Recognizing Scenes Containing Consistent or Inconsistent Objects

Michael L. Mack (michael.mack@vanderbilt.edu)
Department of Psychology, 111 21st Avenue South
Nashville, TN 37240 USA

Thomas J. Palmeri (thomas.j.palmeri@vanderbilt.edu)
Department of Psychology, 111 21st Avenue South
Nashville, TN 37240 USA

Abstract

How does object perception influence scene perception? A recent study of ultrarapid scene categorization (Joubert et al., 2007) reported facilitated scene categorization for scenes with consistent objects compared to scenes with inconsistent objects. One proposal for this consistent-object advantage is that ultrarapid scene categorization is influenced directly by explicit recognition of particular objects in the scene. We instead asked whether a simpler mechanism that relied only on scene categorization without any explicit object recognition could explain the consistent-object advantage. We combined a computational model of scene recognition based on global scene statistics (Oliva & Torralba, 2001) with a diffusion model (Ratcliff, 1978) of perceptual decision making. Simulations show that this model is sufficient to account for the consistent-object advantage. Importantly, this effect need not arise from explicit object recognition, but from the inherent influence certain objects have on the global scene statistics diagnostic for scene categorization.

Keywords: scene categorization; object recognition

Introduction

What is the relationship between scene perception and object perception? Past research has examined how objects are recognized in consistent or inconsistent scenes (e.g., Biederman, Mezzanotte, & Rabinowitz, 1982; Davenport & Potter, 2004; Palmer, 1975). The general finding is that it is easier to recognize objects in semantically consistent scenes, such as recognizing a toaster in a kitchen compared to recognizing a toaster in a bedroom (Davenport & Potter, 2004; Henderson & Hollingworth, 1999; Palmer, 1975).

One proposed mechanism for facilitated recognition of objects contained in consistent scenes is an interacting, dual system account (Davenport, 2007; Davenport & Potter, 2004). At the same time that the object recognition system is extracting information for an object categorization, the scene perception system is extracting evidence for a scene categorization. Object and scene perception systems operate in parallel, sharing information and converging on a full description of the environment, facilitating categorizations that are consistent with one another.

The interacting, dual-system account is supported by evidence for scene perception facilitating object recognition. Of course, the converse should be the case as well. Scene recognition can also be influenced by object perception. Indeed, Davenport and Potter (2004) found that scene categorization was facilitated when the scene contained a consistent object (e.g., a football field with a football player) compared to an inconsistent object (e.g., a football field with a priest).

A recent study (Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007) reported a similar advantage for scenes containing consistent objects versus inconsistent objects in ultrarapid scene categorization. Participants were presented with scenes for only 26ms and performed a go/no-go decision about the scene's superordinate category (natural or man-made). As illustrated in Figure 1, scene images either contained objects consistent with the scenes' category (e.g., an urban street scene with a parked car) or contained objects inconsistent with the scenes' category (e.g., an urban street scene with a large tree). A post-hoc analysis comparing these two types of scenes showed a consistent-object advantage such that participants made fewer errors and were faster to respond when categorizing a scene containing a consistent object. Joubert et al. explained this consistent-object advantage with the interacting, dual-system account: Information extracted by the object recognition system influences the rapid processing and categorization decision by the scene perception system. For a scene containing an inconsistent object, the object information conflicts with the evidence for the scene's category, leading to more errors and slower reaction times.

Previous work has shown that ultrarapid scene categorization is largely determined by coarse, global scene properties (Oliva & Schyns, 1997; Schyns & Oliva, 1994). Furthermore, computational models that represent scenes based on their global spatial structure are sufficient for ultrarapid scene categorization (Oliva & Torralba, 2001). Importantly, such models capture only the diagnostic global features of scenes without explicitly representing any local content of the scene, such as the location, presence, or identity of particular objects (Greene & Oliva, 2009). The feature set used by these models is based on global image statistics calculated across the entire scene, such as the scene's spatial frequency content.

A possibility suggested by the Joubert et al. (2007) results is that ultrarapid scene categorization based on global image properties is influenced in some way by ultrarapid categorization of particular objects in the scene that are either consistent or inconsistent with the scene's category.

We instead asked whether the consistent-object advantage could be explained by a simpler mechanism that relied only

on scene categorization without any explicit object recognition whatsoever.

Consider a forest scene. A small shed in that scene would be considered an inconsistent object. We could replace that shed with a consistent object, say a large bush. The global image statistics of a forest scene with a small shed will only be slightly different from those of a forest scene with a large bush. But they will not be identical. And that's the key. While perhaps quite small, is the difference in image statistics between scenes containing consistent objects versus scenes containing inconsistent objects sufficient to account for the consistent-object advantage? If so, then the consistent-object advantage in ultrarapid scene categorization can be explained by scene categorization alone, without any explicit object recognition.

To explore this possibility, we combined a computational model of scene recognition based solely on global scene statistics (Oliva & Torralba, 2001) with a diffusion model (Ratcliff, 1978) of perceptual decision making. Interpretation of global scene statistics provides evidence that drives a stochastic diffusion of perceptual evidence to a decision threshold. The model aims to explain both response probabilities and reaction time distributions for categorizing scenes containing consistent or inconsistent object. The model includes no explicit object recognition.

This paper is organized as follows: We first attempt a replication of the consistent-object advantage in scene categorization. We then analyze the behavioral data using the pure diffusion model, for reasons that will be made apparent. Finally, we present fits to observed data of our computational model combining a scene categorization front-end with the diffusion model of decision making.

Behavioral Experiment

This experiment attempted to replicate Joubert et al. (2007).

Methods

Participants Fifty Vanderbilt University undergraduate students (twenty-four male; age 18-23 years) participated in the experiment for course credit.

Stimuli The stimuli consisted of color images of naturalistic scenes from an online image database (Oliva & Torralba, 2001). Scene images were divided into categories of natural and man-made environments. The natural scene category included images of beaches, fields, mountains, and forests and the man-made scene category included images of skyscrapers, urban cities, and streets. Two independent observers tagged scenes that contained a salient object that was consistent or inconsistent with the scenes' natural or man-made category (reliability = 0.93). 192 natural scenes (64 with inconsistent objects) and 192 man-made scenes (64 with inconsistent objects) were randomly selected from the database for the experiment. Scene images were presented in color and subtended 10.2°×10.2° of visual angle. Example stimuli are shown in Figure 1.

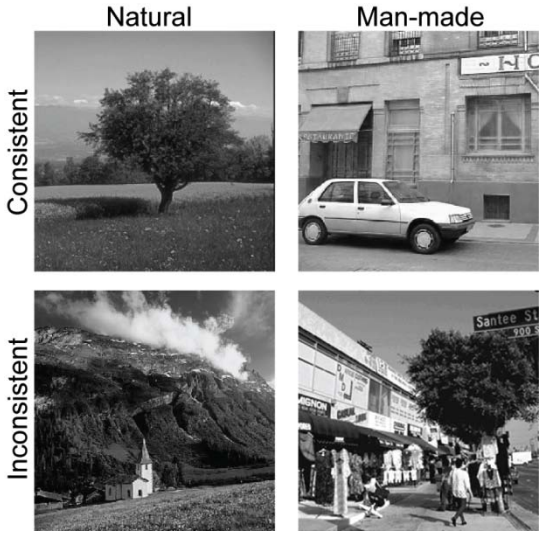

Figure 1: Examples of scene stimuli. Natural scenes (left) and man-made scenes (right) are shown with consistent objects (top) and inconsistent objects (bottom). Color images were used in Experiment 1.

Procedure Participants performed a go/no-go categorization task with target "go" category (natural or man-made scene) randomized for each participant. On each trial, a fixation cross was presented for 500-800ms followed by a brief presentation of the scene image for 26ms. Participants were instructed to press the response key if the scene belonged to the target category and withhold any response otherwise. Responses could be made for 1000ms after onset of the scene image and any responses made after this time window were considered no-go responses. The trial concluded with a 500ms blank period before the next trial began.

The experiment consisted of two blocks of 192 trials with an even split of target and distractor trials. Scene images used as target trials for half of the participants served as distractors for the other half of participants. The entire experiment lasted approximately 25 minutes.

Results

Performance was analyzed separately by target category (natural and man-made) according to accuracy and reaction times for correct responses (see Figure 2). Both target category groups showed a consistent-object effect, with higher accuracy for scenes containing consistent objects compared to inconsistent objects; this effect was larger for the natural scene group (11.6% difference; paired Wilcoxon test: $Z=4.17$, $p<0.001$) than the man-made scene target group (1.4% difference; $Z=2.648$, $p=0.008$). Both groups also showed a consistent-object effect in mean reaction times, with faster responses to scenes containing consistent objects; the effect was larger for the natural scene group (28ms difference; $Z=4.167$, $p<0.001$) than the man-made scene group (10ms difference; $Z=2.435$, $p=0.015$).

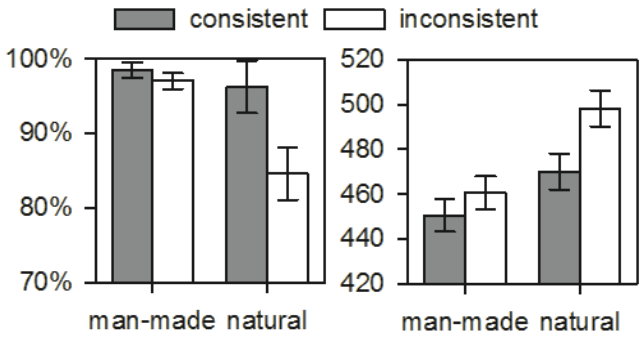

Figure 2: Average accuracy (left) and RT for correct responses (right) for consistent-object scenes (dark columns) and inconsistent-object scenes (light columns). Error bars represent 95% confidence intervals.

Discussion

We replicated the consistent-object advantage found by Joubert et al. (2007). For both man-made or natural scene targets, scenes with consistent objects were categorized faster and with fewer errors than scenes with inconsistent objects. The consistent-object advantage was larger for natural scenes, but this may be explained by stimulus factors as we did not attempt to equate the natural and man-made scene images in terms of visual properties or similarity.

Diffusion Model Analysis

The diffusion model is a well-known model of perceptual decision making (Ratcliff, 1978). Decisions are made through a stochastic accumulation of noisy evidence over time toward a decision threshold (see Figure 3). The rate of accumulation (called the drift rate, *v*) is determined by the quality of the perceptual evidence. Higher quality evidence leads to faster accumulation and faster reaction times. Changing the decision threshold (*a*) affects the tradeoff between speed and accuracy. Overall reaction time is given by the time for the perceptual decision made by the diffusion plus time for non-decision factors (T_{er}), such as stimulus encoding and motor response. Furthermore, in the full diffusion model, variability in drift rate, starting point, and nondecision time can be present and allow for the diffusion model to account for more detailed patterns of reaction time distributions.

The diffusion model is typically applied to two-alternative forced-choice categorization. A recent study compared different versions of the diffusion model to account for go/no-go categorization (Gomez, Ratcliff & Perea, 2007). They tested two versions of the diffusion model, one where evidence accumulates towards a single decision boundary for the "go" response with the other boundary at negative infinity, and another where evidence accumulates to both "go" (explicit response) and "no-go" (no response) boundaries. The two-boundary model was found to provide the best account of behavior associated with several go/no-go categorization tasks (Gomez et al., 2007). Therefore, we modeled the go/no-go scene categorization using a two-boundary diffusion model, with one boundary for a go response and the other boundary for a no-go nonresponse.

Before combining the diffusion model with a scene-recognition front end, we wanted to use the pure diffusion model as a data analysis device in order to pinpoint the source of the consistent-object advantage in accuracy and reaction time. First, the consistent-object advantage could arise from a differences in the time to perceptually process and encode scenes containing consistent versus inconsistent objects, which could be reflected by a difference in the T_{er} parameter. Second, recognizing consistent versus inconsistent objects might bias the decision process, leading to a potential difference in the decision threshold of the accumulation process (the *a* parameter). Third, our hypothesized simple single process account might suggest that the consistent-object advantage will arise from a difference in the quality of the perceptual evidence (the drift rate, *v*) driving the accumulation process.

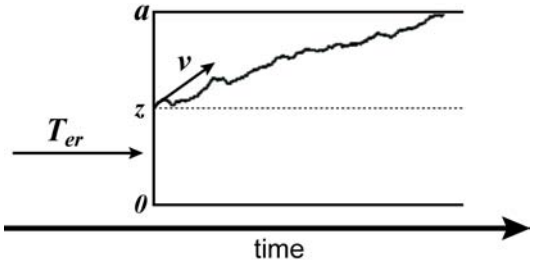

Figure 3: The diffusion model. At starting point *z*, evidence accumulates at drift rate, *v*, towards decision bounds defined by *a* and *0*. Overall reaction time is given by the time of accumulation plus time for non-decision factors (T_{er}).

Model Fitting

The diffusion model was fitted to reaction time distributions using standard techniques (see, Ratcliff & Tuerlinckx, 2002). For each individual participant, RT data for scenes containing consistent versus inconsistent objects were grouped into 6 RT bins defined by the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles. Quantile RTs averaged across participants were then used to generate predicted cumulative distributions of response probabilities (Vandekerckhove & Tuerlinckx, 2007, 2008). Best-fitting model parameters were found using the SIMPLEX method that minimized the Pearson chi-square for the observed versus predicted number of RTs within each RT bin (an additional bin was included in the fitting to count the number of no-go responses). The full diffusion model is defined by seven parameters: starting point of the accumulation process and its variability (z, s_z), decision threshold (a), drift rate and its variability (v, nu), and the nondecision time and its variability (T_{er}, s_t). For our model fits, starting point ($z=a/2$) and its variability, variability of drift rate (nu), and variability of nondecision time (s_t) were held constant across the consistent and inconsistent conditions. We fitted

versions of the diffusion model where the three key parameters, decision threshold (a), nondecision time (T_{er}), and drift rate (v), were either free to vary or were held constant across the consistent and inconsistent conditions.

Results

The variant of the diffusion model with only drift rate as a free parameter provided a significantly better fit to the behavioral data than variants with only nondecision time or decision threshold as a free parameter. Table 1 shows values for the chi-square statistic and the appropriate significance tests for each version of the diffusion model.

Table 1: Diffusion model fits

Free parameters	Chi-square	p
All fixed	5.318	--
a	2.746	0.109 (vs. fixed)
T_{er}	3.645	0.196 (vs. fixed)
$\boldsymbol{v}$	**1.346**	**0.046** (vs. fixed)

Discussion

Diffusion model analyses of the data from Experiment 1 revealed that a model with a separate drift rates for the consistent and inconsistent object condition provided the best account of the behavioral data. Allowing a freely varying nondecision time did not provide a good fit, suggesting that the time necessary for scene encoding was not affected by the consistency of the embedded object. The consistent-object advantage is best accounted for by assuming that the quality of the perceptual evidence is affected by the presence of an inconsistent object.

Scene Categorization Model

To test this further, we extended a successful model of scene categorization (Oliva & Torralba, 2001). Their model is the perceptual front-end that extracts evidence for a scene's category that then drives the diffusion model of decision making. Specifically, the scene categorization model establishes the drift rate of the diffusion process, rather than allowing the drift rate to be a free parameter.

Model Description

We started with a scene categorization model developed by Oliva and Torralba (2001). In this model, scenes are represented by a set of features that describe the global spatial structure of the scene (Greene & Oliva, 2009; Oliva & Torralba, 2001). The feature space, known as the spatial envelope, is defined by measures of global shape properties that are extracted using a bank of Gabor filters of varying spatial scale and orientation.

We followed the procedure outlined by Oliva and Torralba (2001). A bank of Gabor filters spanning four spatial scales and eight orientations were used to extract the scenes' global features. To reduce the dimensionality of the filter responses, each filter output was down-sampled to a lower-resolution (4x4) summary. PCA was then used to further reduce the dimensionality creating a final scene representation consisting of a 50-element vector. Natural versus man-made scene categories were defined by a hyperplane boundary extracted using linear discriminant analysis (see Figure 4).

We used the results of the linear discriminant function to establish the drift rate of the diffusion model for each scene image to be categorized. Specifically, for a given scene, the output of the linear classifier corresponds to the distance of that scene from the boundary separating natural versus man-made scenes. The sign of the distance signifies which category the scene is classified in and the magnitude of the

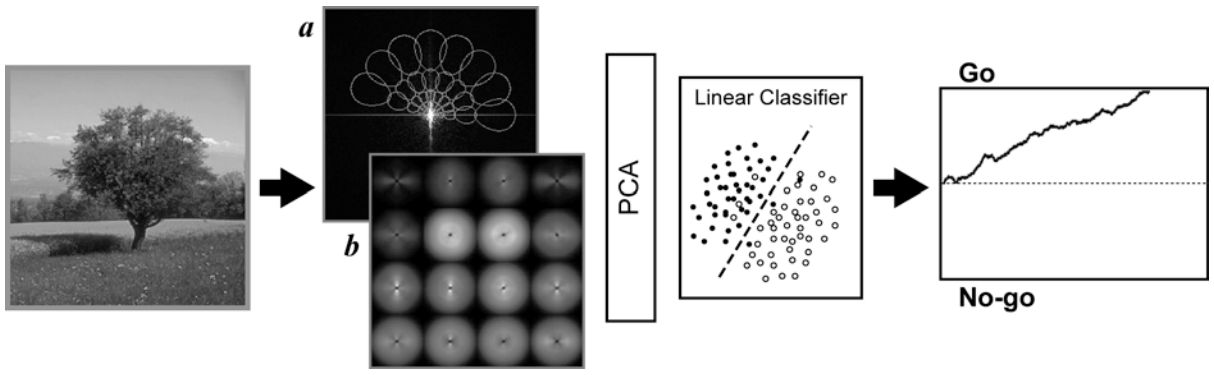

Figure 4: The extended scene categorization model. Scenes are first classified by the scene categorization front-end. The scene's global spatial frequency is extracted with a bank of Gabor filters (*a* - polar plot of global spatial energy, spatial scale and orientation of filters shown by ellipses) and summarized into a low-resolution representation (*b* - subimages in the 4x4 grid show the global energy at that spatial location). Scene representations are projected onto a 50-dimensional principal component space and classified by linear discriminant analysis. The resulting classification value drives a stochastic accumulation of evidence towards go or no-go response boundaries.

distance represents the quality of that classification. Distance is transformed into drift rate with a sigmoid function that includes a scaling parameter. Using that drift rate, the decision process is carried out by the diffusion as a stochastic accumulation of evidence to a threshold.

We want to emphasize that this model assumes no parameters that vary across scenes containing consistent versus inconsistent objects. The scene categorization front-end uses the same discriminant function for scenes with consistent and inconsistent objects. Distance from the discriminant function is transformed into drift rate using the same function for all scenes. The diffusion process determining the time-course of the decision is the same for all scenes. It should also be clear that the model contains no explicit object recognition process. Scenes are represented by global features that capture the scene's spatial frequency structure. The only difference between scenes containing consistent versus inconsistent object is in the global content, not recognition of any individual objects in the scenes.

Simulation Method

First, a set of 200 natural and 200 man-made scene images were randomly selected from the scene database (same as used in Experiment 1) for creating the PCA. A fifty-dimensional principal component space was extracted from these scenes' Gabor-filtered representations and saved for the simulations. Next, a training set consisting of another 100 natural and 100 man-made scenes was randomly selected from the scene database. These scenes were passed through the Gabor filters, projected into the principal component space, and used to train the linear discriminant classifier.

The scene database we used had fewer inconsistent-object scenes compared to consistent-object scenes, since by definition, inconsistent objects are not typically found in those scenes. In order to test an equivalent number of scenes with consistent and inconsistent objects, we randomly selected 500 consistent object scenes and inconsistent object scenes with replacement from the scene database. Scenes used for training were never included in the testing sets. Test trials consisted of first passing a scene through the scene categorization front-end. This stage generated a classification value from the discriminant function that was transformed into a drift rate for the diffusion. The drift rate drove the stochastic accumulation of evidence until a decision threshold (go or no-go) was reached or 1000ms had elapsed (tallied as a no-go response).

The three parameters of the model (drift rate scaling factor, decision threshold, nondecision processing time) were optimized by fitting the predicted reaction time distributions to the observed data using the same procedure used in the earlier diffusion model analysis. We tested the model's performance with both natural and man-made scenes as targets. The entire simulation procedure was repeated with twenty-five separate training and testing sets.

Results

Performance was analyzed separately by target category (natural and man-made) according to accuracy and reaction times for correct responses across simulation repetitions (see Figure 5). The model showed a consistent-object effect only when the target was a natural scene. Accuracy was higher ($Z=4.24$, $p<0.001$) and reaction times were faster ($Z=4.37$, $p<0.001$) for consistent-object scenes compared to inconsistent-object scenes. With man-made scenes as the target, mean differences in both accuracy and reaction time trended in the manner of a consistent-object advantage, but did not reach significance ($Z=0.977$, $p=0.328$; $Z=1.44$, $p=0.15$); recall that the difference observed for human subjects was also quite small.

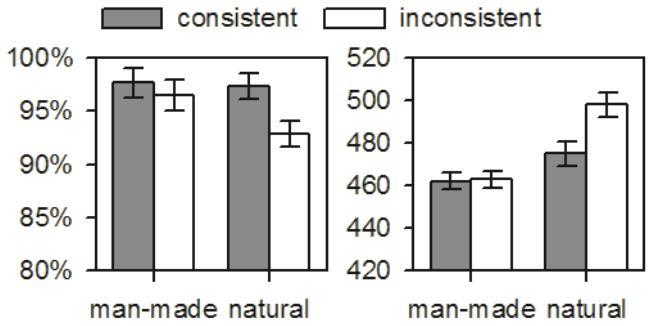

Figure 5: Simulation results. Average accuracy (left) and correct response RTs (right) for consistent (dark columns) and inconsistent (light columns) object scenes.

Discussion

Simulations of ultrarapid natural scene categorization with the extended scene categorization model showed a significant consistent-object advantage for categorizing natural scenes as targets and a small (but not significant) advantage for man-made scenes; this difference across target scene category is qualitatively comparable to what was observed in Experiment 1. These initial simulations suggest that the global features extracted by the perceptual front-end of our model were influenced by the presence of an inconsistent object. This subtle influence may be sufficient to explain the lower accuracy and slower reactions times associated with scenes containing inconsistent objects.

Conclusions

The aim of our work was to test whether the consistent-object advantage observed by Joubert et al. (2007) could be explained using global scene categorization mechanisms without object recognition. By this account, semantically inconsistent objects in scenes can influence the global perceptual evidence diagnostic for scene categorization without any explicit recognition of consistent versus inconsistent objects in the scene.

Consistent with this simple scene categorization account, we presented evidence from a diffusion model analysis that

suggests a difference in the quality of the perceptual evidence available from scenes containing consistent versus inconsistent objects. Furthermore, we showed that a scene categorization model coupled with the diffusion model accounts well for the consistent-object advantage. Instead of distinct scene and object perception systems operating in parallel and competing or cooperating for categorization, the consistent-object advantage can be explained by a single scene perception system that interprets the global statistics found in natural scenes.

It is important to place our findings in their appropriate context. First, we are not arguing that explicit recognition of objects never matters for scene categorization. It goes without saying that fully understanding the environments we encounter during our everyday visual experience requires successful object recognition. However, in the case of ultrarapid ultrasuperordinate scene categorization, we have shown that explicit representation and recognition of objects in those scenes is not necessary to account for the influence of consistent or inconsistent objects. Second, it goes without saying that this demonstration is evidence of sufficiency and not necessity. Further converging evidence is needed to know whether mechanisms described in our model underlie ultrarapid scene categorization in humans.

The computational model we proposed extends a current class of successful scene categorization models to predict both response probabilities and reaction times. This model offers a richer description of scene categorization by accounting for the time course of the perceptual decision. Further behavioral research and application of this model is necessary to better understand the underlying mechanisms of scene categorization and to characterize the relationship between scene and object perception.

Acknowledgments

This work was supported by a grant from the James S. McDonnell Foundation, NSF grant HSD-DHBS05, and by the Temporal Dynamics of Learning Center (NSF Science of Learning Centers grant SBE-0542013).

References

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143-177.

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition, 35*(3), 393-401.

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

Greene, M., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58*(2), 137-176.

Gomez, P., Ratcliff, R., Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General, 136*(3), 389-413.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*(1), 243-271.

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.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72-107.

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.

Palmer, S. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*(5), 519-526.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195-200.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*(6), 1011-1026.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavioral Research Methods, 40*(1), 61-72.