Behavior Research Methods (2019) 51:1179–1186
https://doi.org/10.3758/s13428-018-1053-4

Establishing reference scales for scene naturalness and openness

Naturalness and openness scales

Hanshu Zhang[1] · Joseph W. Houpt[1] · Assaf Harel[1]

Published online: 29 May 2018

Abstract
A key question in the field of scene perception is what information people use when making decisions about images of scenes. A significant body of evidence has indicated the importance of global properties of a scene image. Ideally, well-controlled, real-world images would be used to examine the influence of these properties on perception. Unfortunately, real-world images are generally complex and impractical to control. In the current research, we elicit ratings of naturalness and openness from a large number of subjects using Amazon Mechanic Turk. Subjects were asked to indicate which of a randomly chosen pair of scene images was more representative of a global property. A score and rank for each image was then estimated based on those comparisons using the Bradley–Terry–Luce model. These ranked images offer the opportunity to exercise control over the global scene properties in stimulus set drawn from complex real-world images. This will allow a deeper exploration of the relationship between global scene properties and behavioral and neural responses.

Keywords Scene perception · Global properties · Rating references

Introduction

Research on scene perception has shown that rapid scene categorization may be based on global properties related to the scene structure and affordances (Greene and Oliva, 2009b; see also, Oliva and Torralba, 2001; Torralba and Oliva; 2003). Behavioral research indicates that these global properties are processed faster than basic-level categories and are potentially identified prior to basic-level categorization (Greene & Oliva, 2009a; Loschky & Larson, 2010). Neurophysiological studies also indicate a structured relationship between neural activity and global scene properties. EEG analyses indicate global properties affect perceptual processing early (Groen, Ghebreab, Lamme, & Scholte, 2013; Groen, Ghebreab, Prins, Lamme, & Scholte, 2016; Harel, Groen, Kravitz, Deouell, & Baker, 2016). fMRI research has demonstrated that the parahippocampal place area (PPA) responds preferentially to pictures of scenes, landmarks, and spatial layouts. Moreover, studies found that scene representations in PPA are mainly based on spatial layout information but not scene category per se (Kravitz, Peng, & Baker, 2011; Park, Brady, Greene, & Oliva, 2011), suggesting the global property of "openness" plays a particularly important role in the scene responsive brain area. Previous literature has focused on various aspects of "openness" including spatial enclosure (Greene & Oliva, 2009b), spatial expanse (Kravitz et al., 2011) and spatial boundary (Lowe, Gallivan, Ferber, & Cant, 2016; Park et al., 2011).[1]

Unlike categories, which by definition require a scene to be a member or not, global properties are more continuous in nature. In other words, it is often assumed that global properties should be represented by a ratio or ordinal scale

✉ Hanshu Zhang
zhang.180@wright.edu

[1] Department of Psychology, Wright State University, Dayton, OH 45435, USA

[1] Previous research has defined "openness" in different ways. For example, Greene and Oliva (2009b) suggested that the "degree of openness represents the magnitude of spatial enclosure whereas the degree of expansion refers to the perspective of the spatial layout of the scene"(p. 140). Similarly, Kravitz et al. (2011) used the term *"expanse"* to describe open and closed scenes, defined by whether the scene implied the viewer was in an enclosed space. Yet another related definition is *"Spatial Boundary"*. Park et al. (2011) described spatial boundary as "expansive and open to the horizon or closed and bounded by frontal and lateral surfaces" (see also Lowe et al., 2016). Given these multitude of definitions, in the current research, we aimed to understand how participants rated "openness" their own, without imposing on them any definitions or criteria.

rather than a classificatory scale. Even theories of strict category membership often allow for a latent scale that is mapped probabilistically into a category (e.g., Ashby & Townsend, 1986; Nosofsky, 1986). However, in most research thus far, the distinction has been glossed over by selecting scene images from basic-level categories that typify global properties. For example, for naturalness, a forest may be used; for openness, a calm lake or ocean scene may be used; for manmade, a cityscape may be used, and so on. While this choice of stimuli may reflect practical experimental needs rather than theoretical ones, as it may reduce noise in participants' responses regarding the global properties, it also limits the ability to examine how people use these global properties in general. For example, performance variation across different global properties could be due to differences in perceptual discriminability of the basic-level categories (Banno and Saiki, 2015; Sofer et al., 2015). If nothing else, the use of these "good representations" inflates human observers' performance (Ehinger, Xiao, Torralba, & Oliva, 2011; Torralbo et al., 2009). It remains to be studied whether existing conclusions about global properties will hold when scenes that are less extreme representatives of those properties are used as stimuli.

Following Greene and Oliva (2009b, 2010), we take the position that most, if not all, global properties should be treated as continuous scale properties. For example, a golf course is more manmade than a forest, but less manmade than a restaurant. As such, our goal is to establish an image set that can be used to manipulate the most commonly studied global properties, the manmade–natural dimension and the open–closed dimension (Banno & Saiki, 2015; Harel et al., 2016; Harel, Kravitz, & Baker, 2013; Kravitz et al., 2011; Loschky & Larson, 2010; Park et al., 2011; Sofer et al., 2015; Torralbo et al., 2009).

In the current research, we use the Bradley–Terry–Luce model (BTL; Bradley & Terry, 1952; Suppes & Zinnes, 1963 pp. 48–54) to estimate a derived ratio scale representation of four global properties: "open", "closed", "natural", and "manmade". This scale will allow for research that better aligns with the definition of global properties posited by Greene and Oliva (2009b).

In previous studies, derived global property scales have been estimated, although with a either smaller set of scenes or fewer judgments. For example, Ross and Oliva (2010) asked human observers to give ratings of global properties (depth, openness, and perspectives) in a scene relative to a reference scale defined by a set of preselected images. In Kravitz et al. (2011), participants judged which of a pair of scenes was either more open, more natural, or more distant, and derived a ranking scale for each attribute using Elo rating (Elo, 1978). We extended these investigations in two ways. First, we used a large number of participants and ratings per participant by deploying our task on Amazon Mechanical Turk. Second, we used images from a large database so that the full image set contained a large number of images distributed across a wide range of basic-level categories.

Our task focused on comparing pairs of images (i.e., "Which of these images is more natural?") rather than giving scale rankings or comparison to a fixed exemplar. We used the outcomes of these paired comparisons to derive an ordinal ranking and a ratio score for each scene on the global properties: "open", "closed", "manmade", and "natural". We included three validation checks for the resultant scales. First, we verified that the extremes and midpoints of each scale corresponded with the expected degree of each property. Second, we verified that opposing ranking scales were anti-correlated (e.g., an image with a high "natural" score should have a low "manmade" score). Finally, we compared the BTL model to a baseline model using the AIC.

Methods

Materials Images were selected from the Scene Understanding (SUN) database (Xiao et al., 2010). The SUN database consists of more than 100,000 images classified into 397 basic-level categories. Example images from the database are shown in Fig. 1.

Scene images were selected to be 1024 × 768 pixels. Following the definition of a scene given by Oliva and Torralba (2001), we avoided scenes that included a prominent object, or ones in which objects occupy the majority of the scene image. Examples of the type of image that was excluded are shown in Fig. 2. This resulting collection included 7035 images describing 174 basic-level categories for building the ranking scales.

Participants Human observers (N = 1099) were recruited on Amazon Mechanic Turk (Mturk, https://www.mturk.com/mturk/welcome). Mturk is a crowdsourcing web service that coordinates the supply and the demand of tasks that require human intelligence to complete (Paolacci et al., 2010). We did not ask participants on Mturk to report their demographic information. A real-time demographic information can be retrieved online at http://demographics.mturk-tracker.com/#/gender/all (Difallah, Catasta, Demartini, Ipeirotis, & Cudrè-Mauroux, 2010; Ipeirotis, 2015)

Participants on Mturk were asked to complete a task (HITs on Mturk, an acronym for Human Intelligence Tasks) about information processing of global properties. The task set several eligibility criteria for participants: located in the U.S.; numbers of HITs approved were greater than 50; the HIT approval rate for all requesters was more than 95%.

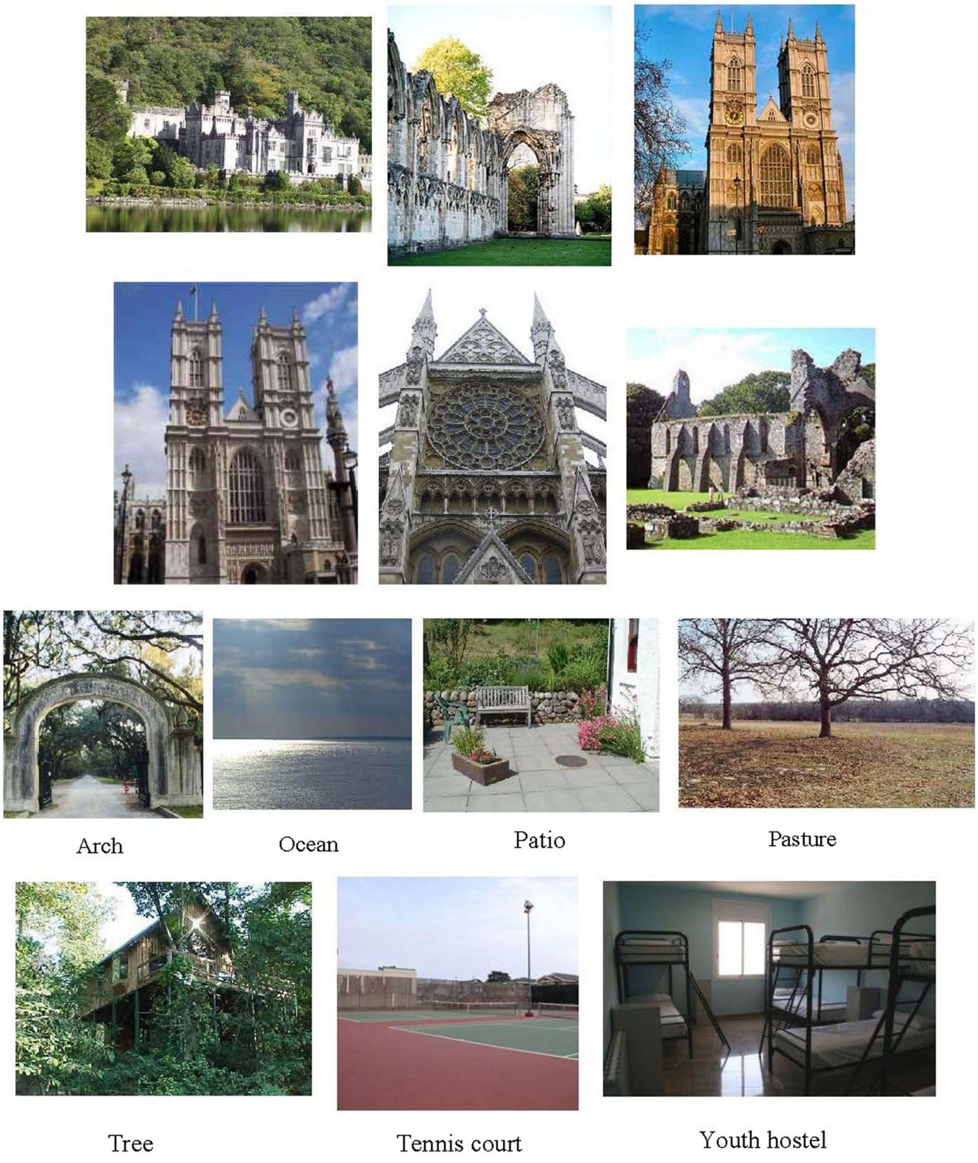


Fig. 1 Different Abbey images (*top*) and different scene categories (*bottom*) in SUN database

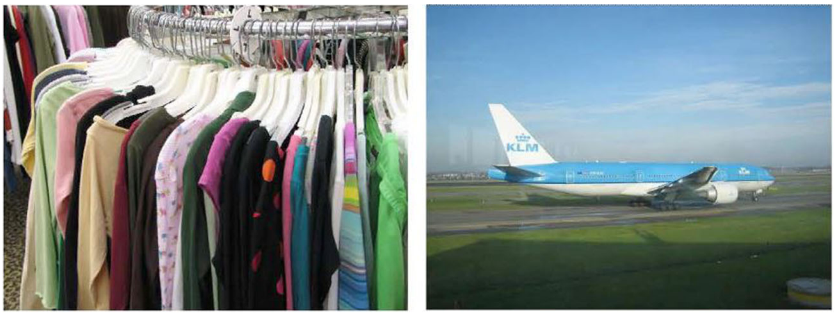


Fig. 2 The example of images excluded from the stimuli: clothes store (*left*) and runway (*right*). The image on the left focuses on the shirts more than a "clothing store" scene. Similarly, viewers may see the image on the right as single object, "airplane," rather than the basic-level scene category "runway"

These criteria were to ensure that all the participants had no difficulty in understanding English and were familiar with Mturk.

Procedure The initial screen informed participants about the study and required their consent to participate before advancing to the main experiment. The initial screen included four images that each demonstrated a global property, "open", "closed", "natural", and "manmade". There was no restriction on the number of times a participant could complete the task, although each time they were required to complete the consent form. Participants received $1.75 as compensation for their participation.

We deployed four versions of the task. In one version of the task, participants were asked to select from two side-by-side images "Which scene is more natural?" (see Fig. 3). The three other versions of the task were identical but with "natural" replaced by one of "manmade", "open", or "closed". After selecting one of two images, the participant needed to click a button to proceed to the next trial. Each participant was assigned only one tested global property per task. As we did not preclude participants from running multiple times, it is possible that some participants completed multiple versions.

Participants were instructed to choose the image that better matched the global property between the two displayed images. The initial trial was fixed to include extremes of the ranking scales to help orient participants to the task. Images were randomly selected from the 7035 scene images on subsequent trials. There were 450 scene pairings in each task, which took participants approximately 15–25 min to finish. For most participants, "catch trials" were included every 30 trials to verify whether they paid attention to the task.[2] Figure 4 shows example images that were used in catch trials. In the "open" and "closed" tasks, catch trials were comparisons between a randomly selected image from the basic level category of "beach" and "bedroom". Participants were supposed to select "beach" for "open" and "bedroom" for "closed". "Broadleaf-forest/dining room" images were chosen for "natural/manmade" catch trials, "broadleaf" was considered as a better answer in "natural" while "dining room" fitted the "manmade" description better. There were 15 catch trials in total for each question type. Participants who failed more than five trials were excluded from the final data analysis.

Nearly all participants were engaged enough in the task to pass our catch-trial criterion. Of the 278 participants who participated in the "manmade" task, 17 participants were excluded. Eighteen of 259 participants in the "closed" were excluded. Among those for whom we had catch trials in "natural" and "open" tasks, three of 141 participants in "natural" task and seven of 140 participants in "open" task were excluded. In total, responses from 1055 participants ("manmade": 261; "closed": 241; "natural": 279; "open": 274) were included in the data analysis.

Analyses

Following Kravitz et al. (2011), we initially attempted to use Elo and the other two rating systems that extended Elo, Glicko and Stephen (Stephenson & Sonas, 2016) to derive the global property scales. However, the assumptions made by those ranking systems that were specific to chess did not necessarily apply to the global scene properties. Furthermore, as Kravitz et al. (2011) pointed out, the order in which comparisons are entered into the model impacts the results of Elo. Ultimately, we chose the BTL model for deriving the scales for each property.

The BTL model assumes that the derived scores for both images in a comparison, ν_a, ν_b, map to the probability that the first image is chosen, π_{ab}, by the equation,

$$\pi_{ab} = \frac{\nu_a}{\nu_a + \nu_b}. \tag{1}$$

For example, if image a has a derived "openness" score of $\nu_a = .5$ and image b has a derived "openness" score of $\nu_b = 1$, then the probability that image a is chosen as more open is $\pi_{ab} = .5/1.5 = 1/3$.

If the scores are allowed to be negative, then the derived scale is a *generalized* ratio scale because the exact probabilities could result form multiplying all of the scores by -1 (Suppes & Zinnes, 1963 p. 51). Thus, the standard is to force the scores to be positive using an exponential transformation, $\lambda_a = \exp(\nu_a)$, so that the derived scale is a ratio scale.

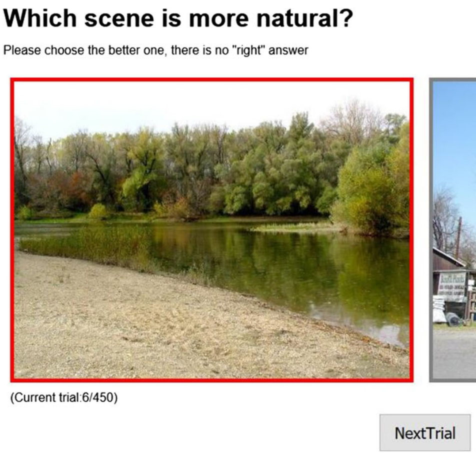
To estimate the scores for each scale, we used the BradleyTerry2 (Turner & Firth, 2012) in R (R Development Core Team, 2011) which implements a penalized quasi-likelihood algorithm from Breslow and Clayton (1993).

Results

The scene images were ranked in accordance with their estimated abilities by the BTL model. The results of Mturk subjective ratings as well as ranking scales for four global properties are available online at https://github.com/jhoupt/SceneCategorization.

Figure 5 shows images with the highest, lowest, and approximately average scores for each scale. These images indicate that the derived scores have good face validity. The images with high "natural" scores include outdoor images

[2] The task was not originally deployed with catch trials, however it quickly became clear that some participants did not actively engage in the task.



Which scene is more natural?

Please choose the better one, there is no "right" answer

(Current trial:6/450)

Fig. 3 A trial from "natural" task question. Participants were asked to click on the image which they thought best answered the question, then clicked the button for the next trial. A *red box* appeared around an image once it was clicked

of natural environments such as mountains and plains and images while low "natural" scores are indoors scenes with no windows or plants pictured. Images with high "manmade" scores include city scopes and other clearly manmade structures, whereas images with low "manmade" scores depict natural scenes. Images with high "open" and low "closed" scores are expansive landscape images and ocean views while images with low "open" and high "closed" scores are small, indoor rooms.

Spearman correlation tests indicated that opposing scales were highly anti-correlated, as expected: "natural" and "manmade", $r = -.86$, $p < .01$; and "open" and "closed" $r = -.93$, $p < .01$. As can be observed in Fig. 5, less natural scenes images included more manmade elements. Likewise, less manmade scenes included more natural elements. Interestingly, "natural" and "open" scores were highly correlated, $r = .83$, $p < .01$, as were "manmade" and "closed" scores, $r = .77$, $p < .01$.

To quantitatively evaluate the how well the BTL model accounted for the data, we developed a simpler, baseline model and calculated the Akaike Information Criterion (AIC) for both baseline and the BTL. For the baseline model, we modeled the proportion of times each image was chosen in a task as a binomial random variable with the

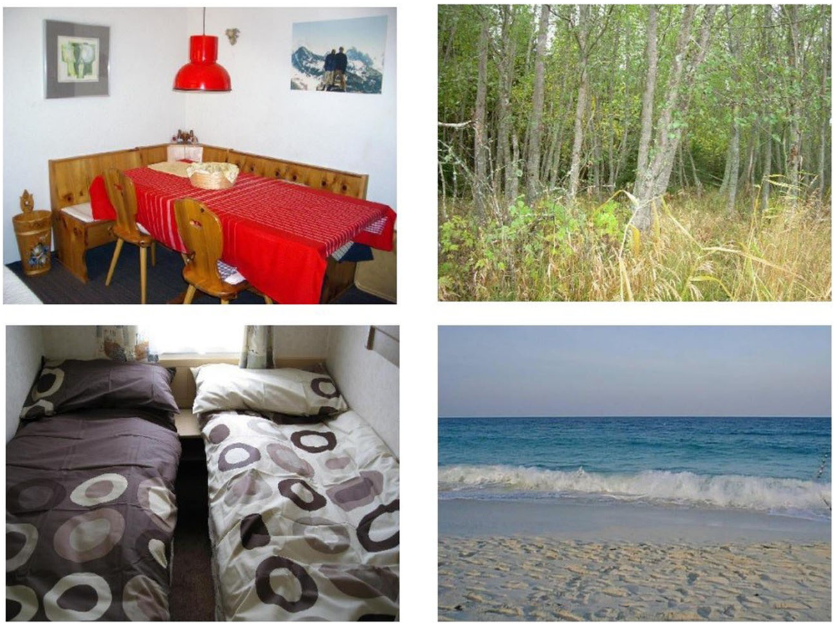


Fig. 4 Catch trials for inspection of whether participants followed the instruction or not. Different "dining room" and "broadleaf" images were chosen for "natural" and "manmade" question (*top*). Likewise, different "bedroom" and "beach" images were selected for "open" and "closed" test (*bottom*)

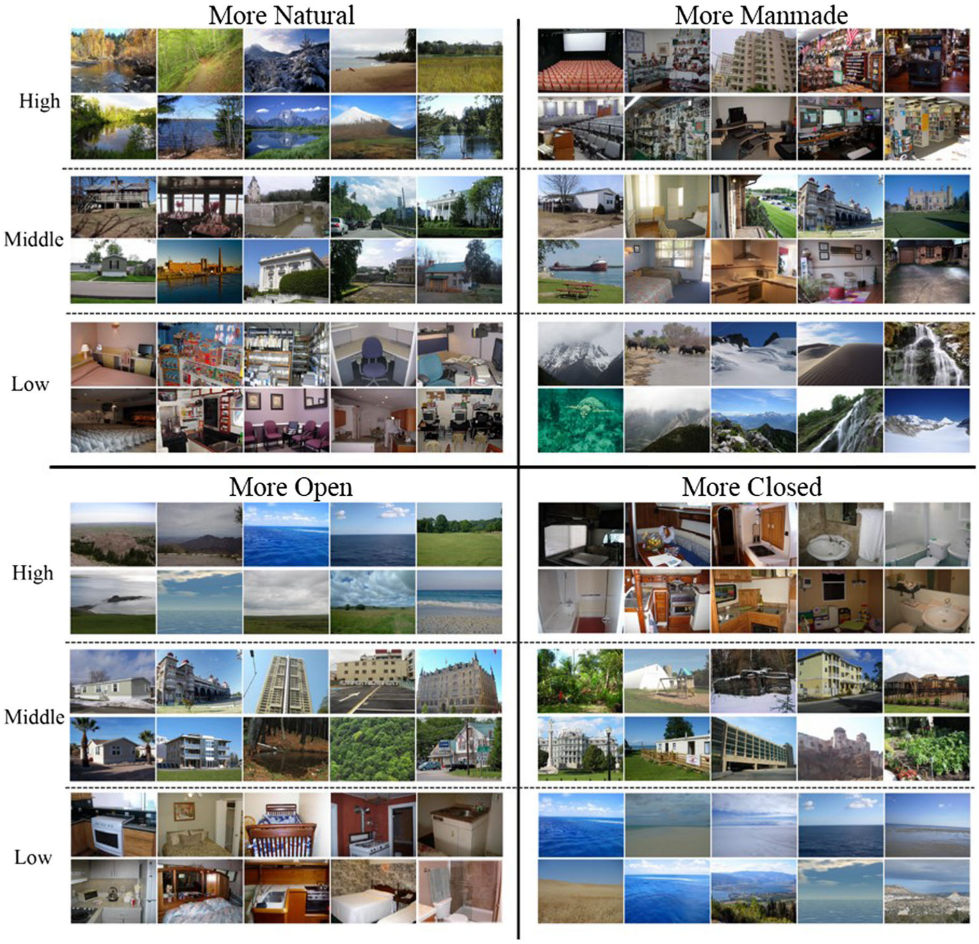


Fig. 5 Reference ranking scales based on fitting the Bradley–Terry–Luce model (Turner & Firth, 2012). From top to down in each scale, *different blocks* stand for dimensions of representative descriptions in "high", "middle", and "low" separately

probability parameter given by the proportion of trials on which that image was chosen. This approach effectively counted each comparison twice because the likelihood was included separately for each of the two images in the comparison. Hence, we then divided the summed likelihood across binomial likelihood of each image by two to calculate the AIC. This model had an advantage in that the binomial probability was the maximum likelihood value for each image, but a disadvantage that it did not account for the competing image on any trial. As such, the comparison gave a conservative indicator of the value of considering the alternative image when determining the likelihood that an image would be chosen. AIC values for the baseline and the BTL model are reported in Table 1.

Discussion

A key question in the field of scene perception is what information people use when making decisions about images of scenes (Malcolm, Groen, & Baker, 2016). Unfortunately, it is impractical to control real-world scene imagery, particularly global properties of the scene, in the same way that it is possible in artificial scenes. In the current work, we attempted to ameliorate this difficulty by creating a database of images ranked on degrees of naturalness and openness. Participants were not explicitly instructed on the definition of "openness" or "naturalness", which allowed us to examine the perception of global properties without imposing any given a-priori definition.

We also found that outdoor images were rated more open than indoor images, in line with a definition of

Table 1 The fitness of model comparison

	AIC			
Model	Natural	Manmade	Open	Closed
Baseline	148771	129684	142106	122419
BTL	*99370*	*89680*	*95217*	*76749*

Note. The better fit is in italics

openness as representing the magnitude of spatial enclosure (i.e., decreasing when boundary elements are increased, see Greene & Oliva, 2009b). Inherent in our approach was the assumption that these global properties of scenes should be treated as ratio scales (cf., Suppes & Zinnes, 1963 p. 50), although the ordinal ranking implied by the ratio scale was the main focus of the analysis. The rankings have four important advantages: (1) the evidence was collected from a large sample of subjective judgments that were presumably more diverse on the basis of demographic information retrieved online than a standard psychology subject pool; (2) images were drawn from a well-known, relatively large, scene image database; (3) a wide variety of basic-level categories were represented; (4) the images included all had a high enough resolution to be presented full screen on a normal-sized monitor.

Future directions

In many studies of global scene properties, researchers have used highly representative images of these properties, treating these essentially continuous properties as categorical. By establishing a metric based on human judgments, we have enabled stronger tests because researchers will be able to examine *metric* relationships between the global properties and both neural and behavioral data. For example, the recent work by Harel et al. (2016) indicating an effect of open versus closed global properties on P2 could be extended to examine whether the degree of openness modulates the P2 magnitude. Similarly, response times and strength of preference can be tested for a functional dependence on the score or rank of an image on a queried global property.

More specifically, recent research has indicated scene categorization may not follow a serial process from high-level to basic-level (or the reverse). Instead, the categorization process may be driven by attributes of perceptual similarities (Banno & Saiki, 2015) or discriminabilities (Sofer et al., 2015). One of the most powerful tools for discriminating serial and parallel cognitive processes is Systems Factorial Technology (Townsend & Nozawa, 1995; Houpt, Blaha, McIntire, Havig, & Townsend, 2013; Little, Altieri, Fifić, & Yang, 2017), however the approach requires that each dimension of interest be selectively influenced. In the current study, 174 basic-level scene categories were sorted by global properties based on large sample of subjective judgments, offering the possibility of manipulating basic-level category without changing the degree of a global property and manipulating the degree of the global property without changing the basic-level category. Hence, these scales provide the foundation for applying systems factorial technology to investigating serial versus parallel processing of scene categorization across levels.

Author Note Preliminary versions of this work were presented at the 2016 Meeting of the Psychonomic Society.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Banno, H., & Saiki, J. (2015). The processing speed of scene categorization at multiple levels of description: The superordinate advantage revisited. *Perception*, *44*, 269–288.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., & Cudré-Mauroux, P. (2015). The dynamics of micro-task crowdsourcing: The case of Amazon MTurk. In *Proceedings of the 24th international conference on world wide web*, (pp. 238–247).

Ehinger, K. A., Xiao, J., Torralba, A., & Oliva, A. (2011). Estimating scene typicality from human ratings and image features. In *Proceedings of the 33rd annual conference of the Cognitive Science Society*. Boston: Cognitive Science Society.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Greene, M. R., & Oliva, A. (2009a). The briefest of glances the time course of natural scene understanding. *Psychological Science*, *20*, 464–472.

Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–176.

Greene, M. R., & Oliva, A. (2010). High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1430–1442.

Groen, I. I., Ghebreab, S., Lamme, V. A., & Scholte, H. S. (2016). The time course of natural scene perception with reduced attention. *Journal of Neurophysiology*, *115*, 931–946.

Groen, I. I., Ghebreab, S., Prins, H., Lamme, V. A., & Scholte, H. S. (2013). From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, *33*, 18814–18824.

Harel, A., Groen, I. I., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *Eneuro*, *3*(5), ENEURO–0139. https://doi.org/10.1523/ENEURO.0139-16.2016

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cerebral Cortex*, *23*(4), 947–957.

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2013). Systems factorial technology with R. *Behavior Research Methods*, *46*, 307–330.

Ipeirotis, P. G. (2010). Analyzing the Amazon Mechanical Turk marketplace. XRDS: Crossroads. *The ACM Magazine for Students*, *17*(2), 16–21.

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience*, *31*, 7322–7333.

Little, D. R., Altieri, N., Fifić, M., & Yang, C.-T. (Eds.) (2017). *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms*. London: Elsevier.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*, 513–536.

Lowe, M. X., Gallivan, J. P., Ferber, S., & Cant, J. S. (2016). Feature diagnosticity and task context shape activity in human scene-selective cortex. *NeuroImage*, *125*, 681–692.

Malcolm, G. L., Groen, I. I., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, *20*, 843–856.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

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

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, *31*, 1333–1340.

R Development Core Team (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0).

Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision, 10*(1). https://doi.org/10.1167/10.1.2

Sofer, I., Crouzet, S. M., & Serre, T. (2015). Explaining the timing of natural scene understanding with a computational model of perceptual categorization. PLoS Computational Biology. Retrieved September 3, from http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004456

Stephenson, A., & Sonas, J. (2016). Playerratings: Dynamic updating methods for player ratings estimation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=PlayerRatings (R package version 1.0-1).

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In Luce, R. D., Bush, R. R., & Galanter, E. (Eds.) *Handbook of mathematical psychology* (Vol. 1, pp. 1–76).

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.

Torralbo, A., Chai, B., Caddigan, E., Walther, D., Beck, D., & Fei-Fei, L. (2009). Categorization of good and bad examples of natural scene categories. *Journal of Vision, 9*(8). https://doi.org/10.1167/9.8.940

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial and coactive theories. *Journal of Mathematical Psychology*, *39*, 321–360.

Turner, H., & Firth, D. (2012). Bradley–Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, *48*(9), 1–21. Retrieved from http://www.jstatsoft.org/v48/i09/

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3485–3492). San Francisco.