Comparing rapid scene categorization of aerial and terrestrial views: A new perspective on scene gist

Lester C. Loschky Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA

Ryan V. Ringer Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA

Katrina Ellis Department of Psychology, Florida Institute of Technology, Melbourne, FL, USA

Bruce C. Hansen Department of Psychology & Neuroscience Program, Colgate University, Hamilton, NY, USA

Scene gist, a viewer's holistic representation of a scene from a single eye fixation, has been extensively studied for terrestrial views, but not for aerial views. We compared rapid scene categorization of both views in three experiments to determine the degree to which diagnostic information is view dependent versus view independent. We found large differences in observers' ability to rapidly categorize aerial and terrestrial scene views, consistent with the idea that scene gist recognition is viewpoint dependent. In addition, computational modeling showed that training models on one view (aerial or terrestrial) led to poor performance on the other view, thereby providing further evidence of viewpoint dependence as a function of available information. Importantly, we found that rapid categorization of terrestrial views (but not aerial views) was strongly interfered with by image rotation, further suggesting that terrestrial-view scene gist recognition is viewpoint dependent, with aerial-view scene recognition being viewpoint independent. Furthermore, rotation-invariant texture images synthesized from aerial views of scenes were twice as recognizable as those synthesized from terrestrial views of scenes (which were at chance), providing further evidence that diagnostic information for rapid scene categorization of aerial views is viewpoint invariant. We discuss the results within a perceptual-expertise framework that distinguishes between configural and featural processing, where terrestrial views are more effectively processed due to their predictable view-dependent configurations whereas aerial views are processed less effectively due to reliance on view-independent features.

Introduction

Every day, satellites gather countless images of life on earth, as seen in applications such as Google Earth and Bing Maps. Intuitively, it is clear that satellite images (referred to here as *aerial views*) are perceptually very different from the *terrestrial views* we see from the ground. The typical person surely has far less experience recognizing aerial views than terrestrial views. Thus, a comparison of how viewers process the gist of each type of view should shed light on general scene gist processing.[1] We define the theoretical construct of *scene gist* as a holistic semantic representation of a scene, rapidly acquired within the time frame of a single eye fixation, and we have operationalized it here as rapid scene categorization at the basic level. Our study derives insights by using our current understanding of how viewers rapidly categorize terrestrial views and testing the degree to which that enables us to understand how viewers rapidly categorize aerial views. These tests show some critical similarities, such as the importance of the natural/man-made distinction in determining the pattern of confusions across scene categories, which is found across the extreme viewpoint change between aerial and terrestrial views. This is, therefore, a universal element of scene gist recognition, which has not previously been shown in such stark relief. However, we have also found interesting, previously unknown, different processes used across the two views. Specifically, we have found

Citation: Loschky, L. C., Ringer, R. V., Ellis, K., & Hanson, 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, http://www.journalofvision.org/content/15/6/11, doi:10.1167/15.6.11.

doi: 10.1167/15.6.11 Received January 30, 2014; published May 18, 2015 ISSN 1534-7362

that rapid categorization of terrestrial views, but not aerial views, is viewpoint dependent, as shown by a rotation effect for the former but not the latter. We argue that this pattern of results is due to the impact of the *gravitational frame* (the fact that we usually view scenes with our bodies aligned vertically with gravity and our head above our feet) in recognizing terrestrial views. This highlights an overlooked aspect of terrestrial scene gist recognition at the core of our day-to-day perception and interaction with our environment. We also underline the importance of configural information for terrestrial scene gist recognition, based on the both the just-noted rotation effect and the finding that repetitive homogeneous pattern information (texture), which lacks recognizable configurations, is more useful for rapidly categorizing aerial views than terrestrial views of scenes. Each of these insights about rapid terrestrial-view categorization was found only by virtue of comparing aerial and terrestrial views. We discuss each of these points in greater detail later.

Our rationale for comparing rapid scene categorization for aerial and terrestrial views is analogous to that of studies that have sought to determine which processes involved in rapid object recognition are viewpoint dependent or independent (Biederman, 1987; Biederman & Gerhardstein, 1995; Foster & Gilson, 2002; Hayward, 2003; Palmeri & Cottrell, 2009; Tarr & Bülthoff, 1998; Tarr, Williams, Hayward, & Gauthier, 1998). For object recognition, findings of viewpoint dependence have been widely reported (Biederman & Gerhardstein, 1995; Palmer, Rosch, & Chase, 1981; Tarr et al., 1998), but the explanations of such viewpoint effects have differed in terms of whether object representations in memory are accessed through stored holistic views of objects (Tarr & Bülthoff, 1998) or through critical features, particularly line junctions, which then make up viewpoint-independent structural description representations (Biederman, 1987; Biederman & Gerhardstein, 1995). Subsequent work has converged on the idea that both sorts of representations may play important roles in object recognition (Foster & Gilson, 2002; Hayward, 2003).

Importantly, investigations of these issues in the realm of rapid scene categorization have only recently begun. Two recent studies investigated the role of viewpoint in determining both humans' and computational models' ability to accurately categorize terrestrial scenes (though without any time limitations; Ehinger & Oliva, 2011; Xiao, Ehinger, Oliva, & Torralba, 2012). These studies used panoramic photographs of terrestrial scenes and found that, within such 360° panoramic views (on the azimuth plane), there were specific consistently preferred smaller views (e.g., normal 65° camera views). Furthermore, those scene categories with the most reliably preferred views were also the most reliably categorized (Ehinger & Oliva, 2011). Ehinger and Oliva argue that such viewpoint dependence, as with object recognition, is consistent with a view-based theory of the perception and representation of scene categories.

Notably, those who have argued against viewpoint dependence in object recognition have argued instead for the importance of viewpoint-independent structural descriptions. In fact, another pair of recent studies has argued for the use of structural descriptions in scene-gist recognition. One found that line drawings of scenes were just as accurately recognized, and able to be decoded by a computational model from fMRI signals in scene-selective brain areas, as full-color photographs of the same scenes (Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). Furthermore, computational modeling using line junctions (e.g., T or Y) produced similar patterns of rapid categorization results to those of human viewers, and when the line junctions were decoupled, human categorization performance plummeted (Walther & Shen, 2014). The authors suggested that these results were consistent with scene gist recognition depending on structural descriptions.

From this discussion, it is clear that the viewpoint dependence versus independence of scene gist recognition is becoming an important issue. However, others have argued that what is critical in determining whether one finds viewpoint dependence or independence is what information is diagnostic for a task and whether that information is available, and that this applies equally to both object and scene categorization (Schyns, 1998). Information is *diagnostic* to the extent that it is perceptible and useful for a given task, and it is available. Thus, in the current study we ask to what degree we find viewpoint dependence or independence in rapid scene categorization of aerial and terrestrial views, and whether that can be explained fully in terms of the availability of certain types of information from each view for the given task. Alternatively, must one invoke different underlying representations or processes to explain differences in rapid scene categorization across the aerial and terrestrial views?

We begin our investigation by noting that compared to terrestrial scenes, aerial views are novel to human viewers—we did not evolve to recognize them, we have only had exposure to them since 1858 (*History of aerial photography*, n.d.), and most people have relatively little exposure to them in their daily lives. Nevertheless, common experience in looking at satellite photographs shows that viewers can categorize aerial views to some degree (Lloyd, Hodgson, & Stokes, 2002). Together, these two observations suggest that viewers recognize aerial views using certain basic information sources and visual processes also utilized during terrestrial-scene categorization, analogous to recognizing a familiar object from a novel viewpoint. Therefore, a comparison of fundamental similarities and differences in the rapid

categorization of aerial and terrestrial views can enable us to identify those basic information sources and processes. Similarities between rapid categorization of aerial views and terrestrial views should highlight *common* information sources and processing mechanisms that are employed regardless of viewpoint (i.e., ones that are very generally applicable for scene gist recognition). Conversely, the extent to which there are differences in the rapid scene categorization of the two viewpoints will provide important constraints on our understanding of scene gist recognition in general. Critically, if viewers are worse at rapidly categorizing aerial than terrestrial views, we must explain *why* that is the case. Simply saying that viewers lack experience with such views begs the question. Instead, a detailed consideration of the types of information available from each viewpoint, together with carefully planned manipulations of both types of views, may enable us to explain *why* aerial and terrestrial views differ in their rapid recognizability. Thus, a careful comparison of viewers' ability to recognize aerial and terrestrial scenes can allow us to draw important new insights into factors affecting both rapid scene categorization in general and rapid terrestrial- and aerial-view categorization in particular.

Overview of the current study

Perhaps the most fundamental characteristic of scene gist recognition is that it occurs rapidly after very short amounts of processing time, so we asked whether rapid scene categorization of aerial and terrestrial views differs in this regard. To answer this question, we used a standard psychophysical manipulation of visual processing time via visual masking stimulus onset asynchronies (SOAs; for reviews, see Breitmeyer & Ogmen, 2000, 2006; Enns & Di Lollo, 2000; Ogmen & Breitmeyer, 2006).[2] Early backward-masking studies of rapid scene categorization found that, for terrestrial views, scene categorization is highly accurate after roughly 100 ms of processing time (i.e., a 100-ms SOA between target and mask; Biederman, Rabinowitz, Glass, & Stacy, 1974; Potter, 1976). However, subsequent research has demonstrated that above-chance accuracy in scene categorization (and animal detection in scenes) occurs after an SOA as small as 12–24 ms, with an inflection point in the SOA function between 40 and 70 ms (Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005; Loschky, Hansen, Sethi, & Pydimari, 2010; Loschky et al., 2007). The incredible speed of scene categorization (and animal detection in scenes) has generated great interest (Kirchner & Thorpe, 2006; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001) and led us to ask whether categorization for aerial scenes is as rapid as for terrestrial scenes.

Based on the conjecture that aerial scenes are *alien* to most human observers, due either to human evolution or to the typical person's visual experience, we would expect that aerial views of scenes would take longer to categorize than terrestrial views. Although a handful of studies have investigated attention and eye movements in aerial scenes (Davies, Tompkinson, Donnelly, Gordon, & Cave, 2006; Lansdale, Underwood, & Davies, 2010; Pannasch, Helmert, Hansen, Larson, & Loschky, 2014; Zelinsky & Schmidt, 2009), to our knowledge the only previous study to investigate human scene categorization for aerial views was by Lloyd et al. (2002). Specifically, they compared the ability of geographers and nongeographers to categorize land use of aerial scenes and found that geographers (who had more experience with such images) were more accurate and had faster reaction times than nongeographers. However, the researchers did not limit viewers' processing time (through backward masking or any other means), nor did they compare rapid scene categorization performance for aerial views and terrestrial views, thus leaving the time course of processing the gist of these two types of views as an open question.

To address this question, Experiment 1 was designed to compare the time course of rapid scene categorization between aerial and terrestrial views. We found that aerial scenes could be rapidly categorized at levels well above chance when participants were given processing times (SOAs) associated with scene gist recognition. However, at these processing durations, viewers' performance with aerial views was far worse than with terrestrial views, and it failed to improve for processing times between 200 and 300 ms, suggesting that rapid scene categorization of aerial views may be limited by data rather than processing (Norman & Bobrow, 1975). If so, the question is then whether aerial views of scenes are missing important information found in terrestrial views.

We therefore asked what sources of information are used in rapidly recognizing the gist of aerial versus terrestrial scenes. Previous research on terrestrial-view scene gist recognition has shown that one such information source is the set of oriented spatial-frequency contrasts available for discriminating between images, similar to that measured in the spatial envelope model (Oliva & Torralba, 2001; Torralba & Oliva, 2003). The spatial envelope computational model has been repeatedly shown to accurately classify scenes into a large number of different categories based on amplitude-spectrum characteristics (i.e., global contrast distribution across spatial frequency and orientation). Additionally, recent studies of humans' terrestrial scene gist recognition have begun analyzing viewers' confusions between categories during rapid

scene categorization to find systematic errors, and using them to learn more about the factors, such as amplitude-spectrum characteristics, that underlie rapid scene categorization (Fei-Fei, Iyer, Koch, & Perona, 2007; Greene & Oliva, 2009; Walther, Caddigan, Fei-Fei, & Beck, 2009).

Experiment 1 used confusion matrices for rapid scene categorization performance of aerial and terrestrial views in order to assess underlying similarities and differences in their processing. The results showed considerable similarities between both types of views, which were further quantified through multidimensional scaling in terms of two-dimensional categorical similarity spaces. Those results suggested the existence of fundamental diagnostic information and/or processes used to categorize both aerial and terrestrial views of scenes. Nevertheless, a computational model inspired by the spatial envelope model was shown to more accurately categorize terrestrial scenes than aerial scenes, providing support for the idea that aerial scenes may be missing certain information in the amplitude spectrum that is available in terrestrial scenes. Furthermore, when such computational models were trained on terrestrial views and tested on aerial views (or vice versa), the models' performance dropped nearly to chance, showing that the amplitude-spectrum information available for distinguishing these scene categories is viewpoint dependent.

This discussion raises the question of whether there are other information sources missing from aerial views of scenes that are available in terrestrial views. We propose that one such information source is what we will call *the gravitational frame*, namely the environmental constraint that we usually see the world with our head above our feet (rather than, e.g., lying on our side or being upside down). Thus, the gravitational frame provides an inherent limit on the range of image orientations for terrestrial views—most terrestrial views in long-term memory should have been seen from within the constraints of the gravitational frame. Conversely, no such constraints based on the gravitational frame would seem to exist for aerial views of scenes, which would be appear to be isotropically symmetrical in terms of viewpoint rotation on the unit circle—that is, similar views should be found after each degree of rotation. Even if aerial views are normally seen with, for example, geographic north being upwards, the orientation of the photographed scenes with respect to the geographic north will not necessarily obey any particular set of rules. This should particularly be true for most natural scenes. Thus, in learning to recognize, for example, what a city or a golf course looks like from an aerial view, the problem of many-to-one matching (Ullman, 1989) should be exacerbated for aerial views relative to terrestrial views due to this lack of constraint by the gravitational frame.

Experiment 2 investigated this novel question regarding viewpoint dependence in scene gist recognition by measuring the effects of rotating aerial and terrestrial scenes on viewers' rapid scene categorization. The results showed that, as the view of terrestrial scenes was increasingly rotated away from the normal upright viewing orientation, rapid categorization was increasingly disrupted, but this was not the case for aerial views of scenes. These results are therefore consistent with the idea that the gravitational frame, by imposing constraints on how terrestrial views are seen, can be thought of as a source of information used to recognize terrestrial views of scenes that is absent from aerial views. Furthermore, as will be discussed later, the results are consistent with viewers' using the configuration of scenes, and to a lesser extent the dominant orientation of image spatial frequencies, as sources of information in rapidly categorizing terrestrial scenes. These results suggest that terrestrial scene gist recognition is to some degree viewpoint-dependent based upon differential information availability and also possibly different memory representations.

Interestingly, these findings also suggest that for categorizing aerial views, observers must use some form of information that is rotation invariant. One such structure-based information source is homogeneous pattern information across the image, namely texture, which is inherently rotation invariant. Vijayaraj et al. (2008) reported large differences between the image statistics of aerial and terrestrial views of scenes. In particular, their structural analyses indicated that aerial-view *textures* differed across basic level categories, thereby suggesting that aerial-view texture properties could provide useful information for artificial vision systems to categorize aerial scenes (Alonso et al., 2007; Bhagavathy, Newsam, & Manjunath, 2002; Graesser et al., 2012). However, while texture properties may aid in aerial-scene classification by artificial vision systems, it remains unclear whether such information can be used by humans when classifying aerial views. The Portilla and Simoncelli (2000) texture synthesis model, which finds homogeneous repeated patterns in a target image and then generates a synthesized image containing such patterns, has been shown to produce texture patterns (synthesized from terrestrial views) that are of relatively little use for recognizing terrestrial scenes (Loschky et al., 2010). This raises the final question explored in the current study: whether texture information is more diagnostic for the rapid categorization of aerial views of scenes than terrestrial views. If so, then this would point to another interesting difference between aerial and terrestrial scene gist recognition, which would provide constraints on theories of scene gist recognition in general and aerial and terrestrial scene gist recognition in particular.

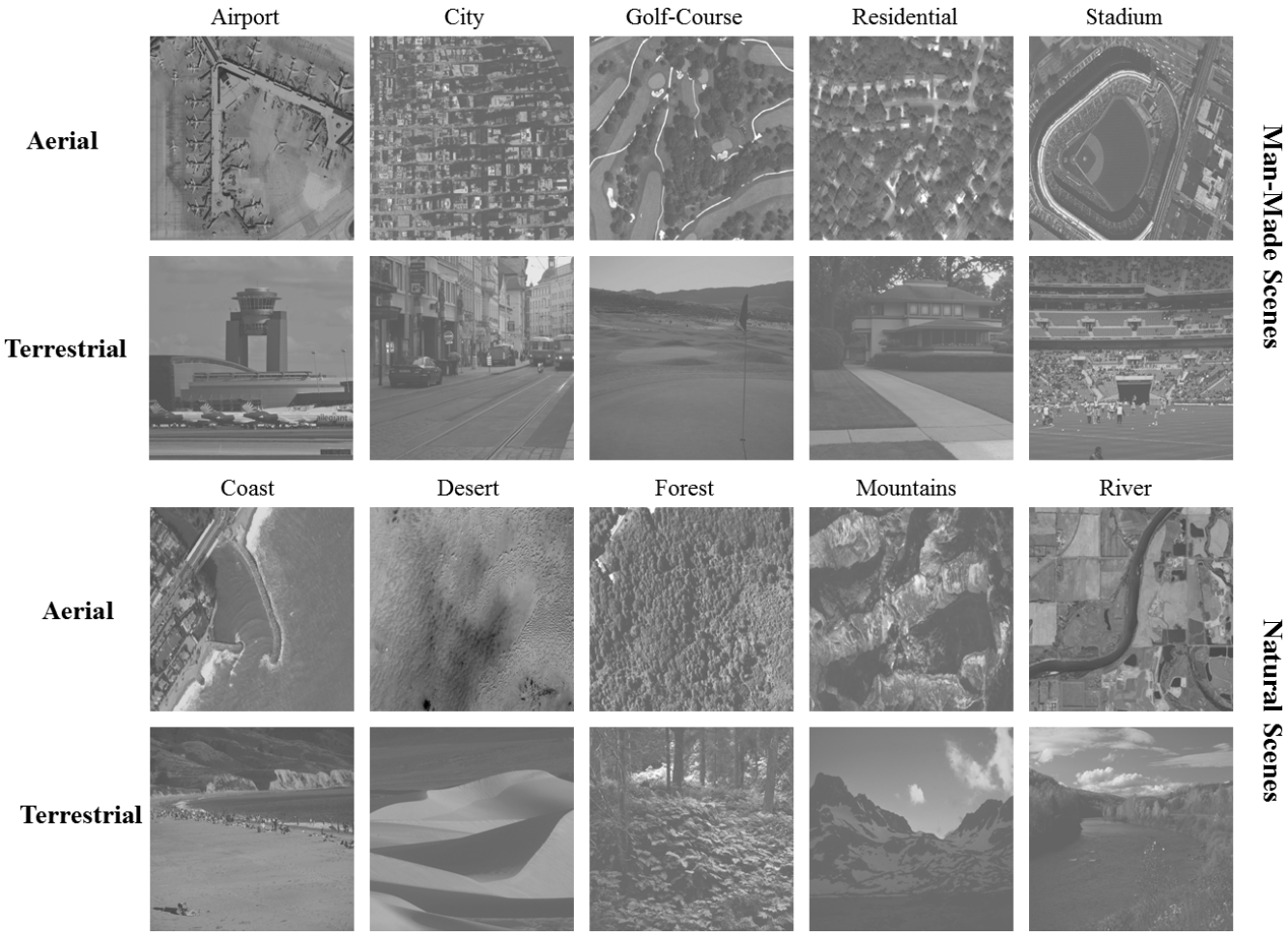


Figure 1. Sample stimuli comparing aerial and terrestrial views of natural and man-made scenes. Example images shown here were selected based on having accuracy equal to the mean accuracy for their respective categories.

Experiment 3 investigated whether there is a difference in the diagnosticity of texture information for recognizing aerial and terrestrial scenes, and found that that was indeed the case. Rapid categorization of texture images generated from terrestrial scenes was at chance at both short and long processing times, but categorization of texture images generated from aerial views of scenes was significantly above chance at longer processing times. Thus, these results suggest that rotation-invariant texture is at least somewhat diagnostic for rapidly categorizing scenes from aerial views, which further points to viewpoint independence for scene gist recognition of aerial views.

Experiment 1

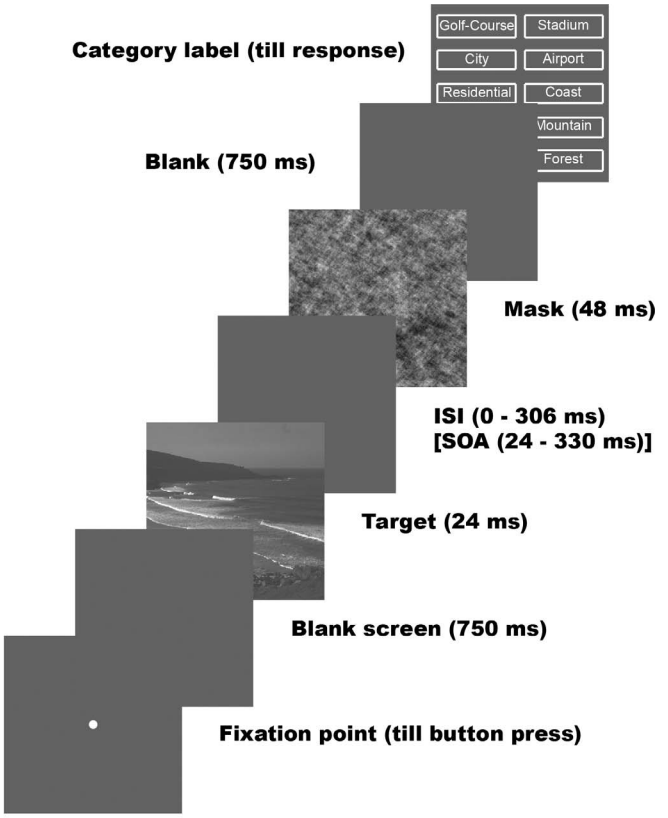
Method

Participants

Twenty-seven Kansas State University introductory psychology students (19 female, eight male) participated for course credit (age: $M = 18.63$, $SD = 0.63$). All subjects were tested to confirm normal or correct-to-normal ($\leq$20/30) vision. Institutional Review Board–approved written informed consent was obtained.

Stimuli

The target images were 320 aerial and 320 terrestrial grayscale photographs of 10 scene categories: five "natural" categories (coast, desert, forest, mountain, and river) and 5 "man-made" categories (airport, city, golf course, residential, and stadium)—thus, 32 images in each basic level category. Examples are shown in Figure 1. Terrestrial photos were collected from the Corel image database and the Internet, and aerial photos were from Google Earth. Because Torralba and Oliva (2002) have shown that mean viewing distance systematically varies across categories of terrestrial scenes, we similarly varied viewing height for aerial views. For example, Torralba and Oliva estimated that average photographic images of buildings are taken from closer viewing distances than average photographs of mountains by roughly an order of magnitude.



Trial Schematic

Figure 2. Schematic showing the sequence of displays in a trial. Trials began with the fixation point and ended with an all-alternative forced-choice response selection screen.

Importantly, exactly the same principle seems to apply to viewing height (i.e., vertical distance) for aerial-scene views. In a pilot study, we asked two viewers to select 30 "good" views f 3-D-modeled mountains and stadiums in Google Earth and record the "eye altitudes" for each selected view. The results (60 mountain view and 60 stadium views) showed that the average eye altitudes of "good" aerial views of mountains ($M = 14{,}126$ m, $SD = 10{,}883$ m) were 17.21 times as high as "good" aerial views of stadiums ($M = 821$ m, $SD = 551$ m). Thus, in the current study, all aerial images were collected under instructions to select Google Earth images with viewing heights that were "best" for recognizing that view's respective scene category. Our assumption was that this would produce the viewing distance for each image that would provide the "best view" of the aerial scene given unlimited processing time.

Backward masks were created by fully phase-randomizing the 640 target photographs, and then all 1,280 target and mask images were normalized for luminance (127 grayscale value) and root-mean-square (RMS) contrast (0.18; for details, see Hansen & Hess, 2007). The size of the images was 736×736 pixels. Using chin rests, participants viewed the photographs at a distance of 53 cm (25.5° visual angle) on Samsung SyncMaster 957 MBS monitors running at 85 Hz and 1024×768 pixel resolution. Additionally, a Spyder2Pro light meter was used to calibrate all monitors to the same luminance (maximum = 80.8 cd/m^2, minimum = 0.430 cd/m^2, gamma = 2.2).

Design and procedure

The study was a 2 (view: aerial vs. terrestrial) × 5 (SOA: 24, 70, 212, or 330 ms) × 10 (scene category: five natural, five man-made) within-subjects design. Aerial and terrestrial scenes were presented in separate blocks, with the order of aerial- and terrestrial-scene blocks counterbalanced across all subjects. SOA was counterbalanced across all levels of view and scene category.

Participants began by reading brief instructions and were allowed to ask questions. They then completed a familiarization task in which they saw 50 labeled scene images (five per category) for 3 s each from their view condition, to familiarize them with the image categories in the experiment. Additionally, they completed 50 practice trials before participating in the experiment. Images used in the familiarization and practice tasks were not used in the experimental trials.

Figure 2 shows a trial schematic. Participants initiated each experimental trial by first fixating on a white fixation dot in the center of a neutral gray screen (equal to the mean luminance of the target and mask images) and clicking a mouse button. Then, 750 ms later, the target image was flashed for 24 ms, followed by a neutral gray-screen interstimulus interval of 0, 46, 188, or 306 ms (i.e., SOAs of 24, 70, 212, and 330 ms) followed by a mask flashed for 48 ms (for a 1:2 target:mask duration ratio). The mask was followed by a neutral gray screen for 750 ms, and then all 10 scene-category labels were presented on the screen in a 5×2 vertical grid (i.e., a 10-alternative forced choice), from which participants selected the appropriate category using the mouse. For each participant, the locations of the labels were randomized for each trial to avoid contaminating the results by any bias toward responding to a favored location (e.g., top right corner).

Results

Prior to our main analyses, we removed any subjects whose overall accuracy was more than two standard deviations below the mean percentage correct, resulting in the removal of one subject. Linear mixed-effects modeling, using the lme4 library (Bates, Maechler, Bolker, & Walker, 2014) in R (version 3.02), was then conducted to determine the fixed effects of view (aerial vs. terrestrial) and processing time (SOA) on accuracy

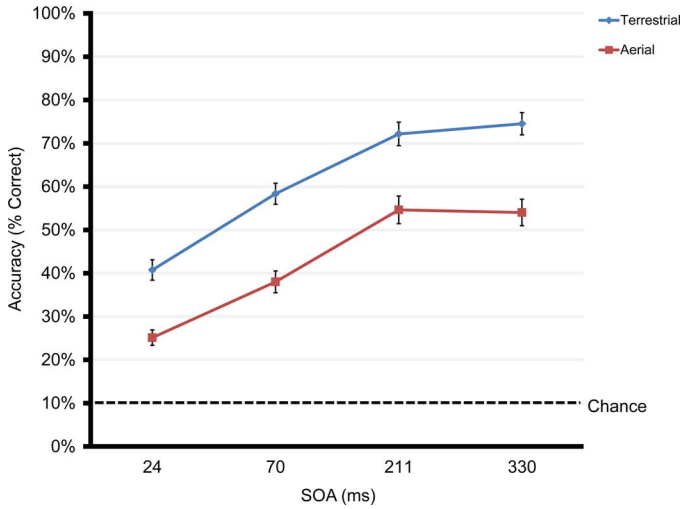


Figure 3. Rapid scene categorization accuracy as a function of view (aerial vs. terrestrial) and processing time (SOA in ms). Error bars = standard error of the mean. Dashed line = chance performance.

from the random effect of the participants.[3] However, because accuracy (e.g., proportion correct) does not necessarily conform to the assumptions of a Gaussian distribution, the mean scores for each subject by factor interaction were logit transformed to improve the fit of the data (Cohen & Cohen, 1983). Model selection consisted of constructing five models with identical fixed-effects structures (SOA × View) and varying random-effects structures, where the model was able to vary across participants as a function of mean accuracy, processing time, view, the additive effects of processing time and view, and the interaction of processing time and view. The model comparisons revealed that the variability across processing time and view was not sufficient to be included in the random-effects structure of the model, and thus the model selected included only variability across participants' mean scores as a random-effects component (Bayesian information criterion [BIC] = 315.50).

As shown in Figure 3, the results are consistent with our hypothesis that viewers would be more accurate in rapidly categorizing terrestrial views ($M = 0.63$, $SD = 0.178$)[4] than aerial views ($M = 0.44$, $SD = 0.176$), $F(1, 25) = 230.87$, $p < 0.001$, $f^2 = 1.49$.[5] This difference was quite dramatic, though performance at the shortest SOA (24 ms) was significantly above chance (i.e., 10% in our 10-alternative forced-choice measure) in both the aerial, $t(25) = 11.90$, $p < 0.001$, and terrestrial, $t(25) = 18.86$, $p < 0.001$, view conditions. There was also a significant main effect for processing time, $F(1, 25) = 376.05$, $p < 0.001$, $f^2 = 1.48$, but the rate of change in accuracy across processing time was unaffected by viewpoint, which was verified statistically by a nonsignificant interaction between these factors (Viewpoint × Processing time), $F(1, 25) = 2.04$, $p = 0.230$, n.s. Thus the effects of viewpoint and processing time were independent, with participants showing nearly identical Accuracy × SOA slopes between 24 and 212 ms SOA (terrestrial = 0.157/ms; Aerial = 0.152/ms) and both views reaching asymptote at 212 ms SOA (212–330-ms SOA slopes: terrestrial = –0.0044/ms; aerial = 0.0189/ms). This suggests that the *rate* of scene-category information extraction occurred in a consistent, possibly automated, fashion for both viewpoints. Nevertheless, Figure 3 also shows that that for terrestrial views, a 330-ms SOA (equal processing time to a single eye fixation) was sufficient to quite accurately categorize scenes (Biederman et al., 1974; Eckstein, Drescher, & Shimozaki, 2006; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky et al., 2007; Potter, 1976; Torralba, Oliva, Castelhano, & Henderson, 2006), but for aerial views, 330 ms produced far worse performance. Indeed, when viewers had processed an aerial scene for 330 ms, their accuracy was slightly worse than when they had processed a terrestrial scene for 70 ms. This result suggests that the problem in rapid categorization of aerial scenes may involve more than a simple lack of processing time; instead, it may be that aerial views of scenes lack critical information that is available in terrestrial views. Thus, it is important to determine to what degree aerial and terrestrial scenes share information that is diagnostic of their basic level category, as well as what diagnostic information is uniquely available to one or the other view.

We next constructed SOA-averaged confusion matrices (i.e., we made confusion matrices for each SOA and then averaged them) for both aerial- and terrestrial-view categorization performance (see Figure 4). Collapsing across the main diagonals of Figure 4 gives an overall accuracy of 44% for aerial views versus 63% for terrestrial views (with averaged off-diagonal error rates of 6.2% and 4.1% for aerial and terrestrial views, respectively). Regarding the off-diagonal confusions, the most general similarity between Figure 4A and B takes place in the bottom right quadrant, showing overall higher confusions among the "natural" scene categories, and the upper right and lower left quadrants, which show relatively lower rates of confusions between "natural" and "man-made" categories. One of the few such confusions is a tendency for the "man-made" golf-course category to be confused with several "natural" scene categories (top right quadrant, middle row).

In order to quantitatively validate these observations, we carried out a multidimensional scaling (MDS) analysis of the SOA-averaged confusion matrices shown in Figure 4.[6] In order to ensure that the MDS analyses for aerial and terrestrial views were only on the basis of confusions, we omitted the main diagonals (i.e., correct responses) from the analyses. We assumed that confusions could be interpreted as similarity

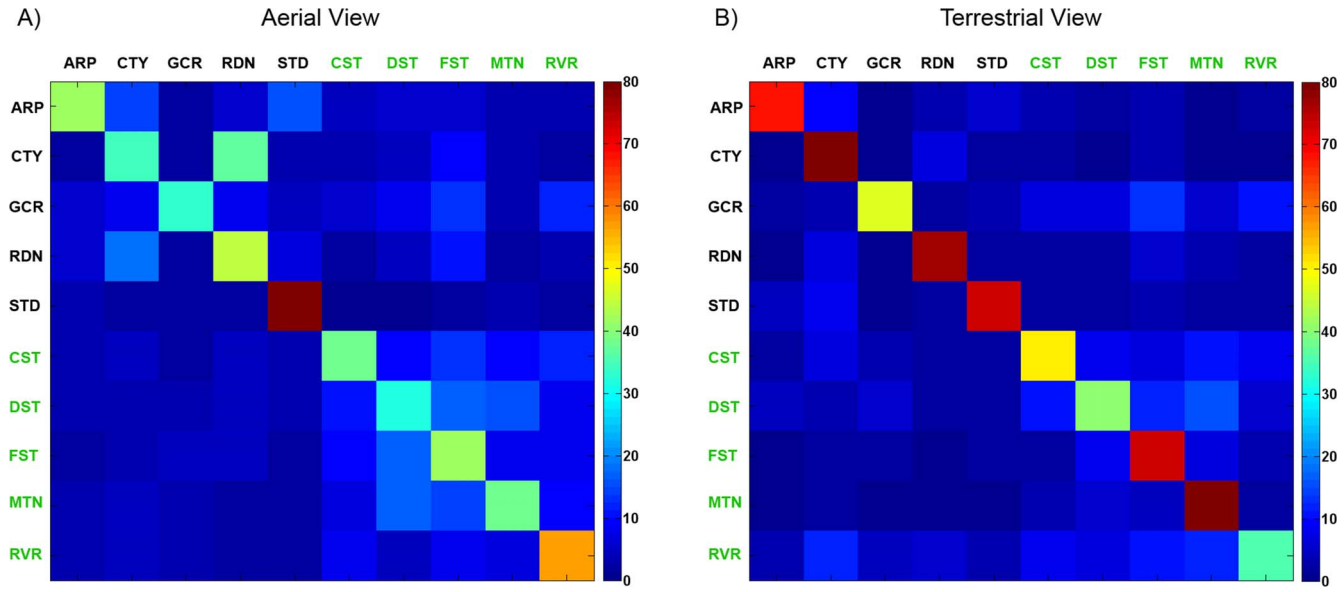


Figure 4. Confusion matrices for aerial (A) and terrestrial (B) views. Rows represent target image category and columns represent average responses made for each response category. Categories have been grouped by man-made (black) and natural (green), arranged in alphabetical order. The color scale is in percentages of each response for a given image category, with the main diagonal representing correct responses and off-diagonals being confusions. ARP: airport; CTY: city; GCR: golf course; RDN: residential; STD: stadium; CST: coast; DST: desert; FST: forest; MTN: mountain; RVR: river.

measures, with more perceptually similar categories being more frequently confused with each other. We also assumed that confusions were symmetrical between categories of images and responses, and took the average confusions for symmetrical off-diagonal cells (e.g., mountain:forest vs. forest:mountain) to create our similarity metric. Note that this assumption ignores the fact that categorization errors are not symmetric about the main diagonals. Given these assumptions, the MDS analysis should separate scene categories into clusters based on confusions (i.e., coded as similarities for this analysis). The resulting two-factor solutions for aerial scenes and terrestrial scenes are shown in Figure 5. As our qualitative observations drawn from Figure 4 suggest, the "natural" category scenes cluster together for both viewpoints. Interestingly, the "man-made" golf-course category also clusters with the "natural" scene categories for both viewpoints, which is consistent with the fact that such imagery is composed of natural landscape content. The remaining "man-made" categories (airport, stadium, city, and residential) are split out in separate parts of the factor space. The MDS analysis, along with the confusion matrices shown in Figure 4, suggests that there may be important similarities in the rapid categorization of both aerial and terrestrial scene views, which may be due to sharing diagnostic information.

One of the clearest differences between the aerial and terrestrial factor spaces shown in Figure 5 is the river category. For terrestrial views, it is closely clustered with the other "natural" category scenes, but for aerial views, it is quite separate from the other "natural" categories. This may be because in aerial views, rivers are very distinct in terms of their highly diagnostic snakelike curving-line feature, whereas for terrestrial views, rivers are actually quite difficult to identify relative to other "natural" categories. The latter difficulty may be because terrestrial rivers share foliage features with forests, mountains, and golf courses, and because their defining feature of a narrowly bounded stream of water is not always easy to distinguish from other bodies of water (e.g., lakes or even coasts). In addition, the MDS analysis shows that, for the aerial views, the stadium category is treated separately from all other categories. In contrast, for terrestrial views, the stadium category is relatively more similar to other "man-made" categories, particularly airport and residential. It is likely that the stadium category is distinct among aerial views because stadiums can be recognized by a single distinct shape (e.g., oval for football stadiums, fan-shaped for baseball stadiums, etc.). Conversely, from a terrestrial viewpoint, stadiums look like many other sorts of buildings. In sum, aerial views of rivers and stadiums have less feature overlap with other "natural" and "man-made" categories than terrestrial views of rivers and stadiums do (Gosselin & Schyns, 2001; Schyns, 1998). Thus, in these two instances, highly diagnostic information is actually more available in the aerial views than terrestrial views.

To measure *general* confusion similarities between aerial and terrestrial views, we calculated the correlation between confusion matrices across the two views.

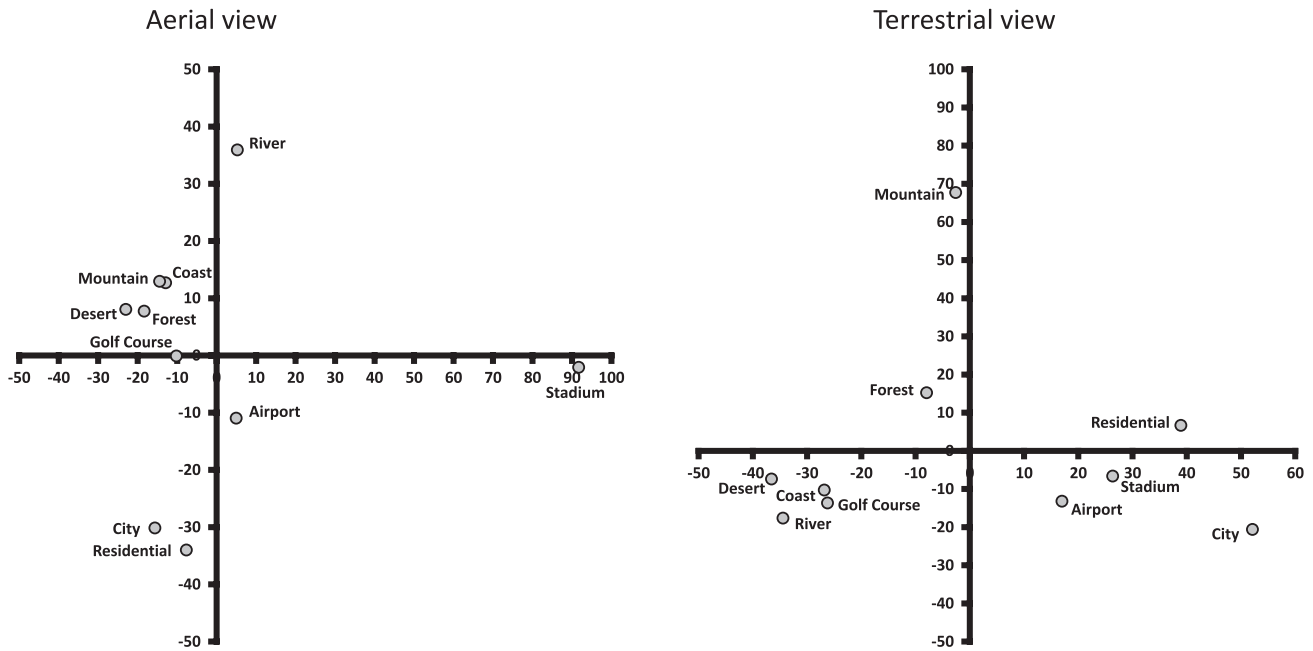


Figure 5. Multidimensional scaling two-factor solutions for aerial and terrestrial views of scenes, based on their respective confusion matrices (excluding the main diagonal).

As in the MDS analysis, the main diagonals (i.e., correct responses) were omitted from these analyses. The confusion matrices provided a Pearson's $r(88) = 0.562$, $p < 0.001$, 95% confidence interval (CI) [0.40, 0.69]; however, there was strong heterogeneity of variance in the data. Because the data were non-normally distributed, we used Spearman's r to determine the rank order correlation coefficient between SOA-averaged aerial and terrestrial confusion matrices. This produced an $r(88) = 0.628$, $p < 0.001$) 95% CI [0.47, 0.73], suggesting a reasonably high degree of similarity in confusions made across the two viewpoints.

We also calculated Spearman correlations between aerial- and terrestrial-view-based performance at each SOA (results shown in Table 1). Table 1 shows a clear nonlinear trend—as processing time increased from 24 to 212 ms SOA, the confusion matrix correlations decreased, then from 330 ms SOA, the correlation increased to the level observed at 24 ms SOA. A possible explanation for this pattern is that, at the earliest processing time of 24 ms, both views use lower level perceptual information, resulting in similar confusions. However, the superior performance for terrestrial views suggests that viewers begin to use higher level information at the intermediate SOA of 70 ms, reducing the correlation between views. Finally, performance for aerial views begins to use such higher level information at the 330-ms SOA (which is equal to the typical fixation duration for terrestrial scenes; Rayner, 1998), increasing the correlation between views. Thus, the confusion matrices for the two views suggest similar information extraction patterns but at different processing times.

One possible interpretation of the reasonably high correlation between confusions across aerial and terrestrial views at the typical-fixation-duration processing time is that we may rely on similar diagnostic image-statistical information (e.g., texture; Vijayaraj et al., 2008) to categorize scenes across both viewpoints. Nevertheless, the significant disparity in accuracy between aerial and terrestrial scenes suggests that such information used to categorize scenes may be better suited to terrestrial views of scenes rather than aerial views. Alternatively, this pattern of results may have little to do with the image-statistical information

	SOA			
	24 ms (no interstimulus interval)	70 ms	212 ms	330 ms
Spearman's *r*	0.615	0.546	0.466	0.647
95% CI	[0.647, 0.729]	[0.383, 0.677]	[0.287, 0.614]	[0.507, 0.753]

Table 1. Spearman correlations between aerial and terrestrial confusion matrices at each SOA.

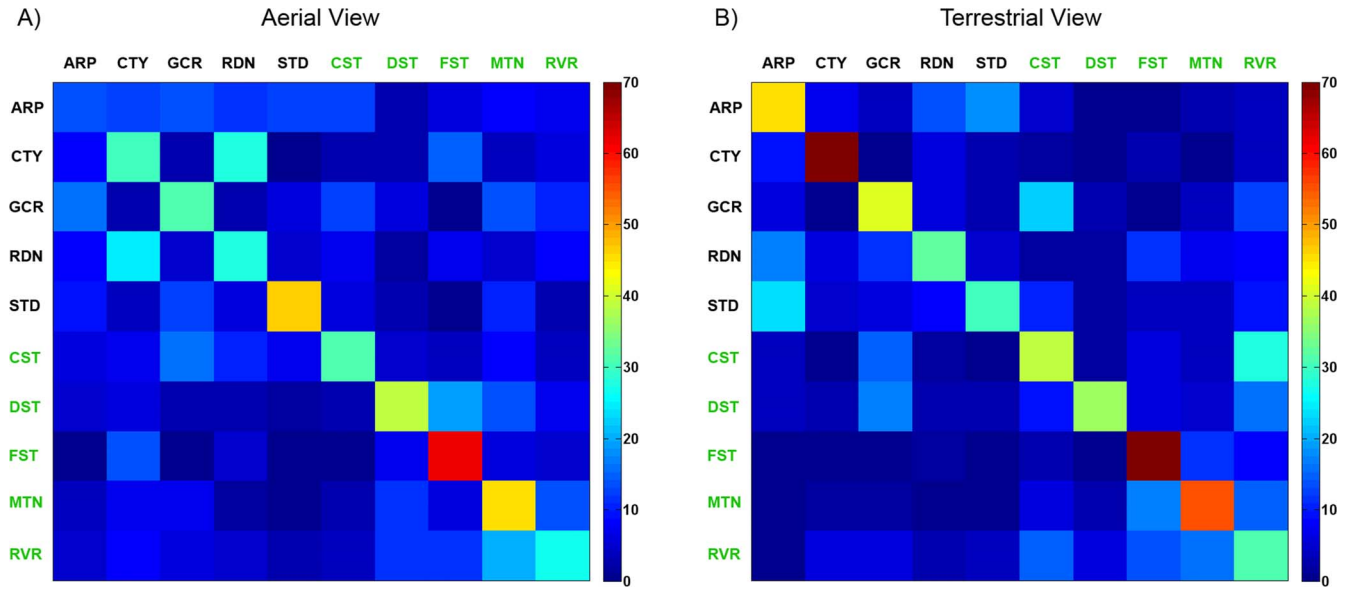


Figure 6. Linear discriminant classification confusion matrices for aerial (A) and terrestrial (B) views. Rows represent target image category and columns represent average response classifications made for each category. Categories have been grouped by man-made (black) and natural (green), arranged in alphabetical order. The color scale is in percentages of each response for a given image category, with the main diagonal representing correct responses and off-diagonals being confusions. ARP: airport; CTY: city; GCR: golf course; RDN: residential; STD: stadium; CST: coast; DST: desert; FST: forest; MTN: mountain; RVR: river.

available for each view. Therefore, to provide insight regarding the possible role of iage-statistical information for some of the confusion similarities mentioned, we employed a type of standard spatial envelope model that was inspired by the original spatial envelope approach (e.g., Torralba & Oliva, 2003) for classifying aerial or terrestrial views, based on orientation characteristics in the amplitude spectrum. Each scene within a given category and view was modeled as a vector of output contrast energies from simulated visual filters (i.e., log Gabors; Hansen & Hess, 2007) centered on six different spatial frequencies (i.e., 0.5, 1.0, 2.0, 4.0, 8.0, and 12.0 c/°) at one of 12 different orientations (from 0° to 165°, in steps of 15°). That is, each image's power spectrum was filtered with a given log Gabor in the Fourier domain, then inverse Fourier transformed to the spatial (i.e., image pixel) domain, then passed through a filter response threshold routine (e.g., Hansen & Hess, 2007) and summed. This was repeated for each filter. Thus, each image was represented as a gated contrast energy distribution across both scale and orientation as captured by simulated visual filters. Multidimensional linear discriminant classifiers were then trained on 16 randomly selected output contrast vectors from scenes from each basic level category within either aerial or terrestrial views (i.e., a 10-dimensional discriminant classifier for each viewpoint was constructed). It is worth noting that the use of multidimensional discriminant classifiers was necessary to produce 10×10 confusion matrices. The classifiers were then tested on the remaining 16 output vectors for each category.

In order to cancel out random classification results, we simulated this process 100 times, each time randomly sampling 16 different output vectors from each image category for the training set, with the remaining set used for testing classification accuracy. Simulated classification performance (via multidimensional linear discriminant classifiers) was then separately averaged across all 100 simulations for aerial and terrestrial views and plotted in confusion-matrix format. The process was carried out iteratively for each filter's central spatial frequency in order to assess which scales produced the highest classification accuracies. This iterative analysis resulted in the 4–8-c/° filter sets producing the highest accuracies, a finding that is consistent with previous work (Hansen & Hess, 2007). The confusion matrices from the filter sets *tuned* to 4–8 c/° are shown in Figure 6.

Importantly, our version of the spatial envelope model was worse at categorizing aerial views than terrestrial views. Collapsing across the main diagonals of Figure 6A and B gives an overall accuracy of 36% for simulated aerial classification versus 46% for simulated terrestrial classification, both of which were nevertheless well above chance (10% in our 10-alternative forced-choice measure), with averaged off-diagonal error rates of 7% and 6%, respectively. This result is consistent with the earlier stated hypothesis that aerial views may be data limited relative to terrestrial views, here defined in terms of information

	SOA			
	24 ms (no interstimulus intervalt)	70 ms	212 ms	330 ms
Aerial	0.266 [0.076, 0.460]	0.260 [0.055, 0.443]	0.213 [0.007, 0.403]	0.358 [0.163, 0.526]
Terrestrial	0.352 [0.157, 0.521]	0.368 [0.174, 0.535]	0.345 [0.148, 0.515]	0.415 [0.227, 0.572]

Table 2. Spearman correlations between human and model classification confusion matrices at each SOA. *Notes*: For all correlations, $p < 0.05$. Bracketed values = 95% CI.

available from the amplitude spectrum for discriminating scene categories and operationalized in terms of our version of the spatial envelope model. In addition, the well-above-chance performance of both models suggests that such available information could potentially be used to discriminate among scene categories from both views. However, the lower performance by the model than by human observers (aerial: humans: 44%, model: 36%; terrestrial: humans: 63%, model: 46%) also suggests that people are not limited to using the same available information and processes involved in our version of the spatial envelope model.

To evaluate the degree to which the model captured human performance patterns, we calculated Spearman's r between the human and model confusion matrices for the aerial and terrestrial views, which produced a somewhat higher correlation for the terrestrial views than the aerial views. For the aerial views, the rank order correlation between the human confusion matrix (Figure 4A) and the linear discriminant classification matrix (based on contrast energy vectors; Figure 6A) was $r(88) = 0.279$, $p = 0.008$, 95% CI [0.076, 0.460] (main diagonals excluded). For the terrestrial views, the same procedure was carried out for the human confusion matrix (Figure 4B) and the linear discriminant classification confusion matrix (based on contrast energy vectors; Figure 6B), which resulted in $r(88) = 0.427$, $p < 0.001$, 95% CI [0.242, 0.583] (main diagonals excluded). The lower correlation between human and model confusion matrices for the aerial views is consistent with the finding that the model performed worse for the aerial views. Together, these results suggest that there is less diagnostic amplitude information available in aerial views than terrestrial views, and that aerial-view categorization may therefore require greater use of other information sources.

The same analysis was also carried out for human versus model confusion matrices at each SOA (results shown in Table 2). The correlations between the confusion matrices for the multidimensional linear discriminant classifier (trained and tested on components sampled from the amplitude spectrum) and the SOA-averaged human data suggest that viewers' confusions for terrestrial views (as noted, $r = 0.43$) were more influenced by amplitude-spectrum image statistics than was the case for the aerial views (again, $r = 0.28$). A similar trend across correlations was also apparent at each SOA, as shown in Table 2, such that as processing time increased, humans' confusions became more similar to the amplitude-spectra-trained classifier confusions. Given that all correlations were statistically significant, the similarity in confusion-matrix patterns shown in Figures 4 and 6 (and Table 1) suggests the hypothesis that observers used similar amplitude-spectrum characteristics to rapidly categorize images from both views.

To test the hypothesis that similar diagnostic image-statistical information is used to categorize scenes across both viewpoints, we trained the multidimensional linear discriminant classifier models on filter output vectors from one view (e.g., terrestrial) and tested them on filter output vectors of the other (e.g., aerial). All other aspects of the modeling procedure were identical to those described perviously. The confusion matrices for these analyses are given in the Supplementary Figure S1. The models performed at near-chance levels in categorizing scene images. For the terrestrial-trained, aerial-tested model, the averaged main diagonal accuracy was 13.8% (averaged error = 9%); and for the aerial-trained, terrestrial-tested model, it was 15.8% (averaged error = 9%). This suggests that the moderate correlation between confusion matrices across views was *not* due to using similar diagnostic image-statistical information to categorize scenes from both views. Instead, the correlation across views may have been due to observers using similar visual processing routines (e.g., Ullman, 1984) to categorize the differing available information within each view, resulting in similar confusions.

Discussion

Experiment 1 was an initial exploration of the time courses of rapid scene categorization for aerial and terrestrial views of scenes. Consistent with the idea that aerial views are novel to most viewers, we found that rapid scene categorization of aerial views was considerably more difficult than for terrestrial views. The results in Figure 3 show that 330 ms of processing time

(equal to a single fixation) was sufficient to rapidly, accurately categorize terrestrial views (Biederman et al., 1974; Eckstein et al., 2006; Joubert et al., 2007; Loschky et al., 2007; Potter, 1976; Torralba et al., 2006) but not aerial views, whose performance at 330 ms SOA was slightly worse than that for terrestrial views at 70 ms SOA. Thus, the relative difficulty in rapidly categorizing aerial views may be due *not* to a lack of processing time (at least within the time frame of a single eye fixation) but instead to a lack of *critical diagnostic information* that is available in terrestrial views. This conclusion from behavioral data was buttressed by our computational analysis of linear discriminant classifications of the images based on their amplitude-spectrum characteristics, which also produced superior performance for terrestrial views compared to aerial views.

This reasoning follows from Norman and Bobrow's (1975) theory of data-limited versus resource-limited processes. For example, in a difficult search task, allocating more processing resources and taking more time generally increases performance. However, in situations where sensory data are limited (such as visual search for a target in noise), performance after a certain point will reach asymptote regardless of the amount of processing time allowed. Thus, in our rapid scene categorization task, if the problem in categorizing aerial views were simply a lack of processing time, performance would be the same between aerial and terrestrial views given sufficient processing time. However, within the limits of a single fixation (330 ms processing time; Rayner, 1998), which is typically understood to be the amount of time needed to acquire scene gist (Biederman et al., 1974; Eckstein et al., 2006; Joubert et al., 2007; Loschky et al., 2007; Potter, 1976; Torralba et al., 2006), this was not true for aerial scenes. Thus, the data are consistent with the idea that aerial views may be somewhat data limited. Furthermore, the correlations between aerial- and terrestrial-view confusion matrices (Table 1) suggest that the information extracted at intermediate points in processing time (e.g., 70–212 ms SOA) may have differed across views. Finally, linear discriminant models trained on amplitude-spectrum information from one view (aerial or terrestrial) produced chance categorization of scenes from the other view, suggesting that the amplitude information available for categorizing each view is different. Importantly, this is evidence of view-dependence of the information used by the spatial envelope model to categorize scenes, which is at least partially explainable in terms of a lack of certain diagnostic information from aerial views.

However, Experiment 1 also suggested that there may be important underlying shared processes during rapid scene categorization of both aerial and terrestrial views. First, as shown by the nearly identical Accuracy $\times$ SOA slopes, it seems that scene-category information extraction occurred at a consistent, possibly automated, rate regardless of viewpoint. Furthermore, the Spearman's r correlation ($r = 0.628$) between the SOA-averaged confusion matrices for aerial and terrestrial views (i.e., Figure 4), along with the results from the MDS analysis (Figure 5), suggests an important underlying commonality in discriminating among scenes from both views. In particular, diagnostic information present in the "natural" scene categories, though likely different in nature between views, appeared to be utilized in a similar manner regardless of view, resulting in similar confusion errors made between those categories across views. The majority of the "man-made" scene categories (airport, city, residential, and stadium) were seldom confused with the "natural" scene categories, again independent of viewpoint. Thus, while rapid aerial-scene categorization performance may suffer from an apparent data limitation as discussed already, for both views the information that is available seems to be extracted at a constant rate, and viewers make similar broad distinctions (e.g., natural vs. man-made scene categories). The confusion similarities may have been due to similar visual processing routines used by observers to discriminate scene categories from both aerial and terrestrial views, based on the different view-specific amplitude-spectrum characteristics available from each view.

In addition, that available diagnostic amplitude information lent itself more to terrestrial-scene categorization, as shown by the better accuracy for classifiers trained only on amplitude-spectrum characteristics for terrestrial views than aerial views. This latter result was consistent with human categorization performance. This suggests that the overall accuracy difference in human performance for aerial and terrestrial views may be due to a failure of amplitude-spectral relationships to reliably convey category distinctions for aerial views. In sum, the results show a strong degree of viewpoint dependence in scene gist recognition, which is likely due to differences in the availability of diagnostic information from each view.

Experiment 2

The results of Experiment 1 provided evidence for the use of view-dependent amplitude-spectrum characteristics for both aerial and terrestrial gist recognition but attenuated for aerial views, resulting in a dramatic difference in categorization accuracy (i.e., Figure 3). However, another potentially important cue, which is only partially available in the amplitude spectrum, is the gravitational frame, namely the constraint imposed

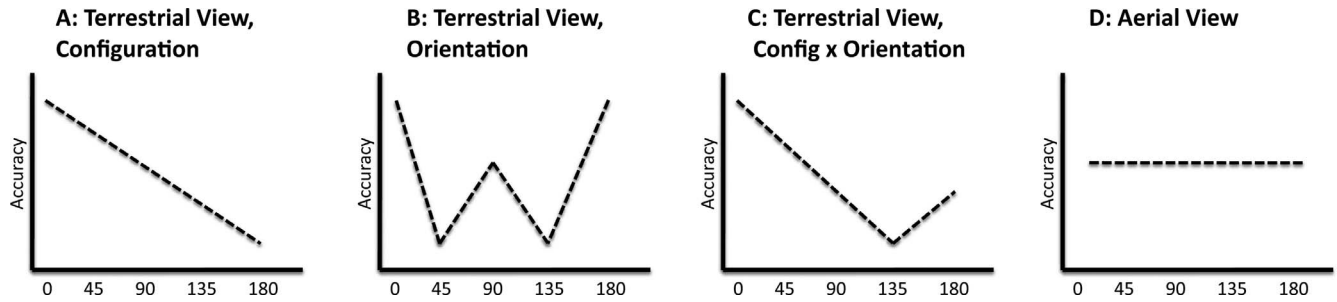


Figure 7. Hypothetical rapid scene categorization as a function of rotation. (A) The configuration hypothesis: Performance monotonically drops with increasing rotation away from the canonical viewpoint (0° rotation). (B) The orientation-bias hypothesis: Performance is poorest at the oblique rotations (45° and 135°), intermediate at 90° where vertical and horizontal orientations are reversed, and best when the vertical horizontal orientation bias is maintained at 0° and 180° rotations. (C) The configuration/orientation hypothesis: Performance decreases steadily from the 0° rotation until it hits bottom at 135°, where both orientation bias and configuration information are most altered. Rotations of 90° and 180° should be somewhat better, and have equal performance. (D) Without an identifiable upright orientation, aerial scenes should be unaffected by rotation.

by gravity on human views of terrestrial scenes. This is clearly present for terrestrial views of scenes (Gregory & McCloskey, 2010; Haji-Khamneh & Harris, 2010; Harris, Jenkin, Dyde, & Jenkin, 2011) but seems inapplicable to categorizing aerial views. If so, this would be further evidence of viewpoint dependence in scene-gist recognition for terrestrial views but not for aerial views, again based on differential availability of diagnostic information. This constraint on the orientations from which terrestrial views, but not aerial views, are seen may influence the learning of the many-to-one mapping of scene views to scene categories. Specifically, the more constrained learning for terrestrial views could allow humans to recruit processes that rely more on the scene configuration for categorizing terrestrial scenes. Such processes might allow for more efficient gist recognition. Since aerial views apparently lack such gravitational-frame constraints on scene configurations, such efficient processing would never be achieved, thus leading to the lower overall accuracy regardless of processing time (i.e., Figure 3).

To test this hypothesis, we investigated the effect of scene rotation on rapidly categorizing aerial and terrestrial scenes. If terrestrial scene gist recognition is constrained by the gravitational frame, then rotation should have a significant impact on rapid terrestrial-scene categorization accuracy. Conversely, we hypothesized that rotation should not affect rapid aerial-scene categorization (due to lack of constraints from the gravitational frame). However, image rotation violates not only the constraints imposed by the gravitational frame but also the global distribution of luminance contrast across orientations (amplitude spectrum), which was shown to be utilized (in part) in the rapid categorization of aerial and terrestrial scenes in Experiment 1. Thus, image rotation sets up four primary competing hypotheses (illustrated in Figure 7).

First, if rotation disrupts rapid scene categorization for terrestrial views (Figure 7A through C) but not aerial views (Figure 7D), it would be consistent with the hypothesis that the diagnostic information removed from terrestrial scenes by rotation is the same information that is always lacking in aerial views, namely processing constraints due to the gravitational frame.

Second, if the configuration of a scene (i.e., its *layout*) referenced to the gravitational frame is highly diagnostic in rapidly categorizing terrestrial views, then categorization accuracy should monotonically decrease as rotation from the canonical 0° view increases (Figure 7A). For example, a terrestrial beach scene in which the upper half of the image is the sky and the lower half is sea and sand would look quite different if the sky were on the bottom and the sea and sand were on the top; likewise if all of these image elements were oriented vertically. We will call this the configuration hypothesis.

Alternatively, increasing degrees of rotation may nonmonotonically affect rapid terrestrial-scene categorization (Figure 7B). Specifically, if the distribution of global oriented amplitude (within scenes) is highly diagnostic for rapidly categorizing terrestrial views of scenes (e.g., Kaping, Tzvetanov, & Treue, 2007; Oliva & Torralba, 2001), then 180° rotations should be less disruptive of scene categorization than 90° rotations, and even less disruptive than 45° and 135° rotations (Figure 7B). For example, a terrestrial view of a beach scene would contain predominately horizontal information, which would remain horizontal after a 180° rotation but would become predominately vertical after a 90° rotation. Meanwhile, oblique rotations (45° and 135°) would be most disruptive due to the predominance of horizontal and vertical orientations compared to oblique orientations in real-world scenes (Hansen &

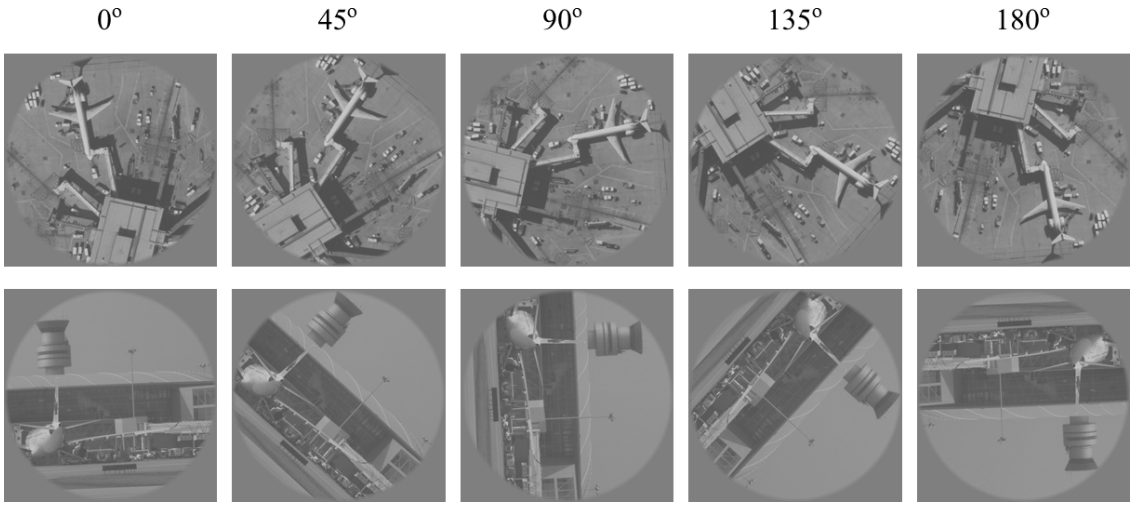


Figure 8. Examples of rotated images for aerial and terrestrial views used in Experiment 2.

Essock, 2004). We will call this the orientation-bias hypothesis.

Finally, if there is an interaction between scene configuration and oriented amplitude in scenes, then we should expect the greatest decrement in performance to occur when both of these parameters are the furthest from the 0° upright viewpoint (Figure 7C). Rotating a scene 135° would produce the greatest impairment of performance, because it would rotate the scene layout from upright within the gravitational frame by three quarters of the distance to 180° while simultaneously transforming the original vertical and horizontal orientations to oblique orientations. The least disruptive rotation should be at 45°, because it is the closest to the 0° gravitational-frame upright orientation but is lacking the horizontal and vertical bias typical of most real-world scenes. We will call this the configuration/orientation hypothesis.

Method

Participants

Thirty-two Kansas State University introductory psychology students (19 female, 13 male) participated for course credit (age: $M = 19.12$, $SD = 1.11$). All had normal or corrected-to-normal ($\leq$20/30) vision. Institutional Review Board–approved written informed consent was obtained.

Stimuli

As in Experiment 1, stimulus images began as square at 736 × 736 pixels and then were windowed with an edge-ramped (ramped to mean luminance) circular aperture (736 pixels in diameter). This was done in order to avoid biasing interpretation of the upright in the gravitational frame when the images were rotated by 45°, 90°, 135°, or 180° (see Figure 8). Scenes were RMS-contrast normalized to a target RMS contrast of 0.18, and normalization was only on the basis of information contained within the circular image window. In addition, all masking images generated from the target images were rotated by the same angle as the targets in a given rotation condition. A total of 300 images were used for each view (aerial or terrestrial), making six images per scene category per rotation per view per subject. Each scene base image was randomly assigned to a given degree of rotation for each subject, and each base image was shown only once to any given subject.

Design and procedure

The study was a 2 (view: aerial vs. terrestrial) × 2 (SOA: 24 vs. 318 ms) × 10 (scene category: five natural, five man-made) × 5 (rotation: 0° [upright], 45°, 90°, 135°, or 180° [inverted]) within-subjects design. The procedures were the same as in Experiment 1, with the exceptions that there were only two interstimulus intervals, 0 and 294 ms (i.e., SOAs of 24 and 318 ms) and that mask duration was reduced to 24 ms for a 1:1 target:mask duration ratio. We used a weaker target:-mask ratio than in Experiment 1 in order to allow for better performance, because the task was more difficult due to the image rotation manipulation. Pilot testing indicated that a 318-ms masking SOA (roughly equal to the average fixation duration in free scene viewing) produced equal performance to a no-mask condition.

Results

Prior to the main analyses, data from three participants were excluded from the analysis due to low

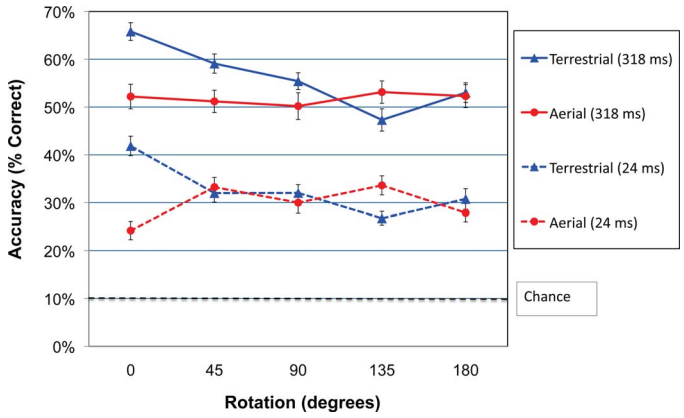


Figure 9. Rapid scene categorization accuracy as a function of view (aerial vs. terrestrial), rotation (0°, 45°, 90°, 135°, or 180°), and processing time (SOA in ms). Error bars = standard error of the mean. Dashed line = chance performance.

performance scores (i.e., >2 standard deviations below the mean of the group). We then conducted linear mixed-effects modeling to determine fixed effects of scene view (aerial vs. terrestrial), rotation, and processing time (SOA) on accuracy from the random effect of the participants (Bates et al., 2014). As in Experiment 1, participants' accuracy means were logit-transformed prior to conducting these tests (Cohen & Cohen, 1983). Confusion matrices for all conditions are given in Supplementary Figure S2.

First, model selection procedures used in Experiment 1 were repeated for Experiment 2. Twelve different models with identical fixed effects (Rotation × SOA × View) and varying random-effects structures (e.g., additive or interactive random effects structures of rotation, SOA, and view) were run and then compared on the basis of model-fit indices (Akaike information criterion and BIC). The model with the lowest BIC contained view as the only necessary random-effects component (BIC = 822.46). As shown in Figure 9, later fixed-effects tests reveal that there was a significant effect of SOA, $F(1, 28) = 824.05$, $p < 0.001$, $f^2 = 1.88$, which did not differ across the two views, $F(1, 28) = 0.207$, $p = 0.877$. There was a small but significant difference between aerial and terrestrial scenes, $F(1, 28) = 9.85$, $p = 0.004$, $f^2 = 0.20$, such that accuracy for terrestrial views was greater than for aerial views. The attenuated effect of scene view observed in Experiment 2 when compared to Experiment 1 ($f^2 = 1.49$) can likely be attributed to the significant effect of rotation, $F(1, 28) = 20.99$, $p < 0.001$, $f^2 = 0.29$, and the strong interaction between view and rotation, $F(1, 28) = 52.15$, $p < 0.001$, $f^2 = 0.47$, due to the fact that rapid categorization of terrestrial views of scenes was significantly affected by rotation, $F(1, 28) = 74.74$, $p < 0.001$, $f^2 = 0.80$, whereas that was not the case for aerial views, $F(1, 28) = 3.26$, $p = 0.082$. This is consistent with the hypothesis that gist recognition for terrestrial scenes is impaired in the absence of the predictable scene layout afforded by the upright scene orientation, which is further evidence that terrestrial scene gist recognition is viewpoint-dependent. There were no significant interactions between SOA and rotation for either aerial or terrestrial scenes—$F(1, 28) = 0.733$, $p = 0.39$, and $F(1, 28) = 1.00$, $p = 0.33$, respectively— and no significant three-way interaction between view, SOA, and rotation, $F(1, 28) = 0.018$, $p = 0.506$.

We conducted a more fine-grained test of the three alternative hypotheses by constructing mixed-effects models to determine the optimal polynomial function to describe the terrestrial data. More specifically, we conducted polynomial regressions of orders 1, 2, and 4 on the assumptions that the configuration hypothesis (Figure 7A) would be fitted best by a simple linear equation, the orientation-bias hypothesis (Figure 7B) by a quartic function, and the configuration/orientation hypothesis (Figure 7C) by a quadratic function. Each of the three polynomial fits was constructed with two different variations, in which the random-effects structure included either the subject mean or the Subject Mean × Rotation interaction, resulting in six mixed-effects models (shown in Table 3). Allowing the residual error to be differentially accounted for by the random-effects structures produces a clear account of how consistent the fixed effects of rotation were across subjects. All models were constructed in R (version 3.0.2) using the lme4 library (Bates et al., 2014).

Fit type	Random-effects structure	AIC	BIC
Linear	Subject mean	646.61	661.29
Linear	Subject mean × Rotation	650.53	672.55
Quadratic	Subject mean	640.60	658.95
Quadratic	Subject mean × Rotation	644.52	670.21
Quartic	Subject mean	641.71	667.40
Quartic	Subject mean × Rotation	645.63	678.66

Table 3. Polynomial regression model selection results for each of the six models generated from the terrestrial accuracy results. *Notes*: The strict configuration hypothesis is represented by the linear fit, the configuration/orientation hypothesis by the quadratic function, and the strict orientation-bias hypothesis by the quartic fit. Models were evaluated in terms of the Akaike and Bayesian information criterion (AIC and BIC) fit indices, with the lowest values representing the optimal model.

As shown in Table 3, the data conform to a quadratic function where the intercept varies across subjects, but the regression slope does not. Thus, while there was variability in terms of overall accuracy across participants, the slopes across levels of rotation were relatively consistent. In terms of the fixed effect of rotation on accuracy, we see a significant negative slope for the first term in the regression equation, $B = -0.001$, $t(28) = -4.43$, $p < 0.001$, and a positive slope for the second term, $B = 3.4 \times 10^{-5}$, $t(28) = 2.84$, $p = 0.008$. When compared to the alternative models, the simple quadratic model achieved the lowest values for our model fit indices (Akaike information criterion = 640.60; BIC = 658.95). Thus this is the model in which the data are best fitted and is the least likely model to overfit the data to the model.

Discussion

The goal of Experiment 2 was to explore the idea that aerial scenes lack particular spatial diagnostic cues referenced to the gravitational frame, which are found in terrestrial scenes. We tested that hypothesis by rotating both terrestrial and aerial views, based on the prediction that rotation would have a greater disruptive effect on terrestrial scenes than aerial scenes, and the results were consistent with that prediction. The results support the hypothesis that rotation disrupted the gravity-based coordinate frame for terrestrial views, removing useful information for identifying them that was always missing in aerial views. Thus, the lack of the diagnostic gravitational-frame constraint may constitute a data limitation of aerial views of scenes. In addition, the differential effects of image rotation on rapid categorization of terrestrial and aerial views is further evidence of view-dependence for terrestrial-view scene gist recognition and view-independence for aerial-view gist recognition.

Another goal of Experiment 2 was to test three alternative hypotheses about the types of information that are diagnostic for rapidly categorizing terrestrial views, namely the configuration/layout of a scene versus the dominant-orientation information in a scene or a combination of the two. The results were consistent with scene configuration being more diagnostic than dominant orientation but also with an interaction between the two cues when categorizing terrestrial scenes. Overall, increasing deviations from upright referenced to the gravitational frame created increasingly greater difficulty in categorizing terrestrial scenes, consistent with scene configuration/layout being highly diagnostic. However, the greatest difficulty was not for scenes rotated 180° from upright, but instead for those rotated 135° from upright, which is an oblique angle. Because oblique angles are those least commonly occurring in terrestrial natural scene views, this points to scenes' dominant orientation being somewhat diagnostic as well.

The terrestrial-view rotation data also provide insight into the results of several previous conflicting studies that investigated the effects of rotation, most often inversion (a 180° rotation), on terrestrial-scene perception. Some studies have found evidence of an inversion effect. Kelley, Chun, and Chua (2003) found that differences in change detection in natural scenes, which depended on whether the change was meaningful or not, disappeared when the scenes were inverted. Likewise, Walther et al. (2009) found that basic-level scene categorization was significantly worse for inverted scenes than upright scenes. Conversely, Guyonneau, Kirchner, and Thorpe (2006) found that animal detection in natural scenes was virtually unaffected by 16 different degrees of rotation. Rieger, Köchy, Schalk, Grüschow, and Heinze (2008) found that object discrimination in natural scene pairs was unaffected by scene inversion (180° rotation) but was inhibited by intermediate (e.g., 45°, 90°, and 135°) orientation changes. The results from our Experiment 2 provide support for the former studies showing inversion-based performance decrements. The contradictory findings from the latter studies may be because the categorization tasks were more object centered, whereas here (and in the former studies) the tasks were more scene centered.

Our results also allowed us to test a further hypothesis regarding the time course of processing dominant-orientation versus configuration/layout information in terrestrial scenes. Specifically, global oriented amplitude in scenes is predominately processed in primary visual cortex (Blasdel, 1992; Bonhoeffer & Grinvald, 1991; De Valois, Yund, & Hepler, 1982; Hubel, Wiesel, & Stryker, 1978; Shapley, Hawken, & Ringach, 2003). Conversely, specific configurations of local orientations seem likely to be processed later in the ventral stream (Oliva & Torralba, 2006)—for example, in the parahippocampal place area and the lateral occipital complex, which have been shown to respond differentially to different scene configurations (Park, Brady, Greene, & Oliva, 2011; Walther et al., 2009). Thus, one might hypothesize that at early processing times, the seemingly simpler dominant-orientation information would be more important, whereas at later processing times, the seemingly more complex configuration/layout information would be more important. However, this hypothesis was not supported by our results, because the effects of rotation on rapid scene categorization of terrestrial views were similar between early (24 ms) and late (330 ms) processing times. Instead, these results are consistent with the hypothesis that scene configuration/layout is diagnostic even at very early processing times.

In sum, the results of Experiment 2 show that rapid scene categorization is strongly view dependent for terrestrial views but view independent for aerial views. Specifically, within that framework, the results suggest that an important constraint used to rapidly categorize terrestrial views of scenes is the gravitational frame, which is absent in aerial views of scenes and thus may constitute a data limitation for them and increase the difficulty in categorizing them. This latter conclusion raises the interesting question of just what image properties might be diagnostic for recognizing the gist of an aerial scene, and this question was investigated in Experiment 3.

Experiment 3

The image-statistical analyses reported in Experiment 1 suggest that humans may, in part, be utilizing global oriented amplitude-spectrum information to rapidly categorize aerial scene views. However, the results of Experiment 2 show that whatever cues are diagnostic for people to rapidly categorize aerial scenes must be rotation invariant (i.e., viewpoint independent), which argues that global oriented-amplitude information is not diagnostic for rapidly categorizing aerial scenes (and only modestly diagnostic for terrestrial scenes). Therefore, it may be that specific configurations of local orientations (e.g., image texture) are diagnostic for aerial-scene categorization.

Some recent image-statistical analyses of aerial-scene classification have pointed to the potential utility of texture (Vijayaraj et al., 2008), and other researchers have argued that texture information is sufficient for classifying terrestrial scenes (Fei-Fei & Perona, 2005; Renninger & Malik, 2004). Conversely, our previous research (Loschky et al., 2010) has shown that texture is of very limited utility in recognizing terrestrial scenes. Similar to our previous study (Loschky et al., 2010), one direct way to test the diagnosticity of texture information for both aerial and terrestrial scene gist recognition is to ask viewers to rapidly categorize texture images generated from aerial and terrestrial views of scenes. Portilla and Simoncelli (2000) provide a well-known computational model of texture. Their model identifies and statistically characterizes homogeneous, repeated local patterns in images and then iteratively coerces random-noise images to share the statistical characterization of the modeled texture. The texture synthesis algorithm destroys many of the pictorial depth cues in a scene image (e.g., linear perspective, texture gradient, height in the field) during the process of modeling and synthesizing repeated two-dimensional patterns on the picture plane. Given that recognizing the gist of terrestrial views may rely, in part, on such depth information (Greene & Oliva, 2009; Torralba & Oliva, 2002), this may partly explain why texture images derived from terrestrial views are difficult to recognize (Loschky et al., 2010). Conversely, such pictorial depth cues appear to be either limited or entirely missing in aerial views. Based on this reasoning, we hypothesized that texture should be more diagnostic for rapidly categorizing textures derived from aerial views than from terrestrial views of scenes.

Texture patterns are processed up through middle vision in the ventral visual stream (e.g., the lateral occipital complex; Hiramatsu, Goda, & Komatsu, 2011) and have been argued to be processed very quickly or preattentively (Julesz, 1981; Landy & Graham, 2004). This suggests the hypothesis that texture information may be extracted early in processing. Conversely, if recognizing the gist of a scene on the basis of texture information is a strategic and effortful process, then it would suggest the alternative hypothesis that the rapid categorization of scene texture images requires longer processing times.

To test these hypotheses, Experiment 3 used the Portilla and Simoncelli texture synthesis algorithm (2000) to create textures from both aerial and terrestrial views of scenes and compared their rapid categorization accuracy (in terms of the original scene categories from which they were generated). The experiment also compared performance with short and long masking SOAs to determine whether there were differences in processing such texture information at early versus late processing times.

Method

Participants

Thirty-seven Kansas State University introductory psychology students (17 female, 20 male) participated for course credit (age: $M = 19.7$, $SD = 2.13$). All participants had normal or corrected-to-normal ($\geq$20/30) vision. Institutional Review Board–approved written informed consent was obtained.

Stimuli

The majority of the original scene images used were taken from Experiments 1 and 2, with a few additional images collected from the Internet and Google Earth. Half of the stimuli in the experiment were texture images synthesized from the original images using Portilla and Simoncelli's (2000) texture synthesis algorithm and were matched for mean luminance and RMS contrast as in Experiments 1 and 2. Example original and synthesized texture images are shown in Figure 10.

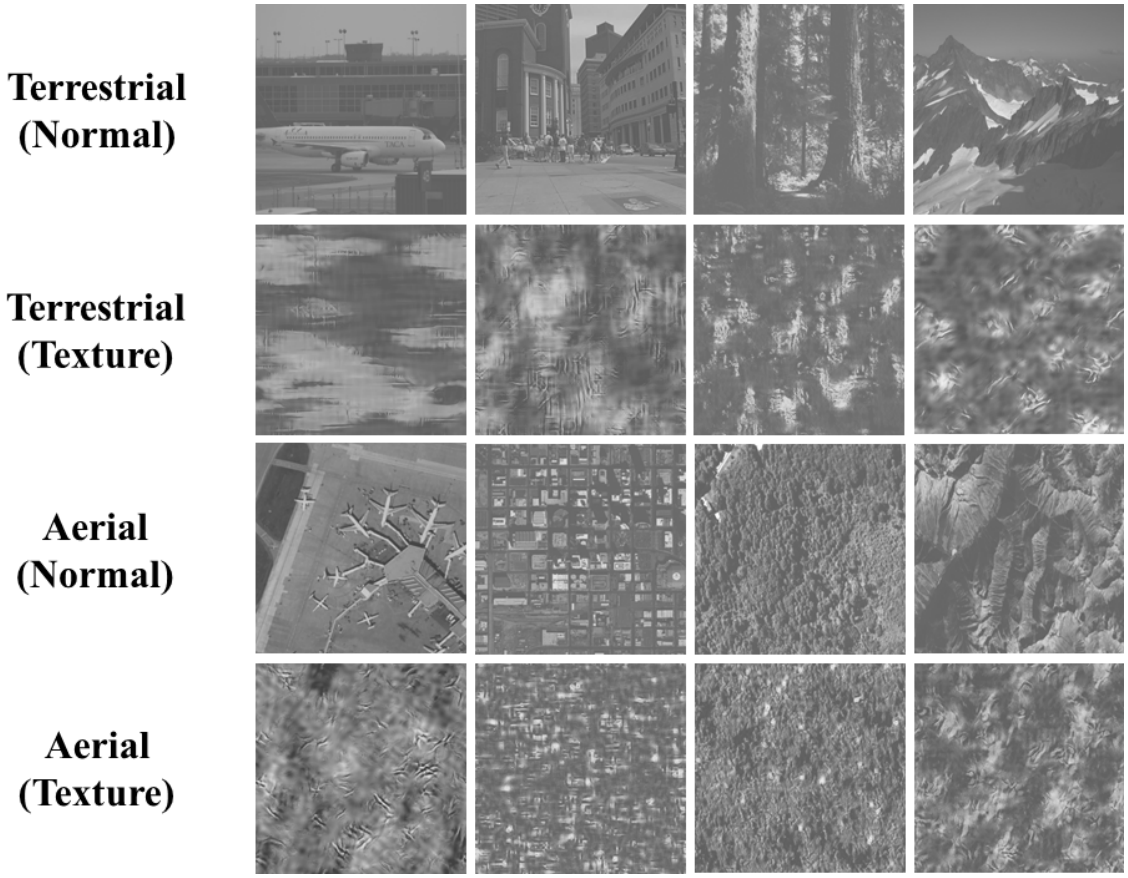


Figure 10. Examples of original scenes and texture images for aerial and terrestrial views used in Experiment 3.

Design and procedure

The design and procedure in Experiment 3 were identical to those in Experiments 1 and 2, with the following exceptions: The experiment had a 2 (view: aerial vs. terrestrial) × 2 (image type: original vs. texture) × 2 (SOA: 35 vs. 330 ms) within-subjects factorial design. The shortest masking SOA was increased from 24 ms (in Experiments 1 and 2) to 35 ms in the current experiment to avoid a floor effect for the texture images. The experiment contained 640 trials in two blocks of 320 trials each, one of aerial views and the other of terrestrial views, with block order counterbalanced across participants. Half of the images in each block were textures, randomly intermixed with the original images. Participants were not explicitly informed about the existence of the texture images, in order to avoid the possibility of their developing different strategies for recognizing them. However, the participants' written instructions explained that some of the images might seem "garbled" or more difficult to distinguish than others, and that they should simply go with their "best guess" if they were unsure of how to respond. In the familiarization and practice trials, the aerial and terrestrial views of scenes were presented in separate blocks, with the terrestrial scenes being shown first because they were easier to recognize. Participants were given 10-min breaks between blocks to reduce fatigue.

Results

Prior to the main analyses, data from three participants were excluded due to low performance scores (i.e., >2 standard deviations below the mean of the group).

Figure 11 shows the confusion matrices for all eight conditions. Perhaps the most striking feature in Figure 11 is the presence of lighter colored vertical bands in the texture-condition confusion matrices (E–H, in the lower row of matrices). Those are primarily in the desert, forest, and mountain response columns, but also to a lesser degree in the river response category for the 330-ms terrestrial texture condition. These brighter vertical stripes represent higher response rates for these response categories, and thus graphically illustrate response biases. Importantly, however, such response biases, shown by brighter vertical stripes, are not apparent in the original-image-condition confusion matrices (A–D, in the upper row of matrices), consistent with the confusion matrices for Experiments 1 and 2 (Figure 4A and B and Supplementary Figures

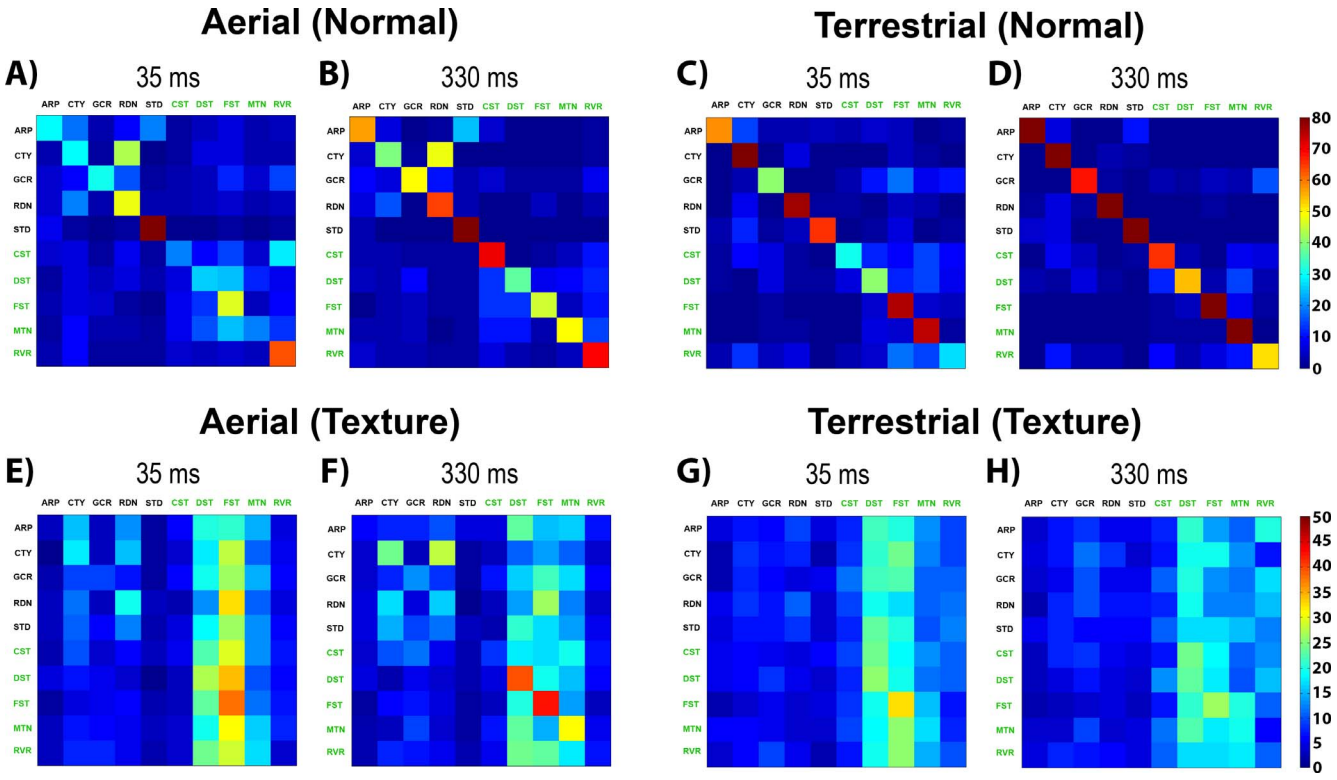


Figure 11. Standard confusion matrices for original and texture images, for aerial and terrestrial views, at short (35 ms) and long (330 ms) SOAs. Within each confusion matrix, rows represent target image category and columns represent average responses made for each response category. Categories have been grouped by man-made (black) and natural (green), arranged in alphabetical order. The color scale is in percentages of each response for a given image category, with the main diagonal representing correct responses and off-diagonals being confusions. ARP: airport; CTY: city; GCR: golf course; RDN: residential; STD: stadium; CST: coast; DST: desert; FST: forest; MTN: mountain; RVR: river.

S1 and S2). The response biases in the texture condition should artificially increase the hit rates for the biased categories and decrease the hit rates for the remaining categories.

In order to eliminate the effects of response bias from the accuracy rates in Experiment 3, we calculated the *response-rate normalized accuracy* using the following formula from Hansen and Loschky (2013): $p(\text{correct response}|\text{response} = X_{\text{cat}})/p(\text{response} = X_{\text{cat}})$, where $X_{\text{cat}} =$ one of the 10 categories. As applied to the confusion matrices, this involved dividing each column in the raw response matrices by the total number of responses for that column and multiplying the matrix by the scalar equal to the presentation rate for all cells (which was constant). The resultant response-rate normalized confusion matrices are shown in Figure 12. Visual comparison of the standard confusion matrices in Figure 11 with the response-rate normalized versions in Figure 12 suggests that the primary difference is in terms of eliminating the response biases found in Figure 11 for the texture images (the brighter vertical stripes for E–H in the lower row). Further inspection of Figure 12 suggests several other observations: (a) The texture-image confusion matrices (E–H) show less differentiation (indicated by lower color contrast) than the original-image confusion matrices (A–D), indicating generally worse performance for the texture images; (b) the terrestrial texture images show the least well-defined main diagonals, indicating the worst accuracy overall; and (c) the four aerial confusion matrices (A, B, E, and F) show systematic confusions in their upper left quadrant between the city and residential categories, in both the original and texture images (as previously shown in Figure 4A).

We also analyzed the Spearman rank order correlations among confusion matrices as a function of view (aerial vs. terrestrial) and image type (original vs. texture) for the 330-ms SOA condition, as shown in Table 4. In general, these correlations suggest the degree to which similar diagnostic information or processing routines (or both) are used across views. The largest correlation, $r(88) = 0.599$, $p < 0.001$, is moderate and is between the original terrestrial and original aerial views. This replicates the results of Experiment 1, which we interpreted as likely suggesting similar processing routines. The correlation between the textures synthesized from aerial and terrestrial views, $r(88) = 0.239$, $p = 0.023$, is considerably smaller.

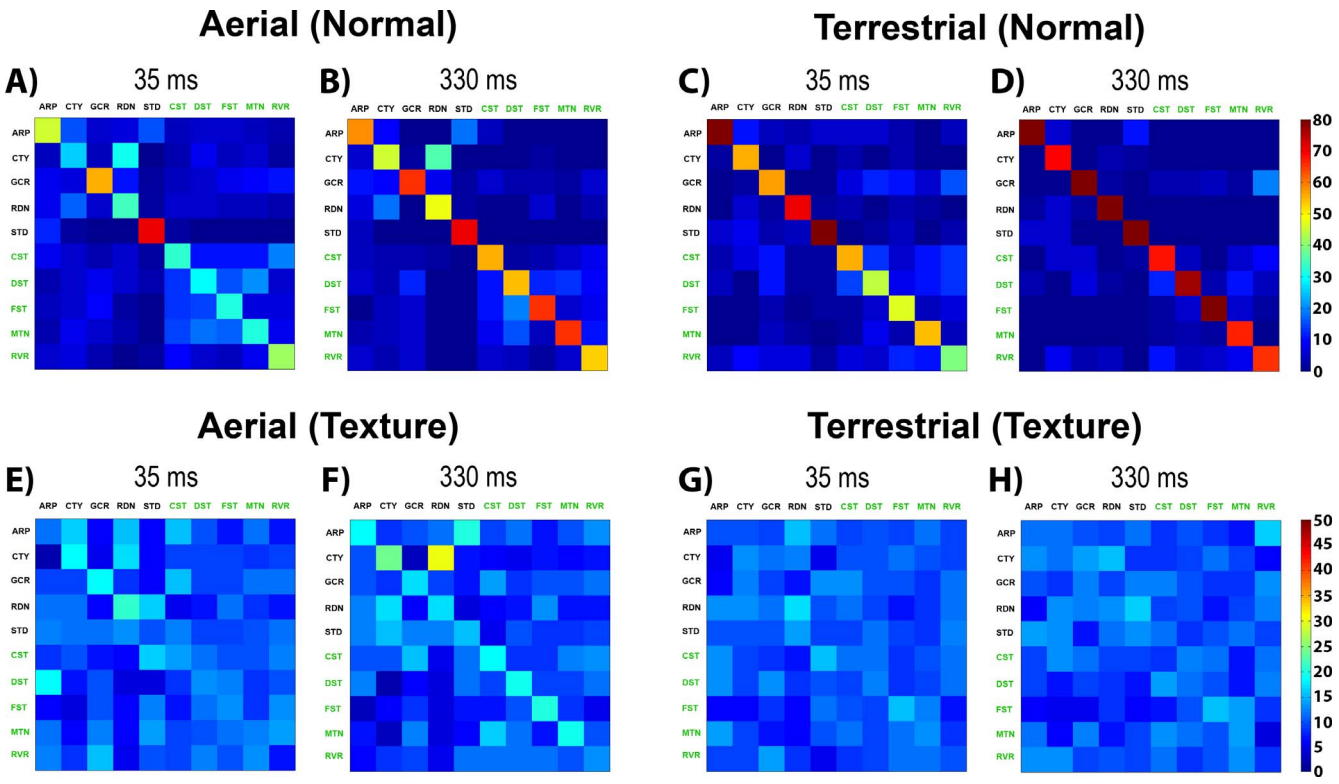


Figure 12. Response-rate normalized versions of the confusion matrices shown in Figure 11.

Slightly larger and of greater interest is correlation between the aerial original and aerial texture images, $r(88) = 0.298$, $p = 0.004$. Whether this is due to shared diagnostic texture information being used across views or shared processing routines is unclear.

We next conducted a 2 (aerial vs. terrestrial) × 2 (original vs. texture) × 2 (long vs. short SOA) repeated-measures regression on the response-rate normalized accuracy data, which were logit transformed. As in Experiments 1 and 2, 12 linear mixed-effects models with identical fixed-effects structures and varying random-effects structures were constructed and compared. A comparison of BIC values demonstrated that the optimal model contained image type (texture vs. normal) as the sole random-effects component that varied across participants (BIC = 458.19). As shown in Figure 13 and suggested by Figure 12, we replicated the terrestrial-view advantage found in Experiment 1, $F(1, 33) = 9.46$, $p = 0.004$, $f^2 = 0.18$, and the expected effect of SOA, $F(1, 33) = 74.48$, $p < 0.001$, $f^2 = 0.519$. We also replicated the finding of Loschky et al. (2010) that rapid scene categorization was much better for original scenes than for the texture images synthesized from them, $F(1, 33) = 696.34$, $p < 0.001$, $f^2 = 1.60$. However, our primary interest was regarding the possible interaction between view (aerial vs. terrestrial) and image type (original vs. texture), which was statistically significant, $F(1, 33) = 153.49$, $p < 0.001$, $f^2 = 0.749$. Specifically, Figure 13 shows that synthesized textures derived from aerial views were more recognizable as members of their scene categories than were textures derived from terrestrial views. This is consistent with the hypothesis that texture information is more useful for recognizing the gist of aerial views than it is for recognizing terrestrial views of scenes. However, Figure 13 also shows a significant three-way interaction

	Aerial original	Terrestrial original	Aerial texture	Terrestrial texture
Aerial original	—	0.599* [0.447, 0.717]	0.298* [0.097, 0.476]	0.164 [−0.044, 0.359]
Terrestrial original		—	0.430* [0.244, 0.585]	0.270* [0.066, 0.452]
Aerial texture			—	0.239* [0.034, 0.425]
Terrestrial texture				—

Table 4. Confusion-matrix Spearman rank order correlations for the 330-ms SOA condition as a function of view (aerial vs. terrestrial) and image type (original vs. texture) in Experiment 3. *Notes*: Correlations reflect response-rate normalized category frequencies, where response rates are equal to *p*(Category Response|Target Category)/*p*(Category Response Frequency Total). *$p < 0.01$ level. Bracketed values = 95% CI.

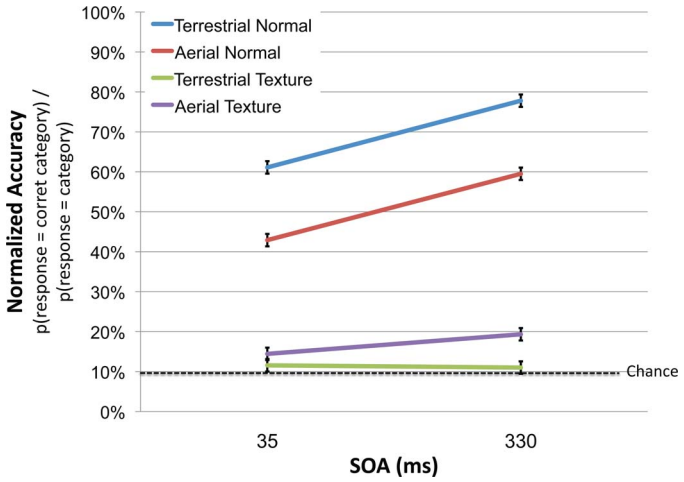


Figure 13. Response-rate normalized accuracy as a function of view (aerial vs. terrestrial), viewing time (SOA in ms), and image type (texture vs. normal). Error bars = standard error of the mean. Dashed line = chance performance.

between view (aerial vs. terrestrial), image type (original vs. texture), and SOA (35 vs. 330 ms), $F(1, 33) = 4.51$, $p = 0.04$, $f^2 = 0.113$. Specifically, there was no significant increase in rapid categorization accuracy for the terrestrial texture images between long and short SOAs, $t(33) = 1.45$, $p = 0.16$, whereas accuracy for aerial texture images increased significantly from short to long SOAs, $t(33) = 2.99$, $p = 0.005$. This significant interaction demonstrates that whatever information is diagnostic for rapidly categorizing aerial texture images becomes available relatively late and may involve relatively effortful processes.

Discussion

Experiment 3 investigated the diagnosticity of rotation-invariant texture information for rapidly categorizing aerial views of scenes. The result at 330 ms SOA showing accuracy nearly twice as high for aerial-scene textures as for terrestrial-scene textures (Figure 13) provides direct human behavioral evidence in support of the hypothesis that texture information is more useful for recognizing aerial views than terrestrial views of scenes. Nevertheless, overall, aerial-scene texture categorization was very poor, ranging from chance to a little under twice the chance rate. In addition, texture information required the processing duration of a full eye fixation (330 ms SOA) to achieve the twice-chance level. Finally, the relatively small correlation (Table 4) between the confusion matrices for categorizing aerial-scene textures and original aerial scenes suggests that the diagnostic information or processing routines involved in categorizing aerial-scene textures, as implemented by the Portilla and Simoncelli (2000) model, are only weakly similar to those involved in categorizing actual aerial scenes.

For rapidly categorizing terrestrial views of scenes, the rotation-invariant information provided by the scene texture images was even less useful and failed to improve as processing time increased from 35 ms up to 330 ms SOA (a full eye fixation). This may be because rapidly categorizing terrestrial images requires global configuration or layout information, which is referenced to the gravitational frame and is destroyed by repetition of local configurations distributed throughout the synthesized texture image (as happens in generating texture images using the Portilla and Simoncelli algorithm). Indeed, given that the confusion-matrix correlation for aerial original versus texture scenes was not significantly greater than its terrestrial counterpart, we speculate that the loss of useful information by the texture algorithm was not limited to global configurations but may also have included loss of local features as well, which seem plausibly diagnostic for recognizing aerial scenes.

In sum, while Experiment 3 showed that texture information is relatively more diagnostic for rapidly categorizing aerial views than terrestrial views, other forms of information must be far more diagnostic for rapidly categorizing both terrestrial and aerial views.

General discussion

The current study investigated the similarities and differences between aerial and terrestrial scene gist recognition as a means of investigating the degree of viewpoint dependence of scene processing, thereby illuminating the diagnostic information sources and visual processes underlying each. We now discuss the theoretical implications of our results, remaining questions, and limitations of the current study.

As noted in our Introduction, one way of explaining viewpoint-dependence versus independence, for object or scene gist recognition, is in terms of the diagnostic information for the task, which according to Schyns (1998) is a function of the available information and task constraints. In the current study, the task for our participants was to determine the basic-level scene category among 10 alternatives. Importantly, numerous recent studies have shown that a fundamental distinction in parsing scene categories is between natural and man-made scenes (Greene & Oliva, 2009; Loschky & Larson, 2010; Rousselet, Joubert, & Fabre-Thorpe, 2005). The MDS results from Experiment 1 indicate that this distinction is important for both aerial and terrestrial views, with the majority of the "natural" basic-level categories clustering together. Thus, this key distinction in rapid scene categorization

seems, to some extent, to be viewpoint-independent. Within Schyns's (1998) diagnostic recognition framework, this implies that the information used to discriminate between natural and man-made scenes is available from both viewpoints and is highly task relevant for rapid scene categorization. However, the results of all three of our experiments show an advantage for terrestrial views over aerial views, indicating a strong degree of viewpoint-dependence in rapid scene categorization in general. According to the diagnostic recognition framework, this fundamental result suggests that task-relevant information for rapid scene categorization is more readily available from terrestrial views of scenes than from aerial views. Such an argument is consistent with our earlier claim that aerial views are more data-limited than terrestrial views.

Related to this, it is important to acknowledge that the results we have found are, as in all studies of categorization, determined to a certain degree by the diagnosticity of the information available in our scenes for the specific category sets we used. A good example of this is the fact that a follow-up study to the current experiments, done with pigeons (Kirkpatrick, Sears, Hansen, & Loschky, 2014), did a similar image-statistical analysis to that reported in Experiment 1 and, contrary to the results of the current study, showed greater scene-category discrimination for aerial views than terrestrial views. However, that experiment only included two scene categories: coasts and mountains. Conversely, the current study compared 10 scene categories, including coasts and mountains, and found that in this larger sample of scene categories, terrestrial views were more discriminable than aerial views. Nevertheless, and interestingly, the current study found that the "river" and "stadium" categories were more accurately recognized from an aerial view than a terrestrial view. As before, we can explain this result in terms of the task-relevant information being more readily available from the aerial views. For instance, river scenes from aerial views are more likely to contain highly diagnostic snakelike curves than terrestrial views, and likewise, an aerial view of a stadium shows the overall shape (e.g., a baseball diamond), which is highly diagnostic but is not easily seen from a terrestrial view.

All of these results can be explained purely in terms of the diagnostic information available to the viewer for the task of rapid scene categorization, without making inferences about "the format of [scene] representations in memory" (Schyns, 1998, p. 166). However, we believe that the scene-rotation results of Experiment 2 cannot be fully explained in that way. Those results showed that rotating terrestrial views of scenes away from the canonical orientation greatly disrupted rapid scene categorization but did not affect aerial-view rapid categorization, which was uniformly lower in accuracy. We have argued that this is evidence of the diagnosticity of configural information for categorizing terrestrial-scene views. However, in order to explain *how* the configuration (or layout) of, say, a coast image is changed by a 180° rotation, one needs to establish a reference frame (e.g., the gravitational frame) relative to which the rotation occurs. We next explain the difference in effect of rotation between aerial and terrestrial views by assuming that aerial views *lack* such a gravitational reference frame.

Thus far, our argument is completely framed in terms of task-relevant diagnostic information (i.e., configuration information is diagnostic for terrestrial scenes, and the gravitational reference frame constitutes missing diagnostic information for aerial scenes). However, we believe this explanation is incomplete, because it does not explain why either changing a scene's configuration (by rotating it relative to the gravitational frame) or completely removing a gravitational reference frame should make rapid scene categorization more difficult. We believe that the answer to these questions requires specifying the format of scene memory representations. For example, a simple explanation is in terms of constraining the process of matching memory representations of scenes with the retinotopically mapped perceptual activation pattern. If memory representations for scenes are derived from experience (whether personal experience or evolutionarily encoded), such that their configurations are largely fixed relative to the gravitational frame, then matching retinotopic activation patterns to memory representations will be greatly constrained (Valentine, 1988; Zelinsky & Schmidt, 2009). Thus, if one has well-established memory representations that are constrained in this way, rotating images and thus changing their retinotopic configuration should make the matching process more difficult, as we showed for terrestrial views.[7] However, if memory representations are not so constrained (as we may assume is the case for aerial views), then the process of searching for a match between the retinotopic pattern and a memory representation will be less constrained, and thus more difficult, regardless of the scene orientation. Likewise, learning aerial scene categories would be predicted to be more difficult than learning terrestrial categories, precisely because aerial scenes should be highly variable in terms of their orientation relative to the gravitational frame (Shiffrin & Schneider, 1977).

Theories of expert object recognition have often argued that configural processing lies at the heart of rapid, accurate object recognition (Palmeri & Cottrell, 2009). Configural processing means that within a given class of objects, there is an invariant or predictable structure that is inherent in all exemplars that facilitates the rapid encoding of simple, low-level information.

When the configuration is violated (e.g., by rotation), matching such perceptual information to the memory representation is impaired. However, this effect is only found when the object classes share a similar configuration and individuals within a class can be described by second-order relational information among parts (Diamond & Carey, 1986). Clearly, terrestrial views of scenes from the same or similar categories do have similar configurations or layouts (Oliva & Torralba, 2006; Sanocki, 2003; Schyns & Oliva, 1994), while aerial scenes would seem less so. Visual expertise is famously viewpoint-dependent (Diamond & Carey, 1986; Maurer, Grand, & Mondloch, 2002; Rossion, Gauthier, Goffaux, Tarr, & Crommelinck, 2002; Valentine, 1988), and given that aerial views have no canonical viewpoint from which one could expect to see them, the likelihood of aerial views fitting into a view-dependent "expert" processing framework is quite low. If we add to this the fact that we have found further evidence for other data limitations in aerial views (e.g., in terms of available amplitude-spectrum information), it strengthens the conclusion that it may be impossible to ever achieve similar levels of expertise for aerial views compared to terrestrial views.

One of the most famous examples of perceptual expertise producing strong configural processing, and large decrements with image rotation, is in face recognition. Thus, explanations of the effects of inversion on face recognition may provide further insights into the effects of scene inversion. Research has shown that face inversion disrupts configural processing but not featural processing (Diamond & Carey, 1986; Farah, Tanaka, & Drain, 1995; Maurer et al., 2002; Rossion et al., 2002; Valentine, 1988). In fact, Farah et al. (1995) showed that such disruption of configural processing by image inversion is not limited to faces but can also be found with random arrays of dots. If such a dot array was processed as a configuration, its recognition was disrupted by inversion, but if it was color-coded such that the array could be parsed into individual parts, or features, its recognition was less affected by inversion. An analogy to such findings suggests that terrestrial scenes are processed in terms of their configuration or layout, which is consistent with our previous conclusions, whereas aerial scenes may instead be processed in terms of their parts or features. This hypothesis should be tested in further research.

It is important to note that the similarity between scene- and face-inversion effects is likely limited by a number of factors. For this reason, we must be cautious in arguing for similarities between scene and face processing. Although both scenes and faces are "mono-oriented"—our experience with both is almost always with a predictable, canonical orientation due to the gravitational frame—faces and scenes differ in that faces are typically identified at the subordinate level (Maurer et al., 2002), unlike scenes. Furthermore, faces and scenes differ in the degree to which their first-order configural relations are constrained (Diamond & Carey, 1986). Specifically, all faces have the same first-order configural relations (two eyes above a nose, which is above a mouth); whereas, at most, all terrestrial views of scenes have a visible or implied horizon with the ground below it and the sky above it. None of these configural constraints, referenced to the gravitational frame, apply to aerial views of scenes. Further work should investigate the extent to which different types of configural processes that have been identified with face processing (i.e., first-order, holistic, and second-order; Maurer et al., 2002) operate in processing terrestrial views of scenes.

Experiment 3 showed that rotation-invariant texture features, captured by the Portilla and Simoncelli (2000) texture algorithm, was roughly twice as useful for recognizing aerial scenes as for recognizing terrestrial scenes, but only for scene categories consisting largely of homogeneous repeated patterns (e.g., forests, deserts, or mountains). Nevertheless, the results also showed that such texture information is of quite limited use even for categorizing aerial views, and that it requires considerable processing time. Therefore, while Experiment 2 clearly demonstrated that aerial views must be categorized using rotation-invariant information, Experiment 3 indicated that it is not simply homogeneous pattern information. One such possibility would be local heterogeneous configurations (e.g., the shapes of airplanes at an airport, or the shapes of stadiums). Another possible diagnostic information source would be larger rotation-invariant configurations. For example, the general configuration of coasts from an aerial view consists of two regions, land and sea, each with a different luminance and texture dividing the image. A third possibility is that whatever monocular/pictorial depth cues are available in aerial scenes (e.g., cast shadows, linear perspective of structures off of the central axis of vision) may contribute rotation-invariant information for rapid aerial-scene categorization. Neither local heterogeneous configurations, such as planes at an airport, nor larger heterogeneous configurations, such as coasts, nor monocular/pictorial depth cues, such as cast shadows, are preserved by the Portilla and Simoncelli (2000) texture algorithm. This may explain why scene textures derived from aerial views of airports and coasts were unrecognizable.

Nevertheless, the current study does not clearly falsify the hypothesis that texture information may be used to recognize either aerial or terrestrial views of scenes. Rather, it falsifies that hypothesis as operationalized in terms of categorizing the output of the Portilla and Simoncelli (2000) algorithm (most clearly

for terrestrial views). While that model is widely regarded as successful in generating texture images, it seems different in various ways from the conception and operationalization of texture previously argued to be important for rapid scene categorization (Fei-Fei & Perona, 2005; Renninger & Malik, 2004). The latter studies operationalized texture as the output of a bank of filters composed of "textons" (unique configurations of Gabor patches), with matching between analyzed input images and memory representations being operationalized in terms of histogram matching. Note that such an operationalization of texture does not require repetition of homogenous patterns, which is central to the definition of texture in the Portilla and Simoncelli (2000) model.[8]

Overall, this study shows that in scene gist recognition, viewpoint matters, and that this can be largely explained in terms of differential availability of diagnostic information. However, the results also suggest that the memory representations for scenes, at least from terrestrial views, are constrained such that they match retinotopic images aligned with the gravitational frame. Importantly, this does not appear to be the case for aerial views. These differences suggest intriguing possibilities involving expertise, which is surely far greater for terrestrial views than aerial views, and how this affects perception of each view. Most importantly, terrestrial views seem to be processed configurally (in terms of scene layout). Such configural processing of terrestrial scenes may be necessary for more computationally difficult scene perception tasks (e.g., navigation, visual search). Configural processing does not appear to occur for aerial views, which may instead be processed in terms of their rotation-invariant features. Indeed, it is entirely possible, even plausible, that there are viewpoint-invariant diagnostic information sources and processing mechanisms involved in terrestrial scene gist recognition, as indicated by the work of Walther and colleagues (Walther et al., 2011; Walther & Shen, 2014), and further research will need to be done to further illuminate these information sources and mechanisms.

Keywords: scene gist, rapid scene categorization, scene classification, viewpoint dependence, viewpoint independence, aerial photography, satellite photography, aerial views, satellite views, terrestrial views, image rotation, image statistics, texture, layout, configuration, time course of perception

Acknowledgments

This study was supported by funding from the Kansas Space Grant Consortium to LCL, RVR, and KE. We would like to thank Michael E. Young for providing statistical advice and suggesting the MDS analysis of the data reported in Experiment 1, Tyler Freeman for suggesting the orientation-bias hypothesis in Experiment 2, and Tera Walton for giving helpful comments for revision. Research in this article was previously presented at the 2010 Annual Meeting of the Vision Sciences Society, with the abstract published in the *Journal of Vision*.

Commercial relationships: none.
Corresponding author: Lester C. Loschky.
Email: loschky@ksu.edu.
Address: Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA.

Footnotes

[1] The term *scene gist* is an important theoretical construct in theories of scene perception (Rayner, Smith, Malcolm, & Henderson, 2009; Wolfe, Vo, Evans, & Greene, 2011) because it has been shown to influence later processes such as 1) *attentional guidance* (Eckstein, Drescher, & Shimozaki, 2006; Gordon, 2004; Torralba, Oliva, Castelhano, & Henderson, 2006), 2) *object recognition* (Bar & Ullman, 1996; Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce & Pollatsek, 1992; Davenport & Potter, 2004; but see Hollingworth & Henderson, 1998), and *long-term memory* for scenes (Brewer & Treyens, 1981; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989). Nevertheless, the theoretical construct *scene gist* implies more than the operational definitions of it, as is the case with virtually every *mentalistic* theoretical construct that has been investigated since the Cognitive Revolution overthrew radical Behaviorism in the 1960s. Therefore, to indicate that the theoretical construct of *scene gist* implies more than we measure, throughout the current paper, we have adhered to the following rule: When discussing the theoretical construct and implications of our study we use the term scene gist; when discussing the method or results we refer to the operational definition we have used, namely *rapid scene categorization*.

[2] Masking SOAs have been shown to strongly influence the time course of brain activity, specifically the amount of time the brain has to integrate sensory information for further processing, as measured in humans by magnetoencephalography (Rieger, Braun, Bulthoff, & Gegenfurtner, 2005) and in macaques by single-cell recordings (Kovacs, Vogels, & Orban, 1995; Rolls, Tovée, & Panzeri, 1999). Nevertheless, it is important to note that the relationship between masking SOAs and the time course of brain activity, while highly systematic, is not one-to-one (VanRullen, 2011).

[3] Upon review, three of the 32 terrestrial airport scenes were found to be indoor scenes rather than outdoor scenes. However, after removing these images and reanalyzing the data, we observed no substantial changes in the F statistics for Experiment 1. Given the lack of change, the original results were maintained for Experiments 1–3.

[4] Although we conducted inferential statistical analyses on logit-transformed data, in reporting descriptive statistics we give the untransformed accuracies for ease of interpretation.

[5] Cohen's f^2 magnitudes for small, medium, and large effect sizes are generally given as 0.10, 0.25, and 0.40, respectively (Cohen, 1988).

[6] We thank Michael E. Young for suggesting and helping with this analysis.

[7] This also raises the interesting question of whether terrestrial-scene categorization of rotated scenes involves mental rotation (e.g., Tarr & Pinker, 1989).

[8] A strong test of the mutual compatibility of these two approaches to operationalizing texture in scenes would be to (1) use the Renninger and Malik (2004) algorithm to classify a set of scene images, (2) use the Portilla and Simoncelli (2000) algorithm to generate textures from those same scene images, and then (3) use the Renninger and Malik algorithm to classify the set of the Portilla and Simoncelli textures. If the Renninger and Malik algorithm performed as well or better with the Portilla and Simoncelli (2000) textures as with original scene images, it would demonstrate clear mutual compatibility between the two operationalizations of texture in scenes. Otherwise, it would indicate a mismatch in how the two models conceptualize texture.

References

Alonso, M. C., & Malpica, J. A. (2008). Classification of multispectral high-resolution satellite imagery using LIDAR elevation data. In G. Bebis, R. Boyle, & B. Parvin (Eds.), *Advances in Visual Computing, 4th International Symposium on Visual Computing, ISVC 2008, Las Vegas, NV, USA, December 2008, Proceedings, Par II* (pp. 85–94). Berlin Heidelburg: Springer-Verlag.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). Lme4: Linear mixed-effects models using Eigen and S4 (Version 1.0-6) [Computer software]. Retrieved from http://CRAN.R-project.org/package=lme4

Bhagavathy, S., Newsam, S., & Manjunath, B. S. (2002). Modeling object classes in aerial images using texture motifs. *Proceedings of the 16th International Conference on Pattern Recognition, 2,* 981–984.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance, 21*(6), 1506–1514.

Biederman, I., Mezzanotte, R., & Rabinowitz, J. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143–177.

Biederman, I., Rabinowitz, J., Glass, A., & Stacy, E. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103,* 597–600.

Blasdel, G. G. (1992). Differential imaging of ocular dominance and orientation selectivity in monkey striate cortex. *The Journal of Neuroscience, 12*(8), 3115–3138.

Bonhoeffer, T., & Grinvald, A. (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature, 353*(6343), 429–431, doi:10.1038/353429a0.

Boyce, S., & Pollatsek, A. (1992). Identification of objects in scenes: The role of scene background in object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 531–543.

Breitmeyer, B. G., & Ogmen, H. (2000). Recent models and findings in visual backward masking: A comparison, review, and update. *Perception & Psychophysics, 62*(8), 1572–1595.

Breitmeyer, B. G., & Ogmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision*. Oxford, UK: Clarendon Press.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*(2), 207–230.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum Associates.

Davenport, J. L., & Potter, M. C. (2004). Scene

consistency in object and background perception. *Psychological Science, 15*(8), 559–564.

Davies, C., Tompkinson, W., Donnelly, N., Gordon, L., & Cave, K. (2006). Visual saliency as an aid to updating digital maps. *Computers in Human Behavior, 22*(4), 672–684, doi:10.1016/j.chb.2005.12.014.

De Valois, R. L., Yund, E. W., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research, 22*(5), 531–544.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*(2), 107–117.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science, 17*(11), 973–980.

Ehinger, K. A., & Oliva, A. (2011, July). *Canonical views of scenes depend on the shape of the space*. Paper presented at the 33rd Annual Cognitive Science Conference, Boston, MA.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4*(9), 345–352.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 628–634.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1):10, 1–29, http://www.journalofvision.org/content/7/1/10, doi:10.1167/7.1.10. [PubMed] [Article]

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In C. Schmid, S. Soalto, & C. Tomasi (Eds.), *Computer vision and pattern recognition, 2005, Vol. 2* (pp. 524–531). Los Alamitos, CA: IEEE Computer Society.

Foster, D. H., & Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 269*(1503), 1939–1947.

Gordon, R. D. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 760–777.

Gosselin, F., & Schyns, P. G. (2001). Why do we SLIP to the basic level? Computational constraints and their implementation. *Psychological Review, 108*(4), 735–758.

Graesser, J., Cheriyadat, A., Vatsavai, R. R., Chandola, V., Long, J., & Bright, E. (2012). Image based characterization of formal and informal neighborhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5*(4), 1164–1176.

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

Gregory, E., & McCloskey, M. (2010). Mirror-image confusions: Implications for representation and processing of object orientation. *Cognition, 116*(1), 110–129, doi:10.1016/j.cognition.2010.04.005.

Guyonneau, R., Kirchner, H., & Thorpe, S. J. (2006). Animals roll around the clock: The rotation invariance of ultrarapid visual processing. *Journal of Vision, 6*(10):1, 1008–1017, http://www.journalofvision.org/content/6/10/1, doi:10.1167/6.10.1. [PubMed] [Article]

Haji-Khamneh, B., & Harris, L. R. (2010). How different types of scenes affect the Subjective Visual Vertical (SVV) and the Perceptual Upright (PU). *Vision Research, 50*(17), 1720–1727. doi:10.1016/j.visres.2010.05.027.

Hansen, B. C., & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision, 4*(12):5, 1044–1060, http://www.journalofvision.org/content/4/12/5, doi:10.1167/4.12.5. [PubMed] [Article]

Hansen, B. C., & Hess, R. F. (2007). Structural sparseness and spatial phase alignment in natural scenes. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 24*(7), 1873–1885.

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):21, 1–21, http://www.journalofvision.org/content/13/13/21, doi:10.1167/13.13.21. [PubMed] [Article]

Harris, L. R., Jenkin, M., Dyde, R. T., & Jenkin, H. (2011). Enhancing visual cues to orientation: Suggestions for space travelers and the elderly. In A. M. Green, C. E. Chapman, J. F. Kalaska, & F. Lepore (Eds.), *Progress in brain research, Vol. 191* (pp. 133–142). Oxford, UK: Elsevier.

Hayward, W. G. (2003). After the viewpoint debate: Where next in object recognition? *Trends in Cognitive Sciences, 7*(10), 425–427, doi:10.1016/j.tics.2003.08.004.

Hiramatsu, C., Goda, N., & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *NeuroImage, 57*(2), 482–494, doi:10.1016/j.neuroimage.2011.04.056.

History of aerial photography. (n.d.). Retrieved April 27, 2015, from http://professionalaerialphotographers.com/content.aspx?page_id=22&club_id=808138&module_id=158950.

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

Hubel, D. H., Wiesel, T. N., & Stryker, M. P. (1978). Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology, 177*(3), 361–379, doi:10.1002/cne.901770302.

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.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature, 290*(5802), 91–97.

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition, 15*(1), 12–19.

Kelley, T. A., Chun, M. M., & Chua, K.-P. (2003). Effects of scene inversion on change detection of targets matched for visual salience. *Journal of Vision, 3*(1):1, 1–5, http://www.journalofvision.org/content/3/1/1, doi:10.1167/3.1.1. [PubMed] [Article]

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46*(11), 1762–1776, doi:10.1016/j.visres.2005.10.002.

Kirkpatrick, K., Sears, T., Hansen, B. C., & Loschky, L. C. (2014). Scene gist categorization by pigeons. *Journal of Experimental Psychology: Animal Behavioral Processes, 40*(2), 162–177, doi:10.1037/xan0000014.

Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward-masking. *Proceedings of the National Academy of Sciences, USA, 92*(12), 5587–5591.

Landy, M. S., & Graham, N. (2004). Visual perception of texture. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1106–1118). Cambridge, MA: MIT Press.

Lansdale, M., Underwood, G., & Davies, C. (2010). Something overlooked? How experts in change detection use visual saliency. *Applied Cognitive Psychology, 24*(2), 213–225, doi:10.1002/acp.1552.

Lloyd, R., Hodgson, M. E., & Stokes, A. (2002). Visual categorization with aerial photographs. *Annals of the Association of American Geographers, 92*(2), 241–266.

Loschky, L. C., Hansen, B. C., Sethi, A., & Pydimari, T. (2010). The role of higher-order image statistics in masking scene gist recognition. *Attention, Perception & Psychophysics, 72*(2), 427–444.

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., 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.

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255–260, doi:10.1016/S1364-6613(02)01903-4.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*(1), 44–64, doi:10.1016/0010-0285(75)90004-3.

Ogmen, H., & Breitmeyer, B. (2006). *The first half second: The microgenesis and temporal dynamics of unconscious and conscious visual processes*. Cambridge, MA: MIT 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, 155,* 23–36.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance IX, Vol. 1* (pp. 135–151). Hillsdale, NJ: Lawrence Erlbaum.

Palmeri, T. J., & Cottrell, G. (2009). Modeling perceptual expertise. In I. A. Gauthier, M. J. Tarr, & D. Bub (Eds.), *Perceptual expertise: Bridging brain and behavior* (pp. 197–244). New York: Oxford University Press.

Pannasch, S., Helmert, J. R., Hansen, B. C., Larson, A. M., & Loschky, L. C. (2014). Commonalities and differences in eye movement behavior when ex-

ploring aerial and terrestrial scenes. In M. Buchroithner, N. Prechtel, & D. Burghardt (Eds.), *Cartography from pole to pole* (pp. 421–430). Berlin: Springer-Verlag.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *The Journal of Neuroscience, 31*(4), 1333–1340, doi:10.1523/JNEUROSCI.3885-10.2011.

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.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–71.

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422.

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.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44,* 2301–2311.

Rieger, J. W., Braun, C., Bülthoff, H. H., & Gegenfurtner, K. R. (2005). The dynamics of visual pattern masking in natural scene processing: A magnetoencephalography study. *Journal of Vision, 5*(3):10, 275–286, http://www.journalofvision.org/content/5/3/10, doi:10.1167/5.3.10. [PubMed] [Article]

Rieger, J. W., Köchy, N., Schalk, F., Grüschow, M., & Heinze, H.-J. (2008). Speed limits: Orientation and semantic context interactions constrain natural scene discrimination dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 34*(1), 56–76, doi:10.1037/0096-1523.34.1.56.

Rolls, E., Tovée, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: Information analysis. *Journal of Cognitive Neuroscience, 11*(3), 300–311.

Rossion, B., Gauthier, I., Goffaux, V., Tarr, M. J., & Crommelinck, M. (2002). Expertise training with novel objects leads to left-lateralized face-like electrophysiological responses. *Psychological Science, 13*(3), 250–257.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition, 12*(6), 852–877.

Sanocki, T. (2003). Representation and perception of spatial layout. *Cognitive Psychology, 47,* 43–86.

Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition, 67*(1–2), 147–179.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5,* 195–200.

Shapley, R., Hawken, M., & Ringach, D. L. (2003). Dynamics of orientation selectivity in the primary visual cortex and the importance of cortical inhibition. *Neuron, 38*(5), 689–699.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*(2), 127–190.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition, 67*(1–2), 1–20.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*(2), 233–282, doi:10.1016/0010-0285(89)90009-1.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience, 1*(4), 275–277.

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. (2002). Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(9), 1226–1238.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14*(3), 391–412.

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.

Ullman, S. (1984). Visual routines. *Cognition, 18*(1–3), 97–159, doi:10.1016/0010-0277(84)90023-4.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition, 32*(3), 193–254, doi:10.1016/0010-0277(89)90036-X.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology, 79*(4), 471–491, doi:10.1111/j.2044-8295.1988.tb02747.x.

VanRullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Frontiers in Psychology, 2,* 1–6.

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.

Vijayaraj, V., Cheriyadat, A. M., Sallee, P., Colder, B., Vatsavai, R. R., Bright, E. A., & Bhaduri, B. L. (2008, October). *Overhead image statistics*. Paper presented at the Applied Image Pattern Recognition Workshop, Washington, D. C., USA.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of Neuroscience, 29*(34), 10573–10581, doi:10.1523/jneurosci.0559-09.2009.

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences, USA, 108*(23), 9661–9666, doi:10.1073/pnas.1015666108.

Walther, D. B., & Shen, D. (2014). Nonaccidental properties underlie human categorization of complex natural scenes. *Psychological Science, 25,* 851–860, doi:10.1177/0956797613512662.

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.

Xiao, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2012, June). *Recognizing scene viewpoint using panoramic place representation*. Paper presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI.

Zelinsky, G. J., & Schmidt, J. (2009). An effect of referential scene constraint on search implies scene segmentation. *Visual Cognition, 17*(6–7), 1004–1028, doi:10.1080/13506280902764315.