Predicting rhesus monkey eye movements during natural-image search

Mark A. Segraves
Department of Neurobiology,
Weinberg College of Arts and Sciences,
Northwestern University, Evanston, IL, USA

Emory Kuo
Department of Neurobiology,
Weinberg College of Arts and Sciences,
Northwestern University, Evanston, IL, USA

Sara Caddigan
Department of Neurobiology,
Weinberg College of Arts and Sciences,
Northwestern University, Evanston, IL, USA

Emily A. Berthiaume
Department of Neurobiology,
Weinberg College of Arts and Sciences,
Northwestern University, Evanston, IL, USA

Konrad P. Kording
Departments of Physical Medicine and Rehabilitation
and Physiology, Feinberg School of Medicine,
Northwestern University, Chicago, IL, USA

There are three prominent factors that can predict human visual-search behavior in natural scenes: the distinctiveness of a location (salience), similarity to the target (relevance), and features of the environment that predict where the object might be (context). We do not currently know how well these factors are able to predict macaque visual search, which matters because it is arguably the most popular model for asking how the brain controls eye movements. Here we trained monkeys to perform the pedestrian search task previously used for human subjects. Salience, relevance, and context models were all predictive of monkey eye fixations and jointly about as precise as for humans. We attempted to disrupt the influence of scene context on search by testing the monkeys with an inverted set of the same images. Surprisingly, the monkeys were able to locate the pedestrian at a rate similar to that for upright images. The best predictions of monkey fixations in searching inverted images were obtained by rotating the results of the model predictions for the original image. The fact that the same models can predict human and monkey search behavior suggests that the monkey can be used as a good model for understanding how the human brain enables natural-scene search.

Introduction

Choosing where to look next is one of our most frequent decisions. There have been substantial advances in understanding the factors that assist in guiding saccades. Recent models can produce good predictions of regions that will be targeted by saccades across natural and artificial stimuli (Koch & Ullman, 1985; Najemnik & Geisler, 2005; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Borji & Itti, 2013; Kümmerer, Theis, & Bethge, 2015). Three known factors can help with the predictions: (a) The salience, or conspicuousness of a point or object in a scene, is based on how different an element is from the rest of the scene in basic stimulus features such as color, contrast, shape, and orientation (Koch & Ullman, 1985). (b) Target relevance, or similarity to the target,

Citation: Segraves, M. A., Kuo, E., Caddigan, S., Berthiaume, E. A., & Kording, K. P. (2017). Predicting rhesus monkey eye movements during natural-image search. *Journal of Vision*, *17*(3):12, 1–17, doi:10.1167/17.3.12.

doi: 10.1167/17.3.12 Received July 8, 2016; published March 29, 2017 ISSN 1534-7362 Copyright 2017 The Authors

can alter the importance of the features for the task at hand; for example, if a human subject is looking for a red target, red features are important (Horowitz et al., 2007). These experimental studies have been complemented by models that detect human figures in natural images (Dalal & Triggs, 2005). (c) Context, enabled by an understanding of the scene, can assist in identifying places where the object is more likely to be found (Neider & Zelinsky, 2006). For instance, a human subject looking for a pen or pencil in an office scene would assign higher weight to a desktop than to bookshelves. Interestingly, some features may contribute to several of these factors (Jansen, Onat, & König, 2009). All these factors must matter during everyday visual search.

We have considerable knowledge about the importance of the three factors from a recent analysis of human search (Ehinger et al., 2009). The authors asked human participants to look for pedestrians in images of urban scenes. They then modeled the behavior combining all three factors, which each contributed considerably to the quality of the predictions. The results could be further improved by using a so-called context oracle, a way of using the judgment of other humans to identify areas of highest contextual structure in each scene. Interestingly, there is something of a confound here: The context model deals with the meaningful context—for example, a pedestrian being on the street—as well as the boring one, where fixations in the middle of the image are more likely (Tatler, 2007; Bindemann, 2010). Humans clearly use all three factors, with each explaining similar amounts of variance, which raises the question of whether other species choose fixations in a similar fashion.

With salience, relevance, and context making such an important contribution to human eye-movement guidance, we sought to investigate the role of these factors in the guidance of rhesus monkey eye movements. With a few notable exceptions (e.g., Einhäuser, Kruse, Hoffmann, & Konig, 2006; Ghazanfar, Nielsen, & Logothetis, 2006; Berg, Boehnke, Marino, Munoz, & Itti, 2009; Shepherd, Steckenfinger, Hasson, & Ghazanfar, 2010; Ramkumar, Fernandes, Kording, & Segraves, 2015; White et al., 2017; Wilming et al., 2017), visual search in nonhuman primates has been investigated using highly artificial stimuli. In order to obtain a full understanding of neural mechanisms guiding visual search in natural environments, it is essential to understand the factors that guide a monkey's search when the monkey views scenes that contain real-world settings. Our goal was to take advantage of the substantial progress that has been made in predicting human saccade targets and apply these quantitative models in an effort to predict the behavior of rhesus monkeys in a naturalistic search task.

Which factors would we expect to be predictive of monkey natural-scene search? It is reasonable to assume that salience—a supposedly innate, bottom-up visual feature—would have a similar influence upon monkey saccades as it does on human saccades. Berg et al. (2009) have shown that salience models are predictive of monkey fixations during viewing of natural and artificial video clips, although prediction of human fixations using salience models was stronger. Likewise, target relevance may play an important role in eye-movement guidance in both humans and monkeys (Fecteau & Munoz, 2006; Henderson, Malcolm, & Schandl, 2009; Ramkumar et al., 2015). However, it is unclear whether monkeys use image context to guide their saccades. One argument against the possibility that monkeys use context is that this might represent a higher cognitive ability that is unique to humans. Further, the use of context may be a learned behavior that humans acquire through experience. For example, humans know to look for pedestrians on sidewalks and not on rooftops, a result of a lifetime of observation. On the other hand, a laboratory monkey has little visual experience in the outside world and may have little or no experience seeing pedestrians in their natural surroundings. However, laboratory monkeys often have access to television, which may make their visual environment similar in some ways to that of typical human subjects. Moreover, the monkeys in this study were trained extensively on pedestrian search before the bulk of our data were acquired, allowing them to (in principle) learn about context by trial and error. In addition, the monkeys are likely to understand basic principles of physics, like gravity, that contribute to the contextual aspect of scene search (Võ & Henderson, 2009). The experiments described here assess the degree to which monkeys use salience, relevance, and context in visual search in comparison to that seen in human behavior.

Here we collected eye-movement data from two monkeys while they participated in a natural-scene search task. We fitted the data with models containing three factors: salience, relevance, and context. We also explored two versions of context: one defined by an algorithm and another defined by human observers. We found that salience, relevance, and context all contribute significantly to the prediction of eye-movements. In an attempt to disrupt image context, we found that the monkeys performed almost as well when viewing inverted images, although adjustments to the relevance and context models were needed to achieve levels of prediction that were similar to those obtained for the viewing of upright images. Lastly, a comparison with human performance reveals that the models are almost as good at predicting monkey eye movements as they are at predicting human ones.

Methods

Animals and surgery

Two female adult rhesus monkeys (*Macaca mulatta*) were used for these experiments, and are identified in this report as M15 and M16. Northwestern University's Animal Care and Use Committee approved all procedures for training, surgery, and experiments performed. These procedures conform to the Association for Research in Vision and Ophthalmology Statement for the Use of Animals in Ophthalmic and Visual Research. Each monkey received preoperative training designed to familiarize it with the experimental setup. This was followed by an aseptic surgery to implant a subconjunctival wire search coil and a titanium receptacle to allow the head to be held stationary during behavioral sessions. Surgical anesthesia was induced with thiopental (5–7 mg/kg intravenously) or propofol (2–6 mg/kg IV) and maintained using isoflurane (1.0%–2.5%) inhaled through an endotracheal tube. Monkey M15's implant included a plastic CILUX recording cylinder aimed at the frontal eye field, for use in experiments not described in this report.

Behavioral paradigms

We used the REX system (Hays, Richmond, & Optican, 1982) based on a PC running QNX, a real-time UNIX operating system, for behavioral control and eye-position monitoring. Visual stimuli were generated by a second, independent graphics process (QNX Photon) running on the same PC, and rear-projected onto a tangent screen in front of the monkey by a CRT video projector (Sony VPH-D50, 75-Hz noninterlaced vertical scan rate, 1024 × 768 resolution). The distance from the monkeys' eyes to the screen was 109 cm, and the projected image size was 48° wide × 36° high.

Both monkeys were trained to perform a calibration task that was run at the beginning of each data-collection session. Eye movements were sampled with either a magnetic search coil or a video eye-tracking system. A red target spot appeared on a gray background at the center, or at 12° eccentricity above, below, left, or right of the center of the screen. This task was used to calibrate the eye coil and eye-tracking camera signals. Monkeys received a water reward after maintaining eye position within a 2° window surrounding the target spot for 500 ms. Monkey M15 was a subject for earlier experiments and had extensive training and experience with the performance of a variety of oculomotor tasks, including visual and memory-guided saccade tasks and a scene-search task that required the monkey to find a target embedded in an image (Phillips & Segraves, 2010; Fernandes, Stevenson, Phillips, Segraves, & Kording, 2013; Glaser et al., 2016). For the search task used in the earlier experiments, the target was an image of a fly superimposed on a variety of natural images including scenes with animals, people, plants, and food. An alpha blending technique was used to embed the fly into the image, making it more difficult to locate. It is important to note that the fly was foreign to all of these images, and its location for each presentation of an image was determined pseudorandomly. Thus, salience and relevance cues might have aided the monkeys' search, but context would have had no influence. Monkey M16 was naïve prior to this study. She received preoperative training, followed by 10 days of postoperative training on the calibration task. For data collection with upright images, both monkeys' eye movements were tracked with a subconjunctival wire search coil, sampled at 1 kHz (Robinson, 1963; Judge, Richmond, & Chu, 1980). For data collection with inverted images, monkey M15's eye movements were tracked by search coil, while monkey M16's were tracked with an infrared eye tracker (ISCAN Inc., Woburn, MA; http://www.iscaninc.com/) at 60 Hz. The output of the infrared eye-tracker system was an analog signal which was calibrated for gain and offset using a signal conditioner (Intronix Technologies Co., Bolton, Ontario, Canada) whose output was then sampled at 1 kHz.

The search task was adopted from the work of Ehinger et al. (2009). We used the "target present" segment of their image set, which included a total of 456 color images containing pedestrians situated in urban scenes at a resolution of 800 × 600 pixels. In addition to one or more pedestrians, every image contained typical real-life combinations of roads, sidewalks, buildings, trees, and cars. Pedestrians subtended an average of 1.8° × 3.6°, corresponding to roughly 31 × 64 pixels. These targets were spatially distributed across the image periphery (target locations ranged from 5.4° to 26° from the screen center; median eccentricity was 17.2°) and were located in each quadrant of the screen with approximately equal frequency. The tasks began with the appearance of a red fixation spot at the center of the screen, and after the monkey fixated the spot, the image was turned on. The monkeys were rewarded when they directed their gaze to, and fixated, the pedestrian for 350–500 ms. The monkeys were allowed up to 20 saccades to find the pedestrian target, after which the image was turned off and a new trial began.

Monkey M15 was used for the development of the pedestrian search task, and received training on the task for 38 daily sessions before the data used in this study were collected. During this time, she was exposed to a subset of 48 images that were not used for the data-collection phase of these experiments. Monkey M16 received just 9 days of training on the task before we began to collect the data included in this report. Her training involved exposure to the same subset of 48 images used for training M15. For the inverted-image task, both monkeys were trained on inverted versions of the same set of 48 images for 7 days before data collection began. The monkeys were then exposed to the remaining images during the data-collection phase.

For each data-collection session, a monkey performed the pedestrian search task with a subset of 102 images. The image for a given trial was selected in a pseudorandom fashion from the group of 102 images used for that day to ensure that the images were presented in roughly equal numbers of trials. We noticed during the training sessions that performance began to decline after performing the task for 120 min, and so we limited data collection to the first 120 min of each behavioral session. Over 1,000 trials were run in each of these daily sessions.

Data analysis

All data obtained in these experiments were analyzed using MATLAB. For each trial, saccades in the eye-movement record were identified using velocity criteria. Endpoints of the saccades were then superimposed on the image that the monkey had been viewing (Figure 1A). Each image was processed using the salience, relevance, context, and combined models of Ehinger et al. (2009; see http://cvcl.mit.edu/SearchModels/). For each model, the parameters were set so that the model identified the region equivalent to 20% of the entire image where saccade endpoints were most likely to be found (Figure 1B through E). For the combined model, we used the same exponent values used for the human data ($\gamma_1 = 0.1$, $\gamma_2 = 0.85$, $\gamma_3 = 0.05$), which define the relative importance of each of the components. Saccade endpoints that fell within the regions predicted by the models were expressed as a percentage of the total number of saccades that landed within the boundaries of the image. Saccade endpoints that fell outside the boundaries of the image were labeled as off-slide fixations, and were not included in calculating model predictions. For each trial's data, we also generated a shuffle control where the saccade endpoints were placed on an image that was selected at random from the stimulus set and subsequently used for comparison to predictions of the models (Figure 2).

Results

Over the course of a 4-day testing period, each monkey viewed a total of 408 images. Completed trials include those where the monkey either found the pedestrian (correct trials) or searched until the limit of 20 saccades was reached (error trials). Monkey M15 completed a total of 4,057 trials with 27,790 total fixations (6.8 fixations/trial), and monkey M16 completed 3,877 trials with a total of 46,292 fixations (11.9 fixations/trial). This data set thus allows us to ask how we can understand the choice of fixations.

It is crucial to know if the monkeys can successfully find pedestrians. One monkey (M15) was experienced at visual-search tasks and completed a higher percentage of correct trials than the more naïve monkey (M16). Monkey M15 correctly located and fixated the pedestrian in 83% of image trials (3,378/4,057). In contrast, monkey M16 performed at a level that correctly fixated the pedestrian in only 54% of the total image trials completed (2,091/3,877).

Despite this difference in success rate for the two monkeys, their behavior shared some common features. For example, the frequency of looking outside of the image boundaries is a measure of the level of motivation to perform the task as well as the difficulty of finding the pedestrian in a particular image. This frequency was remarkably similar between monkeys (M15: 6%, 1,749/27,790; M16: 7%, 3,097/46,292). In addition, for both monkeys the majority of these off-slide fixations occurred during error trials in which the monkey did not locate the pedestrian (M15: 90%, 1,579/1,749; M16: 91%, 2,815/3,097). This similarity in the percentages of off-slide fixations for the two monkeys suggests they shared a similar motivation to correctly perform the task. Monkey M15's more extensive experience with search in natural images may explain the higher percentage of correct trials that she completed. Additionally, M16's performance was at a level similar to her performance in a number of subsequent behavioral testing paradigms, where she typically achieved a lower percentage of correct trials than was the case for M15 (Ramkumar et al., 2015; Glaser et al., 2016; Ramkumar et al., 2016). It seems that the difference in performance is not a categorical difference but simply a lower level of performance for one of the monkeys. As will be seen later, the general ability of the salience, relevance, and context models to predict where these monkeys fixated was also similar with both monkeys. Therefore, both monkeys can find the pedestrian in this task, and both appear to apply similar strategies to succeed.

How well do the models predict fixation choice? A comparison of the percentages of fixations predicted by the models for the entire data set reveals a gradual increase in the percentage of fixations predicted, with

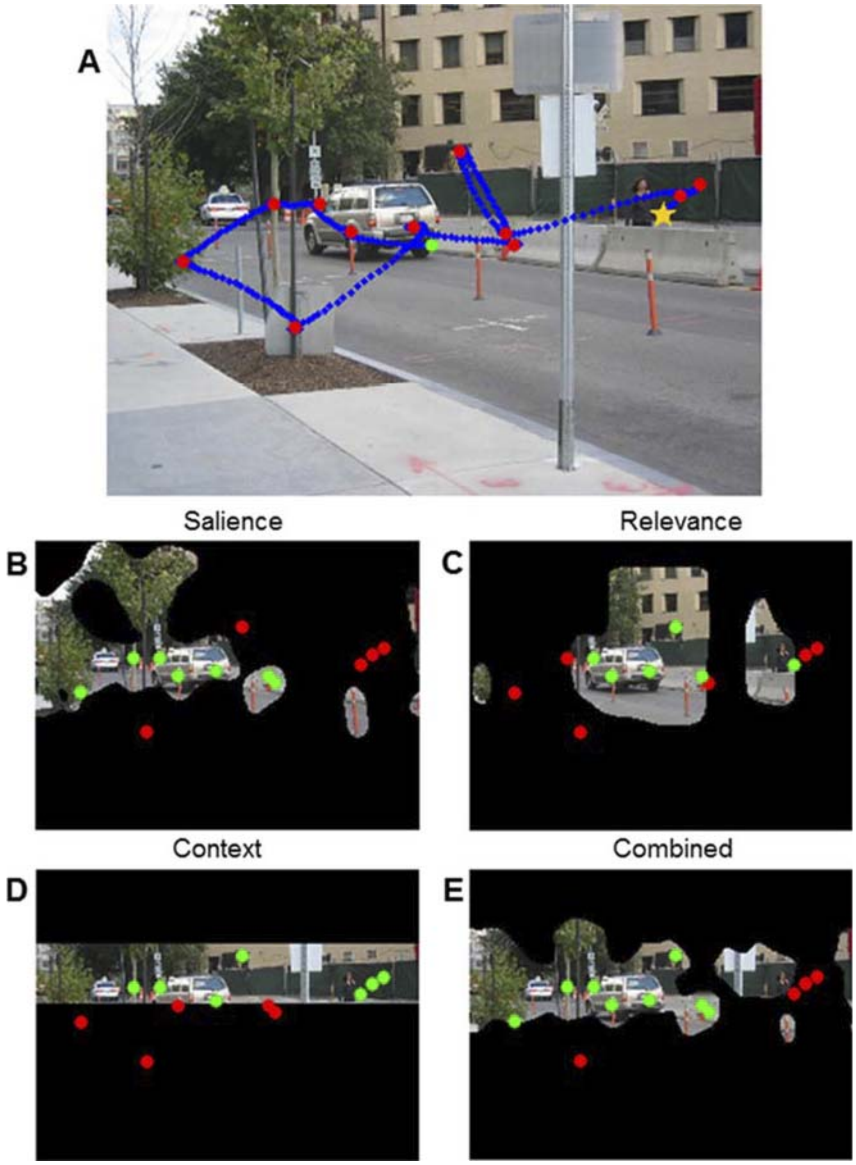

Figure 1. Search behavior and analysis. (A) Example of test image with pedestrian target and a monkey's eye-movement behavior during a single trial. The trial begins with fixation at the green dot in the center of the image and ends when the monkey correctly finds and fixates the pedestrian. Blue dots mark eye position sampled at 1 kHz. Red dots mark saccade endpoints. The gold star marks the location where the monkey captured the pedestrian target. (B–E) Unmasked areas mark 20% of the total image area where salience, relevance, context, and combined models predict that the endpoints of saccades made during searching for pedestrians are most likely to be found. Green dots mark fixations from the trial shown in (A) that fell within the area predicted by the models, and red dots mark fixations located outside of the predicted areas. For this trial, the correspondences of actual saccade endpoints to model predictions were 58% (salience), 50% (relevance), 58% (context), and 67% (combined).

salience being the least predictive, relevance the next most predictive, context the second most predictive, and the combined model the best predictor of saccade endpoints (Figure 3A). Predictions for the more experienced monkey (M15) were slightly higher, but the trend for both monkeys was the same. For the shuffle control, the correspondence between saccade endpoints and model predictions was in the range of only about 20%–30%. The models do a decent job at predicting fixations.

In the pedestrian search task, we should expect that early saccades may be more predictable because the monkey has not yet accumulated information about the visual scene that would bias future saccades. After all, such information could indicate that there are regions of the image where the target cannot be, and the model does not incorporate such terms. Indeed, Velichkovsky and colleagues (Pannasch & Velichkovsky, 2009; Pannasch, Schulz, & Velichkovsky, 2011) have shown that variations in visual processing can affect the outcome of

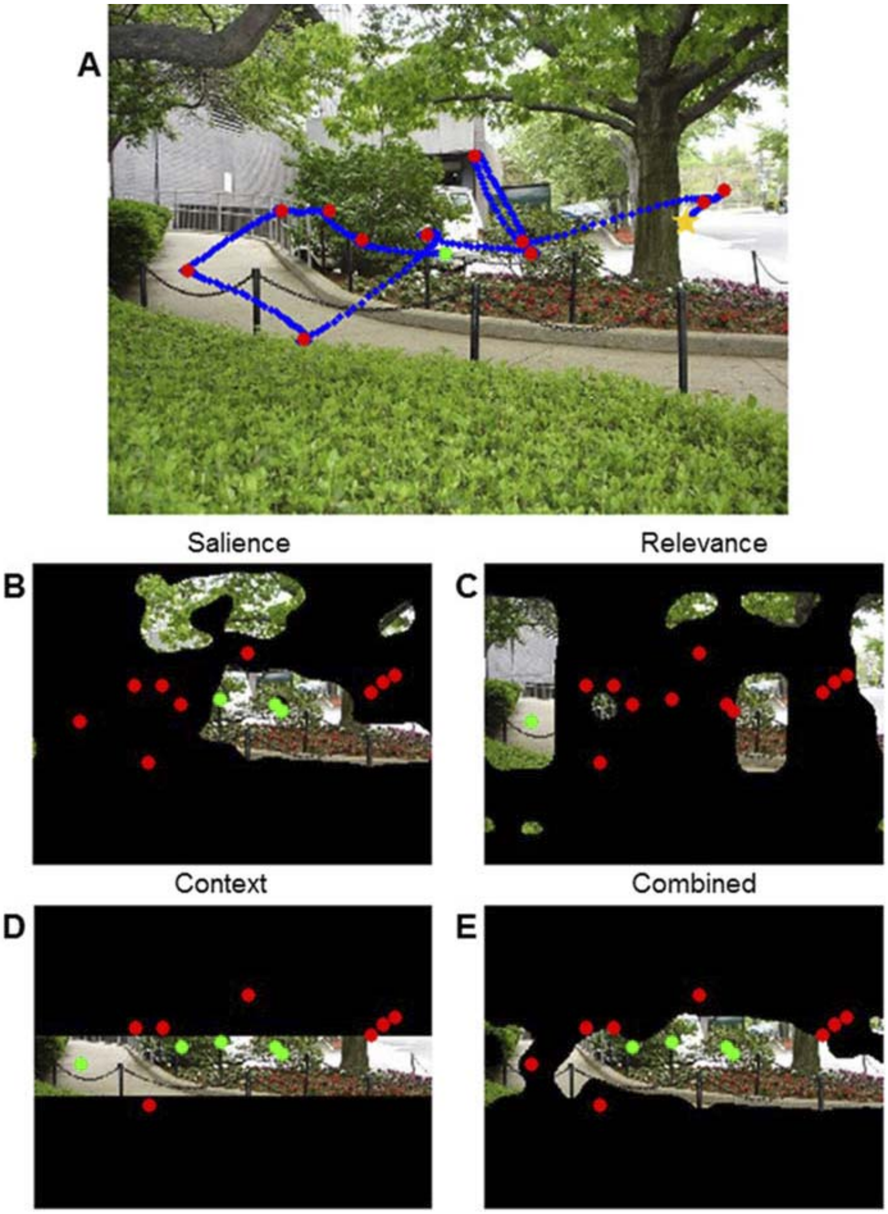

Figure 2. Shuffle control. (A) Eye-movement data from trial shown in Figure 1 placed on an image selected at random from the image library. (B–E) In a manner identical to that of Figure 1, unmasked areas mark 20% of the total area where salience, relevance, context, and combined models predict that the locations of saccadic endpoints made while searching for pedestrians are most likely to be found in this image. The correspondence of saccade endpoints from the Figure 1 trial to model predictions for this randomly selected image were 25% (salience), 8% (relevance), 42% (context), and 33% (combined).

saccade decisions during viewing of images. Likewise, scene complexity, which varied across the image set, affects the efficiency of visual search (Henderson, Chanceaux, & Smith, 2009). The models should have their strongest predictive power during the initial saccades. To allow us to compare the monkey data with those reported by Ehinger et al., we looked at the models' ability to predict the location of the first three fixations. When the endpoints of the first three saccades from both correct and error trials were included, the trend in predictive power of the models was similar to that seen when every saccade made in each trial was considered. However, the magnitude of the predictions of all four models was slightly higher when only the first three saccades were used (Figure 3B). This effect was stronger for the less experienced monkey (M16), which required a higher average number of saccades to find the pedestrian. This resulted in a larger dichotomy between the analyses of all saccades when compared with the analyses of the first three saccades for M16.

To further test the assumption that the predictive power of the models would be greater for earlier

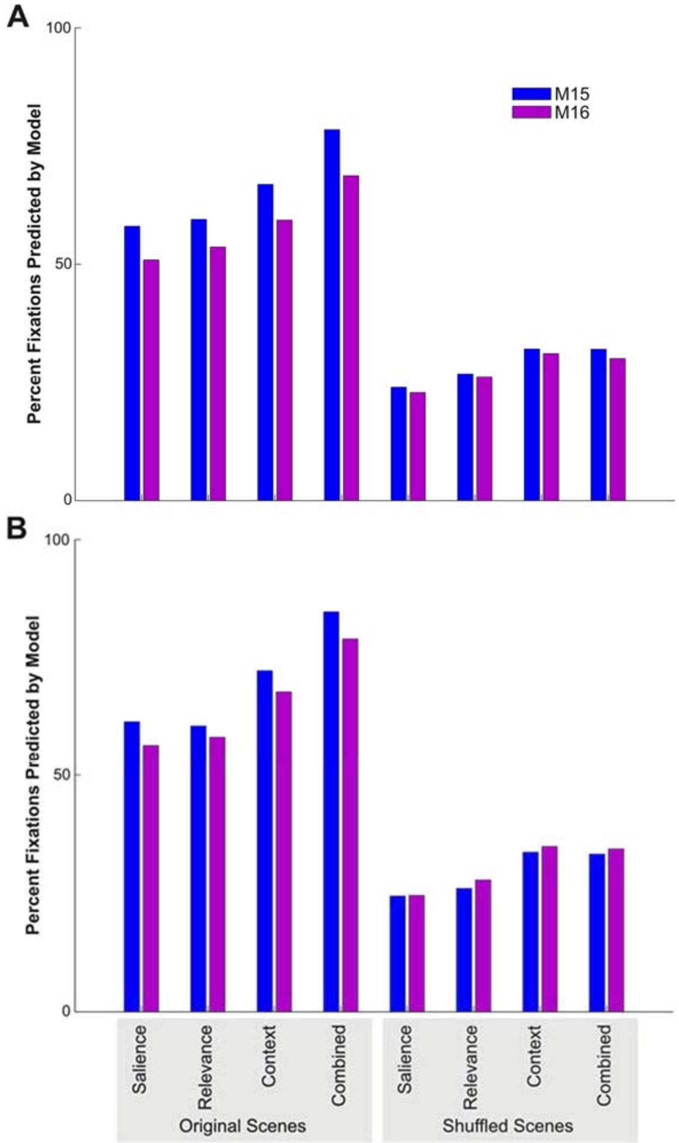

Figure 3. Percentages of fixations predicted by the models for all trials, including both correct and error trials. (A) Percentages predicted by the models for saccade endpoint locations on original scenes and shuffle-control scenes for all saccades that landed within the image boundaries in each trial. (B) Percentages predicted by the models for the first three fixations in each trial. For original- versus shuffled-scene comparisons performed on these data as well as for the data plotted in Figure 5, every comparison with a two-sample *t* test reached significance, $p < 0.0001$. Standard error of the mean for these data were too small to be visible on these bar graphs. Off-slide fixations that landed outside of the boundaries of the image were excluded from this analysis.

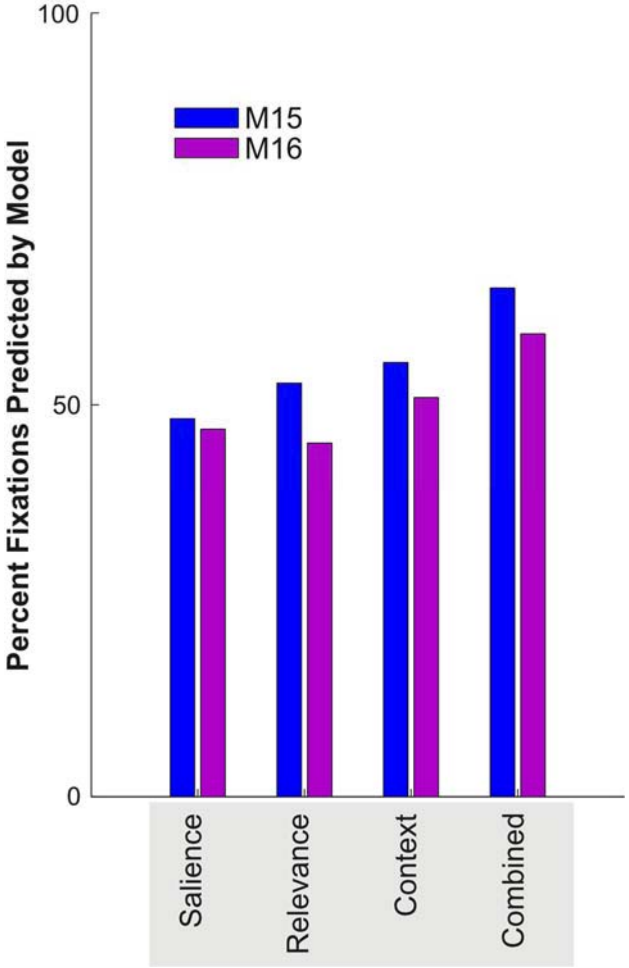

Figure 4. Effect of fixation number: Percentages predicted by the models for saccade endpoint locations on original scenes for all saccades that landed within the image boundaries in each trial. In this case, only fixations that occurred after the third fixation were used to obtain the percentages (Fixations 4–20).

saccades, we compared the ability of the models to predict fixations in trials where there were more than three fixations (Figure 4). The percentages of fixations predicted by the models for the later fixations were clearly reduced in comparison to predictions for the first three fixations (Figure 3B). The predictions plotted in Figure 3B include a number of trials when the monkey located the pedestrian with three or fewer saccades. When we looked only at trials where there were more than three fixations, and compared the percentages of fixations predicted by the models for the first three fixations versus the percentages for Fixations 4–20, we found that the differences were significant ($p < 0.001$) for all models with the exception of the relevance model for monkey M15 ($p = 0.2$). Therefore, early saccades are more predictable than late saccades using these models.

If the models are robust, then they should predict saccades in both successful and unsuccessful trials. We thus compared the data for correct and error trials to determine whether or not the models were predictive only in trials when the monkeys correctly fixated the pedestrian (Figure 5). Predictions are not as high for error trials compared to successful trials, but the trends

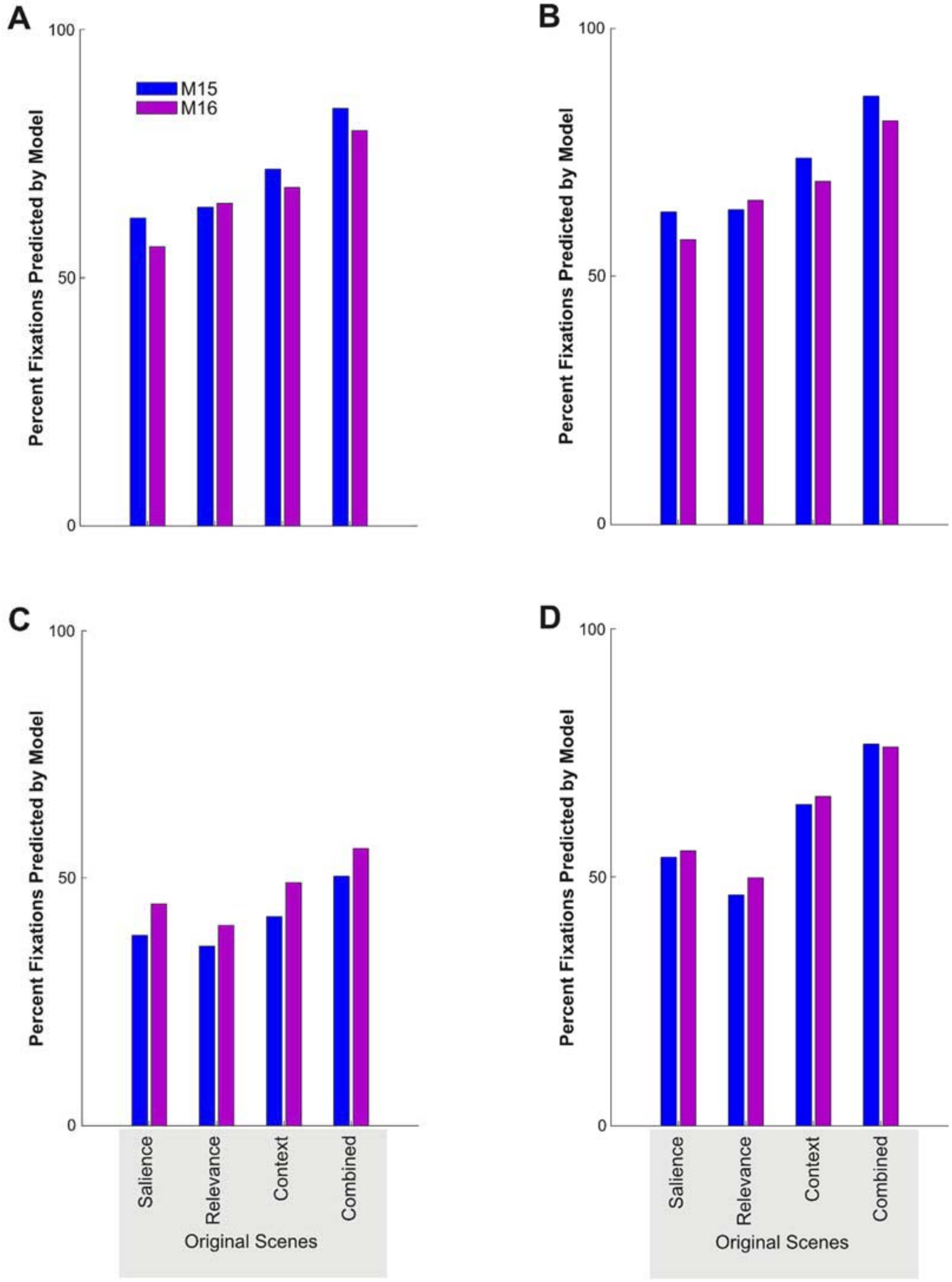

Figure 5. Model predictions for correct and error trials. (A–B) Percentages for saccade endpoint locations for (A) all saccades and (B) the first three saccades landing within the image boundaries in correct trials where the monkeys successfully found and fixated the pedestrian. (C) Percentages for all saccade endpoints in error trials where the monkey failed to find the pedestrian after 20 saccades. (D) Percentages for the first three saccade endpoints in error trials.

between models are preserved (compare Figure 5A and B to Figure 5C and D). For error trials, the prediction percentages were lower, but they were still significantly different from shuffle controls. The models make meaningful predictions even for error trials where the monkey never finds the target.

The observation that the first three saccades are more predictable (Figure 3) could be driven by the inclusion of both successful and unsuccessful trials in the comparison. The increase in model performance when comparing data for the first three fixations to data for all fixations (Figure 3B compared to Figure

3A) is lost by looking only at correct trials (Figure 5B compared to Figure 5A). This suggests that most of the loss of predictive power in looking at all fixations versus the first three can be attributed to error trials where the monkeys reached the limit of 20 saccades without finding the pedestrian. This finding suggests that prediction quality is comodulated by underlying factors that influence success, including scene complexity and spatial bias.

It is known that some of the prediction quality of fixation models comes from the effective modeling of the center bias (Tseng, Carmi, Cameron, Munoz, & Itti, 2009; Bindemann, 2010; Nuthmann & Henderson, 2010; Borji & Tanner, 2016). Shuffle controls can be used to disentangle the effects that are specific to a given image from general location biases. All the models were far better at predicting fixations on the actual images versus the shuffle control (Figure 6). Shuffle-control predictions for the context model are slightly higher than those for the other models (Figure 3), lowering the difference between the shuffled and unshuffled predictions (Figure 6). This should be expected because the context model will naturally include a strong center bias.

In the standard model, the context is provided by a trained computational algorithm; however, as context deals with the meaning of the scene, this is a difficult problem to solve by computational methods alone. Ehinger et al. introduced an additional method by which humans indicate where in the image pedestrians might meaningfully be found. The model obtained was called the context oracle and is better at describing human fixations during pedestrian search. We find that using the human context oracle instead of the computational algorithm for context also leads to better predictions of eye movements for the monkeys. The average predictions for the first three fixations in correct trials were 71% for context versus 87% for context oracle ($p < 0.001$). This suggests that monkeys and humans share a common ability to properly understand the complex context defining a visual scene.

For each daily data-collection session, a monkey performed the pedestrian search task with a subset of 102 images. The image for a given trial was selected in a pseudorandom fashion from the group of 102 images used for that day to ensure that the images were presented in roughly equal numbers of trials. With about 1,000 trials analyzed per session, the monkey saw repeated presentations of each image up to 10 times per session, and we may expect an effect of presentation order. To determine if the effectiveness of the models in predicting fixations remained about the same for repeated presentations of an image, we compared model predictions for the first and fifth presentations of an image (Figure 7). This analysis shows a slight reduction for salience—but not relevance, context, or combined models—between the first and fifth scene presentations. In fact, there was no significant difference ($p > 0.4$) for all four model predictions of first versus fifth presentation of the scene for either monkey. The same result was obtained when we compared the models' ability to predict fixations for Presentations 1–4 to those for Presentations 5–8 (not shown). Presentation order does not seem to be an important driver of search behavior.

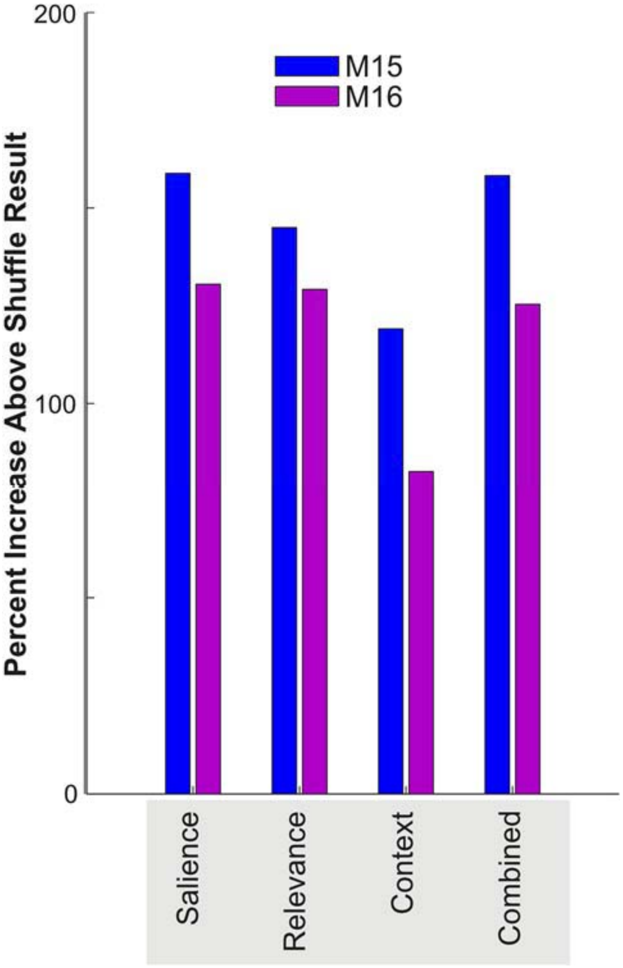

Figure 6. Comparing model predictions for original versus shuffle-control images for correct trials, first three fixations. Values on the y-axis represent the percentage increase of the model prediction for the distribution of fixations on the original search image in comparison to the predictions for the same fixations placed on a shuffle image chosen at random from the image set. If predictions for the original image were equal to predictions for the shuffle image, the percentage increase was zero.

Because context was, somewhat surprisingly, highly effective at predicting monkey fixations, we wondered if there could be ways of disrupting it. We thus tested the monkeys with an inverted set of the same pedestrian-task images (Figure 8). Surprisingly, the monkeys were able to locate the pedestrian at a rate similar to that for

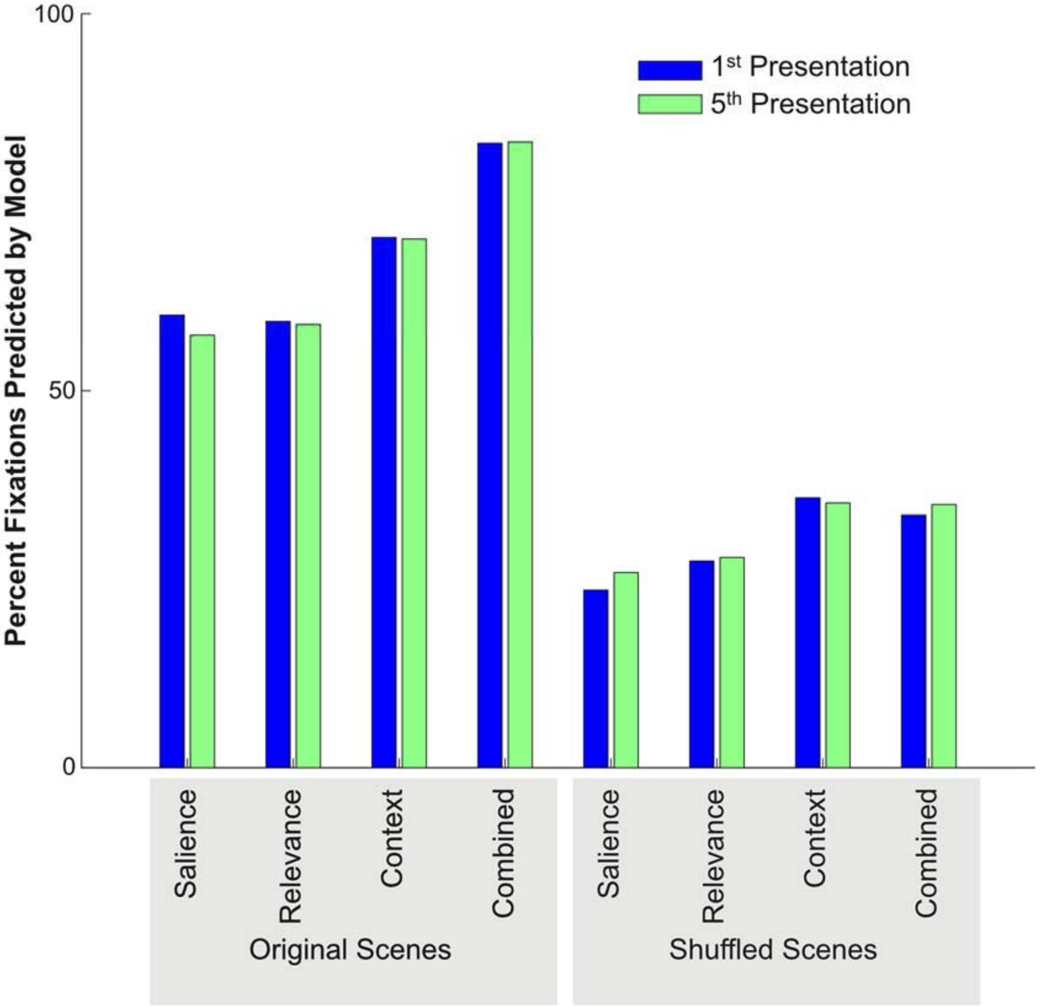

Figure 7. Effect of repeated scene presentations: Percentages of fixations predicted by the models for the first three fixations in all trials. Here, data are divided based upon model predictions when the monkey was presented a scene for the first time (blue) versus when presented for the fifth time (green) during a single data session.

upright images (average correct trials = 69% upright, 64% inverted). To assess the strength of the models in this version of the task, we applied them to the inverted images and looked for correspondence between the model predictions and the locations where the monkeys fixated during the first presentation of each image (Figure 9A). The salience model appeared to perform as well on the inverted images as on the original upright images. The difference between the salience predictions for fixations on the inverted image versus an inverted shuffle control was highly significant for both monkeys ($p < 0.001$). This result could be expected, since although orientation contributes to the salience map, the salience model is not sensitive to the cardinal orientation of the bottom-up visual features that are the basis for its predictive power. In contrast, the relevance model developed for upright images was not successful in predicting fixations on inverted images. Predictions for the relevance model were not significantly different from those for inverted shuffled images (M15: $p > 0.05$; M16: $p > 0.8$). This might be expected since, intuitively, the relevance model searches for objects resembling a human figure with head above arms and torso, above legs. The context model also generated predictions that were not significantly different from the shuffle predictions (M15: $p > 0.20$; M16: $p > 0.05$). This also was not surprising, because context is also disrupted by image inversion. Lastly, the combined model's predictions were small but significantly different from the predictions for shuffled images for both monkeys ($p < 0.001$). This effect was largely due to the inclusion of salience in the combined model. The average percentage of fixations predicted by the combined model for inverted images was substantially less than that predicted by the combined model when the monkeys searched upright images (25% vs. 82%).

Because the relevance and context models were not designed to detect these features in inverted images, we

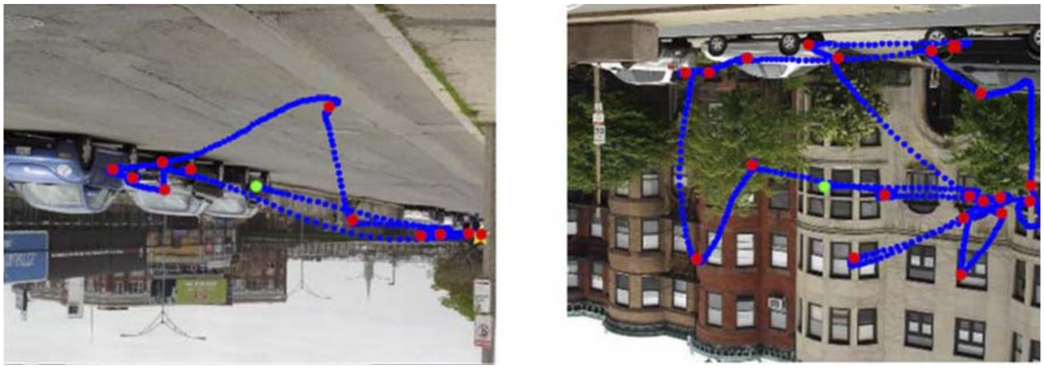

Figure 8. Search for pedestrian in inverted images: Examples of successful and unsuccessful trials where the monkey searched for a pedestrian in an inverted image. As in Figure 1, each trial begins with fixation on the green dot in the center of the image. The trial ends with fixation at the location marked by the gold star when the monkey correctly finds and fixates the pedestrian (left panel), or when the monkey fails to find the pedestrian after making 20 saccades. Blue dots mark eye position sampled at 1 kHz. Red dots mark saccade endpoints.

compared model predictions for the upright versions of the models to the monkeys' fixations on the inverted images. In other words, the predictions for each model were made using the original, upright image, and then we inverted these predictions and compared them to the monkeys' fixations on the inverted images (Figure 9B). This method was successful at predicting the monkeys' fixations on inverted images. However, the percentages of fixations predicted with this method were not as high as those obtained when the monkeys viewed upright images. Predictions were lower for both monkeys, but especially notable for M16 (compare Figure 5B to Figure 9B). Inverting images does not diminish the monkeys' success in finding the pedestrian but does substantially reduce the ability of all models except salience to predict the location of fixations.

With the combination of all three models providing the best predictions of fixations for both upright and inverted images, we looked at the performance of two-source models for comparison (Figure 10). This comparison underlines the fact that each of the models—salience, relevance, and context—make an important contribution to the predictions. The two-source models show substantial increases in percentage of fixations predicted over single-source models (compare to Figure 5B). The percentages of fixations predicted by the salience-plus-relevance and relevance-plus-context models are slightly lower, and significantly different from the three-source model ($p < 0.004$). However, the percentage of fixations predicted by the salience-plus-context model is not significantly different from that of the combined model for either monkey (M15: $p = 0.47$; M16: $p = 0.71$). This suggests that a two-source model that combines context with salience or relevance performs about as well as the three-source model that includes context, salience, and relevance.

How does model performance vary between humans and monkeys? The monkeys performed the pedestrian search task on exactly the same images as in the original study by Ehinger et al., and we applied the same models as used in the human study. This allows us to compare the results of our model fits on monkey behavior with the behavior of humans (Figure 11). Prediction qualities of all the models are surprisingly similar between monkeys and humans, suggesting similar search strategies for both species. In addition, this favors the monkey as an ideal model system if we want to understand how the brain solves the problem of fixation choice in natural scenes.

Discussion

In this study, we have taken visual-search behavior models that have been optimized to predict human search performance and examined the applicability of these models to monkeys performing a pedestrian search task. Using hundreds of images of outdoor real-world environments, the two monkeys searched for a pedestrian in each natural scene. Their eye movements were recorded and we applied the four computational models that Ehinger et al. (2009) constructed from human visual-search behavior in the same search task. Salience, relevance, and context models of human visual-searching behavior, when applied to rhesus-monkey visual-searching behavior, predicted monkey saccades about as well as human saccades.

We used the same way of evaluating the model predictions as Ehinger et al., but one could easily take issue with the implementation. The models estimate, for each pixel of the image, how likely a saccade to that location is. Then they choose the 20% of pixels with the highest expected probability. The percentage prediction estimates are the proportion of trials where the eyes land within those 20% of chosen pixels. This metric is

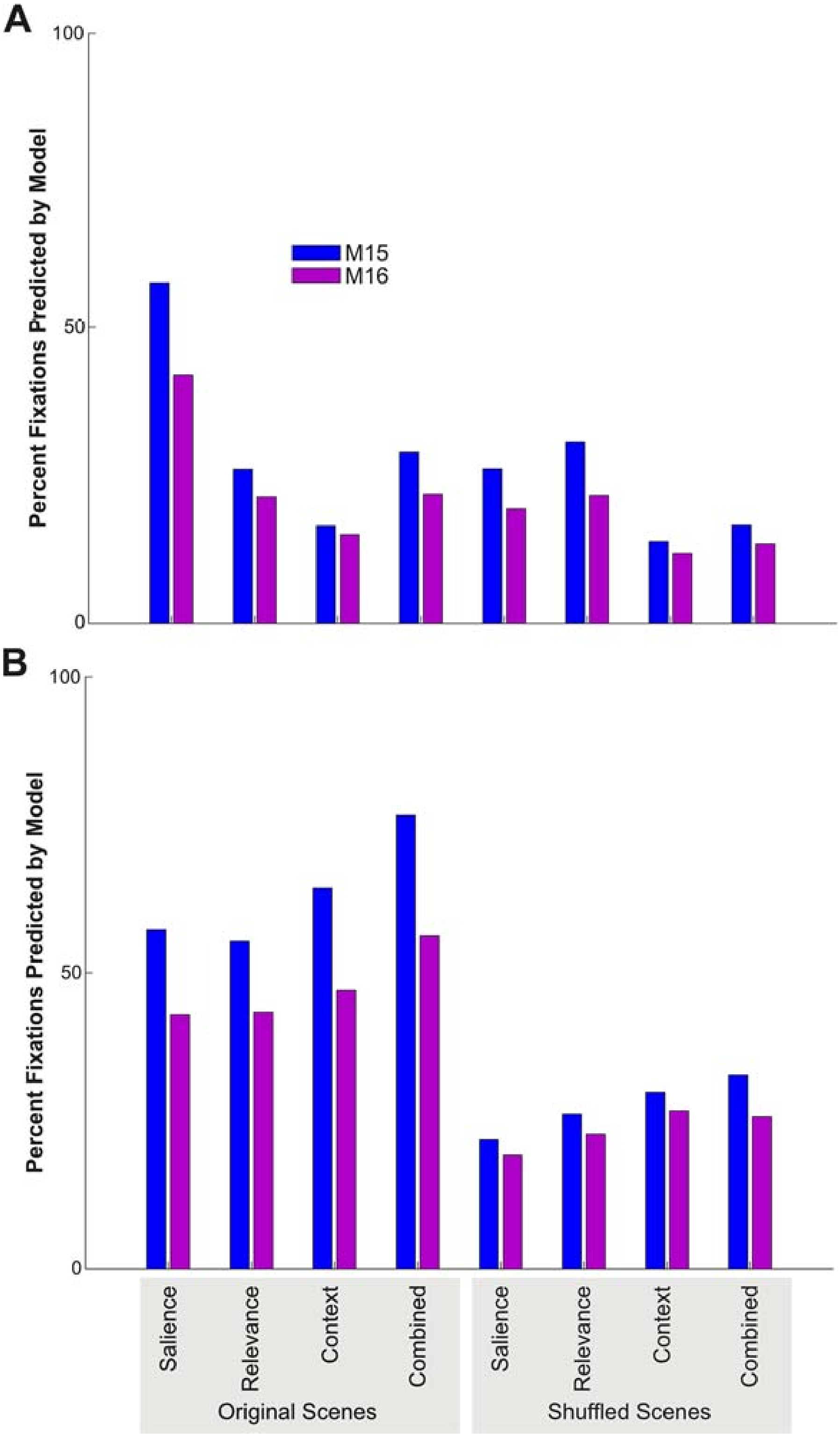

Figure 9. Model predictions for inverted-image search. These data include the first three fixations for trials when the monkeys searched a particular inverted image for the first time. Two methods of analysis were used. (A) Model predictions were made based upon the inverted images, then compared to the locations where the monkeys fixated while searching for the pedestrian on the inverted image. (B) Model predictions were made based upon the original upright orientation of the image. These predictions were then rotated 180° and compared to the monkeys' fixations for inverted images. Although the numbers of trials and fixations were small compared to data obtained for viewing of upright images, the standard error of the mean was always less than 2%.

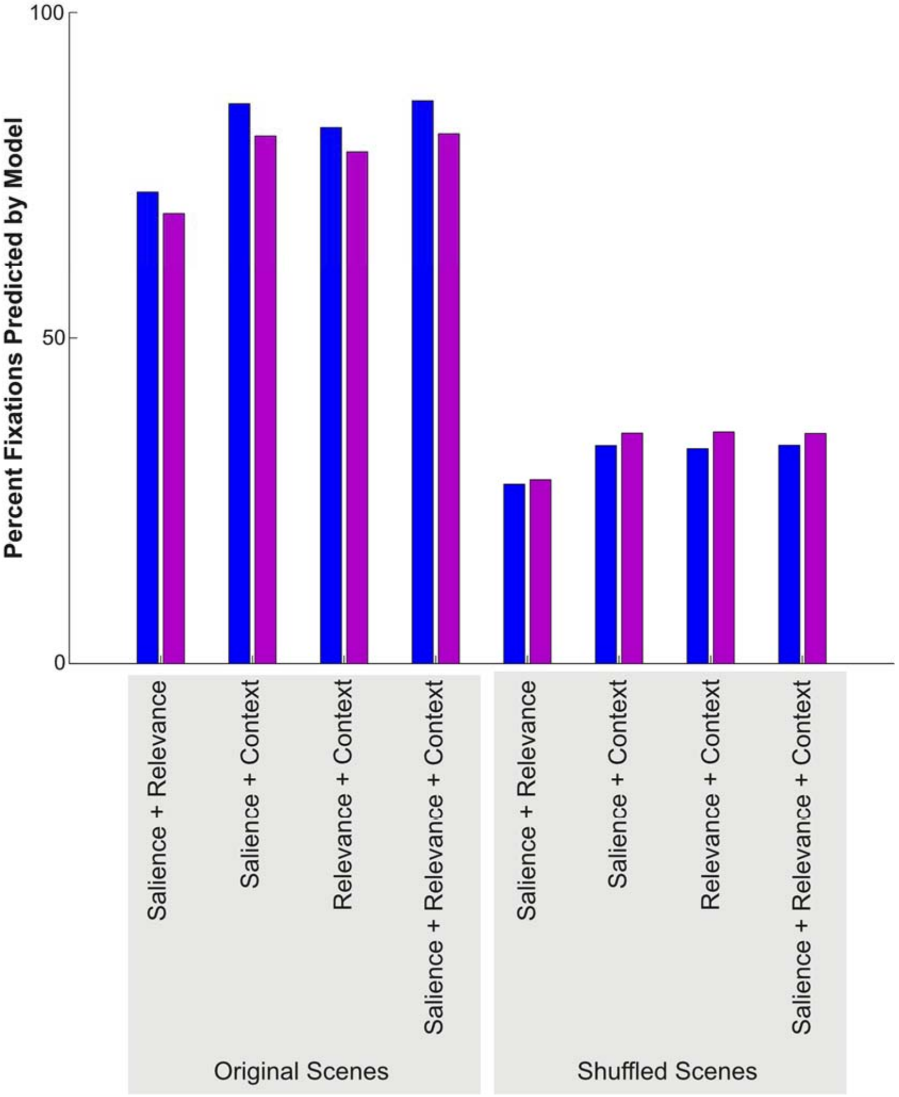

Figure 10. Two- versus three-source model comparison: Percentages of fixations predicted by combinations of two models compared to the combination of all three models. Data are for the first three fixations of correct trials. The same data using single-source model predictions are shown in Figure 5B. The predictions of two-source models shown here represent substantial increases over the predictions of single-source models.

arbitrary, and we primarily used it for comparison to the original study. However, the other metrics in the field have similar problems. The commonly used area under the curve will give a value of 50% for random guesses and basically deals with the trade-off of the proportion of wrongly chosen pixels and rightly chosen pixels. Therefore, it does not seem as if there is an easy way around this issue. For a review about the choice of metric, see Wilming, Betz, Kietzmann, and König (2011). We just have to be aware that even if our models predicted a very high percentage of fixations, this would not imply that we have perfectly understood saccade choice.

The models that we have used clearly omit important variables that influence behavior. There are many factors—for example, planning (Phillips & Segraves, 2010) and optimal visual sampling (Najemnik & Geisler, 2005)—that influence saccade choice. Further evidence of the shortcomings of our current models is that the context oracle developed by Ehinger et al. is a far better predictor of human and monkey fixations than any of the computational models. Developing better models for fixation choice is important, because it both leads to insights into the mechanisms and is practically relevant, for example, in the context of user interfaces (McCormick & Sanders, 1982). However, in the context of this article we wanted to focus on the differential behavior between monkey fixation choice and that of human subjects.

Recent work has shown a new strategy for predicting eye movements that can produce considerably better fixation-choice predictions than conventional ap-

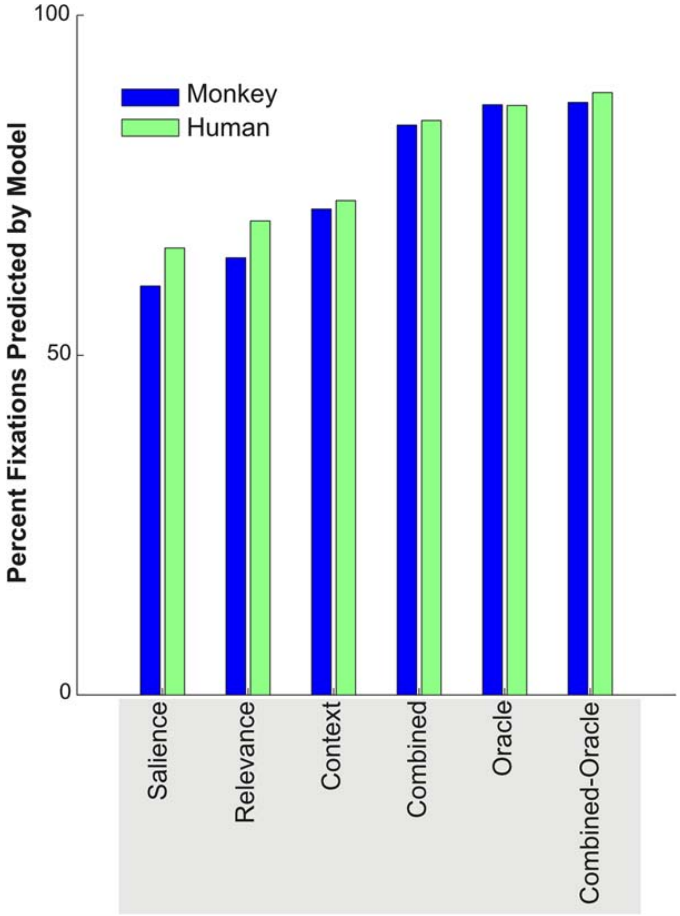

Figure 11. Average model predictions for both monkeys using first three fixations in correct trials (Figure 5B), compared to predictions of human performance for first three fixations with target-present images (Ehinger et al., 2009, table 1).

proaches (Kümmerer et al., 2015). For these approaches, large artificial neural networks are trained to identify objects in the real world. The internal structure of these networks is then used to create regressors for estimating eye movements. These approaches are quite powerful, but they are relatively opaque when it comes to determining which features drive successful predictions. We therefore decided not to compare our results with these models. However, understanding how these models can be so successful is an important task for future work.

The monkeys in our experiment are, arguably, far more motivated than humans at successfully finding the pedestrian. After all, their water intake is controlled, and each successfully found pedestrian gives them a water reward. On the other hand, the typical undergraduates used in human behavioral studies have no such reward incentives. However, these undergraduates are very much motivated by task success, and we know that high scores during psychophysical tasks activate reward-related areas (Vilares, Howard, Fernandes, Gottfried, & Kording, 2012). Still, there is the possibility that higher motivation resulted in elevated performance in the monkeys in comparison to humans. This difference in performance might have narrowed the gap in the model predictions for monkeys versus humans. However, it is not clear that this is detrimental. After all, for a physiological study of fixation choice, we would use motivated monkeys as a stand-in for arguably less motivated human beings.

Monkeys in our experiment saw the same image multiple times: We used a total of 408 images and the monkeys participated in thousands of trials, about 10 trials per image. Although each data set includes a total of four 120-min sessions with a different set of 102 images displayed in each session, there is a concern that the monkey would learn by heart how to find the pedestrians. However, neither of the monkeys seems to have used such a strategy, because performance (based upon model predictions) stayed constant throughout the experiment (Figure 7). In addition, it has been shown that repeated presentation of images to human subjects has limited effects on the distribution of fixations in a free viewing task (Kaspar & König, 2011). However, across images it is possible that the monkeys learned how context works. After all, they do get to see countless images and get to see where they do find the pedestrian. For the interpretation of our study, though, this is irrelevant: Either the monkeys understood context to start with or they quickly learned it from the images. Future work could look at how monkeys learn to properly deal with context.

Apart from replicating the Ehinger et al. study with monkeys, we also introduced the inversion condition. Strikingly, monkey performance was virtually unaffected by inversion. This means that whatever strategy they are using, they must be able to adapt it to a 180° rotation. We observed that if we use the original Ehinger et al. relevance and context models, they do very poorly at predicting the actual fixations. Instead, the monkey behavior can be well predicted by rotating the relevance and context predictions from the original upright image. This might be predicted, because humans can efficiently deal with image rotations even when they are unable to do mental rotation (Farah & Hammond, 1988). In whatever way the monkey represents its fixation choice, it must have the ability to represent or compute with whole-scene rotations. Future research could ask how the brain solves this problem.

In behavioral science, it is important to show generalization. Fitting a model to one behavior will generally be good at describing that behavior but worse at describing similar but related behaviors. Generalization studies are therefore important in asking if the model also works in other situations—for example, differences in gender, socioeconomic status, and so on (Henrich, Heine, & Norenzayan, 2010). Our study shows that the results of the Ehinger et al. study generalize across species.

We had expected that salience and relevance would be important for monkey eye movements (Fecteau & Munoz, 2006), but we had, among ourselves, not been able to agree on our expectations for the context model. We found that the context model explains a great deal of variance for the monkey—about as much as it does for the human. This shows that context matters to the monkeys. Even the human oracle, in which the model is derived from judgments by other humans about where the target might be, is predictive of monkey behavior. It thus seems that understanding visual scenes in terms of the relevant context, the meaning of the scene, can be solved well by monkeys. The origin of this convergence is interesting. Do humans and monkeys share algorithms for dealing with space and physics? Or alternatively, do they share some experiences—for example, watching television, where pedestrians are often shown—which produce the understanding of context? We must also recognize that a variety of other factors may aide the localization of the pedestrian in this task that do not depend directly upon salience, relevance, and context (Biederman, Mezzanotte, & Rabinowitz, 1982; Hollingworth & Henderson, 1998; Castelhano, Mack, & Henderson, 2009). These questions are interesting because they ask how cognitive abilities or viewing experiences are shared between humans and nonhuman primates.

No matter what the origin of this similarity is, our study shows that the algorithm for fixation choice is similar between humans and monkeys. While there exist some differences—for example, in the Ehinger et al. study humans made an average of 3.5 fixations per trial, while monkeys in our study made an average of 9.4 fixations per trial—the current findings suggest that monkeys can be a good proxy to understand how the human brain deals with fixation problems. It also suggests that physiological experiments on monkeys may shed light on the ability of humans to look at what matters to them in cluttered real-world visual scenes.

Conclusions

To understand how the brain guides eye movements during visual search ultimately requires an understanding of how the brain solves the kinds of tasks encountered during everyday life. Studies of human visual search have developed algorithms which are able to predict fixation choices with a high degree of accuracy during search of complex visual scenes. Three important visual features predicting human visual-search behavior are the distinctiveness of a location (salience), similarity to the target (relevance), and features of the environment that predict where the object might be (context). The ideal animal model for studying the brain's control of visual search is the rhesus monkey. In this study, we trained and tested two monkeys on a behavioral task in which they searched for pedestrians in images of urban environments (Ehinger et al., 2009). Monkey eye-movement behavior was then compared to the predictions of the models developed by Ehinger et al. to predict human behavior in this task. The salience, relevance, and context models optimized for human search were all predictive of monkey eye fixations, and a model combining all three was accurate to a level that approached the model's predictions for human behavior. One of the most striking findings was that models of scene context developed for humans were also excellent predictors of monkey fixations. This suggests not only that rhesus monkeys rely upon scene context to guide their search but that monkeys and humans might share similar strategies for incorporating scene context into their search behavior.

Keywords: visual search, real-world scene, eye movement, salience, contextual guidance, target features, behavior

Acknowledgments

We are grateful to the staff of Northwestern University's Center for Comparative Medicine for animal care, and Northwestern University's Instrument Shop for machining. Additionally, we would like to thank K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva for their original work and for making their data set, images, and models openly available to the scientific community. We also thank Daniel Wood for his comments on the manuscript. This work was supported by National Eye Institute Grant EY-021579. The authors have no conflicts of interest, financial or otherwise.

Commercial relationships: none.

Corresponding author: Mark A. Segraves.
Email: m-segraves@northwestern.edu.
Address: Department of Neurobiology, Weinberg College of Arts and Sciences, Northwestern University, Evanston, IL, USA.

References

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision*, *9*(5):19, 1–15, doi:10.1167/9.5.19. [PubMed] [Article]

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

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*, 2577–2587.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 185–207.

Borji, A., & Tanner, J. (2016). Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations. *IEEE Transactions on Neural Networks and Learning Systems*, *27*, 1214–1226.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3):6, 1–15, doi:10.1167/9.3.6. [PubMed] [Article]

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In C. Schmid, S. Soatto, & C. Tomasi (Eds.), *IEEE computer society conference on computer vision and pattern recognition* (pp. 886–893). San Diego, CA: IEEE Computer Society.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*, 945–978.

Einhäuser, W., Kruse, W., Hoffmann, K. P., & Konig, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research*, *46*, 1194–1209.

Farah, M. J., & Hammond, K. M. (1988). Mental rotation and orientation-invariant object recognition: Dissociable processes. *Cognition*, *29*, 29–46.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Science*, *10*, 382–390.

Fernandes, H. L., Stevenson, I. H., Phillips, A. N., Segraves, M. A., & Kording, K. P. (2014). Saliency and saccade encoding in the frontal eye field during natural scene search. *Cerebral Cortex*, *24*, 3232–3245.

Ghazanfar, A. A., Nielsen, K., & Logothetis, N. K. (2006). Eye movements of monkey observers viewing vocalizing conspecifics. *Cognition*, *101*, 515–529.

Glaser, J. I., Wood, D. K., Lawlor, P. N., Ramkumar, P., Kording, K. P., & Segraves, M. A. (2016). The role of expected reward in frontal eye field during natural scene search. *Journal of Neurophysiology*, *116*, 645–657.

Hays, A. V., Richmond, B. J., & Optican, L. M. (1982). A UNIX-based multiple process system for real-time data acquisition and control. In *WESCON conference proceedings* (pp. 2/1-1–2/1-10). Segundo, CA: Electronic Conventions.

Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, *9*(1):32, 1–8, doi:10.1167/9.1.32. [PubMed] [Article]

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin and Review*, *16*, 850–856.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral Brain Sciences*, *33*, 61–83.

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

Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G. A., & Wolfe, J. M. (2007). Tracking unique objects. *Perception and Psychophysics*, *69*, 172–184.

Jansen, L., Onat, S., & König, P. (2009). Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, *9*(1):29, 1–19, doi:10.1167/9.1.29. [PubMed] [Article]

Judge, S. J., Richmond, B. J., & Chu, F. C. (1980). Implantation of magnetic search coils for measurement of eye position: An improved method. *Vision Research*, *20*, 535–538.

Kaspar, K., & König, P. (2011). Overt attention and context factors: The impact of repeated presentations, image type, and individual motivation. *PLoS One*, *6*, e21719.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kümmerer, M., Theis, L., & Bethge, M. (2015). Deep Gaze I: Boosting rediction with feature maps trained on ImageNet. Available at https://arxiv.org/abs/1411.1045

McCormick, E. J., & Sanders, M. S. (1982). *Human factors in engineering and design* (5th ed.). New York: McGraw-Hill.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387–391.

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614–621.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8):20, 1–19, doi:10.1167/10.8.20. [PubMed] [Article]

Pannasch, S., Schulz, J., & Velichkovsky, B. M. (2011). On the control of visual fixation durations in free viewing of complex images. *Attention, Perception, and Psychophysics*, *73*, 1120–1132.

Pannasch, S., & Velichkovsky, B. M. (2009). Distractor effect and saccade amplitudes: Further evidence on different modes of processing in free exploration of visual images. *Visual Cognition*, *17*, 1109–1131.

Phillips, A. N., & Segraves, M. A. (2010). Predictive activity in macaque frontal eye field neurons during natural scene searching. *Journal of Neurophysiology*, *103*, 1238–1252.

Ramkumar, P., Fernandes, H., Kording, K., & Segraves, M. (2015). Modeling peripheral visual acuity enables discovery of gaze strategies at multiple time scales during natural scene search. *Journal of Vision*, *15*(3):19, 1–20, doi:10.1167/15.3.19. [PubMed] [Article]

Ramkumar, P., Lawlor, P. N., Glaser, J. I., Wood, D. K., Phillips, A. N., Segraves, M. A., & Kording, K. P. (2016). Feature-based attention and spatial selection in frontal eye fields during natural scene search. *Journal of Neurophysiology*, *116*, 1328–1343.

Robinson, D. A. (1963). A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Transactions on Bio-Medical Electronics, BME-10*, 137–145.

Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology*, *20*, 649–656.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1–17, doi:10.1167/7.14.4. [PubMed] [Article]

Tseng, P. H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7):4, 1–16, doi:10.1167/9.7.4. [PubMed] [Article]

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, *22*, 1641–1648.

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3):24, 1–15, doi:10.1167/9.3.24. [PubMed] [Article]

White, B. J., Berg, D. J., Kan, J. Y., Marino, R. A., Itti, L., & Munoz, D. P. (2017). Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nature Communications*, *8*, 14263.

Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and limits of models of fixation selection. *PLoS One*, *6*, e24038.

Wilming, N., Kietzmann, T. C., Jutras, M., Xue, C., Treue, S., Buffalo, E., & König, P. (2017). Differential contribution of low and high-level image content to eye movements in monkeys and humans. *Cerebral Cortex*, *27*, 279–293.