Atten Percept Psychophys (2017) 79:2044–2054
DOI 10.3758/s13414-017-1375-9

Visual properties and memorising scenes: Effects of image-space sparseness and uniformity

Jiří Lukavský [1] · Filip Děchtěrenko [1,2]

Published online: 13 July 2017

Abstract Previous studies have demonstrated that humans have a remarkable capacity to memorise a large number of scenes. The research on memorability has shown that memory performance can be predicted by the content of an image. We explored how remembering an image is affected by the image properties within the context of the reference set, including the extent to which it is different from its neighbours (image-space sparseness) and if it belongs to the same category as its neighbours (uniformity). We used a reference set of 2,048 scenes (64 categories), evaluated pairwise scene similarity using deep features from a pretrained convolutional neural network (CNN), and calculated the image-space sparseness and uniformity for each image. We ran three memory experiments, varying the memory workload with experiment length and colour/greyscale presentation. We measured the sensitivity and criterion value changes as a function of image-space sparseness and uniformity. Across all three experiments, we found separate effects of 1) sparseness on memory sensitivity, and 2) uniformity on the recognition criterion. People better remembered (and correctly rejected) images that were more separated from others. People tended to make more false alarms and fewer miss errors in images from categorically uniform portions of the image-space. We propose that both image-space properties affect human decisions when recognising images. Additionally, we found that colour presentation did not yield better memory performance over grayscale images.

Keywords Scene perception · Memory: visual working and short-term memory · Categorization

People have a remarkable capability to remember photographs. This capability is notable in terms of both capacity and fidelity (the number of images that we can remember and the size of the differences that we can distinguish, respectively). People can study thousands of images for a few seconds each and achieve high recognition rates (Standing, 1973; Standing, Conezio, & Haber, 1970; Voss, 2009). Recent studies showed that people are able to encode subtle details. They can distinguish between state changes (Brady, Konkle, Alvarez, & Oliva, 2008) or between a large number of exemplars in a single category (Vogt & Magnussen, 2007). Surprisingly, the information decay for some features over time is not significant. For example, Brady, Konkle, Gill, Oliva, & Alvarez (2013) reported that the long-term memory fidelity for object colour is as high as the fidelity of the working memory.

Recognition performance for scenes (complex photographs with many objects) is comparable to memory for photographs of isolated images (Konkle, Brady, Alvarez, & Oliva, 2010b). The scenes feature multiple objects, which require more resources. On the other hand, the content of the scenes is semantically coherent and spatially constrained, which reduces the variability.

Recent studies on image memorability (Bainbridge, Isola, & Oliva, 2013; Bylinskii, Isola, Bainbridge, Torralba, &

Electronic supplementary material The online version of this article (doi:10.3758/s13414-017-1375-9) contains supplementary material, which is available to authorized users.

✉ Jiří Lukavský
jirilukavsky@gmail.com

[1] Institute of Psychology, The Czech Academy of Sciences, Hybernská 8, 110 00 Prague, Czech Republic

[2] Faculty of Mathematics and Physics, Department of Software and Computer Science Education, Charles University in Prague, Malostranské nám. 25, 118 00 Prague, Czech Republic

Oliva, 2015; Isola, Xiao, Torralba, & Oliva, 2011) have shown interindividual consistency in terms of which images are remembered and which are forgotten. Originally, this consistency was observed in a dataset featuring hundreds of categories (Isola et al., 2011). The consistency also was present at a more fine-grained level when the stimuli were constrained to a narrow selection of categories (Bylinskii et al., 2015). Image memorability can be predicted from the content of the image. Simple low-level features (colour, saturation or number of objects) were poor predictors of which images would be remembered or forgotten. Semantic features, such as the presence of people or interiors, were correlated with higher memorability. Khosla, Raju, Torralba, & Oliva (2015) predicted memorability using a convolutional neuron network (CNN). Eye movement data can further improve prediction in an individual trial in terms of whether an image will be remembered or forgotten (Bylinskii et al., 2015).

To understand what information we store in our memory when memorizing scenes, we may ask how the remembered images interfere with each other. Memorability research has shown that memory performance is dependent upon image content and may be predicted and potentially modelled (Bylinskii et al., 2015; Isola et al., 2011; Khosla et al., 2015). In this manuscript, we assume that images can be projected into an image-space (i.e., representational space). Similar images will have similar representations in the image-space, which will lead to interference and potential memory errors. The latter claim about image similarity has been supported by many studies. Human memory is poorer when the individual is required to distinguish between similar images or other stimuli (Konkle et al., 2010b; Konkle, Brady, Alvarez, & Oliva, 2010a). People also remember distinctive stimuli that stand out from their context more effectively (Bylinskii et al., 2015; Standing, 1973; Watier & Collin, 2012). A similar approach of multidimensional representational space has been proposed in face processing literature (Valentine, 1991).

Konkle et al. (2010a) explored how memory interference is affected by the visual properties of the exemplars. They defined perceptual distinctiveness (variability in terms of colour or shape) and conceptual distinctiveness (few or many different kinds of object within a category). They found that the object categories with conceptually distinctive exemplars were associated with decreased interference in memory as the number of exemplars increased. Perceptual distinctiveness did not affect interference. Their results show observers' capacity to remember visual information in long-term memory depends more on conceptual structure than perceptual distinctiveness. The distinctiveness values were based on observers' ratings (Konkle et al., 2010a). Here, we extend their approach towards scene stimuli. We also replace the observers' ratings of distinctiveness with pairwise similarity measures. Given the current progress in computer image classification and recognition, the computational comparison of similarity is possible and it provides pairwise similarity results for large image sets.

In the image-space model that we apply here, we primarily focus on two properties: *sparseness* and *uniformity* (Fig. 1; for detailed definition, see Experiment 1, *Stimuli/Image-space properties* section). Both properties are derived from the features of a particular image, but they also take the context (the set of reference images) into account. An image comes from a sparsely populated region of the image-space if it is relatively far from its neighbours (versus when it comes from a densely populated region). An image comes from a uniform part of the image-space if it shares its scene category with many of its neighbours. In other words, we are interested in how the proximity of potential distractors and their conceptual similarity affect the memory of an image. An important difference between the proposed properties is that sparseness depends solely on the image-space distance (perceptual similarity), whereas uniformity requires an additional classification scheme among pictures (i.e., categorization). These two measures are not guaranteed to be orthogonal, but they capture potential mechanisms of memory interference: interference via feature similarities (sparseness) and interference via confusion with other images of the same category (uniformity). We tested the effect of these two potential interference mechanisms (sparseness and uniformity) in a series of three experiments and found the effect of sparseness on memory sensitivity, and the effect of uniformity on the recognition criterion.

Experiment 1

We first took a large set of images and evaluated their mutual similarity. In the next step, we asked participants to memorise subsets of these images. Finally, we evaluated how their performance was affected by the similarity measures. The design is similar to that of other large-scale memory experiments (Brady et al., 2008; Konkle et al., 2010b). We used an old/new decision task to assess memory as it is more ecologically valid, because we are rarely allowed to pick among alternatives (Andermane & Bowers, 2015). The common approach (2AFC) also leads to a more complex decision task when the choice of the distractor affects performance (Brady et al., 2008), but we were interested in how image-space properties affect recognition of a single image.

More specifically, we were interested in how the image-space properties affect memory sensitivity and response bias. We measured the similarity of images via the activation of the penultimate layer of a convolutional neuron network (or deep features). Deep features provide a powerful image representation surpassing other features in computer vision. Deep features also are applicable beyond

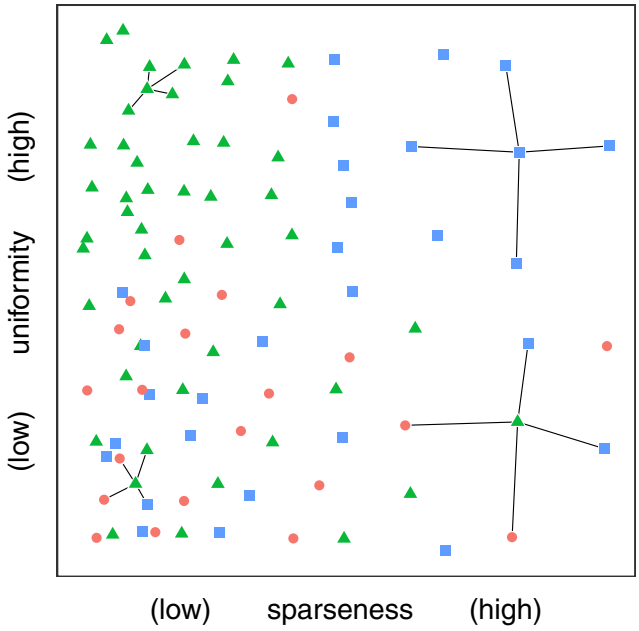

Fig. 1 Schematic illustration of sparseness and uniformity image-space properties. Each point represents an image; different symbols represent different image categories. Lines represent connections to the four most proximal neighbours of the selected images. In this example, the image-space is arranged to highlight high/low values for each property.

the original purpose of the trained network (Razavian, Azizpour, Sullivan, & Carlsson, 2014).

We decided to present the scenes in greyscale to emphasize their structure and layout, because the informativeness of colour is known to vary across scene categories (Oliva & Schyns, 2000). In Experiment 3, we came back to this question and used the colour stimuli.

Method

Participants

Forty university students (34 women) participated in Experiment 1 (age range: 19-47 years, mean: 21.7 years). The participants signed the informed consent form and received course credit for their participation.

Stimuli

We sampled the images from a set of 2,048 scenes (64 categories, 32 images per category). The images were 256 × 256 pixels in size, and we presented them either in greyscale (Experiments 1 and 2) or colour (Experiment 3). The images were selected from the image dataset used by Konkle et al. (2010b). Their dataset contained 64 images per category. We calculated a gist descriptor for each image (Oliva & Torralba, 2001), and we selected 32 images from each category with the smallest distances from the mean category gist[1] to select more typical views within each category.

Measure of image similarity We used deep features distance to measure similarity between images. We obtained the features from a pretrained CNN (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014). This CNN is based on AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) and is trained on 205 scene categories of the Places Database that includes 2.5 million images. To compute similarity, we used the 4096-dimensional features from the response on the last fully connected layer (fc7). We defined the similarity between two images as the Euclidean distance (L2 norm) between the corresponding fc7 activation vectors.

Image-space properties Given a similarity measure, we can project images into a multidimensional space. The distribution of the images in such a space is irregular. We were interested in how the regions of this space vary with respect of two properties: 1) number of images, or how dense or sparse the region is, and 2) categorical uniformity, whether either images of a single category or several categories overlap there. For each image, we defined two properties: sparseness and uniformity. *Sparseness* was defined as the distance of an image from its n-th most proximal neighbour. We transformed all distances to z-scores, so it was possible to compare the values using the mean and variance of observed distances. *Uniformity* was defined as the proportion of images of the same category within the n most proximal neighbours. The particular number of proximal images (n = 31) to inspect was arbitrary. Our choice of 31 was convenient because it yielded the uniformity maximum of 100% if all neighbours were from the same category (31 of 32 images). We tested several other values of n, but the properties were highly correlated.

Procedure

The experimental session consisted of the study part and the test part. The images were presented on a 9.7" iPad, and the experiment lasted approximately 50 minutes.

In the study part, we asked participants to study a sequence of images carefully for later recognition. Additionally, they were asked to touch the screen whenever they noticed that an image had been presented for a second time (vigilance task). We presented 400 greyscale scenes in the study part of Experiment 1. The set consisted of 320 unique scenes (5 per category), and an additional 40 scenes were presented twice

[1] Originally, we intended to measure the similarity with gist descriptors, which provide information on perceptual similarity. During the course of the project, we learnt about deep features, which can capture semantic similarity known to be associated with long-term memory interference (Konkle, Brady, Alvarez, & Oliva, 2010a). For the parallel analysis based on gist-descriptors see Supplement and Discussion.

(0-1 per category). Repeated images were distributed uniformly throughout the study part, and they were set to recur after 3, 15, or 63 intervening items. Each image was presented one at a time for 3 s (subtending approximately 7.2 × 7.2°), and the sequence was interleaved with the presentations of a fixation cross for 800 ms. Feedback (red "X" or green "OK" for 500 ms) was shown only if the participants responded (as in Brady et al., 2008; Konkle et al., 2010b); therefore, they were not notified of miss errors. The participants were allowed to take a short break in the middle of the study part and after the end of the study part (before test part).

In the test part, a single image was shown in the middle of the screen, with labels "new" and "old" on the left and right side, respectively. There was no time limit within which participants had to provide a response. Brief feedback (red "X" or green "OK") was shown at the top of the screen for 500 ms after each response. The test set consisted of 256 scenes (2 old scenes and 2 new foils per category).

Performance measures We evaluated the effects of sparseness and uniformity on sensitivity (d') and response bias (criterion, c). We chose a criterion as a measure of bias, because it is reportedly unaffected by changes in d' (Stanislaw & Todorov, 1999). Given the noise and signal distributions, values of c represent how the criterion shifts relative to the neutral point (over which both distributions cross), and the unit of measurement is the standard deviation. A positive value indicates a more conservative criterion (i.e., miss errors more likely than false alarms).

The recognition test for each participant included only a subgroup of images with respect to the entire image set (256 of 2,048), and we did not have enough recognition data to calculate the per-image sensitivity or criterion. We decided to pool the data of images with similar image-space properties. We divided the main image set of 2,048 images into eight quantiles separately for sparseness and uniformity, with 256 images in each quantile. In the analysis, each quantile was represented with the median of the corresponding property values. Thus, the performance data from the recognition experiment were transformed into eight sensitivity and eight criterion values for each participant.

We used R (R Core Team, 2016) with *lmer* package (Bates, Mächler, Bolker, & Walker, 2015) to perform analyses using linear mixed-effects models. We analysed the effects of sparseness and uniformity separately, and we used a single fixed effect (sparseness or uniformity) as a linear predictor. For random effects, we used the maximal model (Barr, Levy, Scheepers, & Tily, 2013) that included both individual intercepts and slopes. Visual inspection of residual plots did not reveal any obvious departures from homoscedasticity or normality. To assess the validity of the mixed-effects models, we performed likelihood ratio tests to compare the models with fixed effects to the null models with only the random effects (subjects). We rejected results in which the model with fixed effects did not differ significantly from the null model. Throughout the paper, we report both p values for the entire model and p values calculated with Satterthwaite's approximation in *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2014) with 95% confidence intervals for each fixed effect (sparseness, uniformity). In each experiment, we tested four models and the presented p values are not corrected for multiple comparisons. The p values can be compared to adjusted α level 0.0125 (equivalent of $\alpha = 0.05$, Bonferroni correction, 4 tests). We discuss the correction for multiple testing across all three experiments within Experiment 3.

Results

Recognition performance Individual accuracy rates across all participants ranged from 61-92% (mean = 75%), with the respective d' ranging from 0.59 to 2.84 (mean = 1.45). The criterion values ranged from −0.18 to 0.75 (mean = 0.20), showing a significant bias towards positive values (i.e., miss errors) ($t(39) = 7.23$, $p < 0.001$). The correlation between d' and the criterion values for each participant/quantile combination was low (sparseness: $r = -0.003$, $p = 0.962$; uniformity: $r = -0.057$, $p = 0.306$). In the vigilance task, the hit rate was 82% (SEM = 4%) with 3 intervening items, 78% (SEM = 4%) with 15 intervening items, and 69% (SEM = 4%) with 63 intervening items. The false-alarm rate was low (3%, SEM = 0.4%).

Image-space properties Sparseness values ranged from −3.55 to 1.61 (mean = −1.96, SD = 0.62, median = −1.99). The values were significantly lower than zero ($t(2047) = 144.1$, $p < 0.001$, $d = 3.18$). For 11 images only, the sparseness (the distance to the 31st most proximal neighbour) was greater than the mean interimage distance observed in the reference set (i.e., zero). The uniformity values ranged from 0% to 100% (mean = 52%, SD = 24, median = 55%). The Spearman correlation between sparseness and uniformity was significant ($\rho = 0.243$, $p < 0.001$).

We wanted to examine how the values were affected by the choice of parameter values (31 neighbours). We compared the values with alternative calculations (7, 15, 63, and 127 neighbours) and obtained similar results (ρ ranged from +0.875 to +0.959 for sparseness and +0.780 to +0.962 for uniformity).

Image-space properties and memory performance Figure 2 shows how the sensitivity and criterion values varied as functions of sparseness and uniformity. We found that it is easier to recall images that are separated to a greater extent from their neighbours (i.e., images from sparsely populated

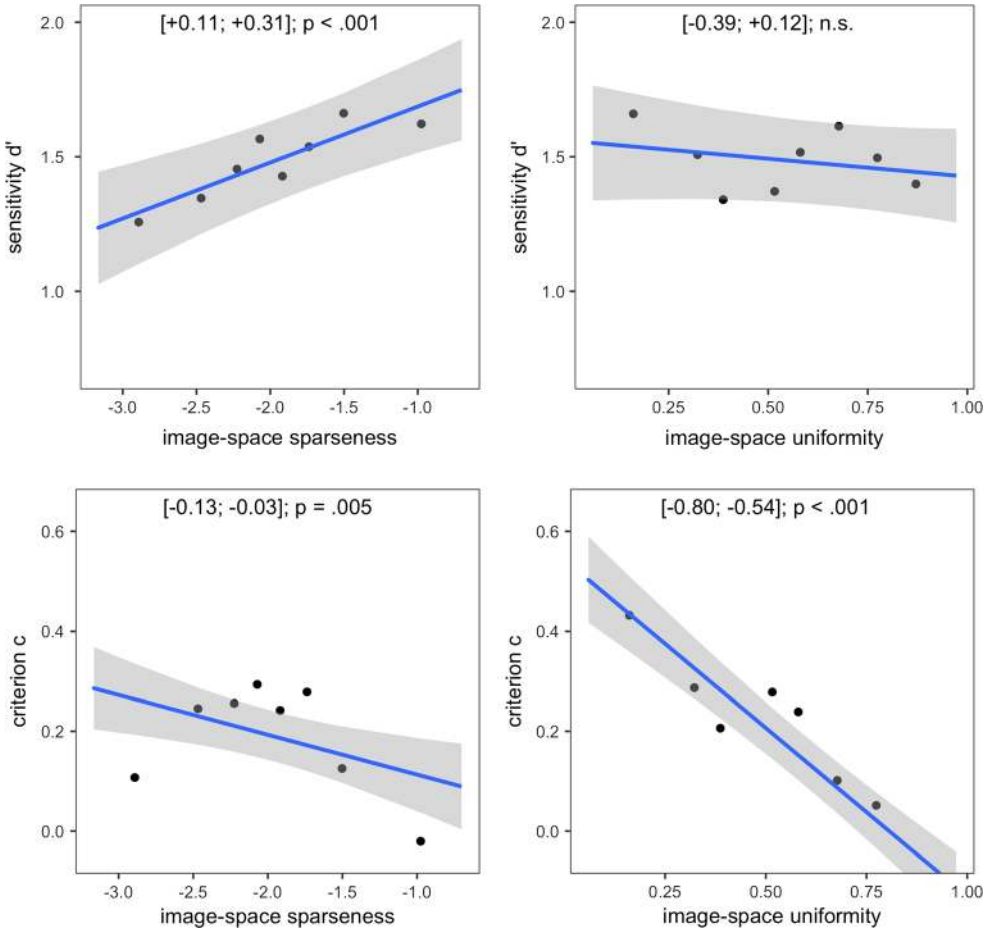

Fig. 2 Performance in a short memory experiment with greyscale images. Sensitivity d' and criterion c as a function of image-space sparseness and uniformity. Confidence intervals for sensitivity/criterion gradients are shown above each plot.

image-space areas, $\chi^2(1) = 15.285$, $p < 0.001$). Sparseness also decreased the criterion ($\chi^2(1) = 8.062$, $p < 0.005$).

The uniformity of the image-space had no effect on the sensitivity ($\chi^2(1) = 1.069$, $p = 0.301$), but it decreased the criterion ($\chi^2(1) = 55.343$, $p < 0.001$). In other words, people made fewer miss errors and more false alarm errors in association with images that had neighbours of the same category. Studied images coming from areas in which categories were intermixed were more often falsely considered to be new.

Discussion

We found a small but statistically significant relationship between image-space sparseness and uniformity. However, despite this correlation, both measures indicated different patterns of performance. Image-space sparseness was associated with higher recognition sensitivity, as more isolated images were easier to remember. Image-space uniformity reduced the criterion. People tended to recognise falsely more typical exemplars of a category. We also observed a smaller but significant effect of sparseness on the recognition criterion.

The participants observed only a portion of the reference set (17.6% in the study part). We calculated the sparseness/uniformity properties relative to the whole image set as estimates how the images relate to our visual experience. We assume that, for example, the image from highly uniform regions of the image-space was compared with our scene-related experience in general, and it was considered to be more typical. In the test part of the experiment, the participants were aware that the corresponding category was present and they falsely claimed that the typical image was previously seen.

We considered two measures of visual similarity: deep features and gist descriptors. Gist descriptors express the distribution of orientation and spatial frequencies in different parts of the image and they have been used for scene classification (Oliva & Torralba, 2001). We do not claim that either deep features or gist descriptor representations are actually used at the brain level, but we assume that representational similarity reflects the statistical regularities in the visual signal and the corresponding difficulty to distinguish between images both for computer vision algorithms and for humans. In the presented experiments, we report the results based on deep features. In the parallel analysis based on the gist descriptors (see

Supplement), we found effects of uniformity, but not of sparseness. This may be attributed to the difference between distances based on gist descriptors and deep features. While gist-based sparseness is based on perceptual similarity, deep features capture the similarity in image semantics. The effect of semantic similarity is in line with the findings that the conceptual/semantic similarity is more predictive of long-term memory performance than perceptual similarity (Konkle et al., 2010a).

Experiment 2

As the next step, we wanted to run a longer experiment to test whether the results would change. We expected two major potential factors involved in a longer experiment: increased difficulty and larger image context.

First, we wanted to check if we see a similar pattern of results in a more difficult experiment. It is possible that the experiment was not sufficiently challenging and that the image similarity was not confusing enough to manifest. To address this issue, we designed a longer experiment with more images to learn (12 vs. 5 images per category in Exp. 1).

Second, we assume that the memory for a particular image is affected by image-space properties (similarity within the reference set). We use the reference set as a sample of our visual experience to estimate the similarity, distance, or overlap of different image categories. The participants never see the entire reference set but only its part (17.6% in the study part of Exp. 1). We assume that, for example, the image from highly uniform regions of the image-space is compared with our scene-related experience in general, and it is considered to be more typical. In the test part of the experiment, the participants falsely claim that the typical image was previously seen. In Experiment 2, we sampled 12 images per category to provide a richer context for participants and remind them better of the variability of each category.

Method

Participants

Thirty-eight university students (36 women) participated in Experiment 2 (age range: 19-26 years, mean: 21.3 years). The participants signed the informed consent form and received course credit for their participation. No participant participated in Experiment 1.

Procedure

In the study part of Experiment 2, we presented 960 greyscale scenes. The set consisted of 768 unique scenes (12 per category), and an additional 96 scenes were presented twice (1-2 per category). Repeated images were set to recur after 3, 15, 63, or 255 intervening items. The test part consisted of 256 images (2 old images and 2 new foils per category, as in Experiment 1). The trial structure was identical to that in Experiment 1. The stimuli were sampled from the same image set used in Experiment 1.

Results

Recognition performance Individual accuracy rates ranged from 51-90% (mean = 70%), showing the longer version was more difficult compared to Experiment 1 ($t(69.4) = 2.47$, $p = 0.016$, Cohen $d = 0.56$). The respective d' values ranged from 0.05 to 2.65 (mean = 1.15). The criterion values ranged from −0.12 to 0.85 (mean = 0.26), showing a significant bias towards miss errors ($t(37) = 8.05$, $p < 0.001$). In the vigilance task, the hit rate was 67% (SEM = 5%) with 3 intervening items, 64% (SEM = 5%) with 15 intervening items, 50% (SEM = 4%) with 63 intervening items, and 41% (SEM = 4%) with 255 intervening items. The false-alarm rate was low (2%, SEM = 0.3%). Hit rates in the vigilance task were lower than in Experiment 1 (all $t > 2.24$, all $p < 0.05$), the difference in false-alarm rates was not significant ($t(75.1) = 1.60$, $p = 0.113$).

Image-space properties and memory performance Similar to Experiment 1, we found a significant effect of image-space sparseness on recognition sensitivity ($\chi^2(1) = 16.794$, $p < 0.001$). On the other hand, the effect of sparseness on the criterion was not significant ($\chi^2(1) = 2.330$, $p = 0.127$). The results are shown in Fig. 3. The result regarding the effect of image-space uniformity was consistent with previous findings. Uniformity decreased the criterion ($\chi^2(1) = 18.956$, $p < 0.001$), and we found no effect on recognition sensitivity ($\chi^2(1) = 3.320$, $p = 0.068$).

In our experiments, we cannot calculate ground truth sparseness or uniformity. We rely on calculations based on the reference set which we consider a sample of our visual experience. The participants saw only part of the reference set, and thus we wanted to check how properties would differ if they were calculated based on the actually presented images. We repeatedly sampled 12 images per category (as in Experiment 2), calculated sparseness and uniformity, and compared them with the values based on the reference set with Spearman correlation coefficient. We found strong relationships for sparseness (median $\rho = 0.980$, 95% confidence interval [CI] [0.970; 0.985], 100 repetitions) and uniformity (median $\rho = 0.933$, 95% CI [0.920; 0.941]). When we sampled 5 images per category (as in Experiment 1), the relationships were still very high

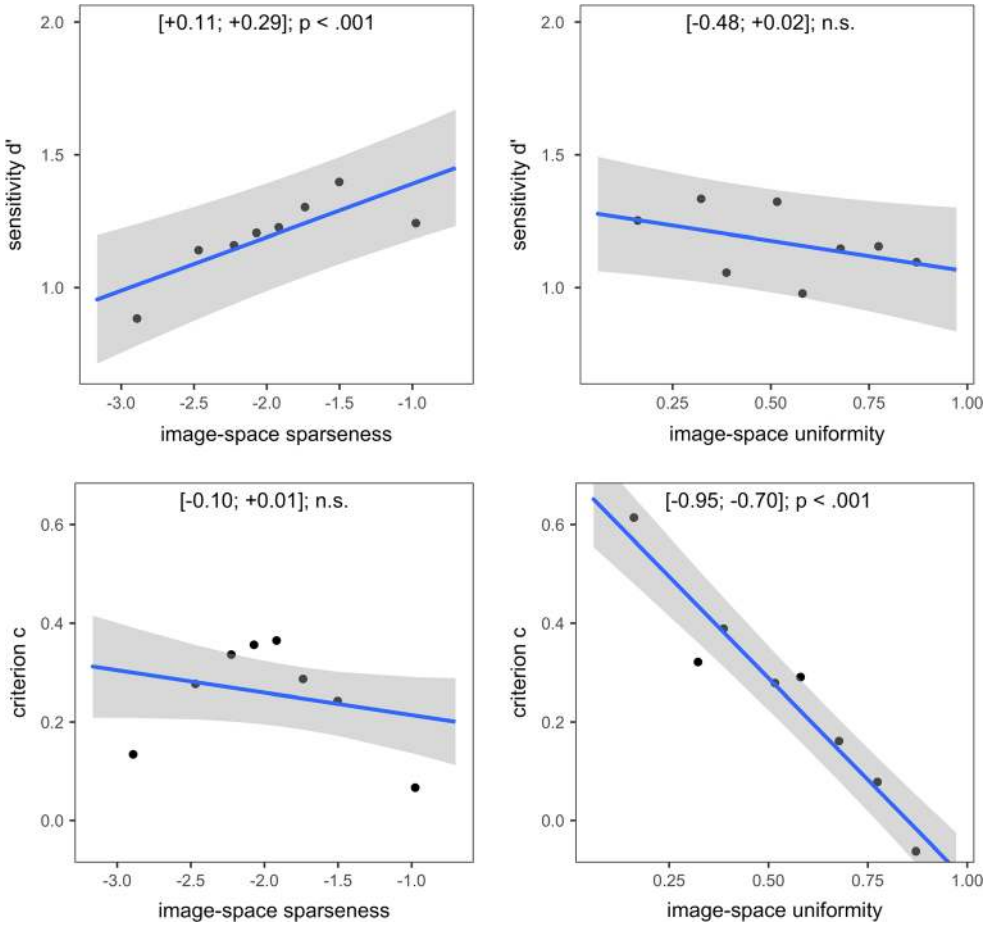

Fig. 3 Performance in a long memory experiment with greyscale images. Sensitivity *d'* and criterion *c* as a function of image-space sparseness and uniformity. Confidence intervals for sensitivity/criterion gradients are shown above each plot.

(sparseness: median $\rho = 0.921$, 95% CI [0.892, 0.949]; uniformity: median $\rho = 0.801$, 95% CI [0.754, 0.844]).

Discussion

The results of Experiment 2 were similar to those of Experiment 1. Image-space sparseness was associated with higher recognition sensitivity, and image-space uniformity reduced the criterion. The effect of sparseness on the recognition criterion was not significant. The experiment showed us that separate effects of sparseness and uniformity are still observed when the experiment is more difficult. In the longer experiment, participants were exposed to a larger portion of the reference set (42.2% in the study part). It could provide them with greater context and remind them of the richness of each category. We showed that image-space properties based on the subset (stimuli of presented to each participant) are good estimates of the image-space properties observed in the reference set especially. These estimates are better when more images are presented (as in Experiment 2). The uniformity is more difficult to estimate, which can be partly caused by our definition. Uniformity was defined as a proportion of images of the same category in the neighbourhood and thus limited to a smaller number of potential values. On the other hand, sparseness was based on pairwise distance, which is a continuous variable irrespective of the number of images.

Experiment 3

In previous experiments, we used greyscale images for two reasons. First, we wanted participants to pay attention to the shapes in the scenes. Second, we wanted the images to be compatible with gist descriptors, which we had used initially for similarity measurements.

Interestingly, it is not clear whether the participants store the images in greyscale or in colour. After seeing a grayscale photograph, they may retain the memories in greyscale. Alternatively, they can attempt to estimate the probable colours, because the colour is diagnostic for some categories (Oliva & Schyns, 2000) and thus easier to infer. The following experiment cannot resolve this question, but we assume the

colour version is easier. With more features provided, participants can better distinguish between similar stimuli.

For Experiment 3, we maintained the structure and length of Experiment 1, but we used colour versions of the images. This approach helps us to test whether the results will hold when we use a slightly different feature space derived from the colour versions of stimuli. Additionally, we expected this experiment to be easier and wanted to test whether the shift in workload would lead to the same pattern of results.

Method

Participants

Forty-four university students (38 women) participated in Experiment 3 (age range: 19-47 years, mean: 21.7 years). The participants signed the informed consent form and received course credit for their participation. No participant participated in either Experiment 1 or 2.

Stimuli

The stimuli were the colour versions of images from the set used in Experiment 1. We recalculated the pairwise distances based on the colour versions and correspondingly updated the sparseness and uniformity values.

Procedure

Both the study and test parts were identical to those of Experiment 1. The participants studied 400 scenes (320 unique scenes, 5 per category; with additional 40 scenes presented twice). Repeated images were distributed uniformly throughout the study part, and they were set to recur after 3, 15, or 63 intervening items. Later, participants were asked to make old/new judgements about 256 scenes (2 old scenes and 2 new foils per category).

Results

Recognition performance Individual accuracy rates across all participants ranged from 62-90% (mean = 76%), and they did not differ from the accuracy rates observed in the grayscale version in Exp. 1 ($t(79.7) = 0.497$, $p = 0.621$, Cohen $d = 0.11$). The respective d' values ranged from 0.65 to 2.74 (mean = 1.55). The criterion values ranged from −0.13 to 0.81 (mean = 0.30), showing a significant bias towards miss errors ($t(43) = 8.90$, $p < 0.001$). In the vigilance task, the hit rate was 83% (SEM = 3%) with 3 intervening items, 80% (SEM = 3%) with 15 intervening items, and 65% (SEM = 3%) with 63 intervening items. The false-alarm rate was low (3%, SEM = 0.3%). Hit rates in the vigilance task were similar to Experiment 1 (all $t < 0.40$, all $p > 0.700$), the difference in false-alarm rates was not significant ($t(81.3) = 1.20$, $p = 0.235$).

Image-space properties Sparseness values calculated for the colour versions ranged from −3.16 to 2.37 (mean = −1.85, SD = 0.59, median = −1.95). The values were lower than the corresponding greyscale-based values found in Experiments 1 and 2 (diff = 0.11, $t(2047) = 13.9$, $p < 0.001$, $d = 0.18$), and both methods yielded similar results ($\rho = +0.825$, $p < 0.001$). The uniformity values ranged from 0% to 100% (mean = 64%, SD = 23, median = 68%). The values were higher than the corresponding greyscale-based values (diff = 0.12, $t(2047) = 32.9$, $p < 0.001$, $d = 0.50$), and both methods yielded similar results ($\rho = +0.761$, $p < 0.001$). The Spearman correlation between sparseness and uniformity was low in the colour versions ($\rho = +0.085$, $p < 0.001$).

Image-space properties and memory performance Figure 4 shows how the sensitivity and criterion values varied as functions of sparseness and uniformity. We found that it is easier to recall images that are separated to a greater extent from their neighbours ($\chi^2(1) = 18.829$, $p < 0.001$). The images from densely populated areas were more likely to be missed ($\chi^2(1) = 18.956$, $p < 0.001$). With respect to uniformity, the images from more uniform areas were more difficult to remember ($\chi^2(1) = 23.89$, $p < 0.001$) and provoked more false-alarm errors ($\chi^2(1) = 58.197$, $p < 0.001$).

Discussion

The effects of sparseness and uniformity that were seen in Experiments 1 and 2 were present in Experiment 3 as well. Additionally, we observed the effect of sparseness on the criterion and the effect of uniformity on sensitivity. In the reported experiments, we analysed the effects separately and performed four statistical tests in each experiment without corrections. If we correct for the multiple hypothesis tests (Bonferroni correction, 12 tests), all relationships between sparseness/sensitivity and uniformity/criterion remain statistically significant. The sparseness/criterion relationship in Exp. 1 will become nonsignificant, and the other two relationships in Experiment 3 will remain significant.

We expected the colour version to be easier, but we found no difference in accuracy rates relative to the grayscale version presented in Exp. 1. The colour presentation did not provide benefit in memorizing and later recognition of the images, which could mean that people estimate the probable colours in grayscale photographs. Alternatively, the colour is not so important for this particular task and stimuli. In some scene perception tasks, no difference between using colour and grayscale stimuli is observed (Meng & Potter, 2008), or the

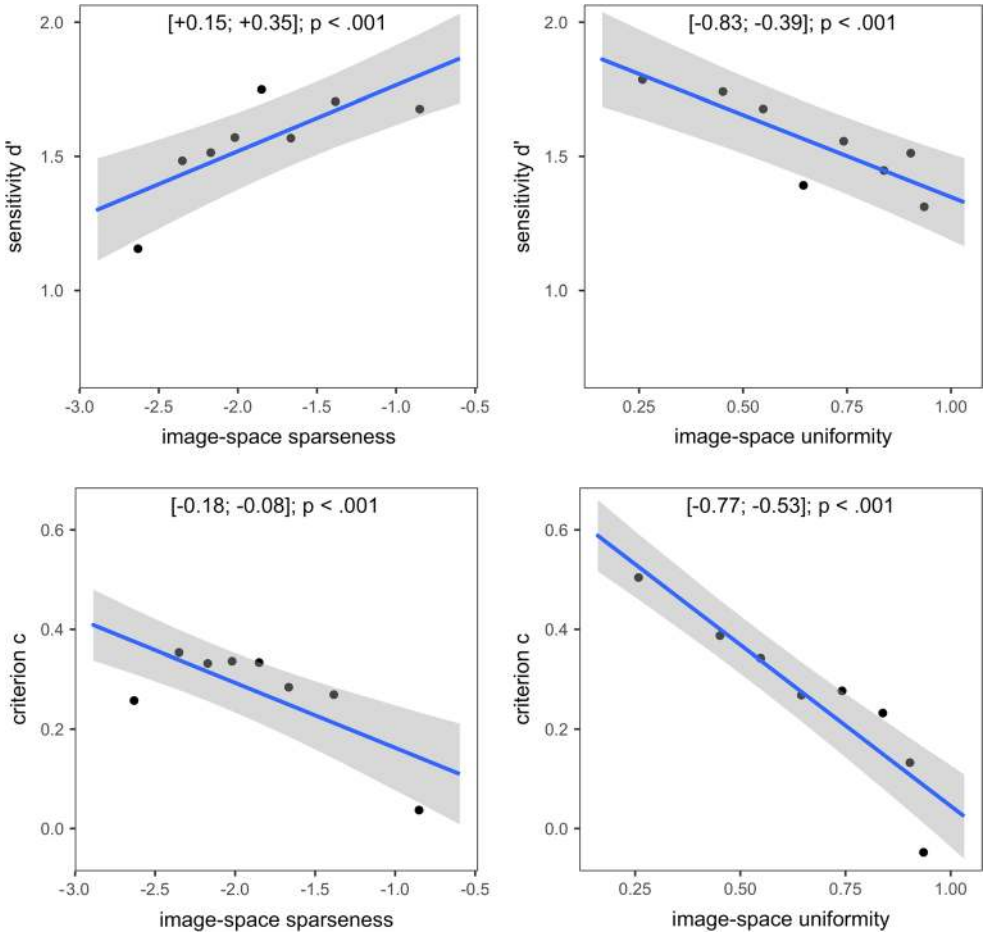

Fig. 4 Performance in a short memory experiment with colour images. Sensitivity *d'* and criterion *c* as a function of image-space sparseness and uniformity. Confidence intervals for sensitivity/criterion gradients are shown above each plot.

grayscale stimuli yield better performance (Nijboer, Kanai, de Haan, & van der Smagt, 2008).

Across all three experiments, we found that miss errors were more likely than false-alarm errors. We used only partial feedback in the study part (we did not warn participants about their miss errors), which could bias their responses. However, other studies showed the same bias toward miss errors despite showing feedback for both miss and false-alarm response (Andermane & Bowers, 2015).

General discussion

We performed three experiments that varied in their difficulty level (number of stimuli) and use of colour or grayscale stimuli, which not only influenced the difficulty but also led to slightly different image features and corresponding levels of image similarity. Across all three experiments, we repeatedly observed a similar pattern. Specifically, people remembered more isolated images more effectively (sparseness increased sensitivity) and were prone to more false alarms and fewer miss errors in more categorically typical images (uniformity decreased bias). Despite the correlation between sparseness and uniformity, the analogous relationships were observed only in Experiment 3 (sparseness/criterion and uniformity/sensitivity, after correction for the multiple tests). We conclude that both image-space properties can affect human memory performance independently.

Images from sparse regions of the image-space have more distinctive features, differ from their neighbours to a greater extent, and are easier to recognize or reject (in case on distractors). Because the participants observed only a random sample from the reference set, these images are different not only from the other images presented to participants but also from other scenes in general. This finding is in accord with many other studies showing increased memory performance for distinctive images (Eysenck, 1979; Nairne, 2006; Standing, 1973). In general, images from sparsely occupied position within the space of representation are easier to retrieve for recognition. The images from highly uniform regions of the image-space were compared with our scene-related experience in general and considered to be more typical. Later, participants falsely claimed that they previously had seen the typical images.

Uniformity can be related to classification certainty. Better recognition of more typical examples could be a result of more general principle of categorization (Rosch, 1978; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Our results are in accord with the assumption that basic-level categories form information-theoretic optimum, maximizing within-category similarity and minimizing between-category similarity (Corter & Gluck, 1992). Thus, the images near the category prototype are similar and easier to confuse with each other, whereas the images near category boundaries are more difficult to categorize. In a categorical search experiment, the images that are further from the decision boundary of an image classifier algorithm guide attention and eye movements more effectively (Maxfield, Stalder, & Zelinsky, 2014).

To measure visual similarity, we used deep features derived from the final fully connected layer of a CNN (Zhou et al., 2014) pretrained on a large dataset of scene stimuli. Initially, we used gist descriptors, and the results are included in the Supplement. To summarise, gist-based uniformity correlated with the criterion and partially with sensitivity (in Experiment 3). Gist-based sparseness was not related to sensitivity or the criterion (after correction). This shows that memory confusion is based to a greater extent on semantic features (the presence of objects) than low-level features (colour or shape), which is in line with previous research (Isola et al., 2011). The distinction between conceptual and perceptual distinctiveness (Konkle et al., 2010a) is consistent with this observation: colour/shape similarity is a poor predictor of memory confusion in general images. We used fc7 features, because we were interested in semantic similarity. Features derived from lower layers of a CNN capture low-level features and would likely yield similar results to the gist. A CNN is not designed to be biologically plausible, although we may draw a loose analogies between a CNN and the human cortex about their processing stages and representations (VanRullen, 2017). Comparative fMRI studies show that representations in the CNN correlate with emerging visual representations in the human brain (Cichy, Khosla, Pantazis, & Oliva, 2017; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016). Semantic similarity corresponds to more processed visual representations found in final CNN layers (in our case, fc7) or the inferotemporal cortex in the brain, whereas perceptual similarity is based on low-level representations in lower layers of a CNN or in the visual cortex.

Contrary to previous studies on image memorability (Bainbridge et al., 2013; Bylinskii et al., 2015; Isola et al., 2011), we did not analyse memory performance per individual image. We sampled images randomly and pooled the data for images with similar image-space properties to obtain more robust estimates of performance. Our focus on the image-space context of a memorised image is similar to that of the research on extrinsic memorability factors. Bylinskii et al. (2015) found that distinctive categories (e.g., cockpit) are prone to the context effect. That is, many images of cockpits look similar; however, when cockpits are presented amongst images from other categories, they become distinctive. This difference is minimal for less distinctive categories (e.g., living room). This conclusion is consistent with our findings on uniformity and its effect on the criterion.

In our experiments, we assumed that the scene images and categories in our reference set were representative. We sampled an equal number of images from each category to reflect the structure of the reference set in the stimuli and provide a similar context to all participants. The balanced sampling was important, because we wanted to include potentially conflicting images to test memory performance adequately for uniformity. In addition to the sole image properties, performance is likely to be affected by participants' experience and exposition (e.g., travel enthusiasts might be better at recognising forests or wilderness tracks) or understanding (Greene, Botros, Beck, & Fei-Fei, 2015), which is not reflected in our experiments.

In summary, we investigated image similarity with respect to a larger reference set. We repeatedly observed that more isolated images are remembered more effectively and more typical images are falsely recognised more often. We propose that both image-space properties affect human decisions when recognising images.

Author Note This research was supported by Czech Science Foundation (GA13-28709S and GA16-07983S) and RVO 68081740; it is part of research programme of Czech Academy of Sciences Strategy AV21.

References

Andermane, N., & Bowers, J. S. (2015). Detailed and gist-like visual memories are forgotten at similar rates over the course of a week. *Psychonomic Bulletin & Review, 22*(5), 1358–1363. doi:10.3758/s13423-015-0800-0

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi: 10.18637/jss.v067.i01

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*(38), 14325–14329. doi:10.1073/pnas.0803390105

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science, 24*(6), 981–990.

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

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(27755). 10.1038/srep27755

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage, 153,* 346–358. doi:10.1016/j.neuroimage.2016.03.063

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin, 111*(2), 291–303. doi:10.1037/0033-2909.111.2.291

Eysenck, M. W. (1979). Depth, elaboration, and distinctiveness. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 89–118). Hillsdale: Erlbaum.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception & Psychophysics, 77*(4), 1239–1251. doi:10.3758/s13414-015-0859-8

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *IEEE Conference on Computer Vision and Pattern Recognition*, 145–152. 10.1109/CVPR.2011.5995721

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2390–2398).

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139*(3), 558–578. doi:10.1037/a0019165

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science, 21*(11), 1551–1556. doi: 10.1177/0956797610385359

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 1097–1105). Curran Associates, Inc.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests in linear mixed effects models. Retrieved from https://CRAN.R-project.org/package=lmerTest

Maxfield, J. T., Stalder, W. D., & Zelinsky, G. J. (2014). Effects of target typicality on categorical search. *Journal of Vision*, *14*(1). 10.1167/14.12.1

Meng, M., & Potter, M. C. (2008). Detecting and remembering pictures with and without visual noise. *Journal of Vision, 8*(9), 7–7. doi:10.1167/8.9.7

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 27–46). New York: Oxford University Press.

Nijboer, T. C. W., Kanai, R., de Haan, E. H. F., & van der Smagt, M. J. (2008). Recognising the forest, but not the trees: An effect of colour on scene perception and recognition. *Consciousness and Cognition, 17*(3), 741–752. doi:10.1016/j.concog.2007.07.008

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*(2), 176–210. doi:10.1006/cogp.1999.0728

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene : A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175. doi: 10.1023/A:1011139631724

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 I.E. Conference on Computer Vision and Pattern Recognition Workshops* (pp. 512–519). Washington, DC, USA: IEEE Computer Society. 10.1109/CVPRW.2014.131

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 189–206). Hillsdale: Erlbaum.

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. doi:10.1016/0010-0285(76)90013-X

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25*(2), 207–222. doi:10.1080/14640747308400340

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science, 19*(2), 73–74. doi:10.3758/BF03337426

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137–149. doi:10.3758/BF03207704

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A, 43*(2), 161–204. doi:10.1080/14640749108400966

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*. 10.3389/fpsyg.2017.00142

Vogt, S., & Magnussen, S. (2007). Long-term memory for 400 pictures on a common theme. *Experimental Psychology, 54*(4), 298–303. doi: 10.1027/1618-3169.54.4.298

Voss, J. L. (2009). Long-term associative memory capacity in man. *Psychonomic Bulletin & Review, 16*(6), 1076–1081. doi:10.3758/PBR.16.6.1076

Watier, N., & Collin, C. (2012). The effects of distinctiveness on memory and metamemory for face-name associations. *Memory, 20*(1), 73–88. doi:10.1080/09658211.2011.637935

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using PLACES Database. *Advances in Neural Information Processing Systems, 27,* 487–495.