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

Lore Goetschalckx [Coresp., 1] , Johan Wagemans [1]

[1] Brain & Cognition, Katholieke Universiteit 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 to study neural and behavioral correlates of memorability while controlling for semantic category.

1 **MemCat: A new category-based image set quantified**

2 **on memorability**

5 Lore Goetschalckx[1], Johan Wagemans[1]

7 [1]Brain & Cognition, KU Leuven, Leuven, Vlaams-Brabant, Belgium

9 Corresponding author:

10 Lore Goetschalckx[1]

11 Tiensestraat 102, Leuven, Vlaams Brabant, 3000, Belgium

12 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 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 to study neural and behavioral correlates of memorability while controlling for semantic category.

Keywords: image memorability, data set, memorability scores, categories, ground truth, visual memory, recognition memory

Introduction

34

35   Photography is booming. Estimates say that over 1.2 trillion images have been captured in 2017

36   (Lee, 2018). That is more than 38K images per second! With a great variability of images out

37   there, a fair amount of research has been devoted to quantifying and understanding how images

38   differ in meaningful ways. Among the studied image properties are, for example, aesthetic

39   appreciation (e.g., Kong, Shen, Lin, Mech, & Fowlkes, 2016), popularity on social media (e.g.,

40   McParlane, Moshfeghi, & Jose, 2014), amount of visual clutter (e.g., Rosenholtz, Li, & Nakano,

41   2007), evoked fixations (e.g., Judd, Ehinger, Durand, & Torralba, 2009), etc.

42   It has been suggested that images also differ systematically in the likelihood of being

43   remembered and recognized later, a property that is referred to as image memorability (Isola,

44   Xiao, Parikh, Torralba, & Oliva, 2014). Using a repeat-detection memory task, in which

45   participants watch a sequence of images and respond whenever they see a repeat of a previously

46   shown image, Isola et al. (2014) assigned 2222 scene images a memorability score based on the

47   proportion of participants recognizing the image upon its repeat. They found that memorability

48   rank scores were highly consistent across participants. In other words, there was a lot of

49   agreement as to which images were remembered and recognized, and which ones were easily

50   forgotten. This suggests that memorability can indeed be considered an intrinsic image property

51   and that whether you will remember a certain image does not only depend on you as an

52   individual, but also on the image itself. The result has also been replicated with a more

53   traditional long-term visual memory task, with a separate study and test phase (Goetschalckx,

54   Moors, & Wagemans, 2018). Moreover, image memorability rankings have been shown to be

55   stable across time (Goetschalckx, Moors, & Wagemans, 2018; Isola et al., 2014), across contexts

56   (Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015), and across encoding types (intentional

57   versus incidental; Goetschalckx, Moors, & Wagemans, in press). Finally, while they might be

58   related to some extent, image memorability does not simply boil down to popularity (Khosla,

59   Raju, Torralba, & Oliva, 2015), aesthetics (Isola et al., 2014; Khosla et al., 2015), interestingness

60   (Isola et al., 2014), or the ability of an image to capture attention (Bainbridge, 2017).

61   The findings spurred new research aimed at understanding and predicting memorability. When it

62   comes to merely predicting, the best results so far are achieved using convolutional neural

63   networks (CNNs; e.g., Khosla et al., 2015). However, when it comes to truly understanding,

64   CNNs have often been critiqued to be black boxes. Nonetheless, Khosla et al.'s (2015) analyses

65   provided some further insight, mostly pointing at differences between broader image categories

66   and content types. For examples, units in the network with the highest correlation with

67   memorability seemed to respond mostly to humans, faces, and objects, while those with the

68   lowest correlation seemed to respond to landscapes and open scenes. Furthermore, the most

69   memorable regions of an image, according to the CNN, often capture people, animals or text.

70   These findings are in line with earlier work (Isola et al., 2014), which showed that the predictive

71   performance of a model trained on mere object statistics (e.g., number of objects) was boosted

72   considerably when the object labels were taken into account (Isola et al., 2014). In addition, a

73   model trained on the overall scene labels alone, already predicted memorability scores with a

74   Spearman's rank correlation of .37 to the ground truth. Memorable images often had labels

75   referring to people, interiors, foregrounds, and human-scaled objects, while labels referring to

76   exteriors, wide-angle vistas, backgrounds, and natural scenes were associated with low image

77   memorability scores. While differences in image categories (or labels) seem to play a large role,

78   they do not explain all the variability in memorability. Interestingly, consistent differences in

79   memorability scores remain even *within* image categories (Bylinskii et al., 2015). Goetschalckx,

80   Moors, Vanmarcke, and Wagemans (2018) for example, have argued that part of that variability

81   might be due to differences in how well the image is organized. Nonetheless, the concept of

82   memorability and its correlates is not yet fully understood to date and further research is required

83   to paint a clearer picture.

84   The current work presents a novel dataset of images quantified on memorability, that can be used

85   in research to achieve this goal (example images in Figure 1). To the best of our knowledge,

86   there were previously five large image sets with memorability scores, three consisting of regular

87   photographs, which is also the focus here: Isola et al. (2014), FIGRIM (Bylinskii et al., 2015),

88   and LaMem (Khosla et al., 2015), and two more specialized sets, which we will not further

89   discuss here: Bainbridge, Isola, and Oliva (2013; face images), and Borkin et al. (2013; data

90   visualizations). For completeness, we also mention a smaller set (850 images) that was used to

91   study which objects in an image are memorable (Dubey, Peterson, Khosla, Yang, & Ghanem,

92   2015). Table 1 compares MemCat to the other large datasets. The comparison is discussed in

93   more detail below.

94   A first feature of the current dataset is that it was designed to be representative for five different

95   broad natural categories and to allow to study memorability differences within semantic

96   categories. The set is characterized by a hierarchy of five broader, but memorability relevant

97   categories, further divided into more fine-grained subcategories. Only the FIGRIM set also offers

98   a category structure, but the number of exemplar images per category was lower: 59-157,

99   compared to 2,000 in the current set. We opted for broad categories to ensure that the whole was

100   still varied and representative enough, while containing a large number of exemplar images per

101   category at the same time. The chosen categories: animal, food, landscape, sports, and vehicle,

102   are memorability relevant in the sense that previous research has shown that they generally differ

103   in memorability. For example, knowing that the presence of people in an image is predictive for

104   memorability (see above), we chose one designated people category, the sports category, and

105   avoided including images with people in other categories. We included an animal category as a

106   non-human animate category, food and vehicle as more object-based categories, and landscape to

107   represent the wide exteriors that are often associated with lower memorability scores.

108   Second, we aimed for a large set, such that it would be suitable for machine learning approaches.

109   With a total of 10,000 images quantified on memorability, the current set is the second largest

110   memorability dataset, after LaMem.

111   Third, we sampled images from existing datasets, such that the image annotations collected

112   there, would also be available for researchers studying memorability. In particular, we searched

113   for images annotated with segmentation masks or at least bounding boxes, reasoning that they

114   may hold some indications of how the image is organized (e.g., where is the subject located),

115   which might be of particular interest when studying memorability within categories and factors

116   other than semantics.

117   In summary, the unique combination of features of MemCat, together with its richness in data,

118   make it a valuable addition to the memorability literature. We review possible uses of this set in

119   the Discussion.

Materials & Methods

121   **Participants.** There were 249 undergraduate psychology students (KU Leuven) who participated

122   in this study in exchange for course credits (216 female, 32 male, 1 other). Four students did not

123   disclose their age and the remainder were aged between 18 and 27 years old ($M = 19.24$, $SD=$

124   0.94). The majority of the participants, however, were recruited through Amazon's Mechanical

125   Turk (AMT) and received a monetary compensation (see further for details). The settings on

126 AMT were chosen such that only workers who indicated to be at least 18 years old and living in

127 the USA could participate. Further eligibility criteria were that the worker had to have an

128 approval rate of at least 95% on previous human intelligence tasks (HITs) and a total number of

129 previously approved HITs of at least 100. A total of 2162 AMT-workers participated in this

130 study (1139 female, 917 male, 4 other, and 102 who did not disclose this information). For the

131 1851 workers who disclosed their age, the reported ages ranged between 18 and 82 years old (*M*

132 = 37.14, *SD* = 11.89). The AMT data collection took place from April 2018 till July 2018. Data

133 collected through AMT has been shown to come from participant samples that are more diverse

134 than student samples and to be comparable in quality and reliability to those collected in the lab

135 (e.g., Buhrmester, Kwang, & Gosling, 2011).

136 **Materials.** MemCat consists of 10,000 images sampled from four previously existing image sets:

137 ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), SUN (Xiao, Hays, Ehinger, Oliva, &

138 Torralba, 2010), and The Open Images Dataset V4 (Kuznetsova et al., 2018). The four source

139 sets were chosen because of their large size (i.e., number of images), the availability of semantic

140 annotations, and the availability of bounding box annotations or more complete segmentation

141 masks for at least a subset of their images. The images selected from the source sets to be

142 included in MemCat belonged to the five broader semantic categories outlined in the

143 Introduction: animal (2,000 images), food (2,000 images), landscape (2,000 images), sports

144 (2,000 images), and vehicle (2,000 images). We explain the different steps in the selection

145 procedure in more detail below.

146 As a first step, we listed at least 20 subcategories for each broader category. The goal was to

147 obtain 2,000 images per category, without including more than 100 exemplar images per

148 subcategory. This was to ensure a reasonable level of variability and to avoid high levels of false

149 alarms in the memory task (see further). The subcategories were then translated to semantic

150 annotations from the source dataset. For example, for the subcategory "bear" (animal), we used

151 COCO images annotated with a "bear" tag and ImageNet images from nodes "American black

152 bear". "brown bear", and "grizzly". An overview of our hierarchy of categories and

153 subcategories, can be found in Figure 2.

154 The second step consisted of automatically sampling exemplar images from the listed

155 subcategories, while satisfying a number of shape restrictions. To avoid that images would stand

156 out because of an extreme aspect ratio, we only sampled images with aspect ratios between 1:2

157   and 2:1. Furthermore, the minimum resolution was set to 62,500 pixels. Finally, to ensure that

158   the images would fit comfortably on most computer monitors, we adopted a maximum height of

159   500 pixels and a maximum width of 800 pixels. However, for SUN and The Open Images

160   Dataset, only a low number of images satisfied the latter two restrictions (they were often too

161   big), which is why we opted to resize (using Hamming interpolation) images from those two

162   source datasets to meet the restrictions. Apart from the shape restriction, we also restricted the

163   sampling to images for which bounding box annotations or more complete segmentation masks

164   were available from the source datasets. Finally, we sampled more images than the target number

165   (2,000 images per broad category), anticipating exclusions in the next step.

166   The third step constituted manual selection work, carried out by the first author, assisted by two

167   student-interns. We manually went through the exemplar images sampled in the previous step,

168   and eliminated images following a number of exclusion rules. The exclusion rules can roughly

169   be divided into two kinds. A first kind of exclusion rule touches upon the quality of the image.

170   We excluded images of poor image quality (e.g., very dark, very much overexposed, blurry,

171   etc.), images that did not convincingly belong to the subcategory they were assigned to,[1] images

172   in greyscale or looking like they were the result of another color filter, images that were not real

173   photographs (e.g., drawings, digitally manipulated images, computer generated images), and

174   collages. A second set of rules concerns factors that could affect the memorability of an image,

175   but were not of interest for the purpose of MemCat. One such factor is text. We excluded images

176   containing large, readable text or text not belonging to the image itself (e.g., date of capture).

177   Furthermore, images depicting remarkably odd scenes (e.g., dog wearing Santa Clause costume)

178   were also excluded. Similarly, we avoided images depicting famous places or people, and images

179   of dead, wounded or fighting animals. Finally, there was one designated "people" category, the

180   sports category, and the presence of people in an image was avoided in all other categories (but

181   we allowed anonymous people in the background in the vehicle category or the presence of a

182   hand in images of the food category). In addition to these exclusion rules, we also tried, to the

183   best of our ability, not to include (near) duplicate images. If the target number of images was not

184   obtained after Step 3, we reverted back to Step 2, if there were still images to sample from, or to

185   Step 1 if we needed to include additional subcategories.

186 Finally, for those categories for which more than the target number of images survived Step 3,

187 there was a fourth step to randomly down-sample the selection to the target number, assigning

188 higher sampling probabilities to images annotated with segmentation masks.

189 In addition to MemCat, we collected 10,000 filler images, that were not quantified on

190 memorability themselves, but were needed in the memory task used to quantify the other images.

191 The filler images were sampled randomly from The Open Images Dataset, but from a different

192 subset to avoid overlap.[2] As these images would function only as filler images, there were fewer

193 restrictions. For example, the images could be of any category, they were allowed to contain text,

194 etc. However, the same shape restrictions were still applied.

195 **Procedure.** Having carefully collected 10,000 images for MemCat, the next step was to quantify

196 them on memorability. Following previous work, this was achieved by presenting the images in

197 an online repeat-detection memory game (Isola et al., 2014; Khosla et al., 2015), in which

198 participants watch a sequence of images and are asked to respond when they recognize a repeat

199 of a previously shown image. Students participating for course credits played the game in the

200 university's computer labs, hosting about 20 students at a time. AMT workers played the game

201 from the comfort of their homes (or whichever location they preferred). Prior to starting the

202 game, all participants were prompted to read through an informed consent page explaining the

203 aims of the study and their rights as participants. They could give their consent by actively

204 ticking a box. The study was approved by SMEC, the Ethical Committee of the Division of

205 Humanities and Social Sciences, KU Leuven, Belgium (approval number: G-2015 08 298).

206 For the task design of the game, we closely followed Khosla et al. (2015), as their version of the

207 game was designed to quantify large numbers of images. We divided the game into blocks of

208 200 trials. On each trial, an image was presented at the center of the browser window for a

209 duration of 600 ms, with an intertrial interval of 800 ms. During this interval, a fixation cross

210 was shown. Sixty-six images were target images, sampled randomly from MemCat, and repeated

211 after 19 to 149 intervening images. Forty-four images were random filler images that were never

212 repeated. Finally, there were 12 additional random filler images that were repeated after 0 to 6

213 intervening images to keep participants attentive and motivated. They are referred to as vigilance

214 trials. Participants could indicate that they recognized a repeat by pressing the space-bar. They

215 did not receive trial-by-trial feedback, but were shown their hit rate as well as number of false

216 alarms at the end of the block. Figure 3 presents a schematic of the game.

217   Each block lasted a little less than 5 minutes. Care was taken to ensure that an image was never

218   repeated more than once and never across blocks. Students were asked to complete as many

219   blocks as they could in one hour, with one bigger, collective break of roughly ten minutes after

220   half an hour, and smaller self-timed breaks between the remainder of the blocks. Most students

221   could complete 8 blocks, but for some groups, slow data uploads at the end of a block resulted in

222   lower numbers. AMT workers could complete one to 16 blocks, were allotted 48h to submit their

223   completed blocks (so, they were allowed to spread the blocks over time), and were paid $0.40

224   per block. To ensure a good quality of the AMT data and to avoid random or disingenuous

225   responses, AMT workers were blocked from playing anymore blocks after two with a dprime

226   lower than 1.5 on the vigilance trials. They were warned the first time this happened.

227   **Memorability Measures.** Following Khosla et al. (2015), we deduced two image memorability

228   measures from the data collected through the repeat-detection memory game. One measure was

229   calculated as the proportion of participants recognizing the image when shown to them for the

230   second time (i.e., the hit rate across participants), as was also done in Isola et al. (2014). The

231   other measure was based on the same principle, but penalized for false alarms (i.e., when

232   participants press the space-bar for the first presentation of the image) in the way proposed in

233   Bainbridge et al. (2013). Rather than H/N_{resp} (first measure), the formula was the following: (H-

234   F)/N_{resp}, where H is the number of participants recognizing the image, F is the number of

235   participants making a false alarm when the image is presented for the first time, and N_{resp} is the

236   total number of participants having been presented with the image. On average, an image's N_{resp}

237   was 99 (after exclusions). Note that the memorability scores have an upper bound of 1 and a

238   lower bound of 0. In theory, (H-F)/N_{resp} could result in a negative score, but in practice it is

239   highly unlikely that there would be more participants making a false alarm for the image than

240   there are participants making a hit.

Results

242   **Participant Performance.** As mentioned, the performance on the easier vigilance trials was

243   taken as an indication of whether participants were playing the memory game in a genuine way.

244   If in a certain block, a participant did not distinguish vigilance repeats from non-repeat trials with

245   a dprime of at least 1.5 (preset performance threshold), that block was excluded from further

246   analyses. The exclusion rate amounted to 3% of all played blocks. Recall, however, that AMT

247   workers were not allowed to play more blocks after two excluded ones.

248   After exclusion, the mean dprime across participants was 2.77 ($SD = 0.56$) for the vigilance

249   repeats, and 2.47 ($SD = 0.50$) for the target repeats. Table *2* summarizes participants' overall

250   performance, collapsing over vigilance and target repeats. Participants generally performed well

251   on the task.

252   **Memorability Scores.** Participants' high performance was also reflected in the average image

253   memorability scores. Figure 4 displays the mean for each of the two memorability measures as a

254   horizontal line ($M_{H/Nresp} = .76$, SD; $M_{(H-F)/Nresp} = .70$). It is comparable to the mean observed in

255   Khosla et al. (2015). In addition, Figure 4 visualizes the distribution of the collected image

256   memorability scores for each of the five broad main categories separately. A simple linear

257   regression revealed that the category explains 43% of the variance in the H/N_{resp} scores and 44%

258   of the variance in the (H-F)/N_{resp}. In line with previous research, the landscape images were on

259   average the least memorable ($M_{H/Nresp} = .60$; $M_{(H-F)/Nresp} = .53$). They were followed by the vehicle

260   images ($M_{H/Nresp} = .76$; $M_{(H-F)/Nresp} = .70$). Somewhat surprisingly, the food images generally came

261   out on top of the ranking ($M_{H/Nresp} = .85$; $M_{(H-F)/Nresp} = .80$), topping the animal $M_{H/Nresp} = .80$; $M_{(H-}$

262     $_{F)/Nresp} = .73$) and sports $M_{H/Nresp} = .78$; $M_{(H-F)/Nresp} = .71$) categories. However, there is still a large

263   degree of variability that is not explained by differences in broad image categories. Indeed,

264   memorability varied considerably within categories as well, with SDs of: .09 (animal; $SD_{(H-}$

265     $_{F)/Nresp} = .09$), .08 (food; $SD_{(H-F)/Nresp} = .08$), .13 (landscape; $SD_{(H-F)/Nresp} = .14$), .09 (sports; $SD_{(H-}$

266     $_{F)/Nresp} = .10$), and .09 (vehicle; $SD_{(H-F)/Nresp} = .09$).

267   Having observed that images from the same broader category indeed still differed in

268   memorability, the next question was whether these differences are consistent across participants.

269   This question taps into the reliability of the memorability measures. Following previous

270   memorability work (e.g., Isola et al., 2014), the consistency was assessed by randomly splitting

271   the participant pool in half, computing the memorability scores for each half separately and

272   determining the Spearman's rank correlation between the two sets of scores. This was repeated

273   for 1,000 splits and the Spearman's rank correlation was averaged across the splits. Figure 5

274   shows the results in function of the mean N_{resp} for each category as well as for the total image set.

275   We first discuss the results for H/ N_{resp} (see Figure 5, left panel). When collapsing over all five

276   categories, the observed mean split-half Spearman's rank correlation with all available responses

277   (N_{resp} = 99, on average) amounted to .78. In comparison, Khosla et al. (2015) reported a mean

278   split-half Spearman's rank correlation of .67 for their LaMem dataset. However, they only

279   collected 80 responses per image. After randomly down-sampling our data to an N_{resp} of 80, we

280   still found a split-half consistency of .73. With the exception of the landscape category, for

281   which we observed a total (i.e., without down-sampling) split-half consistency of .77, the total

282   per category split-half consistency estimates were lower, ranging between .59 and .67. This is

283   possibly due to smaller ranges of memorability scores within those categories (see Figure 4).

284   Note, however, that the split-half consistencies are an underestimate of the reliability of the

285   memorability scores calculated based on the full participant pool. The latter can be estimated

286   from the split-half consistency by means of the Spearman-Brown formula (Brown, 1910;

287   Spearman, 1910). Applying this formula, we found the following final reliabilities for the H/N_{resp}

288   memorability scores : .87 (all), .80 (animal), .75 (food), .87 (landscape), .75 (sports), .78

289   (vehicle).

290   For the (H-F)/N_{resp} memorability scores, we confine the discussion to pointing out that the pattern

291   of results is qualitatively similar, although the final reliabilities are somewhat lower: .86 (all), .74

292   (animal), .71 (food), .85 (landscape), .77 (sports), .71 (vehicle).

293   Finally, after finding that the two image memorability measures were both acceptably reliable,

294   we asked how they compared to each other. In the current dataset, they were highly

295   intercorrelated, as evidenced by a Pearson correlation of .93 when collapsing over all five

296   categories. The per category correlations were: .82 (animal), .90 (food), .91 (landscape), .85

297   (sports), .88 (vehicle).

Discussion

299   We presented a new dataset, MemCat, consisting of a total of 10,000 images, each quantified on

300   memorability using a repeat-detection memory task (first introduced by Isola et al., 2014, version

301   used here based on Khosla et al., 2015). MemCat is the second largest image memorability

302   dataset available, and the largest that is based on a category structure. That is, it is divided into

303   five broader, memorability relevant semantic categories: animal, food, landscape, sports, and

304   vehicle, each with 2,000 exemplar images, which are further divided into subcategories (e.g.,

305   bear, cat, cow). Furthermore, the images were sampled from popular, existing datasets such that

306   additional annotations available there (e.g., segmentations masks or bounding boxes) would also

307   be available to researchers wishing to use MemCat for research aimed at investigating specific

308   factors underlying memorability.

309   Replicating previous research, we found that images differ considerably in memorability and that

310   these differences are highly consistent across participants. Part but not all of this variability can

311   be explained by differences between the five broader semantic categories. Note, however, that

312   this result is correlational in nature, and that one should be cautious drawing causal conclusions.

313   In line with Bylinskii et al. (2015), considerable variability in memorability remained even

314   within the categories. However, the consistency there was somewhat lower, probably because the

315   variance was also lower. When the differences between images become smaller, it becomes

316   harder to reliably and consistently distinguish them. Nevertheless, the consistency estimates per

317   category were still high, indicating that we obtained reliable memorability scores. Finally, we

318   reported results for two different methods to compute memorability scores. One is to compute

319   the hit rate across participants: H/N_{resp}. This was the method used in the original work by Isola et

320   al. (2014). However, in principle, it possible that some images elicit more key presses not

321   because they are truly recognized, but for some other reason (e.g., they seem familiar). That is

322   why Bainbridge et al. (2013) suggested to correct for false alarms (i.e., when participants press

323   the key for the first presentation of an image, when it is not a repeat) by computing $(H-FA)/N_{resp}$.

324   We report both measures for comparison, but note that they lead to a highly similar pattern of

325   results and are also strongly intercorrelated.

326   In what follows, we discuss possible uses of MemCat. Most of what we learned from previous

327   studies about what makes an image memorable is specified in terms of semantic categories or

328   content types (e.g., images of people are more memorable than landscapes). However, a

329   considerable amount of variability was still left unexplained. A primary use of the current dataset

330   is in studies aiming to better understand the factors underlying image memorability. In particular,

331   with 2,000 images for each of five broader categories, it allows to zoom in on variability within

332   categories. This variability is of more interest to practical applications (e.g., advertising,

333   education), because the semantic category or the content type (e.g., a certain product) will often

334   be predefined and it will be a matter of choosing or creating a more memorable depiction of it. In

335   addition to dividing the set into broad semantic categories, we also avoided variability due to

336   other factors already discovered in previous studies (e.g., we excluded images depicting oddities,

337   images containing text or recognizable places or faces), thus creating a set designed to help

338   understand the previously unexplained variability in image memorability.

339   Second, MemCat is also useful as a benchmark for machine learning approaches to automatically

340   predict memorability. Currently, LaMem (Khosla et al., 2015) is most often used, but models can

341   now also be trained and tested on the current dataset. When taking Khosla et al.'s (2015)

342   MemNet-CNN (without retraining), we found that its predictions show a rank correlation of 0.68

343   with the (H-F)/N_{resp} memorability scores in the current set, suggesting that there is room for

344   improvement. Given the category structure in MemCat, one could explore, for the first time,

345   memorability models with one or more layers that are specific to a category. Indeed, it is possible

346   that what makes landscape images memorable is different from what makes animal images

347   memorable.

348   Finally, a third possible use is in neuroscientific studies or psychophysical studies examining

349   effects of memorability. The current set offers a large number of quantified images to choose

350   from. Moreover, it facilitates matching memorability conditions (e.g., high versus low) on

351   semantic category, something that is often done in neuroscientific studies (e.g., Bainbridge,

352   Dilks, & Oliva, 2017; Khaligh-Razavi, Bainbridge, Pantazis, & Oliva, 2016).

353   **Usage.** On the MemCat project page, http://gestaltrevision.be/projects/memcat/, we provide a

354   link to the collection of 10K images as well as links to two data files. One file describes the

355   images that were used and contains columns indicating the image filename in its source dataset,

356   the name of its source dataset, the category (e.g., animal) and subcategory (e.g., bear) we

357   assigned it to, the label that was used to sample it from its source dataset (e.g., American black

358   bear), the current width, the current height, the factor by which it was resized (both the original

359   width and height were multiplied by this factor), the number of hits (H), the number of false

360   alarms (FA), the number of participants it was presented to (N_{resp}), and the two memorability

361   scores. The other file contains the data collected in the repeat-detection memory game. Its

362   columns indicate the participant ID (anonymized), the participant's age, the participant's gender,

363   whether or not they participated through AMT, the block number, the trial number, the image

364   shown, the trial type (target, target repeat, filler, vigilance, vigilance repeat), the participant's

365   response (hit, correct rejection, miss, false alarm), the screen width, and the screen height.

366

367 **Acknowledgments**

368 The authors would like to thank the student-interns who have assisted in the image selection and

369 data collection: Justine Aeyels and Joran Geeraerts, and Christophe Bossens, who is responsible

370 for the technical support in the lab and contributed greatly to the implementation of the repeat-

371 detection memory task on Amazon's Mechanical Turk.

372

373 **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 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

Bainbridge, W. A. (2017). The resiliency of memorability: A predictor of memory separate from attention and priming. *ArXiv*, *1703.07738 [q-bio]*. Retrieved from http://arxiv.org/abs/1703.07738

Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage*, *149*, 141–152. https://doi.org/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. https://doi.org/10.1037/a0033872

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. https://doi.org/10.1109/TVCG.2013.234

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322. https://doi.org/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.

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. https://doi.org/10.1016/j.visres.2015.03.005

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, 248–255. https://doi.org/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. https://doi.org/10.1109/ICCV.2015.130

Goetschalckx, L., Moors, J., & Wagemans, J. (in press). Incidental image memorability. *Memory*.

405   Goetschalckx, L., Moors, P., Vanmarcke, S., & Wagemans, J. (2018). Get the picture?

406   Goodness of image organization contributes to image memorability. *PsyArXiv*.

407   https://doi.org/10.17605/OSF.IO/8ATJ9

408   Goetschalckx, L., Moors, P., & Wagemans, J. (2018). Image memorability across longer time

409   intervals. *Memory*, *26*(5), 581–588. https://doi.org/10.1080/09658211.2017.1383435

410   Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph

411   memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7),

412   1469–1482. https://doi.org/10.1109/TPAMI.2013.200

413   Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans

414   look. *2009 IEEE 12th International Conference on Computer Vision*, 2106–2113.

415   https://doi.org/10.1109/ICCV.2009.5459462

416   Khaligh-Razavi, S.-M., Bainbridge, W. A., Pantazis, D., & Oliva, A. (2016). From what we

417   perceive to what we remember: Characterizing representational dynamics of visual

418   memorability. *BioRxiv*, 049700. https://doi.org/10.1101/049700

419   Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting

420   image memorability at a large scale. *2015 IEEE International Conference on*

421   *Computer Vision (ICCV)*, 2390–2398. https://doi.org/10.1109/ICCV.2015.275

422   Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking

423   network with attributes and content adaptation. *ArXiv CS*, *1606.01621*. Retrieved

424   from http://arxiv.org/abs/1606.01621

425   Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., … Ferrari, V.

426   (2018). The Open Images Dataset V4: Unified image classification, object detection,

427   and visual relationship detection at scale. *ArXiv:1811.00982*.

428   Lee, E. (2018, September 12). Our best photos deserve to be printed [Blog post]. Retrieved

429   from InfoBlog - Insight from InfoTrends website: http://blog.infotrends.com/our-best-

430   photos-deserve-to-be-printed/#more-24362

431   Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., … Dollár, P. (2014).

432   Microsoft COCO: Common Objects in Context. *ArXiv:1405.0312 [Cs]*. Retrieved

433   from http://arxiv.org/abs/1405.0312

434   Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed).

435   Mahwah, NJ: Lawrence Erlbaum Associates.

436   McParlane, P. J., Moshfeghi, Y., & Jose, J. M. (2014). "Nobody comes here anymore, It's too

437   crowded": Predicting image popularity on Flickr. *Proceedings of International*

438   *Conference on Multimedia Retrieval*, 385–391.

439   https://doi.org/10.1145/2578726.2578776

440   Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*,

441   *7*(2), 17–17. https://doi.org/10.1167/7.2.17

442   Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology,*

443   *1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

444   Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-

445   scale scene recognition from abbey to zoo. *2010 IEEE Conference on Computer*

446   *Vision and Pattern Recognition (CVPR)*, 3485–3492.

447   https://doi.org/10.1109/CVPR.2010.5539970

Example images of MemCat.

The memorability score, calculated as the hit rate across participants (H/N$_{resp}$) is indicated in the bottom right corner.

Category hierarchy of MemCat.

PeerJ

MemCat - 10K

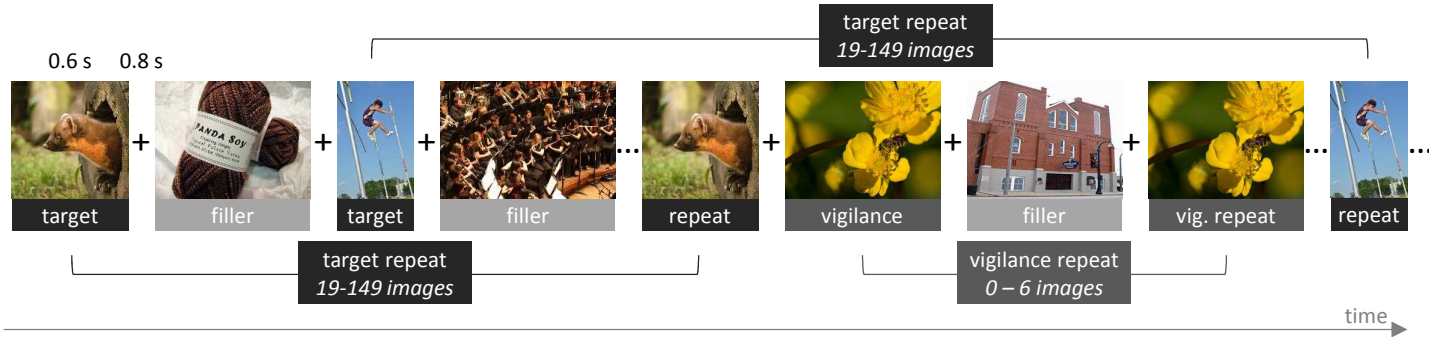
animal - 2K
- bear - 100
- cat - 100
- chicken - 100
- cow - 100
- deer - 99
- dog - 100
- duck - 100
- elephant - 100
- fox - 99
- giraffe - 99
- hamster - 97
- horse - 99
- lion - 100
- marten - 61
- mouse - 48
- pigeon - 99
- rabbit - 99
- sheep - 100
- squirrel - 100
- wild boar - 100
- zebra - 100

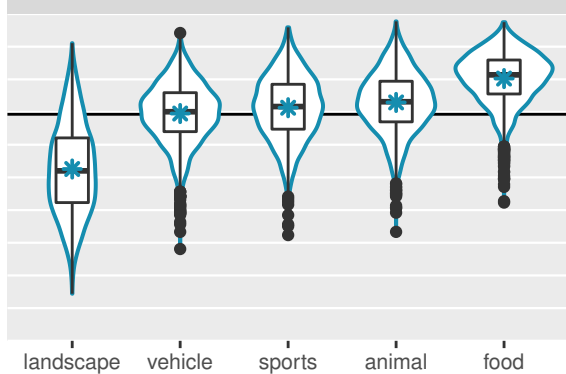
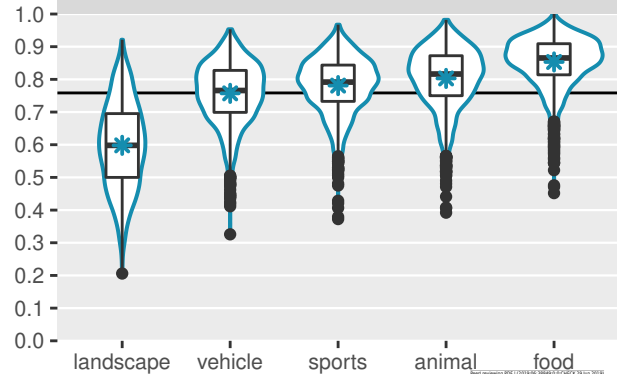
food - 2K
- apple - 31
- bagel - 80
- banana - 82
- barbecue - 45
- bell pepper - 55
- bread - 84
- breakfast - 75
- broccoli - 86
- burrito - 81
- butternut squash - 87
- cake - 87
- champagne - 43
- cheese - 39
- cocktail - 64
- coffee - 61
- donut - 54
- gravy - 73
- hotdog - 82
- juice - 39
- mashed potato - 85
- muffin - 51
- orange - 35
- pasta - 82
- pizza - 84
- salad - 87
- sandwich - 82
- soup - 84
- tea - 49
- yogurt - 63
- zucchini - 50

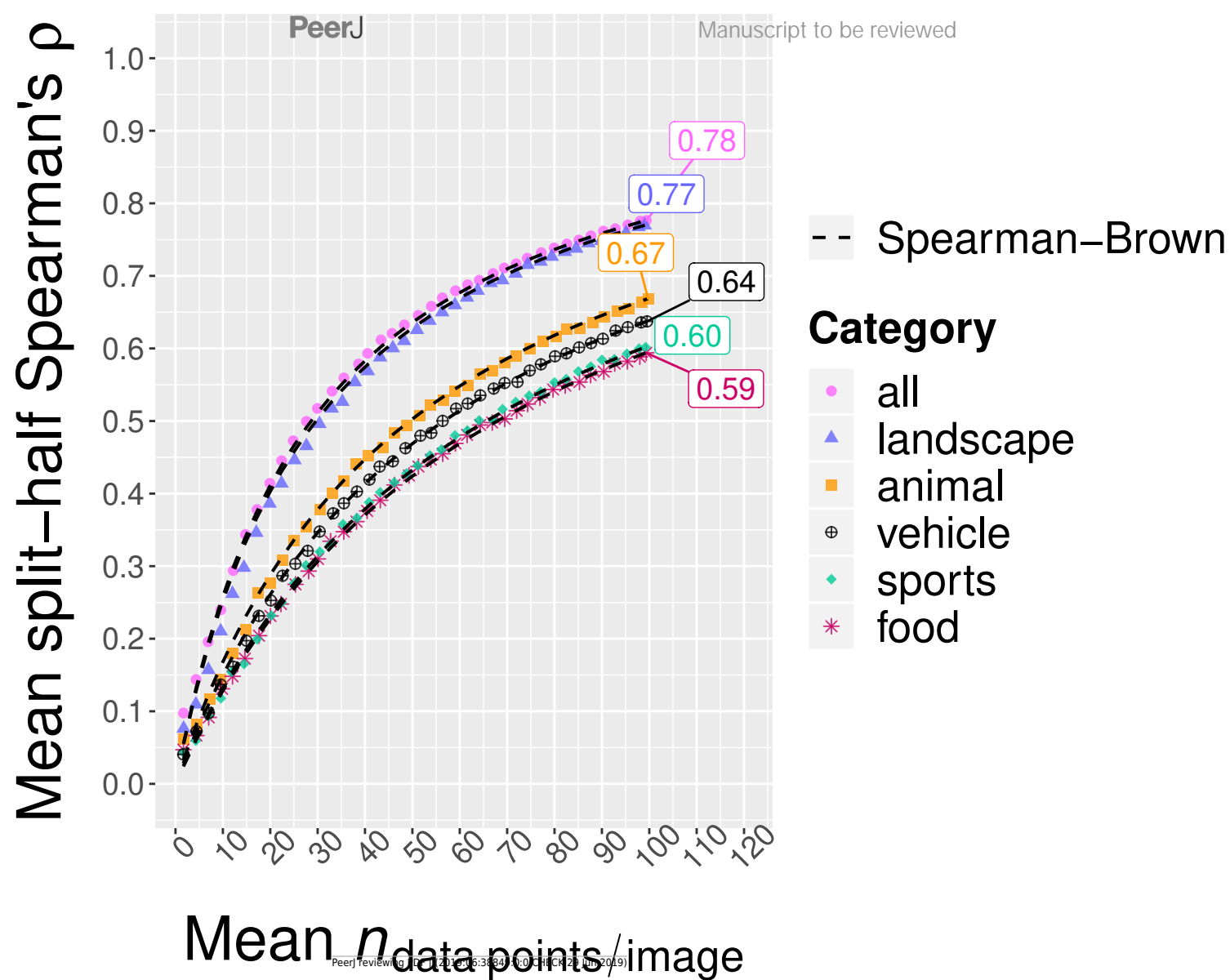
landscape - 2K
- badlands - 95
- barn - 86
- creek - 88
- desert - 79
- field - 96
- forest (broadleaf) - 94
- forest (needleleaf) - 55
- geyser - 95
- icescape - 89
- jungle - 57
- lake - 95
- mangrove - 41
- moor - 4
- mountain - 92
- pasture - 61
- plantation - 96
- river - 94
- savanna - 49
- sea - 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 - 37
- 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

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

Distribution of the collected memorability measures.

The left panel represents memorability scores computed as the hit rate across participants. The right panel 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 2,000 quantified images.

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

Estimates are based on 1,000 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.

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

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