The influence of sequential predictions on scene gist recognition

by

Maverick E. Smith

B.S., Mississippi State University, 2015

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Dr. Lester C. Loschky

Copyright

Abstract

Past research has argued that scene gist, a holistic semantic representation of a scene acquired within a single fixation, is extracted using purely feed-forward mechanisms. Many scene gist recognition studies have presented scenes from multiple categories in randomized sequences. We tested whether rapid scene categorization could be facilitated by priming from sequential expectations. We created more ecologically valid, first-person viewpoint, image sequences, along spatiotemporally connected routes (e.g., an office to a parking lot). Participants identified target scenes at the end of rapid serial visual presentations. Critically, we manipulated whether targets were in coherent or randomized sequences. Target categorization was more accurate in coherent sequences than in randomized sequences. Furthermore, categorization was more accurate for a target following one or more images within the same category than following a switch between categories. Likewise, accuracy was higher for targets more visually similar to their immediately preceding primes. This suggested that prime-to-target visual similarity may explain the coherent sequence advantage. We tested this hypothesis in Experiment 2, which was identical except that target images were removed from the sequences, and participants were asked to predict the scene category of the missing target. Missing images in coherent sequences were more accurately predicted than missing images in randomized sequences, and more predictable images were identified more accurately in Experiment 1. Importantly, partial correlations revealed that image predictability and prime-to-target visual similarity independently contributed to rapid scene gist categorization accuracy suggesting sequential expectations prime and thus facilitate scene recognition processes.

Table of Contents

List of Figures v
List of Tables vi
Acknowledgements vii
Chapter 1 - Introduction 1
Gist of a scene 2
Feed-forward information extraction 3
Perceptual predictions and their influence on information extraction 4
The Scene Perception and Event Comprehension Theory 5
Chapter 2 - Experiment 1 10
Method 10
Participants 10
Stimuli and Design 10
Procedure 14
Experiment 1 Results 16
Exploratory analysis of conceptual priming and image similarity 26
Discussion 31
Chapter 3 - Experiment 2 34
Method 36
Participants 36
Procedure 36
Experiment 2 Results 37
Analysis of image predictability and image similarity 39
Discussion 47
Chapter 4 - General Discussion 49
References 53
Appendix A - Example coherent sequence 61

List of Figures

Figure 1. Trial schematic of the sequence of events within two example trials. 12
Figure 2. Scene categorization accuracy as a function of spatiotemporal coherence. 17
Figure 3. Scene categorization accuracy as a function of within category priming. 20
Figure 4. Scene categorization accuracy as a function of between category priming 22
Figure 5. Scene categorization accuracy as a function of image similarity and coherence 31
Figure 6. Scene categorization accuracy as a function of image predictability 41
Figure 7. Scene categorization accuracy as a function of image similarity and predictability. 44
Figure 8. Example coherent first-person viewpoint sequence. 61

List of Tables

Table 1. Experiment 1: Model comparisons for scene categorization accuracy..... 30
Table 2. Experiment 2: Model comparisons for scene categorization accuracy..... 42
Table 3. Correlations between scene categorization, image prediction, and image similarity..... 46

Acknowledgements

I would like to give a special thanks to my advisor Dr. Lester Loschky for his continued guidance, support, and dedication since I arrived at Kansas State University. I would also like to thank my committee members Dr. Thomas Sanocki (University of South Florida) and Dr. Heather Bailey (Kansas State University) for their help planning and developing this project. Special thanks to the members of the Visual Cognition lab, all of the undergraduate research assistants, and my fellow graduate students, Ryan Ringer, John Hutson, Jared Peterson, and Taylor Simonson for all of their help planning and designing the experiments, creating stimuli, and data collection. I am very thankful for all of the thoughtful comments and feedback they have given me at every stage of this project. Lastly, I would like to thank Dr. Adam Larson (University of Findlay). Even though he was not a member of the Thesis committee, he contributed to the development of this project.

Chapter 1 - Introduction

When navigating through the world, we typically have expectations about the kinds of scenes we will see from one moment to the next. For instance, through repeated experiences of walking from the kitchen to the living room in your and in other people's homes, you may expect that kitchens and living rooms appear near one another. It is possible, but unknown, if we use such knowledge-based expectations when identifying everyday scene categories. Intuitively, any abrupt and unpredictable change in our visual experience should be more difficult to perceive, such as if when expecting to enter a living room from a home kitchen, you instead find yourself in an office cubicle. Despite the surprise that would be elicited by such an unpredictable change, scene perception research has found that when observers are presented with briefly flashed and masked novel pictures of scenes in randomized sequences, they can accurately identify their general meaning or "gist"[1] (Potter, 1976). Here, we define the theoretical construct of *scene gist* as a viewer's holistic semantic representation of a scene which can be acquired within a single eye fixation. We operationalize it in terms of rapid scene categorization at the basic level (Tversky & Hemenway, 1983). We will use the term *scene gist* to refer to the theoretical construct, and *rapid basic level scene categorization* to refer to the measured behavioral outcome. Rapid basic level scene categorization accuracy is typically at ceiling with 100 ms

[1] Scene gist is an important theoretical construct in theories of scene perception (Rayner, Smith, Malcolm, & Henderson, 2009; Wolfe, Vo, Evans, & Greene, 2011). A scene's gist influences 1) attentional selection and visual inspection of a scene (Gordon, 2004; Torralba, Oliva, Castelhano, & Henderson, 2006), 2) object recognition (Bar & Ullman, 1996; Davenport & Potter, 2004) and 3) long term memory for the contents of a scene (Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989). The theoretical construct of gist implies more than the operational definition of it; therefore, when discussing the theoretical construct, we use the term *scene gist*, but when discussing the results, we refer to our operationalization of scene gist, namely *rapid scene categorization* at the basic level (e.g., office, hallway, or parking lot).

flashed and masked stimulus onset asynchronies (SOAs), and the *knee* of the SOA function is typically at 40-50 ms (Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005; Greene & Oliva, 2009; Loschky & Larson, 2010).

Gist of a scene

Scene gist is used to refer to both conceptually and perceptually driven properties of scenes (Oliva, 2005). At a perceptual level, scenes can be described according to their visual properties both in terms of low-level vision (e.g., edges, spatial frequencies, and color) and middle-level vision (e.g., shapes, contours, and spatial layout) (Oliva & Torralba, 2001). At a conceptual level, the meaning of a scene can be reduced to a verbal description of the events that comprise it (e.g., a birthday party) (Potter, 1976) or to a general semantic category label (e.g., a beach) (Loschky & Larson, 2010). Given the speed of gist perception, the information underlying gist recognition is holistic. The layout, or configuration of a scene (e.g., a prototypical beach scene has the sky above the horizon, below which, a diagonal waterline divides the water and sand) contributes to the global visual properties of scenes used to categorize them (Oliva & Torralba, 2001). Such scene layout information and the scene's gravitational frame, which constrains the orientation of its layout (e.g., the sky is above and the sand and water are below), contributes to a scene's gist (Loschky, Ringer, Ellis, & Hansen, 2015). Spectral information can be used to describe the spatial properties of scenes (naturalness, openness, roughness, expansion, ruggedness) and to identify scenes at the superordinate and basic levels (Oliva & Torralba, 2001, 2006). Importantly, images that share perceptual properties are also typically semantically similar. Fast extraction of such global properties make up the perceptual description of a scene and allows for an estimation of its meaning (i.e., conceptual gist) (Ramkumar, Hansen, Pannasch, & Loschky, 2016). Consistent with traditional

views of data-driven perception, the acquisition of conceptual gist may be performed through the sequential build-up of perceptual information as information processed at earlier stages feeds into higher stages until an identification response is made (VanRullen & Thorpe, 2001). Alternatively, there is evidence for considerable amount of overlap in the time course of perceptual processing of scene gist and when information becomes available for a response (Caddigan, Choo, Fei-Fei, & Beck, 2017; Ramkumar et al., 2016; Rao, 1999). Thus, categorization may influence, rather than strictly follow, early visual analysis.

Feed-forward information extraction

As mentioned above, complex natural scenes can be categorized within the first pass of feedforward information flow, from the retina through the higher-level visual cortex, without top-down influences[2] (Bastin et al., 2013; Delorme, Rousselet, Macé, & Fabre-Thorpe, 2004; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Perrinet, Samuelides, & Thorpe, 2004; Serre, Oliva, & Poggio, 2007; VanRullen, 2007). During the first feedforward sweep, a wave front of visually elicited activation travels through the visuomotor system so fast that it is devoid of information from recurrent top-down processing, which develops only after the initial wave. Event-related potentials to target and distractor images presented rapidly begin to diverge by roughly 150 ms post-stimulus onset (Thorpe, Fize, & Marlot, 1996). Studies using magnetoencephalography (MEG) have demonstrated analogous results for rapid scene categorization at the basic level, with a scene's category being distinguishable between 100 –

[2] Visually evoked ERP signals reflect a participants' decision as to the identity of a rapidly presented image (e.g., whether it contains an animal) as early as 150 msecs post-stimulus onset (Thorpe et al., 1996). Though it has been suggested by previous researchers (i.e., VanRullen, 2007), definitive evidence to conclude that rapid visual categorization relies primarily on feedforward processing would be to demonstrate that feedback processing influences perceptual analysis only after the initial feedforward sweep. No such evidence exists to our knowledge.

250 ms post-stimulus onset (Ramkumar et al., 2016). Likewise, studies using artificial neural networks modeled from the constraints of the visual system have found that most of the stimulus-relevant information used in recognition can be extracted from the bottom-up input alone (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Serre et al., 2007). Together, these findings are consistent with models of perception where computations performed by early visual modules feed into conceptual, memory systems unidirectionally, with later conceptual systems only operating after the first feed-forward sweep, and thus not influencing early visual processes (Marr, 1982), which we will call the *feed-forward gist* hypothesis. While informative, these models assign little importance to the many bidirectional connections between neural structures, which may exceed the number of feed-forward connections (Salin & Bullier, 1995). Reciprocal connectivity may provide the infrastructure to support top-down interactions between prediction and perception.

Perceptual predictions and their influence on information extraction

Typically, in studies of rapid scene categorization, scenes from multiple categories are presented in randomized sequences to participants; however, the semantic and structural relationships between the scenes we interact with in our day-to-day lives are not random. In our experience with the world, scene categories change in predictable ways along the paths we take (e.g., walking from your office to a parking lot after a day of work). Hallways, not parking lots, are usually on the other side of office doorways. The visual system may take advantage of such regularities in the environment and make predictions that may influence rapid scene gist categorization, rather than reinterpret the gist of a scene anew on every fixation. No studies have yet investigated this issue. However, a number of separate lines of research speak to it.

Rapid scene categorization is influenced by prior knowledge (e.g., schemas), such that categorization is harder if scenes contain information that violates expectations (e.g., a bolder in a living room as opposed to a bolder on a mountain) (Davenport & Potter, 2004; Greene, Botros, Beck, & Fei-Fei, 2015). There is also work showing the effects of expectations on immediate scene memory, after as short a time as 250 ms, which is similar to the time course of rapid scene categorization. Scene memory has been shown to be influenced by predictions about the surrounding environment (Intraub, 2010; Intraub & Dickinson, 2008). For example, when a scene is briefly presented and then removed, participants will remember seeing more than was shown (e.g., parts of a desk cropped out in an office scene) (Intraub & Dickinson, 2008). However, no one has yet investigated how predictions made prior to viewing a scene image influences rapid scene categorization.

A related and important issue is that of priming of scene perception. Studies have shown perceptual priming between scenes having similar layouts (Sanocki & Epstein, 1997), across different views of the same scene (Castelhano & Pollatsek, 2010), and between sequential pairs of visually similar scenes (Caddigan et al., 2017). Other studies have shown conceptual priming of scene identification by words (Reinitz, Wright, & Loftus, 1989). However, these priming studies have not been done using more ecologically valid spatiotemporally coherent scene sequences such as how we encounter scenes from one moment to the next in our everyday life.

The Scene Perception and Event Comprehension Theory

Our interest in this problem comes from the recently developed theoretical framework of the *Scene Perception & Event Comprehension Theory* (SPECT) (Loschky, Hutson, Smith, Smith, & Magliano, 2018; Loschky, Larson, Smith, & Magliano, submitted). SPECT considers scene gist recognition as one part of the larger problem of perceiving and understanding scenes

and events in the real world, or within the more experimentally tractable context of visual narratives. According to SPECT, the extraction of a scene's gist is essential to understanding an event (i.e., "a segment of time at a given location that is conceived by an observer to have a distinct beginning and an end") (Zacks & Tversky, 2001). SPECT proposes a distinction between front-end and back-end mechanisms. Front-end mechanisms are involved in processing information within single eye fixations, including visual processing of the scene (i.e., information extraction) and selection of what information to process within it (i.e., attentional selection). Back-end mechanisms are involved in processing information in memory across multiple fixations, particularly in constructing the current event model in working memory (i.e., one's online representation of what is happening now). Back-end mechanisms also store event models in episodic long-term memory, and retrieve stored event models, scripts, and schemas from long-term semantic memory, which can influence the construction of the current event model.

Scene gist plays an important role in both front- and back-end processes. The gist of a scene is recognized within a single eye fixation during the process of information extraction (Greene & Oliva, 2009; Larson, 2012). The extracted scene gist then influences back-end processes in the current event model in working memory. According to SPECT, scene gist is very important for laying the foundation of a new event model. Event models include situational information along various event indices (Zwaan & Radvansky, 1998), and the earliest extracted one is likely the location (Larson, 2012), namely the scene category, or gist.

Importantly, SPECT proposes that back-end processes involved in event model construction influence front-end processes involved in information extraction. Suppose a viewer is watching a hand-held video of someone walking from their office to a parking lot. Within the

first fixation, the viewer recognizes the gist of the spatial location of the first scene, for instance, that it is an office. Will the viewer need to extract the same perceptual and conceptual gist on each fixation (i.e., office, office, office, … etc.)? As noted earlier, many authors have argued that scene gist recognition processes are automatic and feed-forward (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Fei-Fei, VanRullen, Koch, & Perona, 2005; Li, VanRullen, Koch, & Perona, 2002). Consistent with this notion, it would be very important to detect when the location changes, since that could indicate that a new event has begun. If the visual system did not extract the gist of the scene anew on every fixation, how would the viewer be able to detect that the location had changed when such a change occurs? This is an assumption consistent with the feed-forward gist hypothesis. Alternatively, it would seem to be a waste of cognitive resources to extract the same gist on consecutive fixations. Thus, it may require fewer processing resources and time to extract the scene location on subsequent fixations after the gist of the initial fixation is extracted so long as the scene category does not change. Research using the flash-preview moving window paradigm has found that providing a preview of a scene prior to a visual search task facilitates search for a target object, suggesting that memory of the scene's gist extracted on the initial fixation is stored in memory across fixations (Hillstrom, Scholey, Liversedge, & Benson, 2012). This suggests the possibility that scene gist extraction may be facilitated on later fixations due to priming from the current event model, which we will call the *scene gist priming hypothesis*. When one has an expectation as to what an upcoming scene category will be, the expectation may conceptually prime representations associated with it, lowering the threshold amount of visual information necessary to identify the scene.

The scene gist priming hypothesis suggests further hypotheses. Scene gist priming could differ based on the number of previous fixations on the same scene, with priming of basic level

scene categorization increasing as the number of fixations on the scene increases. We will call this the *within-scene category priming* hypothesis. Furthermore, we can hypothesize that scene gist priming may occur between spatio-temporally inter-related, but different, basic level scene categories. For instance, if the hand-held video shows the individual with the camera walking to the door of the office, the viewer may predict the scene category on the other side of the office door based on schemas for what scene categories connect with offices and the inferred actions of the individual holding the camera, for example, that it is a hallway. If the viewer's prediction is correct, it could facilitate gist extraction relative to seeing a less predictable scene category, such as a parking lot. We will call this the *between-scene category priming* hypothesis. Between- and within-scene category priming differ in important ways, but both types of priming are consistent with the previously noted point that scenes that share similar features tend to belong to the same scene category (Oliva & Torralba, 2001). Within-scene category priming, across multiple fixations of the same scene category, may be explained by a combination of: 1) perceptual priming due to the overlap in bottom-up processing of similar visual features across different views of the same scene, together with 2) conceptual priming due to the same gist concept shared across multiple fixations (e.g., "office") in the foundation of the event model. However, because between-scene category priming occurs across images that share fewer visual features, the priming should be relatively more conceptual than perceptual. Based on these considerations, we could hypothesize that the degree of within-scene category priming would be greater than between-scene category priming, due to the former priming sharing more perceptual and conceptual information in the event model (e.g., different views of the same office), while the latter priming primarily shares conceptual information (e.g., office and a hallway) though it sometimes shares some perceptual information (e.g., offices and hallways share indoor scene

features such as rectilinearity). We will call this the *within > between-category priming* hypothesis.

Chapter 2 - Experiment 1

To test the above hypotheses, we created 24 spatio-temporally connected image sequences from a starting location to a destination (e.g., going from an office to a parking lot). An example of one of the sequences is provided in the Appendix and all stimuli used in the experiments as well as all data and analyses are available to download at https://osf.io/83sjx/. Target images were presented in either coherent or randomized sequences. Importantly, images in the randomized sequence condition presented the same perceptual information as the coherent condition to the observer, except for their temporal order and the perceived spatial connectedness of scene information between images. But, by randomizing the image order, higher level expectations for subsequent scenes could not be used to influence the ability to categorize an upcoming scene. In coherent sequences, expectations for to-be-presented scenes may prime scene categories resulting in more accurate categorization for scenes presented in coherent than in randomized sequences.

Method

Participants. There were 48 participants (16 females, 32 males, mean age = 19.57) from Kansas State University's undergraduate research pool, who participated in the experiment for course credit. Participants' vision was tested and was 20/30 or better, as measured by the Freiburg Visual Acuity & Contrast Test (FRACT), Landolt C acuity subtest (Bach, 2006). All participants were naïve to the purpose of the experiment and signed an informed consent prior to participating.

Stimuli and Design. Examples of a coherent and a randomized sequence are shown in Figure 1. Figure 1a is a simplified version of Figure 1b. The target is the 3rd picture in both the i) coherent and ii) randomized sequence. Images were photographed from a first-person

viewpoint so that when presented in coherent sequences they appeared as if the observer were "navigating" through the environment from one destination to another, such as from an office to a parking lot. Five categories of images were taken per first-person sequence of spatiotemporally connected scenes. Four hundred and eighty total images were collected by taking screen shot photographs from videos captured at a resolution of 1920 x 1080 pixels on a Canon XA10 HD and a Canon HFM41 HD camcorder. All videos were taken on Kansas State University's campus and the local Manhattan, Kansas metropolitan area. Two hundred and forty images at a resolution of 1024 x 768 were extracted from the on-campus videos, which were composed of 8 different basic level scene categories: 4 indoor categories (office, classroom, hallway, and stairwell) and 4 outdoor categories (parking lot, courtyard, sidewalk, and lawn). Of the 8 categories, two indoor categories (i.e., office, classroom) and two outdoor categories (e.g., parking lot and courtyard) were chosen as destination locations. The remaining four were transitional scene categories (e.g., scene categories presented between destinations). Forward versions of each of the spatiotemporally connected scene sequences (e.g., parking lot to an office, classroom to a courtyard, etc.), and their reverse (e.g., office to a parking lot, courtyard to a classroom, etc.) were taken from each of the four destination locations to each of the others, but not along the same pathways (i.e., the office to the classroom and the classroom to the office sequence were not the same office or classroom and did not contain the same transitional scene categories). The crossing of starting points and destinations, and use of multiple different exemplars for each, was done to ensure that participants would be unable to predict which destination scene category they would appear to be "navigating", given the identity of the first image category on each trial.

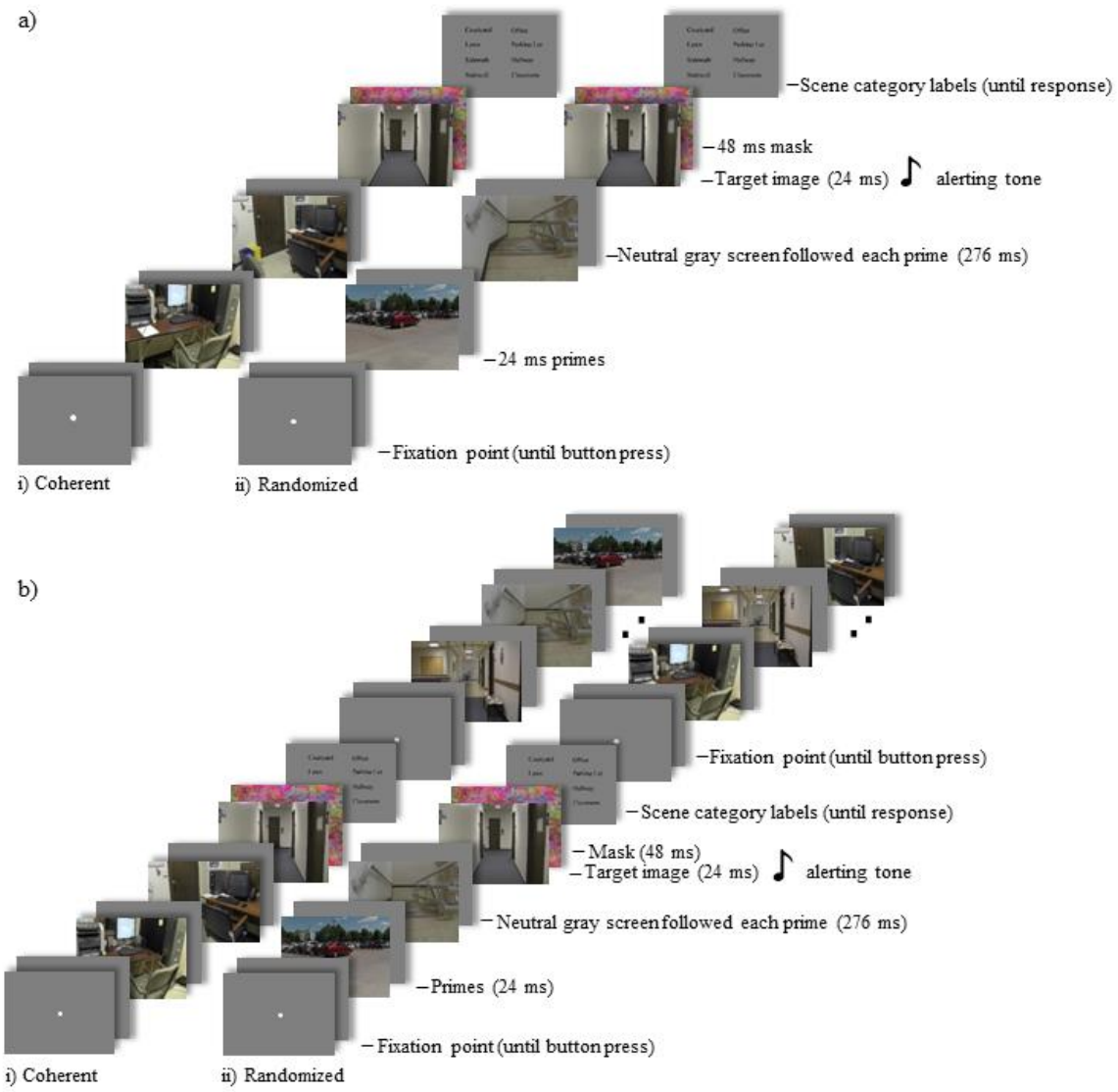

Figure 1. Trial schematic of the sequence of events within two example trials. The target image was flashed and masked for 24 msecs and was shown simultaneously with an attentional alerting tone. Panel a) is a simplified version of a trial, illustrating the order of screens up to the response. The sequence of scenes in i) is coherent beginning with an office and ending with a scene of a parking lot and in ii) the same images are shown in an example randomized sequence. Panel b) shows a more complete version of a trial, showing the continuation of the sequences after the response was made. The continuation was shown so that viewers always saw a full 10 image sequence, regardless of which image in the sequence was the target.

The same number of off-campus as on-campus images were used. There were 8 different off-campus categories: 4 indoor categories (store interior, bedroom, stairwell, and hallway) and 4 outdoor scene categories (park, city center, sidewalk, and alley). Store interior, bedroom, park, and city center were determined as destination locations, and stairwell, hallway, sidewalk, and alley were chosen as transitional categories. Destination categories for the videos that were taken on- versus off-campus were not intermixed into spatiotemporally connected sequences (e.g., a store interior never appeared in the same sequence as an office, nor did it appear as an option in the 8AFC for a sequence taken from on-campus).

Each base sequence we filmed was composed of 20 images, 4 of each scene category (e.g., 4 offices, 4 hallways, 4 stairwells, 4 sidewalks, 4 parking lots). Of the 20 total images in each base sequence, 10 were shown in coherent and 10 in randomized conditions. To reduce predictability, subsequences of 1-3 images from each category were shown to participants before a category shift (e.g., 3 offices, 1 hallway, 2 stairwells, etc.).

Two subsets of 10 images for each base sequence, one subset for the coherent sequence and one for the randomized sequence, were generated independently for each participant to ensure that the same image was not repeated in both conditions for a given participant. We randomly selected which of the 4 images within each category of each spatiotemporally connected sequence each participant was shown. Using the i) coherent sequence in Figure 1b as an example, in the office-to-parking lot sequence, 2 offices were randomly assigned to its coherent version and the remaining 2 office images were assigned to the randomized version. Images within each of the categories were randomly assigned for each participant with two stipulations: 1) an image never repeated in both coherent and randomized versions of the

sequences within participants, and 2) both the coherent and the randomized versions of each sequence each had a total of 10 images. Furthermore, in the coherent condition, if more than 1 image were randomly chosen from a given scene category, they were presented in their coherent spatiotemporal order (e.g., 1, 3, *not* 3, 1). This could only happen by chance in the randomized condition.

All images were presented in color on 17-inch Samsung SyncMaster 957 MBS monitors running at an 85 Hz refresh rate, at a viewing distance of 53.34 cm. Monitors had a resolution of 1024 x 768 pixels and 37.79° x 28.71° of visual angle.

Ten 1/f amplitude color noise masks were used to mask target images. Target-mask pairings were randomized across all trials and participants.

Procedure. The experiment was a 2 (Spatiotemporal coherence: coherent vs. random order) x 2 (Image location: on-campus vs. off-campus) within-subjects design. Image location was treated as a nuisance variable as it was not of interest here. The manipulation of the spatiotemporal coherence of the scenes was blocked and counterbalanced across participants. The order the 24 different sequences were shown in was counterbalanced across participants using a 24 x 24 Williams Latin square (Williams, 1949), with the coherent and randomized versions of the sequences presented in separate blocks. Blocking was done to increase the likelihood that participants would perceive the coherence of spatiotemporally coherent sequences.

Participants completed 10 practice trials to familiarize them with the task and the speed of image presentation. None of the practice images were repeated in the experimental trials. There were 48 total experimental trials per participant: 24 coherent and 24 randomized sequences. Prior to beginning each trial, participants were shown a list of the scene category

labels they were about to view in a random order. As shown in the trial schematic provided in Figure 1a, participants initiated a trial by fixating a dot in the center of a neutral gray background while pressing a button using the computer's mouse. The 10 scene images from each sequence were shown in rapid serial visual presentation (RSVP). Each 10 image sequence included 1 target and 9 primes (or non-target images). The primes/non-target images were given 300 ms of processing time (24 ms presentation + a 276 ms neutral gray inter-stimulus interval [ISI]). We gave participants 300 ms of processing time for primes because that allows images in an RSVP sequence to be identified and stored in conceptual short-term memory (Potter, 1976). The single target image in the sequence was also presented for 24 ms, but was immediately followed (i.e., after a 0 ms ISI) by a 48 ms 1/f noise mask to limit processing time, thus making the target harder to identify (Hansen & Loschky, 2013). Simultaneously with the target onset, we presented a high pitched alerting tone through ear buds, because doing so in an RSVP stream has been shown to reduce the attentional blink (Kranczioch & Thorne, 2013). Pilot testing showed that inserting the alerting tone improved participants' performance in identifying target images in both coherent and randomized versions of each sequence. Immediately following presentation of the target and mask, the participant was shown an 8-alternative forced choice (AFC) array of scene category labels and asked to select the category that matched the target image. The 8-AFC included all 8 categories from the matching on-campus or off-campus base sequence. The location of scene category labels in the 4 x 2 8-AFC grid were randomized on every trial to avoid location-based response biases when guessing (e.g., the top left corner). Thus, reaction time could not be used as a measure of scene facilitation (i.e., each response began with a visual search of the 8-AFC array). The temporal position of the target image within each 10-image sequence was equalized and counter-balanced across participants, so participants could not guess

when the target image would appear in each trial. Since the target image appeared equally often in all 10 sequential positions (including the 1st position), the remainder of the sequence of images were presented immediately after participants made a response as evident in Figure 1b, with processing times for these non-target images being identical to the prime images (i.e., 300 ms). Thus, participants saw the entirety of each 10-image sequence on each trial.

Experiment 1 Results

The analyses of categorization accuracy were conducted using R statistical software (version 3.1.1) with the lme4 library (Bates, Mächler, Bolker, & Walker, 2014). A logistic multilevel model was used to predict the probability of correctly identifying the target scene within each sequence from the predictors of spatiotemporal coherence of the sequence (e.g., coherent vs. randomized), image location (on-campus vs. off-campus), and their interaction. The participant and image intercepts were treated as random effects. Spatiotemporal coherence was effect coded as Coherent = 1 and Randomized = -1. Image location was effect coded as off-campus = 1 and on-campus = -1. Consistent with the *scene gist priming hypothesis*, images presented in coherent sequences (M = 53.5%, SE = 1.47%) were identified more accurately than images presented in randomized sequences (M = 33.85%, SE = 1.39%), $\beta = 0.47$, $z = 9.57$, $p < .0001$. Of less interest to this study, images taken from the on-campus location (M = 46.31, SE = 1.47%) were identified more accurately than images taken from the off-campus location (M =

41.06%, *SE* = 1.45%), β= -0.14, *z* = -2.25, *p* = 0.02. Importantly, the effect of the spatiotemporal coherence was the same for both locations, β= 0.008, *z* = 0.16, *p* = 0.87.

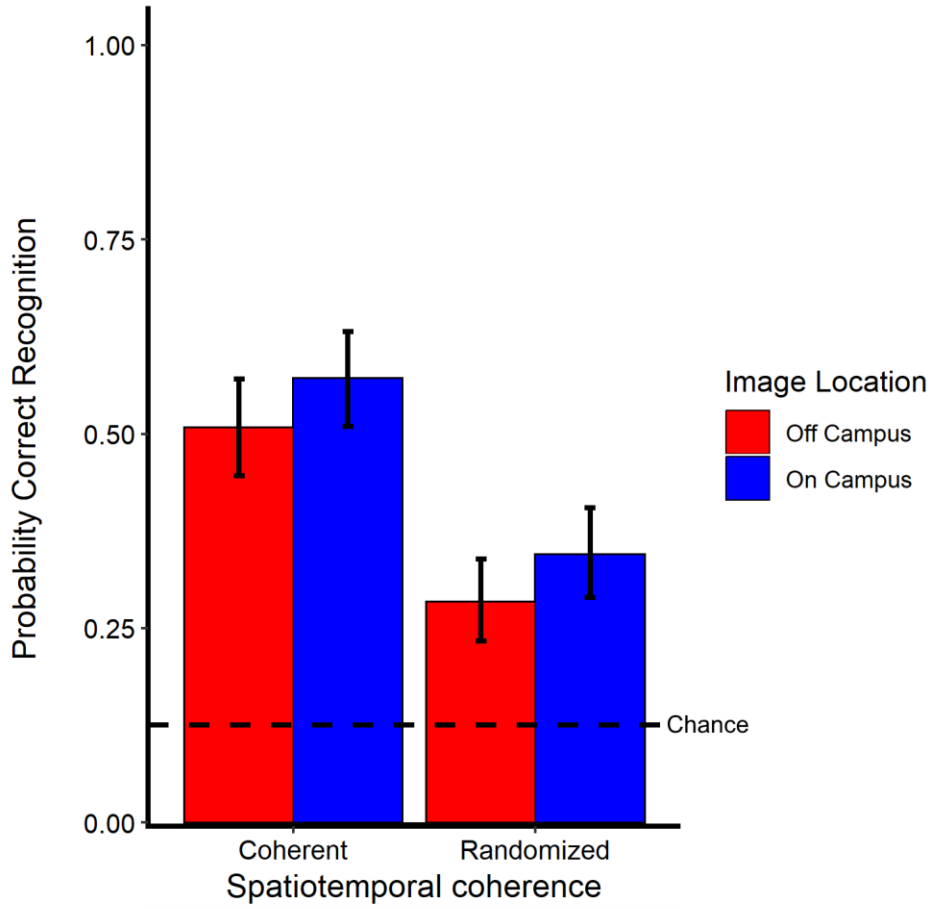

Figure 2. Rapid scene gist categorization accuracy as a function of the spatiotemporal coherence of the scene sequences and image location (on-campus vs. off-campus). Error bars represent 95% confidence intervals around the estimated mean. The probability of correctly identifying the scene by chance on any given trial was 12.5% as represented by the dashed line.

Because sequential expectations for upcoming scenes can be generated both within the same category (e.g., seeing one view of an office may lead one to anticipate another) and between scene categories (e.g., seeing multiple first-person views of someone "navigating"

through an office toward a closed door may prime expectations for a hallway), two separate analyses were conducted. The first analysis was conducted for instances when the target image was the same scene category as the preceding primes (i.e., an office after 0, 1, or 2 views of the same office scene). We refer to this as within-scene category priming. The second analysis was conducted for cases when the immediately prior image to the target was from a different, but contextually related category image to the target scene (e.g., a target hallway after it was preceded by 1, 2, or 3 offices). We refer to this type of priming as between-scene category priming. The number of sequential images of the priming category was treated continuously in both sets of analyses. Using multilevel modeling, one can treat continuous variables as repeated measures (Cohen, Cohen, West, & Aiken, 2003). We also fit two models to examine where asymptotic recognition performance was reached for both within and between category priming. When comparing models, Akiake information criterion (AIC) and Bayesian information criterion (BIC) values (smaller values indicate better model fit) as well as chi square tests of significance were used to assess model fit (Agresti, 2007; Wagenmakers & Farrell, 2004). A difference in AIC and BIC values of 2-6 can be accepted as moderate evidence for the model with the smaller value having the better fit, while differences of 6-10 suggest a strong difference between models, and a difference >10 indicates very strong evidence (Burnham & Anderson, 2004; Raftery, 1995).

As shown in Figure 3, performance improved as the number of prime images from the same category as the target image increased, consistent with the within-scene category priming hypothesis, $\beta = 0.72$, $z = 8.16$, $p < 0.001$, AIC = 1505.40, BIC = 1520.50. Recognition of the target image was better when it was preceded by 2 ($M = 68.63\%$, $SE = 3.26\%$), than when preceded by 1 ($M = 65.49\%$, $SE = 2.48\%$) or 0 ($M = 40.59\%$, $SE = 2.04\%$) primes of the same

category as the target. When perceivers view the same scene category multiple times, facilitation of the scene's gist may increase incrementally until reaching asymptote. To examine where asymptotic recognition of scene gist was reached, a second model was conducted using the log of the number of images that preceded the target. Within-category scene priming was again found, $\beta = 1.28$, $z = 8.58$, $p < 0.001$, AIC = 1498.70, BIC = 1513.80. The model containing the log of within-category scenes was a significantly better model fit than the model containing the linear variable, $\chi^2 = 6.71$, $p < 0.001$, ΔAIC = 6.70, ΔBIC = 6.70. Asymptotic performance was reached when the target image was preceded by 1 image. There is a large amount of overlap in the perceptual and conceptual properties of a scene from one fixation to the next when the scene category repeats across multiple views. This overlap enabled better scene gist categorization.

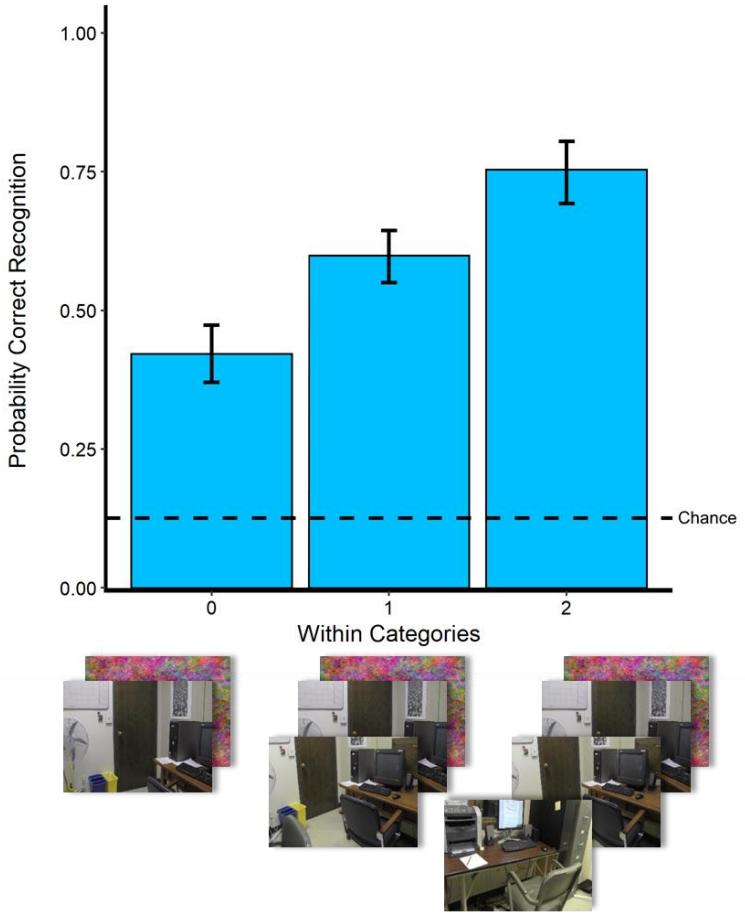

Figure 3. Rapid scene gist categorization accuracy as a function of the number of sequential exposures of the same scene category as the target. The target was preceded by 0, 1, or 2 primes that were of the same category as the target. The target in this illustration is the picture preceding the mask. Error bars represent 95% confidence intervals around the estimated mean probability correct**.**

To test the between-scene-category priming hypothesis, we analyzed the effect of the number of primes from the different category immediately preceding the target image. As the number of primes increased (e.g., a target picture was of a hallway after multiple views of an office), increased between-category scene facilitation was found, $\beta = 0.46$, $z = 2.55$, $p = 0.01$,

AIC = 354.20, BIC = 365.10. As shown in Figure 4, target images of a different category than their primes were more accurately identified if preceded by 3 (M = 42.16%, SE = 4.91%) than if preceded by 2 (M = 29.25%, SE = 4.44%) or 1 (M = 25.00%, SE = 5.00%) primes. To examine the number of scene category images that were viewed before asymptotic performance was reached, a second model was run using the log of between category primes. The log of between category primes significantly predicted scene gist performance $\beta = 0.81$, $z = 2.40$, $p = 0.02$, AIC = 354.95, BIC = 365.89; however, the model containing the logarithmic variable was not found to be significantly different from the first, $\chi^2 = 0$, $p > 0.05$, ΔAIC = 0.75, ΔBIC = 0.79. Unlike the within-scene category priming effect, scene gist facilitation for a target of a different but contextually predictable scene category increased linearly. Together, these results suggest that scene spatiotemporal coherence priming is not limited to a within-category effect (i.e., an effect analogous to repetition priming) as facilitation was found both within and between scene categories. Furthermore, recognition performance was less when the target was preceded by 0 primes than when preceded by 1 or 2 primes that were from the same category as the target. This finding is consistent with the within > between category priming hypothesis, because the amount of facilitation found within scene categories was greater than between scene categories (i.e., if the target was not preceded by any primes that were of the same category as the target, then the prime was of a different but contextually related category).The amount of perceptual and conceptual overlap between two images is greater within than between categories.

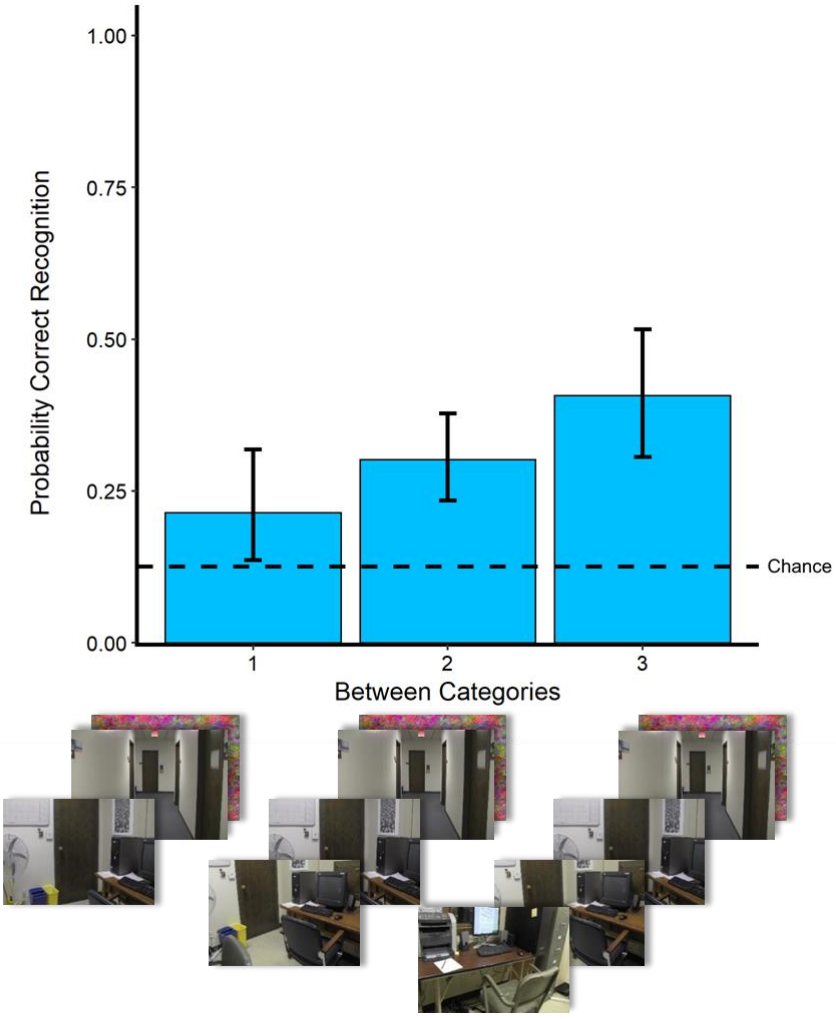

Figure 4. Rapid scene categorization accuracy as a function of the number of primes from the immediately preceding different category of scenes. Here, the target image in the sequence was of a different, but predictable category than the prime images. The target in the example sequence of images is the picture prior to the mask. Error bars represent 95% confidence intervals around the estimated mean probability correct**.**

Target images in coherent sequences were recognized more accurately than targets shown in randomized sequences. Furthermore, we found evidence of both within and between-scene category facilitation. A prediction for a scene category made prior to viewing a scene may

influence its recognition. Alternatively, it is also possible that the effects we observed were not due to facilitation of the current event model on gist perception, but instead were the result of participants' being unable to distinguish the prime from the target category. Primes were processed long enough to both be readily perceived and stored into conceptual short-term memory (Potter, 1976). Unlike primes, targets were briefly flashed and masked to reduce their visibility; therefore, participants may have confused the category of the target with the category of the prime and thus may have responded with the last scene category that they readily perceived and remembered. Such a response bias could explain the accuracy advantage for target scenes in coherent sequences as well as the within-category facilitation effect we observed; however, such a bias in responding to the category prior to the target cannot explain between-scene category priming when the target and prime are from a different scene category. To more carefully examine whether participants were responding with the category label that matched the target, the prime, or one of the other category labels that was neither the prime nor target, a series of multilevel logistic regressions were conducted. If participants confused the prime shown immediately prior to the target with the target itself, then participants should have responded with the prime category label regardless of whether targets were in coherent or randomized sequences. Alternatively, if participants are better able to recognize the target scene in coherent than in randomized sequences, then a response with the target category label should be more likely than a response with a prime or one of the other category labels. To conduct this analysis, we first removed all instances when the target was the first image in a trial in both coherent and randomized sequences. Note, however, that only cases when the prime and target categories *differ* can be used to address this important question. Thus, we also removed all cases when the prime and target categories were *the same*. We then conducted three multilevel logistic

regressions. Only the criterion differed between analyses. In the first analysis, we only included responses with target and prime labels. In the second analysis, we only included responses with target and other non-prime category labels. In the third and final analysis, we only included responses with prime and other category labels. We used the spatiotemporal coherence manipulation as a predictor in all three analyses, and effect coded it as coherent = 1 and randomized = -1. The criterion was the probability of a participants' response (i.e., target vs. prime, target vs. other, and prime vs. other). All models included the random intercept of participant. In the first analysis, a response with the category that matched the target was dummy coded as a 1 and a response with the category that matched the prime image was coded as a 0 (i.e., larger values indicate participants were more likely to respond with the category label that matched the target image). The results showed that when the image sequence was coherent, participants were significantly more like to respond with the category label that matched the target than with the category of the prime, $\beta = 0.17$, $z = 2.30$, $p = 0.02$. In the second analysis, a response with the category of the target was again dummy coded as a 1 and a response with one of the other category labels (excluding the prime category) was coded as a 0. Again, results showed that when sequences were coherent, participants were significantly more likely to respond with the category label that matched the target image than one of the other non-prime scene category labels, $\beta = 0.23$, $z = 3.01$, $p = 0.003$. Lastly, to examine participant errors, a response to the prime was dummy coded as a 1 and a response with any of the other categories was coded as a 0. We found that participants were no more likely in coherent than in randomized sequences to select the scene category that matched the prime versus one of the other category labels, $\beta = 0.05$, $z = 0.73$, $p = 0.47$. Together these results suggest that participants were 1) better able to distinguish the category of the target image from both the

immediately preceding prime and the other scene categories in coherent sequences, and 2) no more likely to select the prime than any of the other scene category labels in coherent versus randomized sequences.

We found that both rapid scene gist categorization performance and prime versus target discriminability were greater when targets were embedded in coherent than in randomized sequences. We also found evidence of within-scene category priming and between-scene category priming, though we observed more facilitation within than between scene categories. The last result is likely produced by both the shared features and conceptual overlap between scenes when the scene category does not change. We considered two possible explanations of this effect as part of an exploratory analysis. One possibility is that expectations for upcoming scenes primed and thus influenced recognition as the contents of the current event model facilitated subsequent scene categorization. An alternative possibility, as mentioned before, is that scenes may have been primed not by their expectation, but by the amount of perceptual featural overlap between images regardless of conceptual influences (though images that share layout information also tend to belong to the same scene category) (Oliva & Torralba, 2001). Repetition priming is well known to enhance perception (Bar & Biederman, 1998). For example, two different views of the same office share overlapping objects, textures, and perhaps even the same spatial layout. Furthermore, even as one navigates from an office into a hallway, many visual features between pairs of images will be more similar (e.g., rectilinear shapes, the presence of right angles, etc.) than for instance when scene categories change in the randomized sequence. When information from an image is extracted, various feature detectors become activated along the ventral visual pathway. If a subsequent scene activates the same, or nearly the same, feature detectors, then the combined activity from the first presentation with the second

may facilitate identification of the second image, producing the within- and between-category priming effect we observed regardless of sequential expectations of a scene's category (Eddy, Schmid, & Holcomb, 2006; Sperber, McCauley, Ragain, & Weil, 1979). There is less featural overlap between scene categories than within the same scene category (Oliva & Torralba, 2001). Scene categorization accuracy was greater within than between scene categories and asymptotic performance was reached after 1 scene image within categories. Importantly, within- and between-scene category priming could occur within the initial sweep of feed-forward activation through the ventral visual pathway as the prime and target share overlapping visual features (Caddigan et al., 2017; Guyader, Chauvin, Peyrin, Hérault, & Marendaz, 2004), including the critically important scene layout (Sanocki & Epstein, 1997). Thus, the extent to which the prime and target activate similar feature detectors may facilitate categorization of the target scene in the RSVP sequence at a relatively low level of visual processing (Bar & Biederman, 1998; Eddy et al., 2006). Therefore, the scene spatiotemporal coherence priming effect we thus observed could have appeared regardless of predictions generated by the event model and conceptual overlap between scene images.

Exploratory analysis of conceptual priming and image similarity. While we hypothesized both perceptual and conceptual priming effects a priori, our method of measuring visual similarity was chosen post-hoc. We estimated similarity using the spatial envelope model available for download at: http://people.csail.mit.edu/torralba/code/spatialenvelope/ (Oliva & Torralba, 2001). Work using this model has demonstrated that spectral information can be used to categorize and describe scenes according to their spatial properties derived from a relatively small set of perceptual dimensions (e.g., roughness, naturalness, openness, expansion, ruggedness). This model has also been used to investigate priming between scenes presented on

different trials (Caddigan et al., 2017). Spectral information was extracted by calculating the response to Gabor filters at three spatial frequencies and eight orientations. Filter responses were concatenated to obtain a feature vector for each image and the Euclidean distance between pairs of images provided a measure of similarity between each target and its immediately preceding prime. We provide examples of image pairs with low and high image similarity in Figure 9 in the Appendix.

We analyzed the effects of prime-to-target image similarity using a linear multilevel model with the random effect of image. We tested priming by examining the perceptual similarity between the target image and its immediately preceding image regardless of scene category. Prior to the analysis, we first took the reciprocal of image similarity so that greater values corresponded to larger similarity between pictures and then we took its natural log due to its positive skew. We removed the target images that were first in a sequence from the analysis, since there was no preceding prime image. Spatiotemporal coherence was again effect coded as Coherent = 1 and Randomized = -1. As predicted, image similarity was significantly higher in coherent (M = 6.82, SE =0.22) than in randomized (M =4.14, SE = 0.10) sequences, β = 0.21, t = 16.84, $p < 0.001$ (larger values indicate more similarity between target and prime). If the similarity between the target and prime is the reason we found higher recognition performance in the coherent compared to the randomized scene sequences, then image similarity would be expected to predict performance above and beyond the manipulation of the spatiotemporal coherence of scene images. Alternatively, if both image similarity and expectations for scene categories influenced recognition performance, then the spatiotemporal coherence would be expected to remain a significant predictor of categorization performance after controlling for the effect of image similarity in a multiple regression.

To test these alternative hypotheses, we fit a series of hierarchical multilevel logistic regressions to recognition performance. Each model contained the random effects of participant and image and only differed in the structure of their fixed effects. Parameters and estimates of each model are in Table 1. We used AIC, BIC, and chi square tests of significance to assess the quality of the statistical models. The first model contained only the scene spatiotemporal coherence as a predictor of recognition performance. Images presented in coherent sequences were identified more accurately than images presented in randomized sequences, $\beta = 0.50$, $z = 9.41$, $p < 0.001$, AIC = 2655.70, BIC = 2678.20. The second model contained only the fixed effect of log of similarity. Similarity significantly predicted recognition performance, $\beta = 0.87$, $z = 9.84$, $p < 0.001$, AIC = 2635.30, BIC = 2657.80. Scene categorization accuracy was greater when the scene immediately preceding the target and the target were similar than when they were not consistent with the hypothesis that perceptual similarity between the prime and target improved categorization performance to the target. In the third model we tested whether the spatiotemporal coherence of the narrative remained a significant predictor of categorization performance when controlling for target-prime similarity. The third model contained both the main effect of log of similarity and the effect of scene spatiotemporal coherence and both uniquely predicted recognition performance (Log of similarity: $\beta = 0.70$, $z = 7.71$, $p < 0.001$; Spatiotemporal coherence: $\beta = 0.37$, $z = 6.68$, $p < 0.001$, AIC = 2591.90, BIC = 2620.00). The third model containing both main effects provided a significantly better model fit than the model containing only the main effect of spatiotemporal coherence, $\chi^2 = 65.83$, $p < 0.001$, ΔAIC = 63.80, ΔBIC = 58.20, thus adding similarity contributed significantly to predicting recognition performance. Importantly, the model containing both main effects also provided a significantly better model fit than the model containing the log of similarity alone, $\chi^2 = 45.36$, $p < 0.001$,

ΔAIC = 43.40 ΔBIC = 37.80. The degree of coherence between scene images predicted unique variance not accounted for by image similarity, suggesting that the amount of perceptual overlap between the prime and target alone did not produce facilitation of scene gist perception. A fourth model contained both main effects and their interaction. Spatiotemporal coherence no longer significantly predicted performance, $\beta = 0.10$, $z = 0.80$, $p = 0.43$, AIC = 2589.00, BIC = 2622.80. However, similarity was still found to significantly predict rapid scene gist categorization performance, $\beta = 0.68$, $z = 7.37$, $p < .001$, AIC = 2589.00, BIC = 2622.80. Furthermore, image similarity's contribution to predicting categorization performance differed when the image was presented in a coherent versus a randomized sequence, $\beta = 0.19$, $z = 2.21$, $p = 0.03$, AIC = 2589.00, BIC = 2622.80. The model containing the interaction between spatiotemporal coherence and the log of similarity provided a significantly better model fit than the model including both main effects alone, $\chi^2 = 4.89$, $p = 0.03$, ΔAIC = 2.90, ΔBIC =2.7; however, the difference in the AIC values for the two models suggests that adding the interaction term only moderately improved model fit, and a comparison of the more conservative BIC estimates revealed that adding the interaction term hurt model fit. As such, the more parsimonious model containing only the main effects was retained and plotted in Figure 5. As shown in the figure, images that had very similar primes embedded in coherent sequences were identified more accurately than images that were very similar but shown in randomized sequences. These results are consistent with hypotheses generated from SPECT that the extent to which facilitation will be found depends upon the degree of spatio-temporal coherence between the contents of the current event model and new scene information.

Table 1. *Model comparisons for scene gist categorization performance*

Model	AIC	BIC	Description	Fixed effects	β	*SE* β	z value	*p* values
Model 1	2655.70	2678.20	Main effect only	Intercept	-0.38	0.09	-4.14	<0.001
				SC	0.50	0.05	9.41	<0.001
Model 2	2635.30	2657.80	Main effect only	Intercept	-1.62	0.16	-10.05	<0.001
				LIS	0.87	0.09	9.84	<0.001
Model 3	2591.90	2620.00	Both main effects	Intercept	-1.38	0.16	-8.46	<0.001
				LIS	0.70	0.09	7.71	<0.001
				SC	0.37	0.06	6.68	<0.001
Model 4	2589.00	2622.80	Both main effects	Intercept	-1.39	0.16	-8.44	<0.001
			and interaction	LIS	0.68	0.09	7.37	<0.001
				SC	0.10	0.13	0.80	0.43
				LIS X SC	0.19	0.08	2.21	0.03

Note: SC = Spatiotemporal coherence of scene sequences, LIS = Log image similarity. The models were conducted by effect coding spatiotemporal coherence as coherent = 1 and randomized = -1.

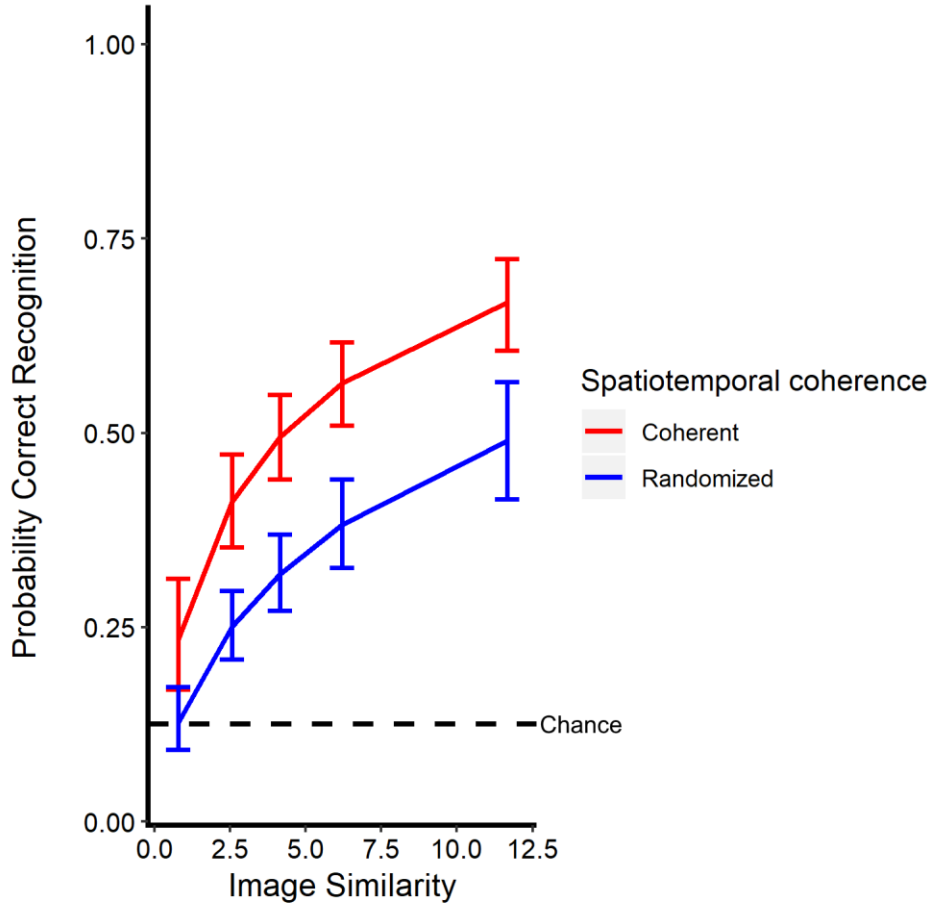

Figure 5. Rapid scene gist categorization accuracy as a function of image similarity and spatiotemporal coherence of scene sequences. Error bars represent 95% confidence intervals around the estimated mean probability correct.

Discussion

The roles that top-down and bottom-up processing play in image recognition remains critical to theories of cognition and perception, and the unique contribution of each remains largely unknown and heavily disputed (Bar, 2004; Bar & Ullman, 1996; Davenport & Potter,

2004; Firestone & Scholl, 2016; Hollingworth & Henderson, 1998)[3]. According to data-driven feed-forward processing accounts, expectations for upcoming scenes should not influence their recognition as perception of scene gist is accomplished by purely feed-forward mechanisms (Bastin et al., 2013; Perrinet et al., 2004; Serre et al., 2007; Thorpe et al., 1996; VanRullen & Thorpe, 2002). If perception is purely feed-forward, then it cannot be contaminated by higher-order processes such as predictions as visual processing is largely driven by the patterns of light hitting the eyes, not by what one expects to see. Nevertheless, in the current experiment, it was found that sequential expectations for scene categories influenced recognition performance, and it did so when the same scene category repeated across multiple views (i.e., Figure 3) as well as when the target image changed to an expected, but different scene category (i.e., Figure 4). Furthermore, priming from similar images based upon shared spectral information between target and prime enhanced accuracy more for targets within coherent sequences than randomized ones (i.e., Figure 5). These results are consistent with a view that perception is the result of a synthesis between bottom-up and top-down factors (Brewer & Loschky, 2005; Bruner, 1957).

Nevertheless, it remains unknown if predictability contributed to rapid scene gist categorization performance above and beyond target and prime similarity. Indeed, one could argue that image similarity may actually have explained all of our results in Experiment 1.

---

[3] Firestone and Scholl (2016) limit their definition of perception to early visual processing or "pure perception," excluding recognition processes (e.g., object recognition, scene category recognition, etc.) from their definition. This is contrary to standard definitions of perception, such as "the act of giving meaning to a detected sensation" (Wolfe et al., 2012, p. 467), "the later steps in the perceptual process, whereby the initial sensory signals are used to represent objects and events so they can be identified, stored in memory, and used in thought and action" (Yantis & Abrams, 2017, p. G-9), or "the acquisition and processing of sensory information in order to see, hear, taste, smell, or feel objects in the world" (Blake & Sekuler, 2006, p. 578) . Furthermore, as mentioned previously, there is data to suggest that processes that result in scene category recognition are activated simultaneously with neural processes that underly the early visual information representation (Ramkumar et al., 2016).

Specifically, we have already noted that image similarity was less for targets and primes between scene categories than within scene categories, which could explain the smaller effect of spatio-temporal coherence for the former than the latter. Furthermore, randomly paired images in the sequences (in the randomized sequences) had lower visual similarity than spatio-temporally connected images in the coherent condition, again potentially explaining our coherence effect. Thus, we need a measure of predictability that is separate from our manipulation of spatio-temporal coherence. Put differently, a remaining question unaddressed by Experiment 1 is whether the images in coherent sequences were, in fact, more predictable than the images in a randomized sequence. Experiment 2 addressed these issues.

Chapter 3 - Experiment 2

Experiment 2 served two purposes. In experiment 2, we explored the extent to which scene categories in coherent and randomized sequences were predictable. One possibility, is that scene categories become predictable as one's understanding for the current event is generated in working memory upon viewing the initial image in a sequence. The presentation of the first scene lays the foundation for the current event model (Loschky et al., 2018; Loschky et al., submitted). Furthermore, consistent with Event Segmentation Theory (EST), event models generate predictions for what will happen next and such conceptual predictions influence online perceptual processing (Zacks, Speer, Swallow, Braver, & Reynolds, 2007). The quality of perceptual predictions critically depends upon whether one's current event model is a good fit to what is happening. After event models are constructed, upcoming scene categories should be more predictable than if no event models are constructed for the scene sequences as would be the case when the order scene images are presented in is random. The second purpose of Experiment 2 was to investigate the extent to which scene category predictability influenced recognition performance found in Experiment 1 and whether it did so uniquely, compared to similarity between target and prime as defined by the amount of overlap in spectral information between the target and its immediately preceding image. Two hypotheses therefore suggest themselves. First, if target images presented in coherent sequences were more predictable in Experiment 1, then prediction accuracy should be greater for target scene categories embedded in spatiotemporally coherent than randomized sequences. Second, if target images in randomized sequences are less predictable, they should be predicted at a rate no better than chance, since participants should have no clue as to which scene will be next when the sequence is

randomized.[4] SPECT would argue that this is because viewers cannot generate a coherent event model when viewing a randomized image sequence. If these two hypotheses are supported, other more interesting hypotheses are suggested. If participants in Experiment 1 used predictions of up-coming items to correctly answer the 8-AFC task, then the predictability of targets in Experiment 2 should significantly explain variance in a regression model of Experiment 1 performance. An important question would then be whether such explained variance due to target predictability, which is purely conceptually driven, would be independent of the variance in recognition explained by visual similarity of the target and its preceding item. Given that items that look alike in terms of Spatial Envelope Model scene descriptors tend to be semantically similar (Oliva & Torralba, 2001), it is also possible that target predictability and prime-to-target perceptual similarity share considerable variance, and therefore the effect of spatiotemporal coherence observed in Experiment 1 may solely be due to the shared activation of feature detectors activated by both the prime and target. However, if the effects of visual similarity are primarily operating at low-level information extraction processes in the front-end, and target predictability effects are primarily due to higher-level back-end event model processes influencing front-end information extraction, then their explained variance in categorization performance may be largely independent. Predictions for an upcoming scene category generated from one's prior understanding of the previous scene categories shown and general knowledge for the kinds of scene categories that tend to follow one another may prime representations associated with the predicted scene, which could influence rapid scene categorization

[4] One way prediction accuracy could greater than chance in the randomized condition would be if one were able to keep track of the scene categories shown in each sequence in the RSVP stream in each trial, and use that to guess the likelihood of seeing the target category using a process of elimination to exclude more frequently seen categories. Because pictures of scenes were presented for only 300 ms, this hypothesis seems unlikely.

performance independent of perceptual similarity between prime and target. Finally, an interesting question is the degree to which consciousness of the coherence manipulation is necessary for prediction accuracy. This question has bearing on whether the effects of image predictability on rapid scene gist categorization are conscious or unconscious.

Method

Participants. Fifty-nine participants from Kansas State University's undergraduate research pool who did not participate in Experiment 1, participated in the experiment for course credit (35 females and 24 males, mean age = 19.63). Participants' vision was tested prior to participation and was 20/30 or better as determined by the Freiburg Visual Acuity and Contrast Test (FrACT) (Bach, 2006). All participants were naïve to the purpose of the experiment and signed an informed consent prior to participating. After participants completed the experiment, they were instructed not to discuss the experiment with friends and fellow classmates, because we asked their awareness of the coherence manipulation at the end of the experiment.

Procedure. The procedure and design was identical to Experiment 1 with the following exceptions: The target image in each sequence was removed and replaced by a neutral gray screen for 24 ms. Images were all presented in the same order as Experiment 1 except that in Experiment 2, participants' task was to identify the scene category that they predicted *would have been* presented prior to the perceptual mask by selecting it from the 8-AFC array of scene category labels. After the experiment, participants were asked a series of questions related to whether they had noticed the differences in coherence between images in the experiment. Participants were asked the following four questions in this order: 1) "Did anyone tell you anything about this study?" 2) "Did you notice anything in the experiment?" 3) "Did you notice anything about the sequence the images were shown in?" and 4) "Did you notice that some of the

images were unexpected?". If participants responded positively to any of the questions, they were asked to type their answers to the questions in a textbox on the computer screen. None of the participants responded that anyone had told them anything about the manipulation of the experiment prior to participating.

Experiment 2 Results

To test the first hypothesis that images presented in coherent sequences were predictable, a logistic multilevel model was used to predict participants' accuracy of predicting the basic level category of the target scene on each trial as a function of the spatiotemporal coherence manipulation. Scene spatiotemporal coherence was effect coded as Coherent = 1 and Randomized = -1 and image location was again effect coded as on-campus = -1 and off-campus = 1. Consistent with our hypothesis, prediction accuracy was found to be reliably greater for images presented in coherent (M = 26.64%, SE = 1.18%) than those in randomized sequences (M = 15.47%, SE = 0.96%), $\beta = 0.35$, $z = 7.13$, $p < 0.001$. Surprisingly, even though the on-campus images were recognized more accurately in Experiment 1 (i.e., Figure 2), prediction accuracy was found to be greater for off-campus (M = 23.38%, SE = 1.13%) than on-campus image sequences (M = 18.73%, SE = 1.04%), $\beta = 0.14$, $z = 2.61$, $p = 0.009$. Spatiotemporal coherence and image location did not interact, $\beta = 0.06$, $z = 1.25$, $p = 0.21$. Importantly, as hypotheseized, prediction accuracy for images presented in randomized sequences was found to be very close to the chance level accuracy of 12.5%. The estimated mean for the off-campus randomized images was equal to 15.63%, 95% CI [12.95% – 18.75%] and it was estimated to be 13.77%, 95% CI [11.25% - 16.74%] for on-campus randomized images. This result is largely consistent with our hypothesis that the predictability of items in the randomized condition should be difficult when the viewer cannot construct a coherent event model to represent the image sequence.

As in Experiment 1, we conducted a series of analyses examining participant responses. Because primes were processed for 300 ms, participants may have responded with the category label that matched the prime immediately prior to the mask, rather than the category label of the target scene. Such a response bias could have produced greater prediction performance for scenes shown in coherent sequences. To assess this possibility, we ran three multilevel logistic regression models with the category label of the participant's response (i.e., target vs. prime, target vs. other, and prime vs. other) as the criterion. As in Experiment 1, we removed all cases when the target was the first image in a sequence and all instances when the category of the prime matched the target. We modeled responses using a fixed effect of spatiotemporal coherence, which was effect coded as coherent = 1 and randomized = -1. As in Experiment 1, the three models only differed in the criterion. If participants made predictions as to the target's identity by selecting the category label that matched the prime regardless of the spatiotemporal coherence of sequences, then responses with the category label of the prime should not differ between coherent and randomized sequences. Alternatively, if participants used the spatiotemporal coherence of the scene images to generate predictions of future scene categories, then participants would be expected to be more likely to select the category label of the target than the prime in coherent sequences. In the first analysis, a response with the target was dummy coded as a 1 and responses with the prime category were coded as a 0. Consistent with the recognition results of Experiment 1, participants successfully discriminated predicted (but unseen) targets from primes (which they did see), $\beta = 0.19$, $z = 2.28$, $p = 0.02$. Thus, even though the target images were never shown in the task, participants were more likely in coherent sequences to respond with the category label that matched the target scene than they were to respond with the label that matched the prime (which they did see). In the second analysis, a

response with the target was dummy coded as a 1 and responses with one of the other (non-prime labels) were coded as a 0. Consistent with the recognition results of Experiment 1, participants were significantly more likely to select the category label that matched the target than one of the other six categories in coherent sequences, $\beta = 0.44$, $z = 6.32$, $p < 0.0001$. Finally, we examined errors by assessing whether participants responded more frequently with either the prime category label or one of the other category labels in coherent versus randomized sequences. Inconsistent with the results of Experiment 1, participants were more likely in coherent sequences to select the label that matched the prime that immediately preceded the unseen but predicted target than one of the other six category labels, $\beta = 0.25$, $z = 3.80$, $p < 0.001$. When participants made errors, they were more likely in coherent than in the randomized sequences to respond with the prime than they were to respond with the label of one of the other six categories. Although this finding is inconsistent with the recognition results from Experiment 1, it is not surprising. Specifically, participants were more likely to see multiple views of the same scene category in a row in the coherent than in randomized sequences. This would make them more likely to predict that they would see the same image as the prime in a coherent sequence than in a randomized sequence.

Analysis of image predictability and image similarity. Participants in Experiments 1 and 2 were shown the same images in the same sequence, with the exception that the target was removed in Experiment 2 and participants' task was to predict what the image would have been if it were present. The design of Experiment 2 therefore allows us to investigate how the predictability of each target scene in the experimental sequences can influence recognition in an independent sample (i.e., Experiment 1). To evaluate the extent to which image predictability contributed to recognition performance independently of prime-to-target image similarity, a

series of multilevel logistic regressions were conducted. Models were compared using AIC and BIC values as well as chi square tests of significance. Model estimates are provided in Table 2. On each trial, a participant in Experiment 2 either correctly predicted the basic level category of the target scene or they did not. In Experiment 1, the same target image was either correctly categorized or it was not. All scene images were presented in the same order in Experiments 1 and 2 with the same target scenes embedded in the same coherent and randomized sequences (i.e., image sequences in Experiment 2 were yoked to those in Experiment 1). To examine how well image predictability in Experiment 2 explains rapid scene categorization accuracy in Experiment 1, image predictability from Experiment 2 was used as a predictor of scene categorization accuracy in Experiment 1. Predictability of each target image in Experiment 2 was effect coded as either the target image was predictable = -1 or not = 1. The image was treated as a random effect at its intercept. Participant was not treated as a random effect because the samples taken for Experiments 1 and 2 were from a different group of participants. For this analysis, data from the last 11 participants (49 - 59) in Experiment 2 were removed so that we could match the number of observations in Experiments 1 and 2, so that each target image within the experiments was both 1) either correctly categorized or not, and 2) was either correctly predicted or not. As shown in Figure 6 and Model 1 in Table 2, image predictability influenced recognition performance such that scene images that were predictable (M = 51.88%, SE = 2.36%) were identified in Experiment 1 more accurately than those that were not (M = 39.50%, SE = 1.22%), β = -0.26, z = -4.45, $p < 0.001$, AIC = 2758.00, BIC = 2774.90. Scene categories that were more predictable from one sample of participants were more accurately categorized in another.

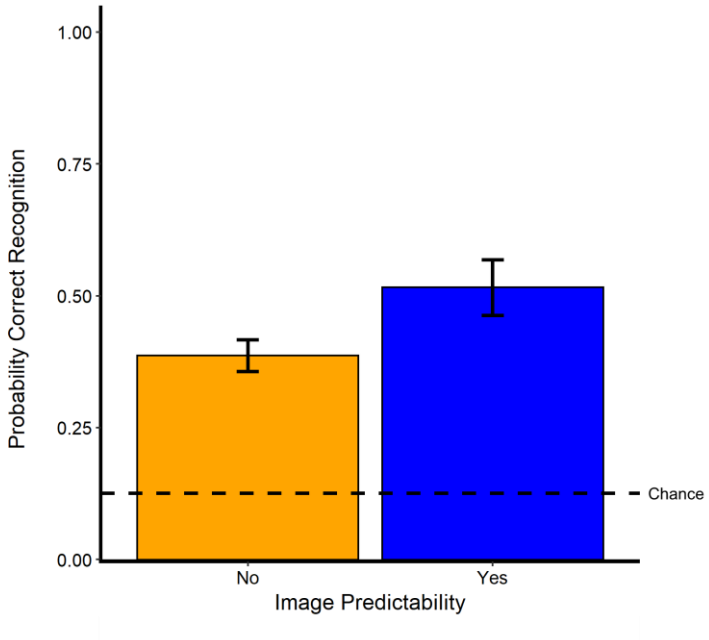

Figure 6. Rapid scene gist categorization accuracy as a function of image predictability. Error bars represent 95% confidence intervals around the estimated mean probability correct.

Table 2. *Model comparisons for recognition performance*.

Model	AIC	BIC	Description	Fixed effects	β	*SE* β	z value	*p* values
Model 1	2758.00	2774.90	Main effect only	Intercept	-0.20	0.07	-3.04	0.002
				IP	-0.26	0.06	-4.45	<0.001
Model 2	2669.40	2686.30	Main effect only	Intercept	-1.53	0.14	-10.98	<0.001
				LIS	0.81	0.08	9.65	<0.001
Model 3	2661.00	2683.50	Both main effects	Intercept	-1.37	0.15	-9.42	<0.001
				LIS	0.78	0.08	9.26	<0.001
				IP	-0.20	0.06	-3.24	0.001
Model 4	2663.00	2691.10	Both main effects	Intercept	-1.35	0.17	-7.859	<0.001
			and interaction	LIS	0.77	0.10	7.57	<0.001
				IP	-0.23	0.16	-1.45	0.15
				LIS * IP	0.02	0.10	0.21	0.84

Note: IP = Image predictability, LIS = Log image similarity

Model 2 contained only the log of image similarity and Model 3 contained both the main effect of log of image similarity and that of image predictability. Consistent with the results of Experiment 1, log image similarity significantly contributed to recognition accuracy, $\beta = 0.81$, $z = 9.65$, $p < 0.001$, AIC = 2669.40, BIC = 2686.30, likely due to the sharing of activation of feature detectors between target and prime. In Model 3, similarity again significantly predicted categorization accuracy, $\beta = 0.78$, $z = 9.26$, $p < 0.001$, AIC = 2661.00, BIC = 2683.50, but, critically, image predictability also remained a significant predictor of categorization accuracy, $\beta = -0.20$, $z = -3.24$, $p = 0.001$, AIC = 2661.00, BIC = 2683.50. A comparison of Model 3, which

contained both main effects and Model 1, which contained only image predictability showed that adding the log of image similarity significantly contributed to recognition performance, χ^2 = 99.00, $p < 0.001$, ΔAIC = 97.00, ΔBIC =91.4. Likewise, comparing Model 3 and Model 2, which contained only the main effect of image similarity, showed that adding image predictability also significantly contributed to recognition accuracy, $\chi^2 = 10.42$, $p = 0.001$, ΔAIC = 8.40, ΔBIC = 2.8. Participants in Experiment 2 never saw the target image, but were asked to identify what it would have been if it were present. The independence between predictability and image similarity suggests that image similarity influences rapid scene categorization by sharing the activation of feature detectors (captured by spatial envelope model scene descriptors) between both prime and target, and is primarily a front-end perceptually driven process. Conversely, image predictability influences recognition performance from the top-down and is purely a back-end event model, conceptually-driven process. In Model 4, rapid scene gist categorization accuracy was fit to both main effects and their interaction. The results showed that image predictability no longer predicted recognition accuracy, $\beta = -0.23$, $z = -1.45$, $p = 0.15$, but image similarity remained a significant predictor, $\beta = 0.77$, $z = 7.57$, $p < 0.001$. However, the interaction between image predictability and similarity was not significant, $\beta = 0.02$, $z = 0.21$, $p = 0.84$, AIC = 2663.00, BIC = 2691.10. A comparison of Model 4, which contained the interaction, with Model 3, which only contained both main effects, showed that they were not significantly different, $\chi^2 = 0.04$, $p = 0.84$, ΔAIC = 2, ΔBIC = 7.60. Furthermore, according to the BIC values, the model containing the interaction, was a worse fit of the data than the model containing only both main effects. For parsimony, we retained Model 3, with both main effects but no interaction term, which is plotted in Figure 7. As shown in Figure 7, images were more accurately recognized when they were visually similar, but regardless of the degree of similarity,

images were more accurately recognized when they were predictable. Furthermore, as with the effect of the spatiotemporal coherence of the multiple scene images (as shown in Figure 5) the rate of change in recognition accuracy as a function of image similarity did not depend upon image predictability. Nevertheless, by treating image predictability categorically, image similarity appears to be the best predictor of categorization performance. As shown in Figure 7, the estimated difference in performance between images that were versus were not predictable was relatively small given the size of the estimated confidence intervals around the mean estimates.

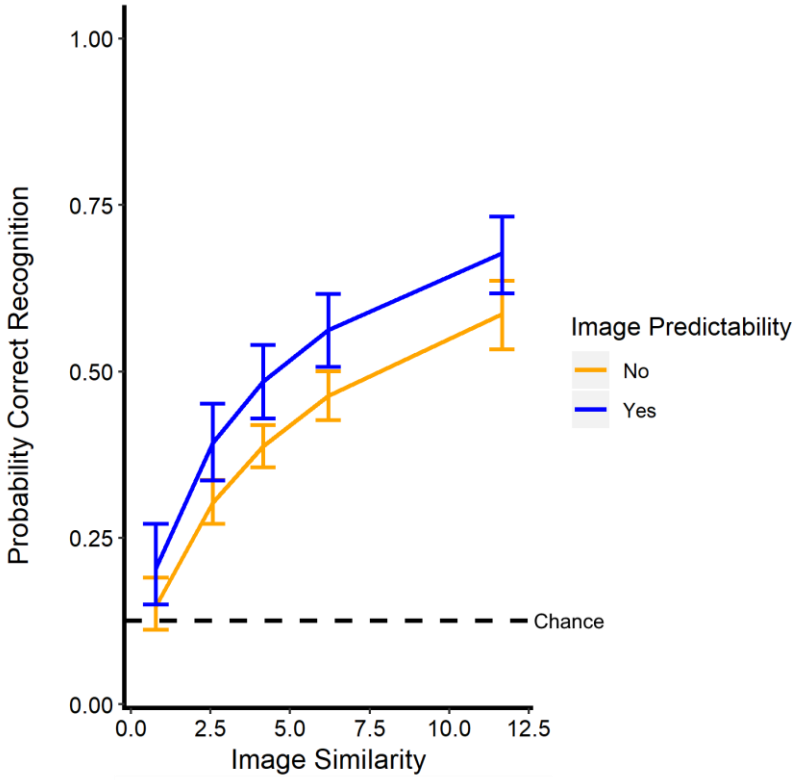

Figure 7. Rapid scene gist categorization accuracy as a function of image similarity and image predictability. Error bars represent 95% confidence intervals around the estimated mean probability correct**.**

Perhaps one reason that the effect of visual similarity was found to be larger than that of image predictability is because the former was a continuous variable, while the latter was categorical. Thus, we made image predictability continuous by calculating the aggregate of image predictability accuracy across all 24 sequences from Experiment 2 and running a series of partial correlations. Across all 59 participants, image predictability accuracy was aggregated for each of the sequences for images presented in coherent and randomized sequences separately. Rapid scene categorization accuracy was also aggregated across all 24 of the sequences using scene gist categorization performance in Experiment 1. This results in each of the 24 randomized and 24 coherent sequences being associated with a value that corresponds to 1) how predictable images were in it and 2) how accurate scene images were categorized in it. Likewise, the average similarity between the target and its immediately preceding prime within each of the 24 sequences was also aggregated separately for both coherent and randomized sequences. Zero order correlations are provided in Table 3. Controlling for prediction accuracy, image similarity was found to significantly correlate with recognition performance, pr (45) = 0.36, p = 0.01. The more similar target images were to their primes, the more accurate rapid scene categorization was even when the effect of image predictability was partialled out. Importantly, though, controlling for image similarity, prediction accuracy was also found to significantly correlate with rapid scene categorization performance, pr (45) = 0.33, p = .02, to almost the same extent as image similarity. These results suggest two independent priming effects at work. One is due to the degree of image feature overlap influencing priming from the bottom up, likely based on the degree to which the prime and target activate similar feature detectors. The other is due to the degree of scene spatiotemporal coherence influencing conceptual priming from the top down, likely based on the degree to which the current event model allows accurate predictions of the

target scene. Importantly, neither type of priming alone can account for the results we found in Experiment 1. Nevertheless, these are correlational analyses, which limits the strength of the causal conclusions we can draw from them.

Table 3. *Zero order correlations between rapid scene categorization, image prediction performance, and image similarity.*

Variables	*M*	*SE*	1	2	3
Categorization accuracy	43.68%	2.05%	-		
Prediction accuracy	21.05%	1.35%	0.4431	-	
Image similarity	5.48	0.12	0.4611	0.3735	-

Note: Image similarity is the reciprocal of the difference between the gist descriptor of the target image and its immediately preceding prime as output from the results of the spatial envelope model (Oliva & Torralba, 2001).

Lastly, we conducted an independent samples *t* test to investigate whether consciously reporting that images were presented in a coherent sequence influenced prediction accuracy. Two raters independently judged from the third question whether each participant reported anything regarding the coherence of the sequence of images. Raters produced strong interrater reliability (Cohen's kappa = 0.98) and resolved discrepancies through thoughtful discussion. Surprisingly, only 38.98% of participants reported that they noticed the coherence of the coherent scene presentations. Surprisingly, we found that noticing the coherence of the images was not a requirement for good prediction performance, t (57) = 1.41, p = 0.17. This surprising result could have a variety of explanations. One possibility is that participants in the experiment

were truly unaware of the spatiotemporal coherence, and awareness is not needed for accurate predictions. An alternative explanation is that the questions we asked were too vague for participants to understand what we were asking them to report.

Discussion

In our day-to-day lives, the scenes we see are typically primed. We rarely, if ever, find ourselves confronted with unexpected scenes. Instead, we tend to either have prior knowledge of the scene category we will encounter from one moment to the next prior to seeing them, or we make (likely unconscious) predictions about what we will see. This is the first study to investigate the extent to which rapid scene categorization could be influenced by sequential expectations of related scene categories. The degree to which bottom-up versus top-down processes interact with online perception has recently found a renewed interest (Firestone & Scholl, 2016). According to SPECT, one's understanding for what is happening in the current event (e.g., navigating through one's office toward the doorway) in the back-end influences information extraction in the front-end (Loschky et al., 2018). In Experiment 1, we found that scenes presented in coherent sequences, which would enable viewers to create a spatiotemporally coherent event model, were identified more accurately than scenes presented in randomized sequences. This result is consistent with the hypothesis generated from SPECT that the contents of the current event model influence front-end information extraction. In Experiment 2, we found that images presented in coherent sequences were more predictable than images presented in randomized sequences, which were at chance. Importantly, predictability of upcoming scenes in the sequence accounted for unique variance in recognition performance in Experiment 1, beyond that of visual similarity between the target and its immediately preceding prime as measured using the Spatial Envelope Model (Oliva & Torralba, 2001). This critical result

suggests that the accuracy advantage for scenes in coherent sequences found in Experiment 1 cannot be explained solely by low-level perceptual processes involved in processing similar features between the target and its immediately preceding image. Being able to predict an upcoming scene category in Experiment 2 was purely conceptual, as participants never saw the target scene. Thus, the finding that image predictability influenced rapid scene categorization performance when controlling for image similarity reflects the influence of the current event model on scene gist extraction. This was true after both small featural changes (e.g., 1st and then the 2nd view of an office) and large featural changes (last view of an office and the 1st view of a hallway).

Chapter 4 - General Discussion

This is the first experiment to investigate how sequential expectations for upcoming scene categories can influence rapid scene categorization. ERP data and computer simulation studies suggest that scene gist perception is an extremely quick and efficient process completed solely by feed-forward mechanisms (Bastin et al., 2013; Perrinet et al., 2004; Serre et al., 2007; VanRullen & Thorpe, 2002). Consistent with such a view, it would be expected that scene categorization may not be influenced by later cognitive processes, such as familiarity (Fabre-Thorpe et al., 2001), and attention (Li et al., 2002), as perception may be the result of more passive processes with information processed at one stage unidirectionally feeding into later stages, not bidirectionally. Consistent with this idea, participants were remarkably good at identifying the gist of scenes presented in randomized sequences (recognition accuracy is well above chance in Figure 2) despite being unable to predict upcoming scenes; therefore, a large part of scene gist perception is driven from the bottom-up. However, in contrast to the feed-forward gist hypothesis, we found that by presenting scenes in coherent, and thus more ecologically valid sequences, expectations for upcoming scenes influenced the processing of scene gist.

Although the findings of the present research provide important new information, future research should examine the extent to which the priming effect we observed is the result of changes in later decision-making stages or early perceptual sensitivity. For instance, if one expects to see a hallway after being shown multiple views of an office, then one's prediction may influence one's response independent of the target image shown (i.e., participants may false alarm and report they saw a hallway regardless of the category of the target). Such an effect would be consistent with prior research that has found that objects in consistent scenes (e.g., a

mixer in a kitchen) are not identified with greater sensitivity than objects in inconsistent scenes (e.g., a mixer in a barnyard), but instead participants have a bias to respond positively to category labels consistent with schematic information available from the presentation of a scene (e.g., chicken in a farmyard) (Hollingworth & Henderson, 1998). As such, scene spatiotemporal coherence priming may be the result of changes in later cognitive processing stages such that expectations influence what participants' *say*, not what they *see*. Alternatively, when a viewer is shown multiple views of an office which appear to be navigating toward a doorway, perceptual predictions of the current event model aided by the activation of schemata information may influence perception of subsequent scenes, resulting in scenes consistent with predictions generated by the current event model being identified with greater sensitivity. Furthermore, even if event model construction processes make subsequent scenes more perceptible, the locus of such an effect remains unknown. There is work showing that predictive coding can influence perception as early as V1 (Muckli et al., 2015). Alternatively, predictions for upcoming scenes may operate at the matching stage of scene identification by lowering the threshold amount of activation in semantic memory needed to match the perceptual input to a stored representation of a variety of stored scene categories in memory (Bar, & Ullman, 1996; Friedman, 1979). Rather than influencing visual processing, priming from expectations may make the relevant semantic memory representations easier to retrieve, which could lower their activation thresholds. Lowering activation thresholds for a constructed visual description of a scene to match a stored description of a scene type in memory, may require less visual information to be extracted to attain a match, thus facilitating recognizing information consistent with expectations (Bar & Ullman, 1996; Friedman, 1979).

SPECT assumes that processes involved in the back-end influence processing in the front-end. One prediction along those lines is that the front-end process of information extraction is influenced by the current event model. *Broad information extraction* applies to the entire scene and includes scene gist recognition. Therefore, one would predict that sensitivity in rapidly categorizing a hallway image after seeing multiple views of an office would be improved. However, it is also possible that the effects of the event model on accuracy in our rapid scene categorization task are due to post-perceptual processes, such as response biases or even intelligent guessing. We rather doubt the intelligent guessing explanation due to the fact that roughly 60% of participants in Experiment 2 were unaware of the coherence manipulation, and those that reported being aware of the coherence manipulation were no better at accurately predicting scene categories than those that did not report awareness. Nevertheless, it may be possible to have unconscious response biases (Engen, 1972). Unfortunately, due to the 8-AFC response procedure used in the current design, assessing response bias is difficult (DeCarlo, 1998). Thus, further work is needed to tease apart the possible mechanisms underlying the effects of predictability and visual similarity on rapid scene categorization, as shown in our study.

One possible method that address these issues comes from previous research, which has shown that perceptual sensitivity (d') in distinguishing real- versus phase-randomized scenes is greater for probable than improbable scenes (Greene et al., 2015), and greater for scenes that are good exemplars of their category than poor exemplars (Caddigan et al., 2017). Thus, in the proposed study, we would replace half of the targets in the coherent and randomized sequences with target images that are phase-randomized, namely images that have the same Fourier amplitude spectrum statistical properties as the targets, but are unrecognizable (Loschky &

Larson, 2008; Loschky et al., 2007). One could then give viewers the task of discriminating real scenes versus phase-randomized scenes, and measure their perceptual sensitivity (d'). The real versus phase-randomized scenes could be embedded in coherent or randomized sequences. Importantly, in contrast to the current study's 8-AFC scene categorization task, there is no plausible reason why expectations regarding the up-coming scene category generated by the coherent sequences should produce a bias to respond either "real" or "phase-randomized" since any scene category, whether predictable or not, is "real." Thus, this method should enable us to establish if the current study's benefits in identifying scenes presented in coherent sequences originated from predictions influencing a) perceptual analysis of the scenes, which would produce an increase in d', or b) a later response-related decision-making stage, which would not influence d'.

References

Agresti, A. (2007). Building and applying logistic regression models. *An introduction to categorical data analysis*, 137-172.

Bach, M. (2006). The Freiburg Visual Acuity Test-Variability unchanged by post-hoc re-analysis. *Graefe's Archive for Clinical and Experimental Ophthalmology, 245*(7), 965-971. doi:10.1007/s00417-006-0474-4

Bacon-Mace, N., Mace, M. J., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research, 45*, 1459-1469.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617.

Bar, M., & Biederman, I. (1998). Subliminal Visual Priming. *Psychological Science, 9*(6), 464-468. doi:10.1111/1467-9280.00086

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception, 25*(3), 343-352.

Bastin, J., Vidal, J. R., Bouvier, S., Perrone-Bertolotti, M., Bénis, D., Kahane, P., . . . Epstein, R. A. (2013). Temporal Components in the Parahippocampal Place Area Revealed by Human Intracerebral Recordings. *33*(24), 10123-10131. doi:10.1523/JNEUROSCI.4646-12.2013 %J The Journal of Neuroscience

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Blake, R., & Sekuler, R. (2006). Perception (5th ed.). New York: McGraw-Hill.

Brewer, W. F., & Loschky, L. (2005). Top-down and bottom-up influences on observation: Evidence from cognitive psychology and the history of science. *Cognitive penetrability of perception: Attention, action, strategies, and bottom-up constraints*, 31-47.

Bruner, J. (1957). On perceptual readiness. *Psychological Review, 64*, 123-152.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research, 33*(2), 261-304.

Caddigan, E., Choo, H., Fei-Fei, L., & Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categories. *Journal of Vision, 17*(1), 21-21.

Castelhano, M. S., & Pollatsek, A. (2010). Extrapolating spatial layout in scene representations. *Memory & Cognition, 38*(8), 1018-1025.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. In: Mahwah, NJ: erlbaum (3rd ed.).

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559-564.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*(2), 186-205.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research, 19*(2), 103-113.

Eddy, M., Schmid, A., & Holcomb, P. J. (2006). Masked repetition priming and event-related brain potentials: A new approach for tracking the time-course of object perception. *Psychophysiology, 43*(6), 564-568.

Engen, T. (1972). The effect of expectation on judgments of odor. *Acta Psychologica, 36*(6), 450-458. doi:https://doi.org/10.1016/0001-6918(72)90025-X

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171-180.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition, 12*(6), 893-924.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*.

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology. General, 108*(3), 316-355.

Gordon, R. D. (2004). Attentional allocation during the perception of scenes. *Journal Of Experimental Psychology-Human Perception And Performance, 30*(4), 760-777. doi:10.1037/0096-1523.30.4.760

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

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20*(4), 464-472.

Guyader, N., Chauvin, A., Peyrin, C., Hérault, J., & Marendaz, C. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude-based process. *Comptes Rendus Biologies, 327*, 313-318.

Hansen, B. C., & Loschky, L. C. (2013). The contribution of amplitude and phase spectra defined scene statistics to the masking of rapid scene categorization. *Journal of Vision, 13*(13), 1–21. doi:10.1167/13.13.21

Hillstrom, A. P., Scholey, H., Liversedge, S. P., & Benson, V. (2012). The effect of the first glimpse at a scene on eye movements during search. *Psychonomic Bulletin & Review, 19*(2), 204-210. doi:10.3758/s13423-011-0205-7

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*(4), 398-415.

Intraub, H. (2010). Rethinking scene perception: A multisource model. In *Psychology of Learning and Motivation* (Vol. 52, pp. 231-264): Elsevier.

Intraub, H., & Dickinson, C. A. (2008). False memory 1/20th of a second later: What the early onset of boundary extension reveals about perception. *Psychological Science, 19*(10), 1007-1014.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47*(26), 3286-3297. doi:10.1016/j.visres.2007.09.013

Kranczioch, C., & Thorne, J. D. (2013). Simultaneous and preceding sounds enhance rapid visual targets: Evidence from the attentional blink. *Advances in Cognitive Psychology, 9*(3), 130.

Larson, A. M. (2012). *Recognizing the setting before reporting the action: Investigating how visual events are mentally constructed from scene images.* (Ph.D. Ph.D. Dissertation), Kansas State University,

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99*(14), 9596-9601. doi:10.1073/pnas.092277599

Loschky, L. C., Hutson, J. P., Smith, M. E., Smith, T. J., & Magliano, J. P. (2018). Viewing Static Visual Narratives Through the Lens of the Scene Perception and Event Comprehension Theory (SPECT). In J. Laubrock, J. Wildfeuer, & A. Dunst (Eds.), *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods* (pp. 217-238): Routledge.

Loschky, L. C., & Larson, A. M. (2008). Localized information is necessary for scene categorization, including the Natural/Man-made distinction. *Journal of Vision, 8*(1), 4:1-9.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made prior to basic-level distinctions in scene gist processing. *Visual Cognition, 18*(4), 513-536.

Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (submitted). The scene perception and event comprehension theory (SPECT). *Topics in Cognitive Science*.

Loschky, L. C., Ringer, R., Ellis, K., & Hansen, B. C. (2015). Comparing rapid scene categorization of aerial and terrestrial views: A new perspective on scene gist. *Journal of Vision, 15*(6:11), 1–29. doi:doi:10.1167/15.6.11

Loschky, L. C., Sethi, A., Simons, D. J., Pydimari, T., Ochs, D., & Corbeille, J. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception & Performance, 33*(6), 1431-1450.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco, CA: W. H. Freeman.

Muckli, L., De Martino, F., Vizioli, L., Petro, Lucy S., Smith, Fraser W., Ugurbil, K., . . . Yacoub, E. (2015). Contextual Feedback to Superficial Layers of V1. *Current Biology, 25*(20), 2690-2695. doi:10.1016/j.cub.2015.08.057

Oliva, A. (2005). Gist of a scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251-256). Burlington, MA: Elsevier Academic Press.

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

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, Special Issue on Visual Perception, 155*, 23-36.

Perrinet, L., Samuelides, M., & Thorpe, S. J. (2004). Sparse spike coding in an asynchronous feed-forward multi-layer neural network using matching pursuit. *Neurocomputing, 57*, 125-134. doi:https://doi.org/10.1016/j.neucom.2004.01.010

Pezdek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal Of Experimental Psychology-Learning Memory And Cognition, 15*(4), 587-595.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning & Memory, 2*(5), 509-522.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111-163.

Ramkumar, P., Hansen, B. C., Pannasch, S., & Loschky, L. C. (2016). Visual information representation and rapid-scene categorization are simultaneous across cortex: An MEG

study. *Neuroimage, 134*, 295-304. doi:http://dx.doi.org/10.1016/j.neuroimage.2016.03.027

Rao, R. P. N. B., D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79 - 87.

Rayner, K., Smith, T. J., Malcolm, G. L., & Henderson, J. M. (2009). Eye movements and visual encoding during scene perception. *Psychological Science, 20*(1), 6-10. doi:10.1111/j.1467-9280.2008.02243.x

Reinitz, M. T., Wright, E., & Loftus, G. R. (1989). Effects of semantic priming on visual encoding of pictures. *Journal of Experimental Psychology: General, 118*(3), 280-297.

Salin, P. A., & Bullier, J. (1995). Corticocortical connections in the visual system: structure and function. *Physiological reviews, 75*(1), 107-154.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science, 8*(5), 374-378.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104*(15), 6424-6429. doi:10.1073/pnas.0700622104

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition, 7*(5), 339-345.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766-786.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15*(1), 121-149.

VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology, 3*(1-2), 167.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13*(4), 454-461.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42*(23), 2593-2615.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*(1), 192-196.

Williams, E. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry, 2*(2), 149-168.

Wolfe, J. M., Vo, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15*(2), 77-84. doi:http://dx.doi.org/10.1016/j.tics.2010.12.001

Wolfe, J.M. et al. (2012). Sensation & Perception (3rd ed). Sunderland, MA: Sinauer.

Yantis, S. & Abrams, R. A. (2017). Sensation & Perception (2nd ed). New York: Worth.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133*(2), 273-293. doi:2007-02367-005 [pii] 10.1037/0033-2909.133.2.273

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*(1), 3-21.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162-185.

Appendix A - Example coherent sequence

Figure 8. Full 20 image coherent first-person viewpoint sequence of scene images beginning in an office (top left) and terminating in a parking lot (bottom right).

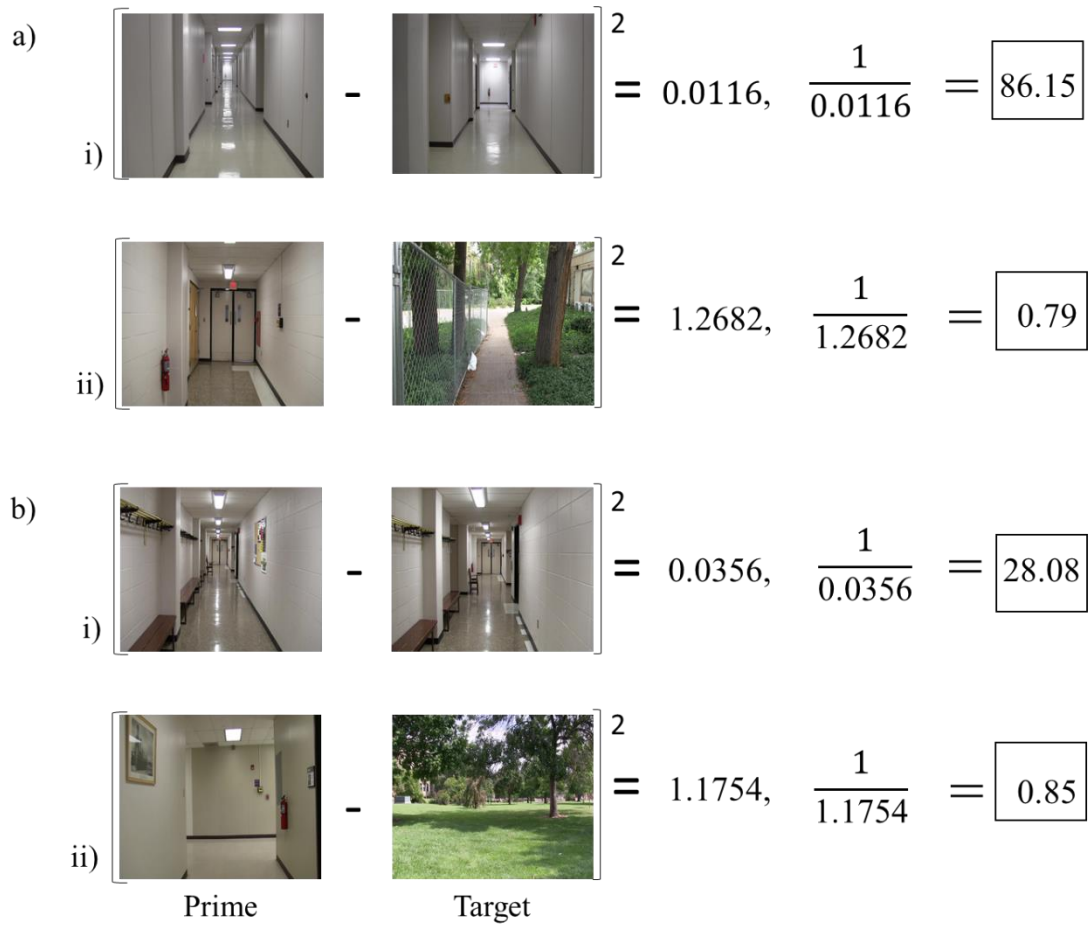

Figure 9. Prime and target image pairs from the a) coherent and b) randomized sequences with the i) highest and ii) lowest values of image similarity. Image similarity was calculated by taking the difference of the gist descriptor between the target and its immediately preceding prime, squaring the difference, and then taking its reciprocal.