and make predictions about it (Medin & Coley, 1998). According to Rosch et al. (1976), categories are represented hierarchically and organized into a taxonomy of different levels of abstraction. The basic level is the best compromise between providing enough information and being cognitively inexpensive. It is also the preferred naming level (e.g., “cat”). Other levels can be superordinate (e.g., “feline” or “mammal”) or subordinate (e.g., “tabby”). Recently, Akagunduz, Bors, and Evans (2019) pointed out that categories are also used to organize memory. More specifically, instead of encoding an image as a mere collection of pixels, we extract visual memory schemas associated with its category (i.e., key regions and objects and their interrelations), along with an image's idiosyncrasies. To map these visual memory schemas, they had participants indicate which image regions made them recognize the image. The resulting maps showed high consistency across participants, suggesting that visual memory schemas partly determine what participants find memorable. Moreover, humans can visually categorize an object depicted in an image very rapidly and accurately (e.g., Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; Fei-Fei, Iyer, Koch, & Perona, 2007; Greene & Oliva, 2009), as well as categorize the image at the level of the whole scene (e.g., Delorme, Richard, & Fabre-Thorpe, 2000; VanRullen & Thorpe, 2001; Xu, Kankanhalli, & Zhao, 2019). Often a single glance suffices. Interestingly, Broers, Potter, and Nieuwenstein (2018) observed enhanced recognition performance for memorable versus non-memorable images in an ultra-rapid serial visual presentation task. Finally, there is also evidence for a category hierarchy in the representations in high-level human visual cortex, with for example clusters for animacy and subclusters for faces and body parts (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cichy, Pantazis, & Oliva, 2014; Kriegeskorte et al., 2008).

While semantic categories (or labels) seem to play a large role in image memorability, they do not explain all the observed variability. Interestingly, consistent differences in memorability scores remain even *within* image categories (Bylinskii et al., 2015). Goetschalckx, Moors, Vanmarcke, and Wagemans (2019) for example, have argued that part of that variability might be due to differences in how well the image is organized. Nonetheless, the concept of memorability and its correlates is not yet fully understood to date and further research is required to paint a clearer picture. The current work presents a novel, category-based dataset of images quantified on memorability, designed for research to achieve this goal (example images in Figure 1).

To the best of our knowledge, there were previously five large image sets with memorability scores, three consisting of regular photographs, which is also the focus here: Isola et al. (2014), FIGRIM (Bylinskii et al., 2015), and LaMem (Khosla et al., 2015), and two more specialized sets, which we will not further discuss here: Bainbridge, Isola, and Oliva (2013; face images), and Borkin et al. (2013; data visualizations). For completeness, we also mention a smaller set (850 images) that was used to study which objects in an image are memorable (Dubey, Peterson, Khosla, Yang, & Ghanem, 2015). Table 1 compares MemCat to the other large datasets. The comparison is discussed in more detail below.

A first feature of the current dataset is its hierarchical category structure. It was designed to be representative for five different broad natural categories and to allow the study of memorability differences within semantic categories. The set is characterized by a hierarchy of five broader categories, further divided into more fine-grained subcategories. Only the FIGRIM set also offers a category structure, but the number of exemplar images per category was lower: 59-157, compared to 2000 in the current set. We opted for broad categories to ensure that the whole was still varied and representative enough, while containing a large number of exemplar images per category at the same time. Moreover, our final choice of categories: animal, food, landscape, sports, and vehicle, was motivated by their relevance for memorability, meaning that they (or related categories) have been observed to differ in their overall memorability in previous research (Isola, Xiao, Parikh, Torralba, & Oliva, 2011; Isola et al., 2014; Aditya Khosla, Raju, Torralba, & Oliva, 2015). For example, knowing that the presence of people in an image is predictive for memorability (see above), we chose one category of images depicting people as the main subject and avoided including images with people in other categories. For this one category, we chose “sports”, because “people” in itself constitutes a category that was too broad in comparison to the other categories and did not lend itself well for a division into subcategories. Furthermore, we included an animal category as a non-human animate category, food and vehicle as more object-based categories, and landscape to represent the wide exteriors that are often associated with lower memorability scores (Isola et al., 2011, 2014; Khosla et al., 2015).

Second, we aimed for a large set, such that it would be suitable for machine learning approaches. With a total of 10,000 images quantified on memorability, the current set is the second largest memorability dataset, after LaMem.

Third, we sampled images from existing datasets, such that the image annotations collected there would also be available for researchers studying memorability. In particular, we searched for images annotated with segmentation masks or at least bounding boxes, reasoning that they may hold some indications of how the image is organized (e.g., where is the subject located), which might be of particular interest when studying memorability within categories and factors other than semantics.

In summary, the unique combination of features of MemCat, together with its richness in data, make it a valuable addition to the memorability. Among the possible uses by memorability researchers are the study of what makes an image memorable beyond its category, a benchmark for machine learning approaches, and a semantically controlled stimulus set for psychophysical or neuroscientific studies about the correlates of memorability (elaborated in the Discussion section). However, given that categorization is a core function of the human mind, MemCat would also appeal to a much broader range of cognitive (neuro)scientists.

Materials & Methods

Participants. There were 249 undergraduate psychology students (KU Leuven) who participated in this study in exchange for course credits (216 female, 32 male, 1 other). Four students did not disclose their age and the remainder were aged between 18 and 27 years old (M = 19.24, SD= 0.94). The majority of the participants, however, were recruited through Amazon's Mechanical Turk (AMT) and received a monetary compensation (see further for details). The settings on AMT were chosen such that only workers who indicated to be at least 18 years old and living in the USA could participate. Further eligibility criteria were that the worker had to have an approval rate of at least 95% on previous human intelligence tasks (HITs) and a total number of previously approved HITs of at least 100. A total of 2162 AMT-workers participated in this study (1139 female, 917 male, 4 other, and 102 who did not disclose this information). For the 1851 workers who disclosed their age, the reported ages ranged between 18 and 82 years old (M = 37.14, SD = 11.89). The AMT data collection took place from April 2018 till July 2018. Data collected through AMT has been shown to come from participant samples that are more diverse than student samples and to be comparable in quality and reliability to those collected in the lab (e.g., Buhrmester, Kwang, & Gosling, 2011).

Materials. MemCat consists of 10,000 images sampled from four previously existing image sets: ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), SUN (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010), and The Open Images Dataset V4 (Kuznetsova et al., 2018). The four source sets were chosen because of their large size (i.e., number of images), the availability of semantic annotations, and the availability of bounding box annotations or more complete segmentation masks for at least a subset of their images. The images selected from the source sets to be included in MemCat belonged to the five broader semantic categories outlined in the Introduction: animal (2000 images), food (2000 images), landscape (2000 images), sports (2000 images), and vehicle (2000 images). We explain the different steps in the selection procedure in more detail below.

As a first step, we listed at least 20 subcategories for each broader category. The goal was to obtain 2000 images per category, without including more than 100 exemplar images per subcategory. This was to ensure a reasonable level of variability and to avoid high levels of false alarms in the memory task (see further). The subcategories were then translated to semantic annotations from the source dataset. For example, for the subcategory “bear” (animal), we used COCO images annotated with a “bear” tag and ImageNet images from nodes “American black bear”. “brown bear”, and “grizzly”. An overview of our hierarchy of categories and subcategories, can be found in Figure 2.

The second step consisted of automatically sampling exemplar images from the listed subcategories, while satisfying a number of shape restrictions. To avoid that images would stand out because of an extreme aspect ratio, we only sampled images with aspect ratios between 1:2 and 2:1. Furthermore, the minimum resolution was set to 62,500 pixels. Finally, to ensure that the images would fit comfortably on most computer monitors, we adopted a maximum height of 500 pixels and a maximum width of 800 pixels. However, for SUN and The Open Images Dataset, only a low number of images satisfied the latter two restrictions (they were often too big), which is why we opted to resize (using Hamming interpolation) images from those two source datasets to meet the restrictions. Apart from the shape restriction, we also restricted the sampling to images for which bounding box annotations or more complete segmentation masks were available from the source datasets. Finally, we sampled more images than the target number (2000 images per broad category), anticipating exclusions in the next step.

The third step constituted manual selection work, carried out by the first author, assisted by two student-interns. We manually went through the exemplar images sampled in the previous step, and eliminated images following a number of exclusion rules. The exclusion rules can roughly be divided into two kinds. A first kind of exclusion rule touches upon the quality of the image. We excluded images of poor image quality (e.g., very dark, very much overexposed, blurry, etc.), images that did not convincingly belong to the subcategory they were assigned to,[1] images in greyscale or looking like they were the result of another color filter, images that were not real photographs (e.g., drawings, digitally manipulated images, computer generated images), and collages. A second set of rules concerns factors that could affect the memorability of an image, but were not of interest for the purpose of MemCat. One such factor is text. We excluded images containing large, readable text or text not belonging to the image itself (e.g., date of capture). Another factor was the presence of people in the image. There was one designated “people” category, the sports category, meaning that every included exemplar image depicted one or more people practicing sports. However, the presence of people was avoided in all other categories (but we allowed anonymous people in the background in the vehicle category or the presence of a hand in images of the food category). Furthermore, images depicting remarkably odd scenes (e.g., dog wearing Santa Clause costume) were also excluded. Similarly, we avoided images depicting famous places or people (e.g., Roger Federer or Cristiano Ronaldo in the sports category), and images of dead, wounded or fighting animals. In addition to these exclusion rules, we also tried, to the best of our ability, not to include (near) duplicate images. If the target number of images was not obtained after Step 3, we reverted back to Step 2, if there were still images to sample from, or to Step 1 if we needed to include additional subcategories.

Finally, for those categories for which more than the target number of images survived Step 3, there was a fourth step to randomly down-sample the selection to the target number, assigning higher sampling probabilities to images annotated with segmentation masks.

In addition to MemCat, we collected 10,000 filler images, that were not quantified on memorability themselves, but were needed in the memory task used to quantify the other images. The filler images were sampled randomly from The Open Images Dataset, but from a different subset to avoid overlap.[2] As these images would function only as filler images, there were fewer restrictions. For example, the images could be of any category, they were allowed to contain text, etc. However, the same shape restrictions were still applied.

Procedure. Having carefully collected 10,000 images for MemCat, the next step was to quantify them on memorability. Following previous work, this was achieved by presenting the images in an online repeat-detection memory game (Isola et al., 2014; Khosla et al., 2015), in which participants watch a sequence of images and are asked to respond when they recognize a repeat of a previously shown image. Students participating for course credits played the game in the university's computer labs, hosting about 20 students at a time. AMT workers played the game from the comfort of their homes (or whichever location they preferred). Prior to starting the game, all participants were prompted to read through an informed consent page explaining the aims of the study and their rights as participants. They could give their consent by actively ticking a box. The study was approved by SMEC, the Ethical Committee of the Division of Humanities and Social Sciences, KU Leuven, Belgium (approval number: G-2015 08 298).

For the task design of the game, we closely followed Khosla et al. (2015), as their version of the game was designed to quantify large numbers of images. We divided the game into blocks of

200 trials. On each trial, an image was presented at the center of the browser window for a duration of 600 ms, with an intertrial interval of 800 ms. During this interval, a fixation cross was shown. Sixty-six images were target images, sampled randomly from MemCat, and repeated after 19 to 149 intervening images. Forty-four images were random filler images that were never repeated. Finally, there were 12 additional random filler images that were repeated after 0 to 6 intervening images to keep participants attentive and motivated. They are referred to as vigilance trials. Participants could indicate that they recognized a repeat by pressing the space-bar. They did not receive trial-by-trial feedback, but were shown their hit rate as well as number of false alarms at the end of the block. Figure 3 presents a schematic of the game.

Each block lasted a little less than 5 min. Care was taken to ensure that an image was never repeated more than once and never across blocks. Students were asked to complete as many blocks as they could in one hour, with one bigger, collective break of roughly 10 min after half an hour, and smaller self-timed breaks between the remainder of the blocks. Most students could complete eight blocks, but for some groups, slow data uploads at the end of a block resulted in lower numbers. AMT workers could complete one to 16 blocks, were allotted 48h to submit their completed blocks (so, they were allowed to spread the blocks over time), and were paid $0.40 per block. To ensure a good quality of the AMT data and to avoid random or disingenuous responses, AMT workers were blocked from playing anymore blocks after two with a d' lower than 1.5 on the vigilance trials. They were warned the first time this happened.

Memorability Measures. We computed two different, but related measures of memorability from the data collected through the repeat-detection memory game. These were the same two measures as used in LaMem, the largest available image memorability dataset yet. As mentioned in the Introduction, one measure is simply the proportion of participants recognizing the image when shown to them for the second time (i.e., the hit rate across participants). This is the "original" memorability measure, as introduced by Isola et al. (2014), also adopted in many other memorability studies (e.g., Bainbridge, Isola, & Oliva, 2013; Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015; Khosla, Raju, Torralba, & Oliva, 2015). The other memorability measure computed for the LaMem images was based on the same principle, but penalized for false alarms (i.e., when participants press the space-bar for the first presentation of the image) in the way proposed by Khosla, Bainbridge, Torralba, and Oliva (2013), who applied it to a dataset of face images (Bainbridge et al., 2013). Rather than H/N_{resp} (first measure), their formula was the following: $(H\text{-}F)/N_{resp}$, where H is the number of participants recognizing the image, F is the number of participants making a false alarm when the image is presented for the first time, and N_{resp} is the total number of participants having been presented with the image. Here, an image's N_{resp} was 99 (after exclusions) on average. Note that the memorability scores have an upper bound of 1 and a lower bound of 0. In theory, $(H\text{-}F)/N_{resp}$ could result in a negative score, but in practice it is highly unlikely that there would be more participants making a false alarm for the image than there are participants making a hit.

Results

Participant Performance. As mentioned, the performance on the easier vigilance trials was taken as an indication of whether participants were playing the memory game in a genuine way. If in a certain block, a participant did not distinguish vigilance repeats from non-repeat trials with a *d′* of at least 1.5 (preset performance threshold), that block was excluded from further analyses. The exclusion rate amounted to 3% of all played blocks. Recall, however, that AMT workers were not allowed to play more blocks after two excluded ones.

After exclusion, the mean *d′* across participants was 2.77 (SD = 0.56) for the vigilance repeats, and 2.47 (SD = 0.50) for the target repeats. Table 2 summarizes participants' overall performance, collapsing over vigilance and target repeats. Participants generally performed well on the task.

Memorability Scores. Participants' high performance was also reflected in the average image memorability scores. Figure 4 displays the mean for each of the two memorability measures as a horizontal line ($M_{H/Nresp}$ = .76, SD; $M_{(H\text{-}F)/Nresp}$ = .70). It is comparable to the mean observed in Khosla et al. (2015). In addition, Figure 4 visualizes the distribution of the collected image memorability scores for each of the five broad main categories separately. A simple linear regression revealed that the category explains 43% of the variance in the H/N_{resp} scores and 44% of the variance in the $(H\text{-}F)/N_{resp.}$ In line with previous research, the landscape images were on average the least memorable ($M_{H/Nresp}$ = .60; $M_{(H\text{-}F)/Nresp}$ = .53). They were followed by the vehicle images ($M_{H/Nresp}$ = .76; $M_{(H\text{-}F)/Nresp}$ = .70). Somewhat surprisingly, the food images generally came out on top of the ranking ($M_{H/Nresp}$ = .85; $M_{(H\text{-}F)/Nresp}$ = .80), topping the animal $M_{H/Nresp}$ = .80; $M_{(H\text{-}F)/Nresp}$ = .73) and sports $M_{H/Nresp}$ = .78; $M_{(H\text{-}F)/Nresp}$ = .71) categories. However, there is still a large degree of variability that is not explained by differences in broad image categories. Indeed, memorability varied considerably within categories as well, with *SDs* of: .09 (animal; $SD_{(H-F)/Nresp} = .09$), .08 (food; $SD_{(H-F)/Nresp} = .08$), .13 (landscape; $SD_{(H-F)/Nresp} = .14$), .09 (sports; $SD_{(H-F)/Nresp} = .10$), and .09 (vehicle; $SD_{(H-F)/Nresp} = .09$).

Having observed that images from the same broader category indeed still differed in memorability, the next question was whether these differences are consistent across participants. This question taps into the reliability of the memorability measures. Following previous memorability work (e.g., Isola et al., 2014), the consistency was assessed by randomly splitting the participant pool in half, computing the memorability scores for each half separately and determining the Spearman's rank correlation between the two sets of scores. This was repeated for 1000 splits and the Spearman's rank correlation was averaged across the splits. Figure 5 shows the results in function of the mean N_{resp} for each category as well as for the total image set.

We first discuss the results for H/ N_{resp} (see Figure 5, left panel). When collapsing over all five categories, the observed mean split-half Spearman's rank correlation with all available responses ($N_{resp} = 99$, on average) amounted to .78. In comparison, Khosla et al. (2015) reported a mean split-half Spearman's rank correlation of .67 for their LaMem dataset. However, they only collected 80 responses per image. After randomly down-sampling our data to an N_{resp} of 80, we still found a split-half consistency of .73. With the exception of the landscape category, for which we observed a total (i.e., without down-sampling) split-half consistency of .77, the total per category split-half consistency estimates were lower, ranging between .59 and .67. This is possibly due to smaller ranges of memorability scores within those categories (see Figure 4).

Note, however, that the split-half consistencies are an underestimate of the reliability of the memorability scores calculated based on the full participant pool. The latter can be estimated from the split-half consistency by means of the Spearman-Brown formula (Brown, 1910; Spearman, 1910). Applying this formula, we found the following final reliabilities for the H/N_{resp} memorability scores : .87 (all), .80 (animal), .75 (food), .87 (landscape), .75 (sports), .78 (vehicle).

For the (H-F)/N_{resp} memorability scores, we confine the discussion to pointing out that the pattern of results is qualitatively similar, although the final reliabilities are somewhat lower: .86 (all), .74 (animal), .71 (food), .85 (landscape), .77 (sports), .71 (vehicle).

Finally, after finding that the two image memorability measures were both acceptably reliable, we asked how they compared to each other. In the current dataset, they were highly intercorrelated, as evidenced by a Pearson correlation of .93 when collapsing over all five categories. The per category correlations were: .82 (animal), .90 (food), .91 (landscape), .85 (sports), .88 (vehicle).

Discussion

We presented a new dataset, MemCat, consisting of a total of 10,000 images, each quantified on memorability using a repeat-detection memory task (first introduced by Isola et al., 2014, version used here based on Khosla et al., 2015). MemCat is the second largest image memorability dataset available, and the largest that is based on a category structure. That is, it is divided into five broader, memorability-relevant semantic categories: animal, food, landscape, sports, and vehicle, each with 2000 exemplar images, which are further divided into subcategories (e.g., bear, cat, cow). Furthermore, the images were sampled from popular, existing datasets such that additional annotations available there (e.g., segmentations masks or bounding boxes) would also be available to researchers wishing to use MemCat for research aimed at investigating specific factors underlying memorability.

Replicating previous research, we found that images differ considerably in memorability and that these differences are highly consistent across participants. Part but not all of this variability can be explained by differences between the five broader semantic categories. Note, however, that this result is correlational in nature, and that one should be cautious drawing causal conclusions. In line with Bylinskii et al. (2015), considerable variability in memorability remained even within the categories. However, the consistency there was somewhat lower, probably because the variance was also lower. When the differences between images become smaller, it becomes harder to reliably and consistently distinguish them. Nevertheless, the consistency estimates per category were still high, indicating that we obtained reliable memorability scores. Finally, we reported results for two different methods to compute memorability scores. One is to compute the hit rate across participants: H/N_{resp}. This was the method used in the original work by Isola et al. (2014). However, in principle, it possible that some images elicit more key presses not because they are truly recognized, but for some other reason (e.g., they seem familiar). That is why Bainbridge et al. (2013) suggested to correct for false alarms (i.e., when participants press the key for the first presentation of an image, when it is not a repeat) by computing $(H\text{-}F)/N_{resp}$. We report both measures for comparison, but note that they lead to a highly similar pattern of results and are also strongly intercorrelated. In what follows, we discuss possible uses of MemCat.

Most of what we learned from previous studies about what makes an image memorable is specified in terms of semantic categories or content types (e.g., images of people are more memorable than landscapes). However, a considerable amount of variability was still left unexplained. A primary use of the current dataset is in studies aiming to better understand the factors underlying image memorability. In particular, with 2000 images for each of five broader categories, it allows to zoom in on variability within categories. This variability is of more interest to practical applications (e.g., advertising, education), because the semantic category or the content type (e.g., a certain product) will often be predefined and it will be a matter of choosing or creating a more memorable depiction of it. In addition to dividing the set into broad semantic categories, we also avoided variability due to other factors already discovered in previous studies (e.g., we excluded images depicting oddities, images containing text or recognizable places or faces), thus creating a set designed to help understand the previously unexplained variability in image memorability.

Second, MemCat is also useful as a benchmark for machine learning approaches to automatically predict memorability. Currently, LaMem (Khosla et al., 2015) is most often used, but models can now also be trained and tested on the current dataset. When taking Khosla et al.'s (2015) MemNet-CNN (without retraining), we found that its predictions show a rank correlation of .68 with the (H-F)/N_{resp} memorability scores in the current set, suggesting that there is room for improvement. Given the category structure in MemCat, one could explore, for the first time, memorability models with one or more layers that are specific to a category. Indeed, it is possible that what makes landscape images memorable is different from what makes animal images memorable.

Finally, a third possible use is in neuroscientific studies or psychophysical studies examining effects of memorability. The current set offers a large number of quantified images to choose from. Moreover, it facilitates matching memorability conditions (e.g., high versus low) on semantic category, something that is often done in neuroscientific studies (e.g., Bainbridge, Dilks, & Oliva, 2017; Khaligh-Razavi, Bainbridge, Pantazis, & Oliva, 2016; Mohsenzadeh, Mullin, Oliva, & Pantazis, 2019).

Usage. On the MemCat project page, http://gestaltrevision.be/projects/memcat/, we provide a link to the collection of 10,000 images as well as links to two data files, all hosted on OSF (also see Additional Information). One file describes the images that were used and contains columns indicating the image filename in its source dataset, the name of its source dataset, the category (e.g., animal) and subcategory (e.g., bear) we assigned it to, the label that was used to sample it from its source dataset (e.g., American black bear), the current width, the current height, the factor by which it was resized (both the original width and height were multiplied by this factor), the number of hits (H), the number of false alarms (FA), the number of participants it was presented to (N_{resp}), and the two memorability scores. The other file contains the data collected in the repeat-detection memory game. Its columns indicate the participant ID (anonymized), the participant's age, the participant's gender, whether or not they participated through AMT, the block number, the trial number, the image shown, the trial type (target, target repeat, filler, vigilance, vigilance repeat), the participant's response (hit, correct rejection, miss, false alarm), the screen width, and the screen height.

Conclusions

With MemCat, we present a large new dataset of 10,000 images fully annotated with ground truth memorability scores collected through an online repeat-detection memory task. It is the second largest memorability dataset to date and the largest with a hierarchical category structure. The results showed that images differ in memorability in ways that are consistent across participants, even within semantic categories. Among other things, MemCat allows the study of which factors might underlie such differences. Its richness in data and unique combination of features will appeal to a broad range of researchers in cognitive science and beyond (e.g., computer vision).

Acknowledgments

The authors would like to thank the student-interns who have assisted in the image selection and data collection: Justine Aeyels and Joran Geeraerts, and Christophe Bossens, who is responsible for the technical support in the lab and contributed greatly to the implementation of the repeat-detection memory task on Amazon's Mechanical Turk.

Endnotes

[1] This could happen, for example, with images from the COCO source dataset. COCO images do not come with a single, overall scene label, but instead come with multiple semantic tags describing what is in the image. For this reason, an image annotated with the tag "cat," for instance, could be more of a living room image that just happens to have a cat sleeping somewhere in a corner in the background.

[2] The source set is presented in three different subsets: train, validation, and test. We sampled from the validation and test subsets for MemCat, and from the train subset for the fillers images used in the memory task.

References

Akagunduz, E., Bors, A., & Evans, K. (2019). Defining image memorability using the visual memory schema. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Advance online publication. doi:10.1109/tpami.2019.2914392

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research, 45*(11), 1459–1469. doi:10.1016/j.visres.2005.01.004

Bainbridge, W. A. (2017). The resiliency of memorability: A predictor of memory separate from attention and priming. *arXiv e-prints.* arXiv: 1703.07738 [q-bio.NC]

Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage*, *149*, 141–152. doi:10.1016/j.neuroimage.2017.01.063

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. doi:10.1037/a0033872

Benitez, J. M., Castro, J. L., & Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, *8*(5), 1156–1164. doi:10.1109/72.623216

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2306–2315. doi:10.1109/TVCG.2013.234

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*(5:4), 1–34. doi:10.1167/11.5.4

Broers, N., Potter, M. C., & Nieuwenstein, M. R. (2017). Enhanced recognition of memorable pictures in ultra-fast RSVP. *Psychonomic Bulletin & Review*, 25(3), 1080–1086. doi: 10.3758/s13423-017-1295-7

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322. doi:10.1111/j.2044-8295.1910.tb00207.x

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*(Part B), 165–178. doi:10.1016/j.visres.2015.03.005

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1–1. doi:10.1167/13.10.1

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. doi:10.1038/nn.3635

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research*, *40*(16), 2187–2200. doi:10.1016/S0042-6989(00)00083-3

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. Red Hook, NY: Curran Associates. doi:10.1109/CVPR.2009.5206848

Dubey, R., Peterson, J., Khosla, A., Yang, M. H., & Ghanem, B. (2015). What makes an object memorable? *2015 IEEE International Conference on Computer Vision (ICCV)*, 1089–1097. doi:10.1109/ICCV.2015.130

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 10. doi:10.1167/7.1.10

Goetschalckx, L., Moors, J., & Wagemans, J. (2019). Incidental image memorability. *Memory, 27*(9), 1273-1282. doi:10.1080/09658211.2019.1652328

Goetschalckx, L., Moors, P., Vanmarcke, S., & Wagemans, J. (2019). Get the picture? Goodness of image organization contributes to image memorability. *Journal of Cognition, 2*(1). doi:10.5334/joc.80

Goetschalckx, L., Moors, P., & Wagemans, J. (2018). Image memorability across longer time intervals. *Memory*, *26*(5), 581–588. https://doi.org/10.1080/09658211.2017.1383435

Greene, M., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. doi:10.1016/j.cogpsych.2008.06.001

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In *2013 IEEE International Conference on Computer* Vision (ICCV). Red Hook, NY: Curran Associates. doi:10.1109/iccv.2013.205

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems (NIPS) 24* (pp. 2429-2437). Red Hook, NY: Curran Associates. doi:10.21236/ada554133

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. doi:10.1109/TPAMI.2013.200

Khaligh-Razavi, S.-M., Bainbridge, W. A., Pantazis, D., & Oliva, A. (2016). From what we perceive to what we remember: Characterizing representational dynamics of visual memorability. *bioRxiv*. doi:10.1101/049700. eprint: https://www.biorxiv.org/content/early/2016/04/22/049700.full.pdf

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2013). Modifying the memorability of face photographs. In 2*013 IEEE International Conference on Computer Vision (ICCV)*. Red Hook, NY: Curran Associates. doi:10.1109/iccv.2013.397

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2390–2398. Red Hook, NY: Curran Associates. doi:10.1109/ICCV.2015.275

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. doi:10.1016/j.neuron.2008.10.043

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., … Ferrari, V. (2018). The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR.* arXiv:1811.00982.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., … Dollár, P. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *2014 European Conference on Computer Vision (ECCV)* (pp. 740–755). Cham: Springer International Publishing. doi:10.1007/978-3-319-10602-1_48

Medin, D. L., & Coley, J. D. (1998). Concepts and categorization. In J. Hochberg (Ed.), *Perception and cognition at century's end* (pp. 403–439). San Diego, CA: Academic Press. doi:10.1016/B978-012301160-2/50015-0

Mohsenzadeh, Y., Mullin, C., Oliva, A., & Pantazis, D. (2019). The perceptual neural trace of memorable unseen scenes. *Scientific Reports, 9*(1), 6033. doi:10.1038/s41598-019-42429-x

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655–668. https://doi.org/10.1068/p3029

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3485–3492). Red Hook, NY: Curran Associates. doi:10.1109/CVPR.2010.5539970

Xu, B., Kankanhalli, M. S., & Zhao, Q. (2019). Ultra-rapid object categorization in real-world scenes with top-down manipulations. *PLOS ONE*, *14*(4), e0214444. https://doi.org/10.1371/journal.pone.0214444

Figure 1

Example images of MemCat.

The memorability score, calculated as the hit rate across participants (H/Nresp) is indicated in the bottom right corner. In line with previous research, images differed consistently in their memorability score, even within semantic categories. MemCat represents five broader semantic categories: animal, food, landscape, sports and vehicle. Each row (A–C) displays exemplar images in that category order.

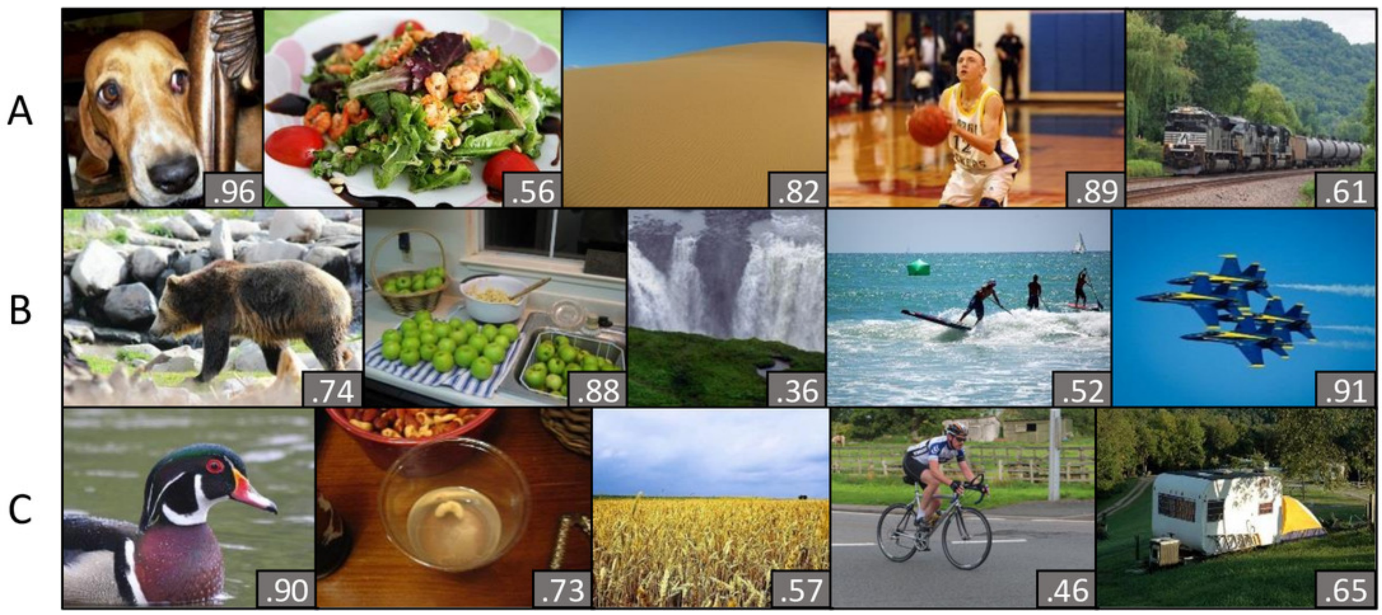


Figure 2

Category hierarchy of MemCat.

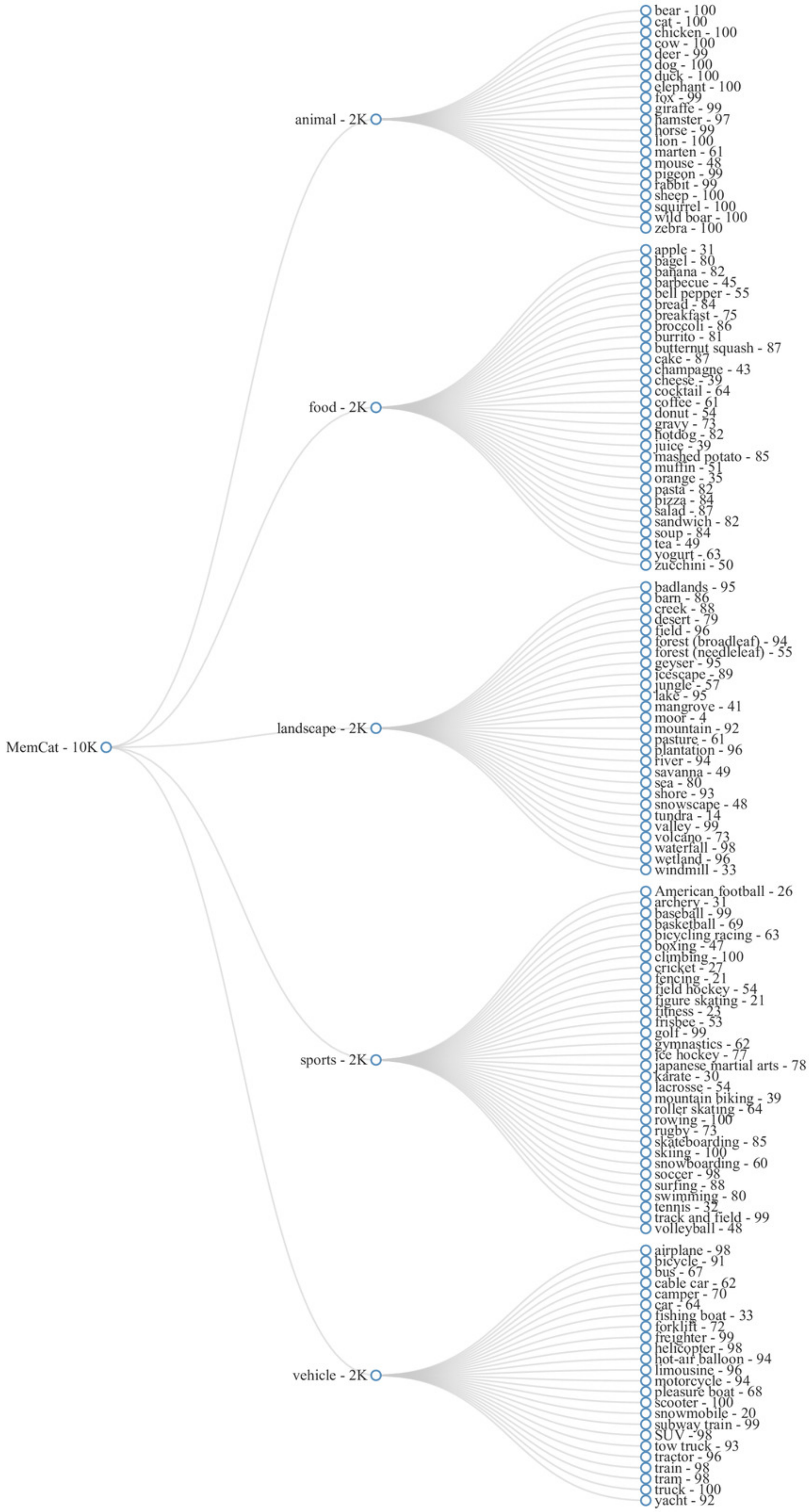

MemCat - 10K
animal - 2K
bear - 100
cat - 100
chicken - 100
cow - 100
deer - 99
dog - 100
duck - 100
elephant - 100
fox - 99
giraffe - 99
hamster - 97
horse - 99
lion - 100
marten - 61
mouse - 48
pigeon - 99
rabbit - 99
sheep - 100
squirrel - 100
wild boar - 100
zebra - 100
food - 2K
apple - 31
bagel - 80
banana - 82
barbecue - 45
bell pepper - 55
bread - 84
breakfast - 75
broccoli - 86
burrito - 81
butternut squash - 87
cake - 87
champagne - 43
cheese - 39
cocktail - 64
coffee - 61
donut - 54
gravy - 73
hotdog - 82
juice - 39
mashed potato - 85
muffin - 51
orange - 35
pasta - 82
pizza - 84
salad - 87
sandwich - 82
soup - 84
tea - 49
yogurt - 63
zucchini - 50
landscape - 2K
badlands - 95
barn - 86
creek - 88
desert - 79
field - 96
forest (broadleaf) - 94
forest (needleleaf) - 55
geyser - 95
icescape - 89
jungle - 57
lake - 95
mangrove - 41
moor - 4
mountain - 92
pasture - 61
plantation - 96
river - 94
savanna - 49
sea - 80
shore - 93
snowscape - 48
tundra - 14
valley - 99
volcano - 73
waterfall - 98
wetland - 96
windmill - 33
sports - 2K
American football - 26
archery - 31
baseball - 99
basketball - 69
bicycling racing - 63
boxing - 47
climbing - 100
cricket - 27
fencing - 21
field hockey - 54
figure skating - 21
fitness - 23
frisbee - 53
golf - 99
gymnastics - 62
ice hockey - 77
japanese martial arts - 78
karate - 30
lacrosse - 54
mountain biking - 39
roller skating - 64
rowing - 100
rugby - 73
skateboarding - 85
skiing - 100
snowboarding - 60
soccer - 98
surfing - 88
swimming - 80
tennis - 32
track and field - 99
volleyball - 48
vehicle - 2K
airplane - 98
bicycle - 91
bus - 67
cable car - 62
camper - 70
car - 64
fishing boat - 33
forklift - 72
freighter - 99
helicopter - 98
hot-air balloon - 94
limousine - 96
motorcycle - 94
pleasure boat - 68
scooter - 100
snowmobile - 20
subway train - 99
SUV - 98
tow truck - 93
tractor - 96
train - 98
tram - 98
truck - 100
yacht - 92

Figure 3

Schematic of our implementation of the repeat-detection memory game first introduced by Isola et al. (2014).

Each image is presented for 600 ms, with an intertrial interval of 800 ms. Participants are instructed to press the space-bar whenever they recognize a repeat of a previously shown image.

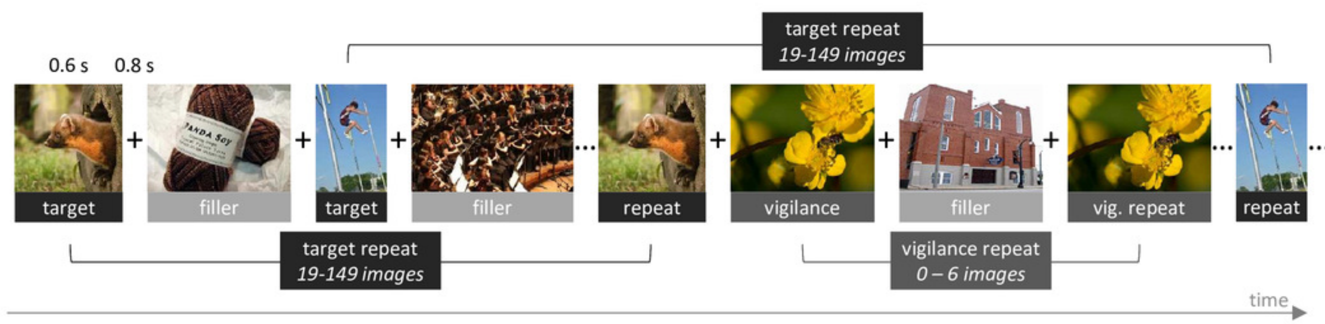


Figure 4

Distribution of the collected memorability measures.

Panel A represents memorability scores computed as the hit rate across participants. Panel B represents scores corrected for false alarms. The horizontal lines indicate the global mean memorability scores. The asterisks represent the mean per category. Each category contains 2000 quantified images. In addition to overall differences across categories, we observed considerable variability in memorability within categories too.

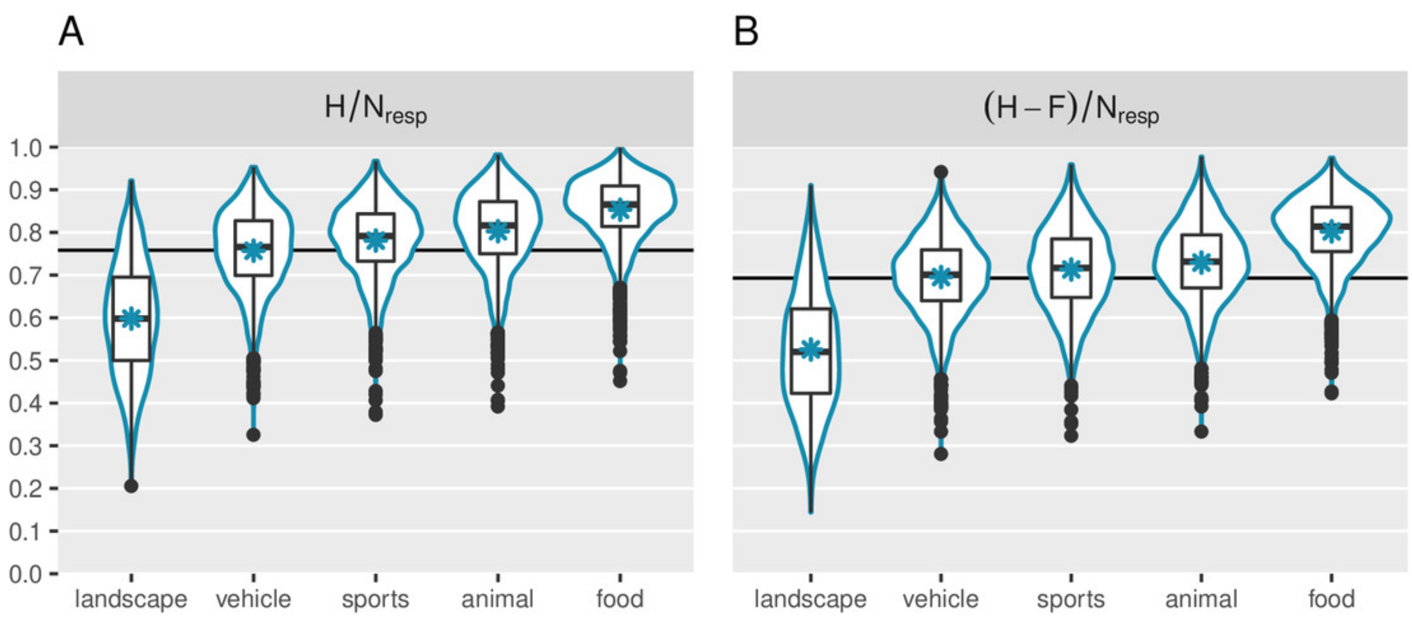
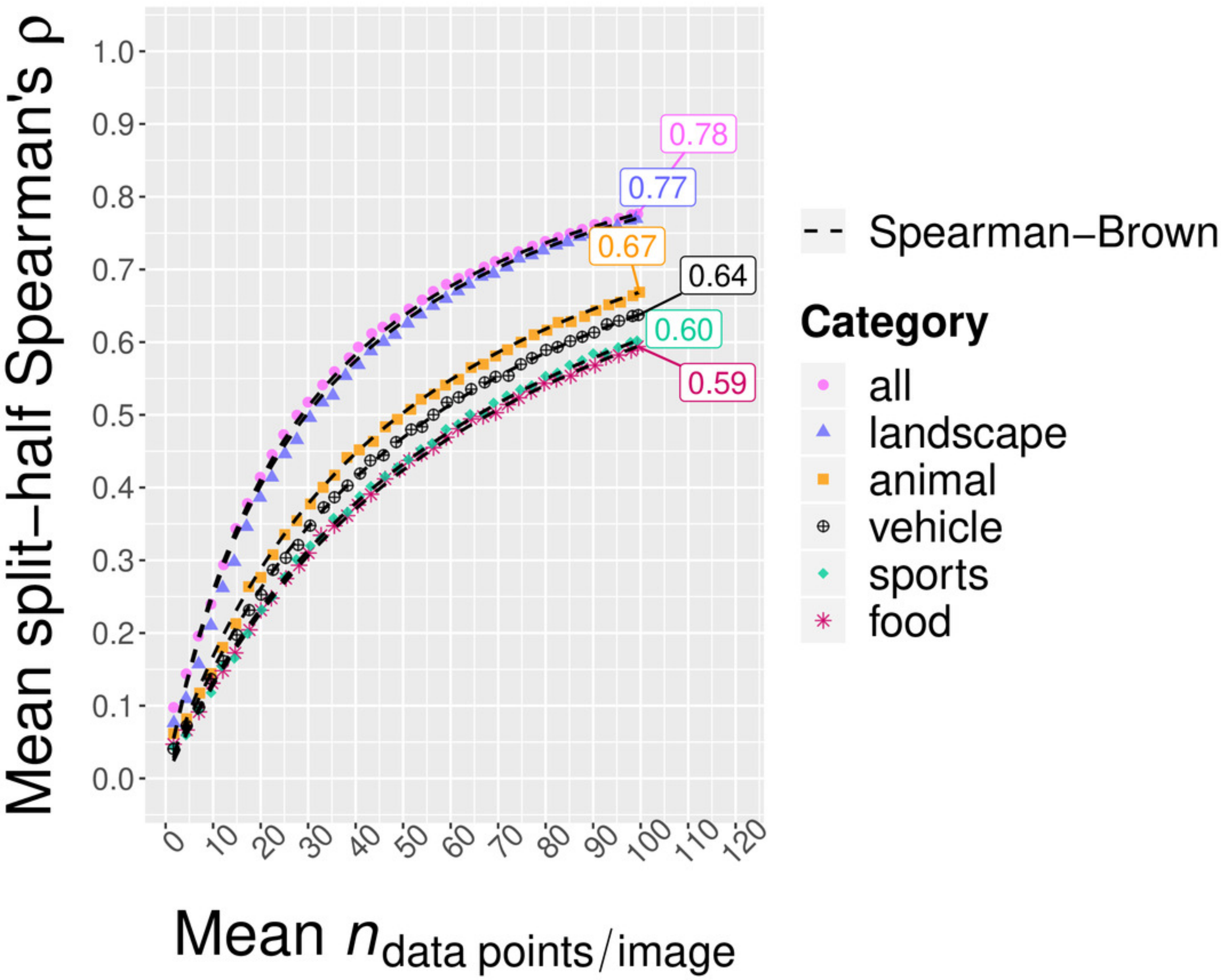


Figure 5

Split-half consistency across participants in function of N_{resp}.

Estimates are based on 1000 random splits. N_{resp} corresponds to the total number of data points for an image, not to the number that goes into one half during the split-half procedure. The dashed line represents predicted consistencies based on the Spearman-Brown formula (Brown, 1910; Spearman, 1910) applied to the observed consistency when N_{resp} is the maximum number of available data points. Even though the consistency was lower when zooming in on a single category compared to the whole set at once (possibly due to a smaller range of scores), images still showed highly consistent differences in memorability within categories.



Mean split−half Spearman's ρ
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
110
120
Mean $n_{\text{data points/image}}$
0.78
0.77
0.67
0.64
0.60
0.59
Spearman−Brown
Category
all
landscape
animal
vehicle
sports
food

Table 1(on next page)

Comparison MemCat to other memorability datasets.

	Isola et al. (2014)	FIGRIM	LaMem	MemCat
Category-based	no	yes	no	yes
Number of quantified images	2222	1754	~60K	10K
Bounding boxes or segmentation data	yes	yes	no	yes

Table 2 (on next page)

Recognition memory performance.

The table presents descriptive statistics across participants (*n* = 2291) for five Signal Detection Theory measures. See Macmillan and Creelman (2005) for an explanation of these measures.

	d’	β	Hit rate	False alarm rate	Prop. correct
Mean	2.50	4.43	.76	.05	.87
Median	2.48	3.00	.79	.04	.88
SD	0.49	5.48	.14	.04	.05
Min	0.69	0.09	.03	.00	.60
Max	4.46	98.26	1.00	.49	.98