2017

Processing global properties in Scene Categorization

Hanshu Zhang
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the Industrial and Organizational Psychology Commons

Repository Citation
https://corescholar.libraries.wright.edu/etd_all/1738

This Thesis is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu, library-corescholar@wright.edu.
PROCESSING GLOBAL PROPERTIES IN SCENE CATEGORIZATION

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science

By

HANSHU ZHANG
B.S. Wuhan University 2010

2017
Wright State University
I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Hanshu Zhang ENTITLED Processing global properties in scene categorization BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Joseph Houpt, Ph.D.
Thesis Director

Scott Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Committee on Final Examination

Scott Watamaniuk, Ph.D.

Assaf Harel, Ph.D.

Robert E.W. Fyffe, Ph.D.
Vice President of Research and Dean of the Graduate School
Abstract

Zhang, Hanshu. M.S., Department of Psychology, Wright State University, 2017. Processing global properties in scene categorization.

The current research examined the role of global properties in human observers’ scene perception. In Experiment 1, comparisons of four global properties (“natural”, “manmade”, “open”, and “closed”) were collected online from a wide range of subjective choices. These answers were analyzed in a pairwise comparison model to generate four standardized reference ranking scales describing the extent to which characteristics can describe scene global properties. In Experiment 2, scene images selected from the reference scales were used to test human’s performance in processing global properties conjunctively. Cognitive modeling indicated that human observers were more efficient in categorizing scene images as “natural and open” but less efficient in classifying scene images as “manmade or closed” than the predicted baseline.
Contents

1 Introduction 1

1.1 Scene Perception Models ................................. 1

1.1.1 Basic-level categorization .............................. 4

1.1.2 Object Interference ................................. 5

1.1.3 Brain Region Activations ............................. 6

1.2 Current Research ............................................. 7

2 Experiment 1: Ranking Scales Task 9

2.1 Methods ...................................................... 10

2.1.1 Scene Images Database ............................... 10

2.1.2 Amazon Mechanic Turk ............................. 12

2.1.3 Procedure ................................................. 12

2.2 Data Analysis .............................................. 15

2.3 Results ...................................................... 16

2.4 Discussion .................................................. 18
List of Figures

1. Examples from SUN database ........................................ 11
2. Excluded examples ..................................................... 12
3. Amazon Mechanic Turk Task Trial ................................. 13
4. Amazon Mechanic Turk Catch Trials .............................. 14
5. Reference Ranking Scales ............................................. 17
6. Simulated Capacity Function ........................................ 21
7. Experiment Stimuli ..................................................... 23
8. Experiment Procedure ................................................ 24
9. Excluded Subject Example .......................................... 26
10. Average Group Choice Probability .............................. 27
11. Average Group Response Time .................................... 29
12. “AND” Capacity Analysis ........................................... 30
13. “OR” Capacity Analysis ............................................. 31
14. Image of “cabin outdoor” as “natural and open” ................. 33
List of Tables

1  Average Group Choice Probability and Response Times . . . . . . . . 32
Acknowledgment

I would like to thank Dr. Joseph Houpt for his tremendous help not only on the current study but for his advising during my graduate study. I also appreciate the help and suggestions from my committee members, Dr. Assaf Harel and Dr. Scott Watamaniuk, and my lab-mates on the current study. I received a lot of warm encouragement from friends and family during my research time. I would not have been able to finish my master degree without their support.
1 Introduction

Imagine that you wake up in your bedroom, and then go to the bathroom to take a shower, and before leaving for work, you will need to get breakfast in the kitchen. And then your next destination would probably be office, classroom or laboratory. Picturing your bedroom, kitchen and office, each of these scenes can correspond to a basic-level scene category. Alternatively, these basic-level scene categories can also be expressed by high-level scene categories that are understandable among human observers. For example, the “kitchen, bedroom and office” listed above are “indoor, manmade, and closed” scenes while it is also easy to come up with some “outdoor, natural, and open” scenes such as “mountain and beach”.

The current study focuses on global features (e.g. “natural”, “open”) instead of the basic-level scene categories (e.g. “forest”, “office”). In particular, my goal was to address the two following questions: Are there reference ranking scales to represent different global features? How do different global features work together in understanding and processing of scene images?

1.1 Scene Perception Models

Henderson and Hollingworth (1999) defined scene as a semantically coherent (and often nameable) view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner. Oliva and Torralba (2001) thought scene as a place in which we can move to differentiate it from “object” and “texture”. In the current research, my discussion of the scene is based on
the idea to take scene as a environment representing spatial relations between objects and backgrounds that constitute a functional meaning. Existing research has shown that people can accurately capture the gist of a scene (a city, a mountain, etc.) within a brief glance. For example, human observers could perform go/no-go categorizations of “sea, mountain, indoor and urban” when scene images were flashed for only 26ms (Rousselet, Joubert, & Fabre-Thorpe, 2005). As an early scene perception model, Biederman (1981) discussed three potential top-down routes to achieve schema from a single glance, given that the scene was not exhaustively processed through all bottom-up levels such as physical parsing and object identification. The first route was through an initial identification of one (or more) discriminable objects like a “pop-out” effect in the scene. The second route was via features that “emerged” as objects that were brought into relation to each other to form a scene. The third route came from spatial integration through semantic information. Biederman’s proposal was one of the scene perceptual schema (or frame) models that focused on object-scene context understanding (Henderson & Hollingworth, 1999, for a review). The models argued that the knowledge of scene categories can be reached through the contextual information such as interactions of a scene type and objects commonly found in the scene. For example, “refrigerator” and “blender” can be associated with “kitchen”.

Another early scene understanding model concentrated on advantages of basic-level categories. B. Tversky and Hemenway (1983) presented evidence for a preference of basic-level in scene categories (e.g. school, home, beach, mountain). The evidence came from subjects’ judgments of the attributes and activities characteristics of the scenes, and labels used for scenes captured in photographs and referred to in sentences. They discussed a hierarchical structure of taxonomy representation in environmental scenes, and argued that more general categories were impoverished for general use while more specific categories were too discriminative to be widely applied. Therefore, the advantages of basic-level scene category in cognition, behavior, and
communication were because of their information-rich bundles of features, primarily parts.

More discussions between basic-level and high-level categorization in scene perception would be discussed in the later section. However, what is meant by basic-level in psychological perceived structure is obscure. A pioneering work done by Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) argued that the basic level of abstraction in a taxonomy was the level at which categories carried the most information, possessed the highest cue validity, and were thus the most differentiated from one another. While some evidence they provided was against by the later research (e.g. Mandler & Bauer, 1988), the general ideal is that perceptual attributes in basic-level concepts showed informativeness and distinctiveness (Murphy, 1991; Murphy & Smith, 1982). In the context of scene perception categorization studies, the distinction between basic-level and high-level is simpler, as the example given in the beginning of the introduction. The focus of current research was not the comparisons between the different levels of scene categorizations nor human expertises performance (e.g. Tanaka & Taylor, 1991).

An alternative scene perception model is a computational framework proposed by Oliva and Torralba (2001). Their model provides a holistic description of the scene where local object information is not considered. The spatial-envelope model descriptions operate over projections of space onto two-dimensional views and capture information about both the layout and texture of surfaces. Instead of describing “forest” as “an environment with trees, bushes, and levels”, the spatial-envelope model would illustrate it as a set of perceptual properties (naturalness, openness, roughness, ruggedness, and expansion): “Forest is an enclosed natural environment with a dense, isotropic texture.” (Oliva, Park, & Konkle, 2011, for a review). Oliva and Torralba (2001) showed that these global properties were highly correlated with
spectral information and spatial arrangement of the scenes, and subjects recruited in the study could also categorize scenes based on these properties. Therefore, the spatial-envelope model offered a meaningful representation of scenes integrating both visual perception and semantic knowledge.

The development of spatial-envelope descriptions presented a new perspective on how global properties consisting of low-level features can provide top-down information in early visual processing (Torralba & Oliva, 2003). This mechanism may account for the phenomenon that people can identify a place as “forest” without having to recognize the “trees” (Greene & Oliva, 2009b). Greene and Oliva (2009b) proposed that rapid scene categorization did not require segmenting a scene into objects, instead used a vocabulary of global and ecological properties that describe spatial and functional aspects of scene space. They selected seven global properties from the previous research (see footnote in Greene & Oliva, 2009b) to reflect the variations of categories while including different levels of scene description. The seven global properties (i.e. openness, expansion, mean depth, temperature, transience, concealment, navigability) offered a guideline for many later discussions in scene processing.

While both global properties and superordinate-levels define high-level categories, there is an important distinction. Superordinate-level categorization sometimes included indoor-outdoor discrimination (e.g., Banno & Saiki, 2015), however indoor-outdoor is not a global property as defined by Greene and Oliva (2009b).

1.1.1 Basic-level categorization

Behavioral studies have shown that people can categorize global properties with high performance. Greene and Oliva (2009a) tested subjects’ categorization performance with different global properties and basic-level categories. They reported that certain global visual information was more easily gleaned from an image than its basic-level
category. Loschky and Larson (2010) presented further data indicating that superordinate-level
categorization required less processing time. They proposed that the “natural/manmade”
distinction was made before basic-level distinctions in scene gist categorization.

Two recent studies concerned about the variances in basic-level scene categories
embedded in the global properties and revisited the advantage of superordinate level
categorization for higher accuracy and faster response times. Banno and Saiki (2015)
controlled the similarities between basic-level categorizations. They suggested that
visual processing in “natural/manmade” categorization was robust and showed strong
primacy. On the contrary, the advantage of indoor/outdoor categorization could
be reversed by using dissimilar basic-level categories of outdoor scenes such as tall
buildings and suburbs. Sofer, Crouzet, and Serre (2015) employed a similar conception
of “similar/dissimilar”, they trained machine-learning classifiers to derive quantitative
measures of task-specific perceptual discriminability based on the distance between
individual images and different categorization boundaries. In that experiment, they
showed it was possible to control the perceptual discriminability to reverse the advantage
of “natural/manmade” scene categorization rendering subjects’ superordinate categorization
arbitrarily slower and less accurate than basic-level categorization.

1.1.2 Object Interference

Davenport and Potter (2004) argued that objects and their background were
processed interactively. Therefore, though the spatial-envelope model did not consider
object information, human observers’ global property processing may still be affected
by contextual processing. When observers were required to answer if the scene was
“natural” or “manmade”, the efficiency of contextual categorization was impaired
by the presence of a salient object in the scene, especially when the object was
incongruent with the context (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007).
Joubert et al. (2007) argued that a manmade object on a natural background or a large biological object (animal, tree, etc.) in a manmade scene were cases where salient objects were incongruent with the context. But considering the discussion in the last section about basic-level categorization performance, the evidence Joubert et al. (2007) presented was merely another aspect showing variances described in scene images that might decrease the validity of global properties categorization advantages.

1.1.3 Brain Region Activations

The structural information of global properties has a robust representation in the scene responsive brain area. It is well known that the parahippocampal place area (PPA) responds preferentially to pictures of scenes, landmarks, and spatial layouts. There are other regions that also respond to scene beyond the PPA. For example, lateral occipital complex (LOC) represents object shape and category. Given that scenes contain objects, LOC is considered to represent the object contents and object interactions in a scene (Park & Chun, 2014, for a review).

Park, Brady, Greene, and Oliva (2011) tested brain region activities using functional magnetic resonance imaging (fMRI) analysis. They found that scene and object areas in the brain represent spatial boundaries (open, closed) and the degree of naturalness (urban, natural) of a scene. Subjects were presented with blocks of scene images combining different spatial boundary and contents. Park et al. (2011) trained linear support machine vector (SVM) to classify each block of scenes shown by the region activities into categories of “closed natural, closed urban, open urban, and open natural”. They found that both the PPA and LOC could be used to accurately classify scenes though they led to different errors: the PPA resulted in more confusion in scenes that have the same spatial boundaries, whereas the LOC led to more confusion in scenes that have the same content, suggesting the consistent evidence of spatial
representations in brain activities.

Another data-driven study reported similar results. Kravitz, Peng, and Baker (2011) indicated that in both PPA and early visual cortex (EVC) regions, spatial rather than semantic factors defined the structure of representations. Early visual cortex commonly refers to visual areas V1, V2, and V3 that encode scene images low-level features such as space, orientation and spatial frequency (Kay, Naselaris, Prenger, & Gallant, 2008, for related discussion). In PPA, descriptions were defined primarily by the expanse (open, closed) and in EVC primarily by distance (near, far). The two studies together suggest that scene representations in PPA are mainly based on spatial layout information but not scene category per se. The spatial information discussion in neuroscience motivated the current study to include openness (open vs. closed) in the discussion of conjunctive global properties processing.

1.2 Current Research

Though Kadar and Ben-Shahar (2012) proposed a hierarchy paradigm in scene gist processing, there was a lack of direct evidence showing how human observers processed different global properties conjunctively. Kadar and Ben-Shahar (2012) argued that the decision of manmade or natural was made first, and only then was followed by more complex decisions. They found that manmade scenes seem to divide between indoor and outdoor scenes, while natural scenes appear to divide between open and closed scenes. However, Sofer et al. (2015) pointed out observed differences in timing across categorization task did not necessarily reflect the fact that some categorization tasks take precedence over others, attenuating the serial processing in scene categorizations assumed by some previous research.

In order to discuss subjects’ performance in processing different global properties conjunctively, the current research explored two global properties of naturalness and
openness. Naturalness represents information of scene contents. And naturalness was the global property that has been discussed mostly in contextual congruence scene processing and superordinate/basic-level categorizations. Openness was selected from the research of scene responsive brain areas in which openness indicates the spatial boundaries information.

In sum, the current research is dedicated to two research goals. Experiment 1 intended to build two subjective reference ranking scales, “natural – manmade” and “open – closed”, of scene images that can describe characteristics representing the dimension changes of global properties. The Experiment 2 compared human observers’ performance in judging different global properties of scene images from ranking scales conjunctively and separately, using capacity coefficient analysis from a cognitive mathematical modeling — Systems Factorial Technology (Townsend & Nozawa, 1995, hereafter SFT), to test processing efficiency of multiple global properties. The hypothesis of Experiment 2 is that visual features of global properties could be extracted by human observers efficiently and accurately.
2 Experiment 1: Ranking Scales Task

Scene images as stimuli played an important role in scene related research. However, there were two limitations of the scene images selected in previous studies. First, a lot of the investigations that discussed human observers' performance of global properties in naturalness tended to used scene images based on basic-level categories. Though Greene and Oliva (2009b) emphasized that the basic-level category label was not the determinant of the global property ranking for any single global property, some studies such as Loschky and Larson (2008, 2010) still used “coast, mountain, forest, and open country” to represent natural scenes and “city center, street, tall building and highway” for manmade scenes. The choices that infer global properties based on basic-level categories might be due to their investigation in comparing human observers’ categorization performance in different levels. Nevertheless, some basic-level categories such as highway, parking lot, and playground were intuitively not easy to infer their naturalness degree if they contain both natural and manmade elements.

Second, most previous studies had a pre-selection for “good representatives” of the global properties they would like to test. This filtering procedure may lead to high accuracy and fast performance in scene categorization task (Ehinger, Xiao, Torralba, & Oliva, 2011; Torralbo et al., 2009). The discussion of perceptual discriminability (Sofer et al., 2015), similar/dissimilar in basic-level categories (Banno & Saiki, 2015), and salient object inference (Joubert et al., 2007) all questioned the generalization of processing advantages of global properties. This set a requirement for the significant
amount of scene images including descriptive variances in the current study.

Greene and Oliva (2009b, 2010) showed that observers could provide normative rankings on global properties with a high degree of consistency. However, their rankings included a relatively small number of images and did not take the indoor scenes into consideration. They defined the non-target description of naturalness as “the scene is a manmade urban environment”. Thus, there is a need for scene images reference scales describing transitions of identifying characteristics within different global categories to a larger extent. Another notable feature is that most previous studies used scene images in 256 × 256 pixels. Previous research has talked that both object size (Fize, Cauchoix, & Fabre-Thorpe, 2011) and gaze allocation (Groen, Silson, & Baker, 2017) could influence scene processing, thus a high-resolution image database is essential in the current study to indicate spatial layout and object content in real-world visual scenes.

2.1 Methods

2.1.1 Scene Images Database

Images were from the Scene Understanding (SUN) database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). The SUN database consists of more than 100,000 images classified into 397 basic-level categories (Figure 1). Scene images were selected to be 1024 × 768 pixels. Since scene was expected to include spatial features (e.g. “a scene is mainly characterized as a place in which we can move”, Oliva & Torralba, 2001), images were excluded base on the author’s personal judgment whether they were representing good spatial relationships or not (e.g. Figure 2) – for example, subjects may more likely take the description as single object “airplane” rather than basic-level scene category “runway”. This collection included 7035 images describing 174 basic-level categories for building the ranking scales.
Figure 1. Different Abbey images (top) and different scene categories (bottom) in SUN database (Xiao et al., 2010).
Figure 2. The example of images excluded from the stimuli: clothes store (left) and runway (right).

2.1.2 Amazon Mechanic Turk

Human observers 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, Chandler, & Ipeirotis, 2010). According to an online Mturk demographic report (HTTP://demographics.mturk-tracker.com/#/gender/all): in the month of September, 2016, approximately 75% of workers on Mturk were from U.S.; the gender of female and male were equally distributed across countries, with more male workers in U.S.; approximately 40% Mturk workers were born between 1980 to 1990, another 20% of workers were born between 1990 to 2000, and 20% of workers were born between 1970 to 1980; the median household income for U.S. based workers was around $50K per year.

2.1.3 Procedure

Subjects 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 eligibilities of subjects: located in the U.S.; numbers of HITs approved were greater than 50; the HIT approval rate for all requesters were more than 95%.
Figure 3. A trial in the Mturk task of “natural” question. Subject needed to click on the image which they thought answered question better, and then clicked the button for next trial.

This ensured all the subjects had no difficulty in understanding English and were familiar with Mturk.

Subjects needed to complete a consent form before the start. There were four images in the consent form showing examples of global properties in “open”, “closed”, “natural”, and “manmade”. There was no restriction for subjects to do the task only once, but each time they participated, they were required to complete the consent form. The compensation was $1.75 after subjects submitted their answers.

Mturk task explored four global properties (i.e. “natural/manmade/open/closed”), and each subject was assigned only one tested global property per task. Each task included 450 trials during which only one global property was queried. The task took about 15-25 minutes to finish.

In each trial, subjects were instructed to choose from a random image pair which the image answered the question in the task (e.g. “which one is more natural”) by clicking the image and then clicking “next trial” for the next pair (Figure 3). Subjects were asked to choose the image that answered the question better. The
task came with a fixed first trial as a hint of where to start. In the second version posted, “catch trials” were added for every 30 trials as an inspection to check if subjects paid attention to the task. In “open/closed” questions, catch trials were different images in “beach/bedroom”. Subjects were supposed to select “beach” for “open” and “bedroom” for “closed”. “Broadleaf/dining room” images were chosen for “natural/manmade” catch trials (Figure 4), “broadleaf” was considered as a better answer in “natural” while “dining room” fitted the “manmade” description more. There were 15 catch trials in total for each question type. Subjects who failed more than 5 trials were excluded from the final data analysis. The results of catch trials showed that the overall performance of subjects were consistent in following the instruction. For example, for 278 subjects who participated in the “manmade” task, only 3 subjects’ selections in catch trials were below the chance. Finally, 1055 subjects (“natural”: 279; “manmade”: 261; “open”: 274; “closed”: 241) in total were
included in the data analysis.

2.2 Data Analysis

Kravitz et al. (2011) used Elo rating for building ranking scales. The Elo rating system was developed for evaluating skills of chess players. Following this idea, Elo and other two rating systems that extended Elo – Glicko and Steph – were used to analyze the data (Stephenson & Sonas, 2016). However, the assumptions made by chess rating systems like all players start with initial performance scores were not necessarily applicable to scene ranking. Kravitz et al. (2011) also pointed out that the orders of comparisons impacted the results of Elo. My second attempt employed support vector machines (SVM). SVM is a powerful tool in machine learning for categorization that can be adapted for ranking (Joachims, 2002). But since data collected on Mturk consisted of more than 100,000 pairings per ranking task, the matrix needs to be hold in memory for representation is about 7 GB, it was not reasonable to complete using personal computer. Thus I turned to seek more efficient methods.

The data analysis finally used the Bradley-Terry model to estimate ranking using R package BradleyTerry2 (Turner & Firth, 2012). The Bradley-Terry model is a logistic model for paired evaluations (cf. Agresti, 2002). Let $\Pi_{ab}$ stands for the probability that $a$ is preferred to $b$, the Bradley-Terry model has an “ability” parameter $\{\lambda_i\}$ for each player (in this case, scene) such that

$$\Pi_{ab} = \exp(\lambda_a)/[\exp(\lambda_a) + \exp(\lambda_b)]$$

(1)

Thus, the probability that $a$ is preferred to $b$ is $\frac{1}{2}$ when ability $\lambda_a = \lambda_b$ and exceeds $\frac{1}{2}$ when ability $\lambda_a > \lambda_b$. The abilities $\{\lambda_i\}$ can be estimated from linear model fits by
using the penalized quasi-likelihood algorithm of Breslow and Clayton (1993). The ability of each player $i$ is related to explanatory variables $x_{i1}, \ldots, x_{ip}$ through a linear predictor with coefficients $\beta_1, \ldots, \beta_p$; the $U_i$ are independent errors. In the current study, the estimated abilities $\{\lambda_i\}$ were used to target global properties rankings.

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_i + U_i$$  \hspace{1cm} (2)

### 2.3 Results

The scene images were ranked in accordance with their estimated abilities by the Bradley-Terry model. The Spearman’s rank correlation coefficient among four different scales showed that the Bradley-Terry model results were consistent between natural and manmade ($r = -0.86, p < 0.01$), and open and closed ($r = -0.93, p < 0.01$). As Figure 5 indicated, when natural scene images were changing from high to low dimension in describing representative characteristics, scene images included more manmade elements. The low dimension in describing representative characteristics of natural scenes was close to high dimension in describing representative characteristics of manmade scenes, revealing the high negative correlations. Interestingly, natural and open were highly correlated ($r = 0.83, p < 0.01$), so were manmade and closed ($r = 0.77, p < 0.01$).
Figure 5. Reference ranking scales generated by Bradley-Terry 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.
2.4 Discussion

Previous research (Greene & Oliva, 2009b, 2010; Kravitz et al., 2011; Ross & Oliva, 2009) explored the subjective rankings in describing global properties but with different restrictions (e.g. dichotomous categorization in “natural – urban”; a limited number of scene images included). The standardized ranking scales built in Experiment 1 consisted of a relatively large number of scene images and a wide range of subjective opinions about how people perceived the scene images. The ratings by human observers may include perceptual attributes that were not considered in previous computational categorization models, providing potential values for future studies comparing human observers and computational model categorization performance. Moreover, the ranking scales covering 174 basic-level categories offer diverse scene images for future exploration of perceptual similarities/dissimilarities in different levels of scene categorization.

This pairwise comparison method in generating rankings was consistent with the previous finding that there was significant covariation between properties (Greene & Oliva, 2009b). Experiment 2 took advantage of individual scene image’s different rankings on global properties scales to examine how human observers process global properties conjunctively.
3 Experiment 2: Capacity Coefficient Task

Existing research has discussed parallel mechanisms in scene understanding from various perspectives. Rousselet, Fabre-Thorpe, and Thorpe (2002) reported that human observers responded to two simultaneously presented scenes as fast as to a single one in a rapid animal versus non-animal categorization task. Rousselet, Thorpe, and Fabre-Thorpe (2004) revisited this parallel processing model in a later study but found that human observers’ performance decreased with increasing number of scenes. It was also believed that people process global features and local features of the scene in parallel as global pathway and local pathway. Although computational model assumed different global properties are processed in parallel (Oliva & Torralba, 2001), there was a lack of exploration in human observers’ performance. Most research comparing human observers’ performance between global properties and basic-level categorization focus on only one single global property, typically naturalness (i.e. manmade and natural). Since each scene image comes with different global properties, it is critical to discuss whether human observers process multiple global properties independently or not rather than simplifying global properties to naturalness instead.

3.1 Methods

3.1.1 Capacity Coefficient

SFT is a powerful tool in exploring the combination of different sources in information processing. Capacity refers to the information throughput characteristics of the system, addressing the question of how much work can be completed in a given
amount of time (Houpt, Blaha, McIntire, Havig, & Townsend, 2014).

The capacity coefficient tests how the processing rate of each source changes as more sources are added, based on an unlimited capacity independent parallel (UCIP) system. There are two different capacity test types: “AND” and “OR”. “OR” capacity tests the first terminating process: it is a ratio of actual performance when all sources of information are presented compared to predicted performance baseline using a cumulative hazard function \( H \). In an “OR” capacity scene task, the capacity coefficient is calculated by the subject’s performance in judging the scene “manmade or closed” versus “manmade”/“closed” separately.

\[
C_{OR}(t) = \frac{H_{manmade-closed}(t)}{H_{manmade}(t) + H_{closed}(t)}
\]

In an “AND” capacity test, subjects are not able to answer the question until they finish processing information from both resources. Like the “OR” capacity test, it is a ratio of actual performance to predicted baseline but using a cumulative reverse hazard function \( K \). In the “AND” capacity scene task, the capacity coefficient is calculated by the subject performance in answering the scene “natural”/“open” separately versus the single condition “natural and open”.

\[
C_{AND}(t) = \frac{K_{natural}(t) + K_{open}(t)}{K_{natural-open}(t)}
\]

Different explanations could be applied to the capacity coefficient result (Figure 6). \( C(t) > 1 \) implies super workload capacity, which means performance is better than predicted by the UCIP model. When \( C(t) < 1 \), it means that at least one hypothesis in capacity, independent or parallel, is violated. Subject’s performance may decrease because of either limited capacity in processing the additional source of
Figure 6. The capacity coefficient results indicate super capacity, unlimited capacity or limited capacity when $C(t) > 1$ (green line), $C(t) = 1$ (blue line), or $C(t) < 1$ (red line).

information, information processing of different sources is dependent on each other, or the subject processes information in serial. In the current study, the capacity test aimed at exploring whether people could process multiple global properties without reduced performance or not.

3.1.2 Subjects

Seventeen students ($M_{age} = 19.65$, $Female = 15$) recruited from Psychology undergraduate classes at Wright State University participated the experiment for class credit.
3.1.3 Stimuli

Based on the scales created in Experiment 1 by the Bradley-Terry Model, different image types in the combination of properties (i.e., “natural-open”, “natural-closed”, “manmade-open”, and “manmade-closed”) were selected. Because of the high correlations in rankings between natural and open, manmade and closed, there were more images of “natural-open”/“manmade-closed” than “manmade-open”/“natural-closed” in a certain range of the scales. The first step was to get the boundaries of rankings in “open” of “manmade-open” and “closed” of “natural-closed” to make sure there were sufficient images for stimuli. The second step selected “natural-open” and “manmade-closed” scene images matching distribution ranges in the ranking scales for images of “manmade-open” and “natural-closed”. Therefore, though more scenes were described as “manmade-closed” and “natural-open” in the ranking database, four image types had the same range of representations in global property rankings scales (Figure 7). All images were 1024 × 768 pixels. Stimuli were displayed in the center of a 20” monitor with a resolution of 1280 × 1080 pixels and a refresh rate of 85 Hz. Images subtended 19.85° × 15.43° of visual angle.

3.1.4 Procedure

Each trial started with a fixation cross varying from 450ms to 500ms, followed by a scene image for 350ms. The screen then turned gray until the end of the trial. The response period was 2000ms for each trial (Figure 8). There were three different blocks: “natural/man-made,” “open/closed”, “natural and open/manmade or closed”. Subjects were instructed to respond following the instruction accurately and quickly as possible. The order of the first two blocks (“natural/manmade” and “open/closed”) was shuffled between subjects. In “natural/manmade” block, subjects were asked to press the left button if the scene was “natural”, press the right button
if the scene was “manmade”. Likewise, in the “open/closed” block, subjects needed to press the left button for “open” images and the right button for “closed” images. These blocks consisted of 192 trials each (96 images repeated twice) and took about 8 minutes. In the third “natural and open/manmade or closed” block, subjects needed to press the left button if the scene was “natural and open”, and press the right button if the scene was “manmade or closed”. The third block consisted of 432 trials. To balance the subjects’ responses in choosing the left and right buttons, 24 “natural and open” images repeated nine times, all the other 72 images repeated three times. The third block took about 18 minutes. Subjects were forced to take a break half way through the block. The entire experiment took about 36 minutes.

*Figure 7. Scene images as stimuli in the capacity coefficient experiment.*
Figure 8. Procedure of each trial in the capacity coefficient task. Each trial started with a fixation cross varying from 450ms to 500ms, followed by a scene image for 350ms. The screen then turned gray and waited for the subject’s response. The respond period was 2000ms for each trial.
3.2 Results

A proportion test checked if subjects’ choices in the “natural/manmade” and “open/closed” blocks matched experimental expectations. Subjects were not wrong if they were inconsistent with selected criteria judging global properties. However, they were expected to choose the statistically different proportion of “natural-open” and “natural-closed” images compared to the proportion of “manmade-open” and “manmade-closed” images in the “natural/manmade” blocks. In the “open/closed” blocks, they should indicate that their selections in “natural-open” and “manmade-open” images were statistically different from “natural-closed” and “manmade-closed” images. Eight subjects who failed the proportion test, showing their data were uninterpretable, were excluded from the data analysis. The example given in Figure 9 shows that in a “natural” block, the subject’s proportion in selecting “natural-open” and “natural-closed” was statistically different from “manmade-open” and “manmade-closed” ($\chi^2 = 104.64, p < 0.01$). However, the proportion selected of “open” block in “natural-open” and “manmade-open” was not statistically discriminable from those selected in “natural-closed” and “manmade-closed” ($\chi^2 = 3.55, p = 0.97$). Finally, the data analysis included nine subjects in total.

3.2.1 Response Choices and Response Times

Subjects’ choice preferences were represented by the proportion of each image type that was chosen in the block. Figure 10 shows the average group level of choice probabilities. Each subject’s choice probability is given in the Appendix. The filled dots stand for the image types that matched selected criteria in the task. For example, in the “natural” block, subjects were expected to choose the “natural-closed” and “natural-open” for “natural” and the other image types for “manmade”.

Bayesian paired $t$-test tested that if subjects were independent in choosing “natural
Figure 9. This subject was excluded from the capacity test data analysis. See text for details.
Figure 10. The average percentage of each image type chosen in different blocks. Black dots mean the subjects’ responses matched the selection criteria. For example, in the “natural” block, subjects were expected to choose “natural-closed” and “natural-open” images.
and open”, and “manmade or closed” based on joint probabilities (Equation 4). The estimated Bayes factor suggested that there was minimal evidence supporting dependent choices between “natural and open” (BF: 1.47), or rather, the model of dependence choices is 1.47 times to occur than a model of independence\(^1\). So were “manmade or closed” selections (BF: 1.70).

\[ P(A \text{ and } B) = P_A \times P_B \quad (5) \]

For response times (Figure 11), ANOVA tests reported that there were statistically significant differences but small effect size between image types, \(F(3, 24) = 3.48, p = .03, \eta^2_G = .08\), and interaction of image types and blocks, \(F(6, 48) = 2.78, p = .02, \eta^2_G = .01\). Bayesian ANOVA test indicated strong evidence in supporting main effect of image types (BF: 16.05). However, compared to the baseline model with only subject factor, there was strong evidence against the explanation for the main effect of block (BF: .0053), main effects of block and image type (BF: .11), and main effects of block and image type with their interaction (BF: .03). Taken together, image type influenced the response times while no clear evidence for the role of block and interaction of image type and block.

### 3.2.2 Capacity Coefficient Analysis

All subjects indicated super capacity in “natural and open” AND capacity test \((M = 8.20, SD = 2.29)\). The group level (Figure 12) was super capacity in \(^t\)-test (Houpt & Townsend, 2012), \(t(8) = 10.73, p < .01\). All subjects were limited capacity in “manmade or closed” OR capacity test \((M = -5.91, SD = 1.77)\), and the group level (Figure 13) was limited capacity, \(t(8) = -10.04, p < .01\).

\(^1\)The common interpretation of Bayes factors criteria can be found at Kass and Raftery (1995).
Figure 11. Average response times of each image type across different conditions.
Figure 12. “Natural and open” capacity results. All subjects ($M = 8.20, SD = 2.29$) and the group level ($t(8) = 10.73, p < 0.01$) were super capacity.
Figure 13. “Manmade or closed” capacity results. All subjects ($M = -5.91, SD = 1.77$) and the group level ($t(8) = -10.04, p < 0.01$) were limited capacity.
Table 1

Average Group Choice Probability and Response Times

<table>
<thead>
<tr>
<th>Task</th>
<th>Image Type</th>
<th>Choice Probability (SD)</th>
<th>Response Times (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>N-C</td>
<td>0.56(0.09)</td>
<td>608.82(284.27)</td>
</tr>
<tr>
<td></td>
<td>N-O</td>
<td>0.60(0.14)</td>
<td>645.09(271.16)</td>
</tr>
<tr>
<td></td>
<td>M-O</td>
<td>0.09(0.13)</td>
<td>587.80(326.81)</td>
</tr>
<tr>
<td></td>
<td>M-C</td>
<td>0.07(0.11)</td>
<td>566.46(218.70)</td>
</tr>
<tr>
<td>Open</td>
<td>N-C</td>
<td>0.77(0.18)</td>
<td>645.50(225.55)</td>
</tr>
<tr>
<td></td>
<td>N-O</td>
<td>0.89(0.07)</td>
<td>583.67(232.70)</td>
</tr>
<tr>
<td></td>
<td>M-O</td>
<td>0.70(0.17)</td>
<td>650.00(313.70)</td>
</tr>
<tr>
<td></td>
<td>M-C</td>
<td>0.23(0.14)</td>
<td>575.18(287.73)</td>
</tr>
<tr>
<td>Natural and Open</td>
<td>N-C</td>
<td>0.59(0.21)</td>
<td>625.59(321.17)</td>
</tr>
<tr>
<td></td>
<td>N-O</td>
<td>0.63(0.14)</td>
<td>619.93(271.81)</td>
</tr>
<tr>
<td></td>
<td>M-O</td>
<td>0.31(0.35)</td>
<td>592.45(268.99)</td>
</tr>
<tr>
<td></td>
<td>M-C</td>
<td>0.21(0.27)</td>
<td>591.20(256.88)</td>
</tr>
</tbody>
</table>

*Note.* Answers matched the selected criteria are in boldface. N-C = Natural-Closed; N-O = Natural-Open; M-O = Manmade-Open; M-C = Manmade-Closed.

### 3.3 Discussion

In the capacity coefficient analysis, subjects answered to “natural and open” faster than the baseline predicted from their responses to “natural” and “open” separately. However, when comparing “manmade or closed” choices to baseline, subjects were limited capacity and responded slower than predicted baseline. The global properties of “natural” and “open” in the ranking scales are highly correlated. There is a possibility that subjects were inclined to take “natural” images as “open”, facilitating their answers that the scene was “natural and open”. On the other side, subjects were slower when deciding the conjunctive criteria of “manmade or closed”.

As outlined in Figure 10, subjects on average were relatively consistent with ranking scales in discriminating “natural” and “manmade”. However, subjects were
Due the fact the that subjects may have different criteria, they might not agree that this image depicts a “natural-closed” scene. This caused the discrepancy between subjects’ response choices and selected criteria.

more likely to categorize “natural” image as “open” when they were required to choose between “open” and “closed”. Therefore, there was a high proportion of “natural-closed” scenes categorized as “open” images. Similarly, “natural-closed” images were frequently judged to be “natural and open”.

There are no absolute correct answers in scene categorization. Figure 10 could indicate subject’s response preferences better than the common used “accuracy”. One potential reason for the discrepancy between the subjects’ choices and ranking selections was that images used in Experiment 2 were mostly in the middle ranges of ranking scales. The selection made subjects’ judgment in the definition of “natural” and “open” more obscure compared to the scenes depicting highly typical characteristics. For example, a scene image “cabin outdoor” (Figure 14) of “natural and open” could either be considered as “natural” because of the “snow” and “trees”, or the subject could also take it as “manmade” for the “cabin” in the near view.
Greene and Oliva (2009a) found that presentation time threshold for 75%-correct performance of global property classification was longest in “openness” task ($M = 47\text{ms}, SD = 4.6$) and shortest in “naturalness” task ($M = 19\text{ms}, SD = 1.9$). However, there was no such dramatically different response times observed in Experiment 2 between “natural” ($M = 602.04\text{ms}, SD = 332.77$) and “open” ($M = 617.59\text{ms}, SD = 329.82$). The different observations may be caused by the fact that Greene and Oliva (2009a) tested scene images with most typical features in these properties.

A hypothesis for the performance discrepancy in “manmade or closed” and “natural and open” capacity analysis is that subjects simplified the task in deciding “natural and open”/“manmade or closed” into “natural”/“manmade” discrimination. As discussed before, capacity coefficient compares actual performance with predicted performance. Assuming subjects were performing only one task, the cumulative hazard function ($H$) of response times in “manmade or closed” OR task $H_{\text{manmade-closed}}(t)$ can be expressed as $H_{\text{manmade}}(t)$. Therefore, $H_{\text{manmade}}(t)$ would always be less than the predicted performance, which is the sum of cumulative hazard function of response times in “manmade” and “closed” respectively, resulting in limited capacity (Equation 5). Likewise, the reverse cumulative hazard function ($K$) of “natural and open” AND task $K_{\text{natural-open}}(t)$ can be replaced with $K_{\text{natural}}(t)$. For the predicted performance took into consideration of response times in “open” task, the ratio would always be larger than 1, leading to super capacity (Equation 6).

$$C_{\text{OR}}(t) = \frac{H_{\text{manmade-closed}}(t)}{H_{\text{manmade}}(t) + H_{\text{closed}}(t)}$$

$$= \frac{H_{\text{manmade}}(t)}{H_{\text{manmade}}(t) + H_{\text{closed}}(t)} < 1$$
A. Tversky and Kahneman (1983) discussed conjunction fallacy. They gave the personality sketches of a fictitious individual “Linda” and asked which of a pair of descriptions best fit. More respondents indicated that “Linda is a bank teller and is active in the feminist movement” was more probable than “Linda is a bank teller”, though the latter had a larger probability for $p(A \& B) \leq p(A)$. Alternatively, a disjunctive rule can be expressed by $p(A|B) \geq p(B)$. When the statement that “Linda is a bank teller” was replaced by “Linda is a bank teller whether or not she is active in the feminist movement”, “banker teller and feminist movement” still ranked as the most likely expression though the two alternative expressions in the replacement are the same.

In the current study, subjects were expected to choose “natural and open” with a lower probabilities compared to their choices of “natural” or “open” images separately. There was no substantial evidence in subjects’ choice probabilities showing that they were choosing “natural and open” dependently from the Bayesian analysis. However, they were prone to describe the scene as “natural and open” rather than neither “natural” nor “open” which maybe due to the “representative description” of “natural and open”. The phenomenon of conjunction fallacy also existed in “manmade or closed” choices. Though it was hard to tell if “manmade or closed” were dependent choices based on Bayesian analysis, subjects tended to choose “manmade” or “closed” with larger probabilities compared to their choices in “manmade or closed” selections.
4 General Discussion

The current research examined the role of global properties in human observers’ scene perception. In Experiment 1, comparisons on four global properties (“natural”, “manmade”, “open”, and “closed”) were collected online from a wide range of subjective choices. These answers were analyzed in a pairwise comparison model to generate four standardized reference ranking scales describing the extent to which characteristics can describe scene global properties. In Experiment 2, scene images selected from the reference scales were used to test human’s performance in processing global properties conjunctively. Cognitive modeling indicated that human observers were more efficient in categorizing scene images as “natural and open” but less efficient in classifying scene images as “manmade or closed” than the predicted baseline.

The reference ranking scales offer a possible solution for the conceptual discrepancy of scene categorizations between the superordinate and global properties. Both Park et al. (2011) and Kravitz et al. (2011) observed that scene representations in PPA were based on spatial information (open/closed). It is notable that they adopted different stimuli: Kravitz et al. (2011) used indoor scenes (i.e. “concert halls” and “living rooms”), while Park et al. (2011) had “natural – urban” categories to exclude indoor scenes. Indoor and outdoor are dichotomous superordinate level categories and might not be suitable breakpoints to polarize global properties of naturalness. The reference ranking scales developed here can provide a holistic view in combining superordinate-level and global properties scene categorizations together.

As discussed before, there was no consistent agreement in the advantage of
basic-level or global properties in scene processing so far. Rosch et al. (1976) speculated
the objects could be identified more rapidly in basic-level categories because the
superordinate-levels were derived by inference from the class membership of the
basic object and that subordinate-levels were derived from observation of attributes –
additional to those needed to perceive the basic level. Recent evidence showed that the
scene categorization may not follow assumption of serial processing: from high-level
to basic-level or verse versa. Instead, the categorization process of scene perception
might be driven by attributes of perceptual similarities (Banno & Saiki, 2015) or
discriminabilities (Sofer et al., 2015). Perceptual distance from multidimensional
scaling analysis (Berman et al., 2014; Kadar & Ben-Shahar, 2012) could be another
aspect to be considered. In Experiment 1, 174 basic-level categories were sorted
by global properties, one of the perceptual attributes assumed to influence scene
processing. Though the number of scene images in different basic-level categories
varied, the ranking scales database provides future studies better manipulations in
testing serial processing or parallel processing of scene categorization across different
levels.

B. Tversky and Hemenway (1984) argued that the knowledge about parts was
particularly salient at the basic-level. Parts of a basic-level category object refer to
both a perceptual entity and to a functional role. In the given examples, the leg of a
chair or the handle of a screwdriver had both particular appearance and function. To
clarify the criteria of parts: attributes considered to be parts, then refer to segments
of whole that are less than wholes; they are judged by a majority of naïve informants
to be parts, and they fit into a has a or is partially made of sentence frame. The
same logic can be applied in information extracting of scene processing: Scenes are
composed of objects.

Biederman (1981) illustrated a model of routes to achieve scene understanding
while his initial goal was to explain the object detection (also see Biederman, 1972). There is a lot of existing research which discusses how contextual information of scene-object guides attention during the real-world visual search by instructing where to search and what to expect (Castelhano & Heaven, 2010; Chun, 2000; Pereira & Castelhano, 2014; Rosenbaum & Jiang, 2013; Spotorno, Malcolm, & Tatler, 2014; Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011). From the global properties perspective, Greene and Oliva (2009b) compared the global properties representation to local region representation. They found that neither a collection of the local region (e.g. 9% sky, 25% rock, and 66% water) nor prominent object (e.g. “grass” in a field) model could predict human rapid scene categorization performance. Therefore, they suggested that the qualia of object perception in a brief glance might be based on an inference of these objects given global scene structure and schema activation.

Greene and Wolfe (2011) reported that global image properties alone were not sufficient in guiding visual search attention. Though Joubert et al. (2007) showed that objects could facilitate or interfere the “manmade/natural” scenes judgment, little is know how objects influence perceived global properties quantitatively and qualitatively. For example, beach is a typical basic-level category depicting naturalness. However, it is hard to guess how a yacht in the scene might decrease the naturalness perception. And even, compared to beach chair on the beach, which scene is more natural? In the current Experiment 1, the reference ranking scales provide future study a potential way exploring the perceptual interactive role of objects in scene processing.

Given that the present study used the scene images as stimuli testing scene processing, some differentiating features might be impossible perceived in the images processing compared to the actual scenes. A recent study by Greene, Baldassano, Esteva, Beck, and Fei-Fei (2016) proposed that a scene’s category may be determined
by the scene’s function or the possibilities for actions within a scene. Alternatively speaking, a kitchen is a kitchen because it is a space that affords cooking, not because it shares objects or other visual features with other kitchens. This draws attention that in the future study, the technology of virtual reality or mixed reality might be a better representation substitute of scene images for human observers to interact in the scene processing.

Experiment 2 stressed the need to explore human observer’s performance on different global properties conjunctively. Greene and Oliva (2010) asked subjects to decide to which pole of the global property the scene image belonged (e.g. small depth/large depth) from a group of ranked scene images. They attempted to test the properties as independently as possible, albeit there was admitted significant covariation between properties. Greene and Oliva (2009b) examined correlations between pairs of global properties they tested. Openness was in high correlation with expansion \( r = 0.75 \) and mean depth \( r = 0.9 \). Because they didn’t include naturalness in the discussion, the correlation between openness and naturalness was not estimated. In the current study, it seems that because naturalness was highly correlated with openness reported in Experiment 1, human observers did not treat naturalness and openness independently. Though it was hard to tell if their choice preference in these two global properties were being dependently processed from the Bayesian analysis, SFT results from their response time analysis indicated that human observers might either simplify judgment on conjunctive global properties into one single dimension, or information processing of naturalness and openness were in fact dependent.

The scene images selected in Experiment 2 were from the same ranges on the reference scales. The underlying assumption is that subjects’ responses to different block questions, not the characteristics describing the global properties in the images,
resulted in the differences in subjects’ performance. In a future study, it would be interesting to see the variations in the human observer’s response times as the descriptive characteristics of global properties changes from the “high” to “low” dimension. Human observers may respond fast at first, then gradually become slower when there are more mixed elements displayed in the scene images and be able to respond fast in the “low” descriptive dimension again when rejecting decision can be made quick. This test can provide a more comprehensive view of human observers’ performance in global properties categorization.

One limitation is that more subjects need to be recruited to test the consistency among subjects’ judgments for next step. Unlike previous research, subjects were not instructed on the definitions of the global properties in the current study. Therefore, subjects’ judgment of scenes’ naturalness and openness were based on their own criteria. In Experiment 1, subjects were recruited online across the country, avoiding the bias caused by life experience on the criteria of the global properties. In Experiment 2, it was impractical to control this factor among university undergraduate students. Moreover, the data analysis dropped a significant proportion of subjects who participated (8/17). It was hard to tell if they had different criteria for their response choices or they did not follow the instructions. Though the group level always indicated consistent capacity coefficient results, the variances for group response choices percentage were still high. There is a need to include more subjects for Experiment 2 and maybe add some other procedures in research design like after-experiment interviews to check if they apply different criteria as instructed. It could be possible that subjects may intuitively take “indoor” as “manmade” and “outdoor” as “natural”, thus indoor-outdoor can be an additional factor to control in tested global properties.
References


Appendices
Figure 1. Choice probability for Subject 1 in capacity coefficient task.
Figure 2. Choice probability for Subject 2 in capacity coefficient task.
Figure 3. Choice probability for Subject 3 in capacity coefficient task.
Figure 4. Choice probability for Subject 4 in capacity coefficient task.
Figure 5. Choice probability for Subject 5 in capacity coefficient task.
Figure 6. Choice probability for Subject 6 in capacity coefficient task.
Figure 7. Choice probability for Subject 7 in capacity coefficient task.
Figure 8. Choice probability for Subject 8 in capacity coefficient task.
Figure 9. Choice probability for Subject 9 in capacity coefficient task.