Leveraging rapid scene perception in attentional learning

Juliana D. Adema (juliana.adema@mail.utoronto.ca)
Shuran Tang (shuran.tang@mail.utoronto.ca)
Nahal Alizadeh Saghati (nahal.alizadehsaghati@mail.utoronto.ca)
Michael L. Mack (michael.mack@utoronto.ca)
Department of Psychology, University of Toronto,
100 St George St, Toronto, ON M5S 3G3 Canada

Abstract

In addition to saliency and goal-based factors, a scene's semantic content has been shown to guide attention in visual search tasks. Here, we ask if this rapidly available guidance signal can be leveraged to learn new attentional strategies. In a variant of the scene preview paradigm (Castelhano & Heaven, 2010), participants searched for targets embedded in real-world scenes with target locations linked to scene gist. We found that activating gist with scene previews significantly increased search efficiency over time in a manner consistent with formal theories of skill acquisition. We combine VGG16 and EBRW to provide a biologically inspired account of the gist preview advantage and its effects on learning in gist-guided attention. Preliminary model results suggest that, when preview information is useful, stimulus features may amplify the similarities and differences between exemplars.

Keywords: visual search; scene gist; guided attention; VGG16; EBRW

Introduction

Humans have a remarkable ability to rapidly categorize a visual scene. With no more than 25 milliseconds of exposure, a scene's conceptual category or gist (e.g., kitchen, forest) is readily perceived by an observer. Why do we have this ability to quickly map visual features to rich conceptual knowledge? One prominent theory (Torralba, Oliva, Castelhano, et al., 2006) suggests scene gist is rapidly perceived in order to guide exploration of our visual environment towards information-rich regions (e.g., countertops in a kitchen). Such gist-based guidance allows for efficient sampling of behaviourally-relevant information contained within the visual scene. Can gist and its associated vast semantic knowledge be flexibly, and usefully, used in new learning? Here, we empirically test the hypothesis that rapidly available guidance signals from scene gist can be leveraged to learn new attentional strategies. Moreover, we combine a formal process model of category learning and skill acquisition, exemplar-based random walk (EBRW), with a well-established deep-learning vision model, VGG16, to provide a comprehensive account of the experimental effect.

Categorization is the abstraction of stimulus features to group similar objects (or scenes) together (Greene & Fei-Fei, 2014). Scene categorization is a rapid (Evans, Horowitz, & Wolfe, 2011; Harel, Groen, Kravitz, et al., 2016; Lowe, Rajsic, Ferber, et al., 2018; Mack, Gauthier, Sadr, et al., 2008; Mack & Palmeri, 2010), and seemingly automatic (Greene & Fei-Fei, 2014; Mack & Palmeri, 2015) visual perceptual process, occurring with less than 100 milliseconds of exposure to a natural scene image. In addition to visual features of a scene, scene gist is an important part of a scene's overall context and has been found to guide attention in visual search (Castelhano & Heaven, 2010; Robbins & Hout, 2020).

A scene's properties, and the higher-order mental states of the observer (i.e., expectations, goals), both influence the allocation of visual attention (Võ & Wolfe, 2013; Yantis, 1996). The understanding of these features is akin to "involuntary" versus "voluntary" direction of attention (Yantis, 1996). Although saliency and top-down factors undoubtably affect spatial attention and search behaviours, other important details that can be extracted from a scene, including its semantic contents and other global properties (Chun & Jiang, 1998) must be accounted for. Work has already been done to advance the notion of scene category as capable of learning-induced transformations. Brockmole and Henderson (2006) showed participants natural scene images with target letter embedded within. Critically, they used novel images (presented only once throughout the search task) as a baseline measurement of reaction time (RT), and repeated images to observe effects of contextual cueing. Their results reflect that target information associated with natural scenes was learned more quickly than targets linked to arbitrary configurations of letters and numbers; the semantic information contained in a scene is supposed to facilitate this process (Brockmole & Henderson, 2006).

Whether scene gist can be the driver of attentional guidance still requires investigation, though recent studies have provided support for this notion (Robbins & Hout, 2020). We hypothesize that initial search strategies will be largely informed by prior knowledge of information-rich areas of scene categories. As trials progress, however, participants will learn that scene gist is linked to specific target locations. That is, the consistent presence of a target in one location is expected to disrupt the typical category-guided search patterns. Fully understanding this process requires a formal model that can account for how perceptual information is used to build new knowledge. The rich literature of cognitive learning models (e.g., Nosofsky & Palmeri, 1997) provides formal accounts of how perceptual information is leveraged in new learning, yet is missing the key perceptual mechanisms that translate visual stimuli into meaningful perceptual components. We explore whether the integration of a vision model and a cognitive model can provide a formal means for understanding attention learning in scene perception.

We use VGG16, a deep convolutional neural network for image recognition (Simonyan & Zisserman, 2014), to obtain stimulus features as input to EBRW. This perceptual frontend allows for the investigation of scene perception together with learning in visual search. VGG16 performance in specific layers has been highly correlated with processing stages in human primary visual cortex (Eberhardt, Cader, & Serre, 2016) (but see Palmerston, Zhou, and Chan (2020) for comments on functional connectivity beyond these areas). Essentially, VGG16 provides a biologically inspired characterization of the perceptual features underlying rapid scene perception.

We combine VGG16 perceptual representations with EBRW (Nosofsky & Palmeri, 1997) to characterize the nature of learning in scene-gist guided attention. EBRW formally defines category learning as a mutual interaction between instance-based memory, selective attention, and random walk evidence accumulation (Nosofsky & Palmeri, 1997). Category decisions are based on comparisons between the current stimulus and stored representations, or exemplars, of the relevant categories (Nosofsky & Palmeri, 1997). If the stimulus is similar to the exemplars of one category, the accumulation of evidence towards the category and subsequent decision will be made relatively quickly, whereas categorical ambiguity would slow the random walk process. Importantly, increased experience should lead to faster performance. Within learning experiments, each trial is encoded into EBRW as an additional exemplar of a category. Here, we conceptualize EBRW such that perceptual representations of scenes, which potentially hold key scene gist information, are encoded as exemplars of target location. Over multiple trials, these category representations associated with target location accumulate to drive faster responses in a manner consistent with the classic notion of a power law of practice.

We had participants perform a novel attentional learning task, finding that activating gist with scene previews significantly increases search efficiency in a manner consistent with formal theories of skill acquisition. As a proof of concept, we provide preliminary analysis of a model combining VGG16 and EBRW to account for scene gist-based learning effects on attention.

Methods

In a variant of the scene preview paradigm (Castelhano & Heaven, 2010), participants searched for targets embedded in real-world scenes with target locations linked to scene gist.

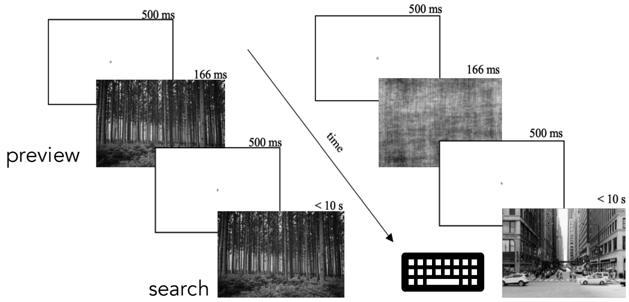

Figure 1. Trial schematic for each preview type condition in the behavioural experiment. After being shown an intact (left) or scrambled (right) preview of the search image, participants searched for a target for up to ten seconds, and identified it using a key press.

Participants

All participants were recruited from an undergraduate subject pool and given course credit as compensation for their participation. All participants had normal or corrected-to-normal vision. This experiment was approved by the research ethics board at the University of Toronto. Participants were briefed and gave their written consent before the experiment began. 43 participants (17-22 years) took part in the experiment. Eleven participants had to be excluded from the data analysis due to insufficient mean accuracy (<0.85) in identifying the target's identity across all trials of the experiment. The resulting sample size was therefore 32 (mean = 18.85 years; 24 female).

Materials

Search and preview images

The search images were 180 photographs of outdoor scenes, obtained from Google Images. 90 images were of forests, representing the 'natural' category. The other 90 images were of city streets, representing the 'manmade' category. All images were resized to 1600 x 1067 pixels and converted to grayscale.

Phase-scrambled images were created by deconstructing all 180 scenes images into constituent magnitude and phase spectra using fast Fourier transform (Sadr & Sinha, 2004), randomizing the phase component, and inverting the Fourier transform of the randomized phase and original magnitude components. Randomising the phase component of an image retains its low-level visual attributes but renders it semantically incoherent (Mack et al., 2008). By presenting these images as previews before the intact search image, participants are, in a sense, still being given "the same image". The difference is that they will not be able to discern any gist information from that preview. Phase scrambling was performed using functions found in the pylab Python library.

Targets

Two targets were used in these experiments: an illustration of a kangaroo, and an illustration of a lizard. Both targets were taken from Keynote for iOS's "shapes", a clipart library accessible in the application. Each image was white with a black outline and presented at 50% opacity. Targets were each presented at a unique location on the right or left side of the screen depending on scene context in each trial of the experiment, such that there are 180 possible target locations. These category-incongruent targets were chosen over category-congruent targets (e.g., a fire hydrant on a city street) to encourage participants to explore the whole scene without being influenced by additional contextual factors implied by a congruent object's identity.

Procedures

There were 180 trials in total, and participants completed every trial. All instructions were provided to the participant onscreen. They were advised to maintain fixation when beginning each trial, including when the preview was presented, and then to find the target in the image as quickly as possible. The first fixation was presented for 500 ms, followed by a preview of the search image for 166.7 ms, after which another fixation was presented for 500 ms. The search image was then shown, with the target, for up to ten seconds or until the participant had signalled that they have found the target with a key press. An intermediate screen was shown until the participant was ready to begin the next trial. Throughout the experiment, targets consistently appeared on either the left or right side of the screen, depending on the trial image's scene category. For example, a participant would always see a target on the right side of a city street search image, and on the left for a forest search image. Premade conditions were counterbalanced for target side and preview type. Participants were not informed of the target location patterns.

Participants were subjected to both types of conditions at once, such that trials of each type were randomly interleaved throughout the entire experiment. The conditions were distinguished by the preview that was shown on each trial: a preview of the search image, or a phase-scrambled version of the search image. The latter preview type eliminated any distinguishable category information from the preview, such that there was no helpful gist information until the search task had started.

Computational model

To investigate how attention guidance is potentially impacted by associative learning between visual scene representations and target location, we conducted simulations with an integrative computational framework that combined a deep-learning vision model (VGG16) with a category learning model (EBRW).

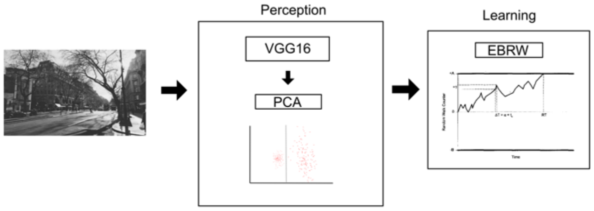

Figure 2. VGG16 and EBRW model. Perceptual representations of scenes were first generated by a Places365-trained VGG16 model and feature vectors from the first fully connected layer (FC1) were extracted. The dimensionality of these feature vectors was reduced with PCA to define a two dimensional perceptual space. Each scene's position along the two dimensions was then inputted to EBRW. Within EBRW, each scene was compared to previously encountered scenes associated with the two target locations. Similarity to these stored scene exemplars drove a drift diffusion evidence accumulation process to predict visual search response times.

Perceptual frontend (VGG16)

To obtain a measure of visual similarity between exemplars, a necessary component of EBRW model, we turned to a well-established deep-learning vision model. Briefly, the image feature extraction was done using VGG16, a deep convolutional neural network consisting of 16 layers (Simonyan & Zisserman, 2014), where the 'learning' layers consist of five convolutional layers and three fully connected layers (Krizhevsky, Sutskever, & Hinton, 2017); this was implemented using the Keras application programming interface (API) and TensorFlow machine learning platform in Python (Zhou, Lapedriza, Khosla, et al., 2018). To summarize, the model works as follows: the images are resized to 224 x 224 pixels, each convolutional layer applies 'filters' to extract image features, and the model predicts the probability of the image belonging to one of the two stimulus categories (i.e., forest or city street) (Krizhevsky et al., 2017).

The perceptual representations across the layers of VGG16 include varying degrees of scene-specific and scene-gist information. With the goal of primarily representing scenes according to scene gist while also maintaining important variability across the scenes, we focused on representations from the fully connected layer FC1. To best link these high-dimensional visual representations onto the cognitive mechanisms of EBRW, we performed dimensionality reduction with PCA (*scikit-learn*) following the approach of similar prior work (Mack & Palmeri, 2010). Specifically, we retained the first two principal components. Although this resulted in a significant reduction of information from the VGG16 representations, the resulting two-dimensional representations retained both scene-gist and scene-specific information and were appropriate for the computational constraints of EBRW simulations.

EBRW

EBRW simulations were based on RTs from all participants for each repetition of each trial type. Image features generated by the VGG16 perceptual frontend served as perceptual representations of each scene image for EBRW. EBRW is defined by five parameters: α, a constant accounting for the time needed to extract a category label for a retrieved instance; *c,* a constant used for transforming distance measures into similarity measures; *w*, a vector of attention weights for the image feature dimensions; and two regression parameters that map the arbitrary units of EBRW accumulation evidence to RT*: k*, a scaling factor for the slope; and μ_r, the y-intercept, or the mean of the RTs.

To simulate the current task, PCA representations of the VGG16 FC1 layer for scenes were stored in category knowledge along with category labels corresponding to target location (right vs. left of screen). On each trial, the current scene representation was compared to previously encountered scene representations associated with the two target category locations to drive a drift diffusion process. The higher a match between the visual features of a scene and previously encountered scenes with the same target location, the faster the predicted visual search. At the end of a trial, the current scene's representation was encoded into the model's category knowledge as a new exemplar for the correct target location, making it available to impact the similarity calculations in subsequent search trials. Across trials, the number of stored exemplars for each target location grew to reflect prior experiences in linking scene category with target locations.

To isolate the potential mechanistic effect of preview type, three separate models with one parameter allowed to vary between preview type (α, *c,* and *w*) were fit to participants' RTs. Parameter optimization was conducted by minimizing the root mean-squared error (RMSE) between observed and model-predicted RTs with differential evolution, a stochastic genetic algorithm-based optimization method (*scipy* v.1.5.2*, differential_evolution*). Since all three models shared the same degrees of freedom in accounting for the behavioural data, the final RMSE of the three models served as an index for comparison. Given the exploratory nature of the VGG+EBRW model for visual search and the relatively noisy nature of the behavioural data, we opted to optimize parameters based on fits to the entire behavioural dataset all at once rather than individual fits specific to each participant. Although this approach precludes characterizing model mechanisms at the level of individual participants, it does provide a well-powered method for identifying likely mechanisms that describe group-level differences between preview types and offers preliminary proof-of-concept evidence for the proposed computational framework.

Results

Experiment

Mean values of the median RTs, per participant, for trials with intact and scrambled previews are 3077 ms (SD = 633.3) and 3074 ms (SD = 651.2), respectively. The effect of learning repetition, preview type, and their interaction on log-transformed RT was estimated with a linear mixed effects regression model using the R 'lme4' package (Bates et al., 2020). Participants were modelled as a random effect.

Participants averaged 87.8% and 85.9% successful searches for intact and phase-scrambled trials respectively. Thus, the type of preview did not affect participants' ability to complete the search. However, search efficiency was impacted by preview type. Search RTs decreased over trials (β=-0.003135, 95% CI [-0.004484, -0.001788], t=-4.559, p=5e-6) but showed no main effect of preview type (β=-0.08955 (95% CI [-0.1931, 0.01381], t=-1.696, p=0.0899). Critically, preview type significantly interacted with trial number such that intact scene previews led to faster search RTs over the course of the experiment relative to phase-scrambled previews (β=0.002286, 95% CI [0.0003540, 0.004222], t=2.316, p=0.0206). This effect suggests that an intact scene preview speeds visual search, potentially through the activation of the scene's gist and its association with target location.

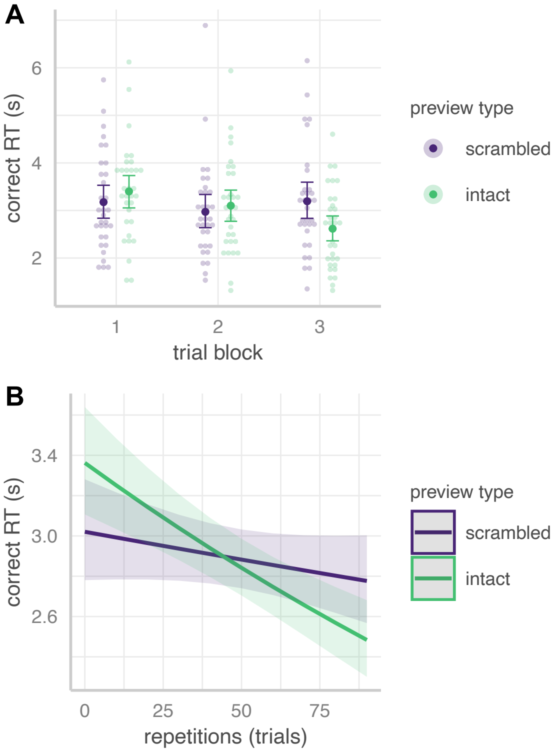

Figure 3. A: Average median RTs for trial blocks of each trial type. Darker dots are group-level means, while lighter dots reflect individual participants' binned median RTs. Error bars represent bootstrapped 95% confidence interval. B: Estimated marginal means from LMER model for RT as a function of trial repetition and preview type. Margins around each line represent 95% confidence bands.

Model simulations

As a preliminary proof of concept analysis, we conducted model simulations with the aim of accounting for group-level trends in the visual search behaviour. First, VGG16 trained on Places365 was used to generate perceptual representations of each scene image. Specifically, we extracted vector representations for the scenes from the first fully connected layer (FC1) and submitted these representations to principal component analysis (PCA) keeping only the first two principal components. Although these components explained only 17.4% of the variance across the layer representations, they captured category differences well. Both categories showed similar within-category cosine distances (forest: μ=0.953, σ=0.085, range=[0.506, 1.315]; street: μ=0.944, σ=0.056, range=[0.631, 1.066]) and larger between category distances (μ=1.065, σ=0.034 range=[0.856, 1.186]). These reduced perceptual representations served as input to EBRW.

We formalized increases in search efficiency over time as a category learning process, specifically EBRW, that linked VGG16 scene representations to target location (i.e., left vs. right side of the search scene). Our implementation followed the standard formulization of EBRW (Nosofsky & Palmeri, 1997); here, we focus on a conceputal overview rather than restating the model equations. On each trial, the summed simliarity between the current search scene's VGG16 representation and the previously stored scenes associated with the two target locations is calculated. The degree that the current scene matches one target location's stored scenes over the other is then used to estimate both the probability of selecting that target location and the response time of the visual search. With each correct trial, the current scene's VGG16 representation and its associated target location is stored in memory. Thus, over time, EBRW builds "category" representations that link perceptual features of scenes to target locations.

As a first step in evaluating this model, we focused on three parameters in EBRW: sensitivity, *c*, which acts as a scaling parameter on the similarity calculation between the current scene and stored exmplars; attention weight, *w*, which biases the contribution of the two perceptual feature dimensions in calculating similarity; and *alpha*, which determines the time needed to retrieve an exemplar from memory. These parameters are linked to distinct mechanisms that each could influence search efficiency over learning by way of an intact scene preview. Three separate models that each each allowed one of these parameters to vary across preview type was fit to all participants' RTs across trial repetitions (Figure 3A). To be clear, this combined VGG16+EBRW model provides a mechanistic account of how visual features are linked to scene gist and how, over the course of multiple experiences, scene gist can be associated with attentional strategies (i.e., attend to the left or right of the scene). This model does not formalize the deployment of overt or covert attention; rather focusing on how regularities in prior experience can lead to more guided attention.

Model analyses revealed that the sensitivity model provided the best fit to behaviour ($RMSE_c = 153.79$; $RMSE_\alpha = 153.91$; $RMSE_w$ = 153.97). Importantly, sensitivity was greater for the intact preview (0.128) than for the scrambled preview (0.089) condition. In EBRW, higher sensitivity acts to sharpen the similarity function. Thus, intact scene previews may serve to activate related stored scene exemplars associated with efficient search strategies. These simulations are preliminary and based on group averages of search behaviour. Model fits to individual participants' data will be a key future analysis to pinpoint the specific model mechanisms that give rise to the greater learning effect observed for intact previews.

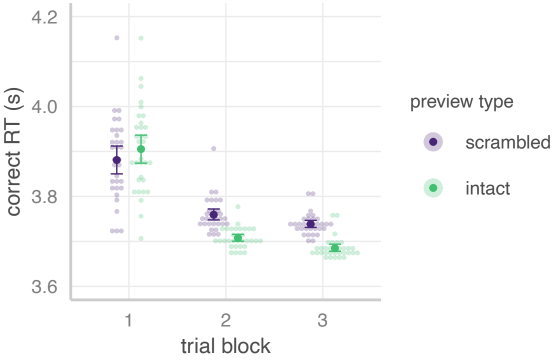

Figure 4. VGG16+EBRW predictions of RT over repetitions of preview types using image features determined by VGG16 with separate sensitivity parameters for preview type. Darker points represent average RT per trial block and lighter points depicted model predictions for each participant. Error bars represent bootstrapped 95% confidence intervals.

Discussion

We show that, while mean RTs were similar for intact and scrambled trials, the rate of learning over trials for intact previews was substantially steeper than the learning rate for scrambled previews (Figure 3). These results are consistent with the hypotheses, such that it seems that there is some component of the intact previews that is more informative than the scrambled previews for the visual search task. It should be noted that many studies have found that performance is best in conditions where scene primes are identical to the search images (Brockmole & Henderson, 2006; Castelhano & Heaven, 2010; Võ & Wolfe, 2013). Some have even argued that semantic gist is informative if and only if visual information is also present (Makovski, 2018). Therefore, it cannot yet be concluded with confidence that gist per se is what caused participants to find the targets faster when presented with the intact previews; it could very well be that the visual information also had a role in driving those observed differences.

Nonetheless, it is worthwhile to investigate the potential mechanisms underlying this intact preview advantage. Although cognitive models provide a rich landscape for characterizing learning, they lack perceptual mechanisms that translate sensory information into perceptual representations; this is especially true for complicated stimuli like real world scenes. Here we provided the first steps towards this challenge by integrating VGG16 and EBRW, finding initial evidence that this approach is viable. In particular, we found that differences in the sensitivity parameter (c) across preview type provided the best account of the behavioural data, and that that parameter had a larger value for the intact condition than the scrambled condition. Higher sensitivity values translate to a sharper activation function, where similar items may appear more similar to each other, and different items may appear more different from one another. Here, that means that intact previews may better activate relevant category knowledge, speeding the retrieval of the newly learned associations between target locations and scene gist (Annis & Palmeri, 2019). These findings are similar to those reported by Annis and Palmeri (2019), such that exemplar distinctiveness, as defined by changes in sensitivity, was correlated with visual expertise. Again, this interpretation is based on preliminary model simulations of group-level behaviour, but it is consistent with both behavioural and model predictions. Future model analyses conducted at the level of individual participants will provide a formal evaluation of the sensitivity hypothesis suggested here.

Our approach offers the best of both words: sophisticated models of vision, and theory-driven mechanisms of human learning. Here, we apply it to questions of novel scene learning to gain perspective on the time course of learning beyond what is observable from behavioural RT data, and, more specifically, to gather information about the potential mechanisms behind the intact preview advantage. Our method also allows us to make mechanistic inferences based on RTs and accuracy alone, extending the usefulness of these measures in studies of learning and attentional guidance.

Given the exploratory nature of our approach, there are some limitations that must be addressed in order to further refine the model. Although visual search performance clearly demonstrates a speeding for intact previews, the trial-by-trial data is quite variable even when looking at group averages. Characterizing individual participant performance with tailored model simulations as opposed to the group average model fits in the current work may provide a better characterization of the underlying mechanisms. Also, VGG16 provides a wealth of perceptual representations across its many layers, only one of which we leverage in the current model implementation. A systematic evaluation of the gist-level and scene-specific information across the VGG16 layers as it relates to visual search performance will be key to understanding both trial-by-trial variability in search and the type of perceptional information that can drive new learning for attention guidance. Ultimately, the modeling framework we propose offers the potential to link the complex nature of naturalistic scene perception to novel learning situations.

In conclusion, it is well established that scene categorization is a rapid, behaviourally relevant process. However, we do not know the extent to which scene gist information is useful in novel learning contexts. The results of the current behavioural experiment together with the modelling framework described here suggest that rapid scene categorization can be associated with new attentional strategies and increase visual search efficiency. These preliminary findings motivate future work that will characterize such learning effects in individuals and determine the nature of rapidly encoded scene information that best drives novel learning.

Acknowledgements

Thanks to Belinda Hoang, Teresa Zhu, and Tanya Iyer for assistance with data collection. This research is supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grants (RGPIN-2017-06753 to MLM), NSERC Canada Graduate Scholarship-Master's (CGS M) to JDA, and Canada Foundation for Innovation and Ontario Research Fund (36601 to MLM).

References

Annis, J., & Palmeri, T. J. (2019). Modeling memory dynamics in visual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(9),* 1599–1618.

Bates, D., Maechler, M., Bolker B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., & Fox, J. (2020). lme4: Linear Mixed-Effects Models using "Eigen" and S4 (1.1-23) [Computer software]. https://CRAN.R-project.org/package=lme4

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13(1), 99–108.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention, Perception & Psychophysics*, *72(5),* 1283–1297.

Chun, M. M., & Jiang, Y. (1998). Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology, 36(1),* 28–71.

Eberhardt, S., Cader, J., & Serre, T. (2016). How Deep is the Feature Analysis underlying Rapid Visual Categorization? ArXiv:1606.01167 [Cs]. http://arxiv.org/abs/1606.01167

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When Categories Collide: Accumulation of Information About Multiple Categories in Rapid Scene Perception. *Psychological Science, 22(6),* 739–746.

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision, 14(1)*, 14–14.

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The Temporal Dynamics of Scene Processing: A Multifaceted EEG Investigation. *Eneuro, 3(5)*, ENEURO.0139-16.2016.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60(6),* 84–90.

Lowe, M. X., Rajsic, J., Ferber, S., & Walther, D. B. (2018). Discriminating scene categories from brain activity within 100 milliseconds. *Cortex, 106*, 275–287.

Mack, M. L., Gauthier, I., Sadr, J., & Palmeri, T. J. (2008). Object detection and basic-level categorization:

Sometimes you know it is there before you know what it is. *Psychonomic Bulletin & Review, 15(1)*, 28–35.

Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision, 10(3)*, 1–11.

Mack, M.L., & Palmeri, T. J. (2015). The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General, 144(3)*, 551–569.

Makovski, T. (2018). Meaning in learning: Contextual cueing relies on objects' visual features and not on objects' meaning. *Memory & Cognition, 46(1)*, 58–67.

Nosofsky, R. M., & Palmeri, T. J. (1997). An Exemplar-Based Random Walk Model of Speeded Classification. *Psychological Review, 104(2)*, 266–300.

Palmerston, J. B., Zhou, Y., & Chan, R. H. M. (2020). Comparing biological and artificial vision systems: Network measures of functional connectivity. *Neuroscience Letters, 739*, 135407.

Robbins, A., & Hout, M. C. (2020). Scene priming provides clues about target appearance that improve attentional guidance during categorical search. *Journal of Experimental Psychology: Human Perception and Performance.*

Sadr, J., & Sinha, P. (2004). Object recognition and Random Image Structure Evolution. *Cognitive Science, 28(2*), 259–

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, San Diego, CA.

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.

Võ, M. L.-H., & Wolfe, J. M. (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition, 126(2),* 198–212.

Yantis, S. (1996). Attentional capture in vision. In A. F. Kramer, M. G. H. Coles, & G. D. Logan (Eds.), Converging operations in the study of visual selective attention (p. 45–76). American Psychological Association.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6),* 1452–1464.