Modeling guidance and recognition in categorical search: Bridging human and computer object detection

Gregory J. Zelinsky Department of Psychology, Stony Brook University, Stony Brook, NY, USA; Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

Yifan Peng Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

Alexander C. Berg Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

Dimitris Samaras Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

Search is commonly described as a repeating cycle of guidance to target-like objects, followed by the recognition of these objects as targets or distractors. Are these indeed separate processes using different visual features? We addressed this question by comparing observer behavior to that of support vector machine (SVM) models trained on guidance and recognition tasks. Observers searched for a categorically defined teddy bear target in four-object arrays. Target-absent trials consisted of random category distractors rated in their visual similarity to teddy bears. Guidance, quantified as first-fixated objects during search, was strongest for targets, followed by target-similar, medium-similarity, and target-dissimilar distractors. False positive errors to first-fixated distractors also decreased with increasing dissimilarity to the target category. To model guidance, nine teddy bear detectors, using features ranging in biological plausibility, were trained on unblurred bears then tested on blurred versions of the same objects appearing in each search display. Guidance estimates were based on target probabilities obtained from these detectors. To model recognition, nine bear/nonbear classifiers, trained and tested on unblurred objects, were used to classify the object that would be fixated first (based on the detector estimates) as a teddy bear or a distractor. Patterns of categorical guidance and recognition accuracy were modeled almost perfectly by an HMAX model in combination with a color histogram feature. We conclude that guidance and recognition in the context of search are not separate processes mediated by different features, and that what the literature knows as guidance is really recognition performed on blurred objects viewed in the visual periphery.

Introduction

Object detection by humans

Object detection by humans, commonly referred to as visual search, is widely believed to consist of at least two processing stages: one that compares a representation of a target to a search scene for the purpose of guiding movements of focal attention and gaze and another that routes visual information at the region of space selected by focal attention to higher brain areas for the purpose of recognizing an object as a target or rejecting it as a distractor. This cycle of guidance and recognition operations then repeats until the target is detected or the display is exhaustively inspected. The present study focuses on the very first pass through this *guidance-recognition cycle*, asking whether this cycle of sequentially repeating guidance and recognition operations is still a useful and necessary way to conceptualize search.

Quite a lot is known about how search is guided to a visual pattern and the information that is used to make

Citation: Zelinsky, G. J., Peng, Y., Berg, A. C., & Samaras, D. (2013). Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *Journal of Vision*, *13*(3):30, 1–20, http://www.journalofvision.org/content/13/3/30, doi:10.1167/13.3.30.

doi: 10.1167/13.3.30 Received December 27, 2012; published October 8, 2013 ISSN 1534-7362

these guidance decisions (for reviews, see Eckstein, 2011; Wolfe, 1998; Zelinsky, 2008). Although individual models differ in their details, the modal model view of this guidance process holds that a target is first represented in terms of a small set of basic visual features, properties like color, orientation, and spatial scale. In the case of a previewed target, this representation is likely derived by weighting these basic features to reflect their presence in the target preview, and their expected presence in the search display (e.g., Chelazzi, Duncan, Miller, & Desimone, 1998; Wolfe, 1994). A similar process has been proposed for categorically defined targets via representations of weighted features residing in long-term memory (Alexander & Zelinsky, 2011; Elazary & Itti, 2010). This target representation is then preattentively compared to a search scene to obtain a map of evidence for the target at each image location, what has been referred to as a *target map* (Zelinsky, 2008). Barring the existence of any top-down contextual cues to the target's likely location (e.g., Brockmole & Henderson, 2006; Eckstein, Drescher, & Shimozaki, 2006; Neider & Zelinsky, 2006; Torralba, Oliva, Castelhano, & Henderson, 2006), this map is then used to prioritize the movements of covert attention and overt gaze to target-like patterns in a scene in a process known as *search guidance*. Even in the absence of useful knowledge of a target's appearance, search might still be guided to potentially important objects based on their salience (e.g., Itti & Koch, 2001; Koch & Ullman, 1985), a measure of local feature contrast in a scene. The features used in the creation of a saliency map are believed to be the same basic visual features underlying target guidance (Navalpakkam & Itti, 2005; Wolfe, 1994), although signals indicating a target are not likely to correspond to those indicating the most salient object (Zelinsky, Zhang, Yu, Chen, & Samaras, 2006).

Following the guidance of attention, and typically gaze, to the most likely target pattern in the search image, the guidance-recognition cycle suggests a transition to the task of determining whether this pattern is a target or a distractor. Important to the present discussion, this component of the search process is usually discussed in the context of object detection and recognition, not search. There is good reason why the search literature has sought to distance itself from the task of recognizing selected objects—object recognition is still an open problem, and theories of visual search would not be able to advance if they had to wait for this problem to be solved. The search literature has historically taken two paths around this object recognition obstacle. One path sets simple detection thresholds on specific feature dimensions and assumes a target is detected if evidence for these features exceeds these threshold values (for a review, see Palmer, Verghese, & Pavel, 2000). A second path avoids the problem altogether and treats the recognition task following guidance as a sort of "black box"; the object at each fixation of attention or gaze is assumed to be recognized, but the process of how this is accomplished is left unspecified (e.g., Itti & Koch, 2000). The first approach is best suited when objects are simple patterns that can be well defined in terms of one or two basic features (but see Zelinsky, 2008); the second approach is usually followed when the objects become more complex and their feature dimensions are unknown (but see Wolfe, 1994).

Freed from the presumed need for simple analyses imposed by preattentive processing, the object recognition literature has hypothesized the existence and use of a wide range of visual features. The broadest cut through this literature separates structural-description theories, those that assume the composition of objects from simpler three-dimensional features (e.g., Biederman, 1987; Biederman & Gerhardstein, 1993; Marr & Nishihara, 1978), from image-description theories, those that assume the extraction of two-dimensional features from information closer to the retinal mosaic (Bülthoff & Edelman, 1992; Poggio & Edelman, 1990; for a review, see Tarr & Bülthoff, 1998). Within the realm of image-based theories, the features that have been suggested for object recognition are often far more complex than the simple features believed to underlie search guidance. This is most evident in the case of face recognition, where it is common to assume features that code the metrical spatial relationships between the regions of a face (Sigala & Logothetis, 2002), or even large-scale features that code these facial configurations directly (Zhang & Cottrell, 2005). Neurophysiological and neurocomputational work in inferior temporal cortex also suggests that the features used for object recognition are considerably more complex than the simple color and edge-based features found in earlier visual areas (Gross, Bender, & Rocha-Miranda, 1969; for extended discussions, see Rolls & Deco, 2002; Tanaka, 1996), with the suggestion that domain-specific neurons code features for specific objects or object classes (Grill-Spector, 2009; Huth, Nishimoto, Vu, & Gallant, 2012; Perrett et al., 1984). It is perhaps not unfair to characterize the object recognition literature as converging on the assumed existence of features or feature weightings dedicated to the recognition of faces (Kanwisher, 2000), scenes (Epstein & Kanwisher, 1998), body parts (Downing, Jiang, Shuman, & Kanwisher, 2001), and various other categories of known (Gauthier, Skudlarski, Gore, & Anderson, 2000; Haxby et al., 2001; Xu, 2005) and novel (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999) objects, but with the specific information coded by these features still largely unknown.

Object detection by computers

The computer vision community doesn't distinguish between guidance and recognition processes as does the behavioral vision communities. Whereas it is important for behavioral vision researchers to describe the prioritization of patterns for object recognition, computer vision researchers dismiss the need for such prioritization, electing instead to simply pass a moving detection window over the entirety of an image. An obvious reason for this differential importance attached to the guidance process is anatomical in origin; humans have a single sensor, the fovea, from which we collect high resolution information from the center of our visual world. Although the eye movement system allows this sensor to be repositioned very quickly, the detection judgment following each eye movement requires an estimated 120 ms in the context of a simple search task (Becker, 2011), and far longer in the context of realistic scene viewing (Wang, Hwang, & Pomplun, 2009). Over the course of multiple fixations, as would be the case when searching visually complex scenes, these individual judgments might add up to significant delays in finding potentially important targets. Search guidance has presumably evolved to minimize these delays by intelligently allocating the fovea to patterns that are most likely to be the target. Computers, of course, are not bound by this limitation, and have at their disposal high resolution information from wherever the detection window is positioned.[1] Moreover, detection judgments by computers are extremely fast by comparison, making the need to prioritize these decisions unnecessary. Although the detection of fleeting targets in video still highlight the need for processing speed (Nascimento & Marques, 2006; Viola & Jones, 2001), computer vision approaches are generally more concerned with the accuracy of detection rather than the duration of each detection operation. Humans are therefore confronted with the need to prioritize detection decisions during search in a way that computers are not.

The behavioral vision and computer vision communities also differ in the types of features that they use to model object detection. Whereas models from a behavioral perspective usually restrict features to those with known biological plausibility, computer vision models are again unconstrained in this regard. The features proposed by computer vision researchers reflect a "whatever works" mentality, which is consistent with their goal of engineering robust systems for the accurate detection of objects from a diverse range of categories, irrespective of whether these systems resemble processes or mechanisms known to exist in behavioral vision systems. Indeed, detecting objects over an increasingly large number of categories has become a sort of contest in this community (Everingham, van Gool, Williams, Winn, & Zisserman, 2012), with each being an opportunity to highlight new features that outperform ones presented at previous challenges. These competitions have fueled enormous progress in the computer vision community over the past decade, with respectable detection performance now possible for hundreds of target categories (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010; Russakovsky, Lin, Yu, & Fei-Fei, 2012; van de Sande, Uijlings, Gevers, & Smeulders, 2011; Vedaldi, Gulshan, Varma, & Zisserman, 2009; Zhu, Chen, Yuille, & Freeman, 2010). A goal of this community is to extend this to tens of thousands of object classes (Deng, Berg, Li, & Fei-Fei, 2010), a number that starts to approximate human object recognition abilities.

Bridging human and computer object detection

The parallel efforts of behavioral and computer vision researchers to understand object detection begs the question of whether each community can help to inform and advance the goals of the other. The human visual system is capable of recognizing many thousands of object categories with high accuracy under often difficult conditions, and in this sense is a goal to which computer vision researchers aspire (absent some annoying limitations, such as a foveated retina). To the extent that the human object detection system can be understood, it may be possible to reverse engineer it to create automated systems with comparable abilities. The potential gains for understanding behavioral vision are as great. Although the human visual system provides an implementation proof that extremely robust object detection is possible, there is still much to learn about how this system actually works. By contrast, computer vision models, although still relatively poor in their object detection abilities compared to humans, are computationally explicit—each provides a sort of blueprint for how an object detection system might be built. Moreover, these systems have been demonstrated to work under naturalistic conditions, which cannot always be said for behaviorally based models. These are goals to which behavioral researchers aspire.

The present work brings together behavioral and computer vision perspectives to address the problem of categorical search—the detection of an object from a target class rather than a specific target from a preview. Although a great deal is known about how specific targets are represented and used to guide search (e.g., Wolfe, 1994; Zelinsky, 2008), the representations and comparison operations underlying categorical guidance are essentially unknown (but see Maxfield & Zelinsky, 2012; Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009). What *is* known is that the method for extracting

features must be different in these two cases; simple visual filters can be used to extract appearance-based features from a previewed target image, but in the case of categorical search there exists no preview from which to extract these features. For this reason the present work will take as a starting point the features, and the methods for learning these features, that have been demonstrated by the computer vision and computational neuroscience communities to enable the detection of realistic object categories. The adoption of this broad scope, one that includes features and methods from both communities, is central to the bridging goal of this study. Our guiding philosophy is that computer vision features, although not built with biological constraints in mind, have proven success in object class detection and should therefore be considered as viable descriptions for how categorical search might be accomplished by humans. Excluding these features on the basis of biological implausibility, particularly at this early stage in theoretical development, would presume a level of understanding of the brain's organization and functional architecture that does not yet exist. Each of these computer vision models, and the features from which they are built, is in a sense a testable hypothesis with the potential to refine and advance object detection theory, without which behavioral and biological research might flounder for years in a sea of data.

A basic question to ask when applying a new set of features to the task of categorical search is whether different features are needed to accomplish the two subtasks of guidance and recognition; are the features used to guide gaze to a categorically defined target the same as the features used to recognize that object as a target once it is fixated? As already noted, the visual search community has not vigorously engaged this question, and in fact has seemed content with the assumption that search guidance and object recognition use different features that are tailored to the specific demands of the different tasks. There is even good reason to suspect why this might be true. By definition, the features used to guide gaze to an object must work on a blurred view of that object as it would be perceived in peripheral vision. The features used for recognition, however, would be expected to exploit high-resolution information about an object when it is available. Color is another example of a potentially asymmetric use of object information. Search guidance has long been known to use color (Rutishauser & Koch, 2007; Williams, 1967; Williams & Reingold, 2001; see also Hwang, Higgins, & Pomplun, 2007, for guidance by color in realistic scenes), presumably because of its relative immunity to the effects of blurring. Object recognition, however, places less importance on the role of color (Biederman & Ju, 1988), with many behavioral and computer vision models of recognition ignoring color altogether (e.g., Edelman & Duvdevani-Bar, 1997; Hummel & Biederman, 1992; Lowe, 2004; Riesenhuber & Poggio, 1999; Torralba, Murphy, Freeman, & Rubin, 2003; Ullman, Vidal-Naquet, & Sali, 2002).

To address the question of whether search guidance and recognition might use the same features, we trained classifiers to find, and then recognize, a target from a real-world object category—teddy bears. Object detectors were trained on unblurred images of teddy bears and random objects, then tested on blurred images of new objects. These are the conditions that exist at the time of guidance, with the detector responses approximating the signal used to guide gaze to an object. Bear/nonbear classifiers were trained on the same unblurred images, but were tested on unblurred images as well. These are the conditions that exist at the time of recognition, after gaze has landed on an object and the task is to determine whether it is a target. Both the object detectors and classifiers were therefore trained on the same images and tested on the same images (different training and testing sets) with the only difference being whether the test images were blurred or not. We then compared model performance to that of humans performing the identical categorical search for teddy bears. To the extent that a single model can predict human performance in both search subtasks, an argument can be made for treating search guidance and recognition as, essentially, a single process. However, finding that these two subtasks require different models would be an argument for treating the guidance and recognition components of search as separate processes that use different features.

We had four specific goals in this study. First, we wanted to see how well a range of SVM (support vector machine) models can predict search guidance to targets. Can a model trained on unblurred teddy bears and random objects predict search guidance to blurred bears? Second, once search is guided to an object, how well might these same models predict human recognition of that object? Can a model trained on unblurred objects recognize other unblurred objects as teddy bears or distractors at human levels? Third, and assuming positive answers to our first two questions, can search guidance and recognition be described using the same set of visual features? This primary goal of our study will inform whether these two search components should continue to be treated as separate processes or whether guidance might be more usefully conceptualized as a preliminary form of recognition performed on blurred objects. Lastly, what sorts of models might these be? Will describing search guidance and recognition require complex features from the computer vision community with questionable biological plausibility, or can one or both of the search subtasks be accomplished using the relatively simple features that are known to exist in early visual cortical areas? One prediction might be that models adopting biologically plausible features

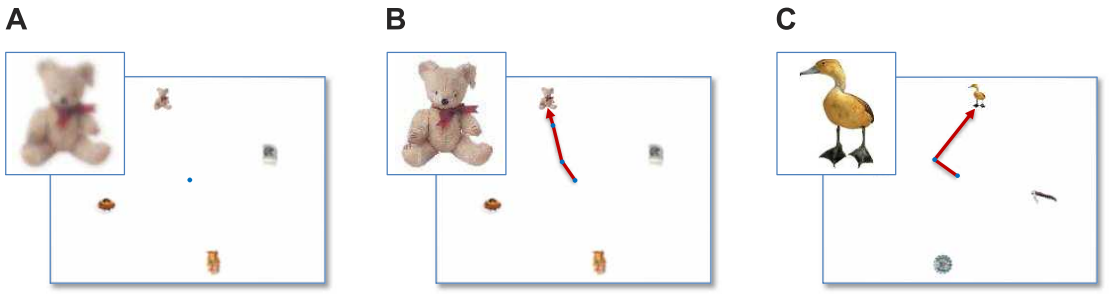

Figure 1. Representative search displays used in the behavioral experiment, illustrating the relationship between object eccentricity and retinal acuity. (A) Target present display, with the teddy bear target shown enlarged at inset. Note that all objects would be perceived as blurred when viewed from a central starting fixation position (blue dot). (B) The same target in the same search display viewed after its fixation; the red arrows and blue dots show representative eye movements and fixations during search. (C) A target-absent trial in which the first fixated object was a high-similarity bearlike distractor, again with representative search behavior.

would best describe human search behavior because these models capture more closely the way that humans represent categorical stimuli. Alternatively, given that human object detection behavior is still far superior to that of existing models, it may be that the best performing model, regardless of its biological plausibility, would also best predict human performance.

Methods

Figure 1 illustrates a core problem facing people as they search. Suppose the task is to determine whether a teddy bear target is present in this object array. Assuming that gaze is located at the center (Figure 1A, blue dot), note that all of the objects are slightly blurred due to retinal acuity limitations. This blurring necessarily reduces one's confidence that a member of the teddy bear class is present. To offset the impact of retinal blur and to boost the confidence of search decisions, people therefore make eye movements to suspected targets. Upon fixating the bear (Figure 1B) this object would no longer be blurred, allowing its confident recognition. Much of the efficiency of search behavior can be attributed to the fact that visual similarity relationships to the target are used to guide these eye movements (Alexander & Zelinsky, 2011, 2012; Duncan & Humphreys, 1989). This is clearest when a target actually appears in the search display, but holds even when the search objects are all nontargets. Presumably, the duck in Figure 1C was fixated first because it was more visually similar to a teddy bear than the other distractors. We conducted an experiment to explore the effects of these similarity relationships on search during the very first pass through the guidance-recognition cycle, taking into account the visual information available to guidance (blurred images) and recognition (unblurred images) and quantifying both behaviors in terms of computationally explicit models.

Behavioral methods

Targets were images of teddy bears, adapted from Cockrill (2001, as described in Yang & Zelinsky, 2009); nontargets were images of objects from a broad range of categories (Hemera Photo-objects collection). Target and nontarget images were single segmented objects displayed in color against a white background. Nontargets were further rated for visual similarity to the teddy bear category using similarity estimates obtained from a previous study (Alexander & Zelinsky, 2011). In that study, each of 142 participants were shown groups of five nonbear objects (500 total) and asked to rank order them based on their visual similarity to teddy bears. From these 71,000 similarity estimates, nontarget objects were divided into *high-*, *medium-*, and *low-similarity* groups relative to the teddy bear target class.

From these groups of 198 bears and 500 similarity-rated nonbears (Figure 2) we constructed three types of search displays, each of which depicted four objects that were normalized to subtend ~2.8° of visual angle and were equally spaced on an imaginary circle surrounding central fixation (8.9° radius). *Target-present* displays (TP) depicted a bear target with three random distractors unrated in their similarity to bears (Figures 1A, B). The two other display types were target-absent (TA) conditions differing in their composition of similarity-rated nontargets. *High-medium* displays (TA-HM) depicted one high-similarity bearlike distractor and three medium-similarity distractors, and *high-medium-low* (TA-HML, Figure 1C) displays depicted one low-similarity nonbearlike distractor, one high-similarity bearlike distractor, and two medium-similarity distractors.

Participants were eight experimentally naïve students from Stony Brook University's human subject pool, all of whom reported normal or corrected to normal visual acuity. Their task was categorical search, meaning that the teddy bear target category

Figure 2. Representative objects used as search stimuli in the behavioral and computational experiments. (A) Teddy bear targets (5 of 44). (B) High-similarity bearlike distractors (5 of 44). (C) Low-similarity nonbearlike distractors (5 of 40).

was designated by instruction; no specific target preview was shown prior to each search display. None of the target or nontarget objects appearing in the search displays repeated throughout the experiment, minimizing the potential for exemplar-based search (Yang & Zelinsky, 2009). Observers were stabilized using a chin and head rest, and eye position was sampled at 500 Hz using an Eyelink II eye tracker (SR Research) with default saccade detection settings. The experiment began with a nine-point calibration routine needed to map eye position to screen coordinates. Calibrations were not accepted until the average error was less than 0.4° and the maximum error was less than 0.9°. Each trial began with the observer fixating the center of a flat-screen computer monitor (ViewSonic P95f+ running at 100 Hz) and pressing a button. A search display then appeared and the task was to make a speeded target present or absent judgment while maintaining accuracy. Judgments were registered by pressing either the left or right shoulder triggers of a GamePad controller. Display type was varied within subject and interleaved over trials. There were 44 practice trials followed by 132 experimental trials, with the experiment lasting approximately 50 minutes.

Computational methods

The computational methods consisted of training nine SVM-based models to distinguish images of teddy bears from random objects, then testing them on a trial-by-trial basis using the identical objects that participants viewed in each of their search displays. All of the models used standard features that already existed in the behavioral and computer vision literatures; it was not our goal to propose new features, but rather to evaluate the potential for existing features to explain guidance and recognition behavior during search. We selected features ranging in biological plausibility and descriptor complexity. The following provides a brief description of each of the features explored in this study.

SIFT-BOW

One of the most common descriptors used for object detection in computer vision is the "scale invariant feature transform" or SIFT feature. Introduced by Lowe (2004), it represents the structure of gradients in a local image patch using 16 spatially distributed histograms of scaled and normalized oriented edge energy. Following the bag-of-words procedure (Csurka, Dance, Fan, Willamowski, & Bray, 2004), we used vector quantized SIFT descriptors, computed in a uniform grid over bounding boxes, to create a vocabulary of 200 visual words. We will refer to models using this 200-dimensional feature as "SIFT with bag-of-words," or SIFT-BOW for short.

SIFT-SPM

A second model was also based on the SIFT feature, but replaced the single spatial histogram that was computed over bounding boxes with a multiscale spatial histogramming technique (Lazebnik, Schmidt, & Ponce, 2006). This spatial pyramid aggregates vector-quantized SIFT features into histograms over increasingly fine subregions. These histograms are then concatenated to form the feature descriptor. Our implementation used a two-layer spatial pyramid—with one spatial histogram at the top layer and four at the lower layer—giving us a 1,000-dimensional (5×200) model that we will refer to as "SIFT with spatial pyramid matching," or SIFT-SPM for short.

V1 Feature

Moving towards more biologically inspired approaches, we implemented the V1 feature model introduced by Pinto, Cox, and DiCarlo (2008). This model applies a bank of Gabor filters to an image in order to approximate simple cell responses in primary visual cortex, but uses different sized kernels and local normalization to make the technique tolerant to variation in contrast and size. Principle components analysis is used to reduce the dimensionality of this feature from 86,400 to 635 prior to use in classification.

HMAX

We also implemented a four-layer HMAX model designed to capture the initial feed-forward visual processing known to be performed by simple and complex cells in primary visual cortex (Serre, Wolf, & Poggio, 2005). In this basic version of the model, the responses of simple cells (S1), again approximated by a bank of Gabor filters applied to an image, are pooled by complex cells (C1) using a local maximum operation, allowing limited invariance to changes in position and scale. Prototype patches are sampled from the C1 responses in training images. The maximum response in a window for each C1 prototype forms the C2 feature for that window. In our implementation we used a bank of Gabor filters with 16 scales and eight orientations, and extracted 1,000 C1 patches from positive training samples for use as prototypes for classification.

Color

The SIFT, V1, and HMAX features do not represent color information, but color is known to be an important feature for guiding search (Motter & Belky, 1998; Rutishauser & Koch, 2007; Williams & Reingold, 2001). We therefore implemented a simple color histogram feature (Swain & Ballard, 1991). This was defined in the DKL color space (Derrington, Krauskopf, & Lennie, 1984), which approximates the sensitivity of short-, medium-, and long-wavelength cone receptors, and captures the luminance and color opponent properties of double opponent cells. The color histogram feature used 10 evenly spaced bins over three channels, luminance, red-green, and blue-yellow, each normalized to the [0, 1] range. The procedure for computing this feature first required converting images from RGB space to DKL space using the conversion described in Hwang, Higgins, and Pomplun (2009). Next, we built an image pyramid using three scales per image, and from each layer we sampled 24 × 24 pixel image patches where each patch was separated by 12 pixels. A color histogram was computed for each sampled patch. We then randomly selected from positive training samples 250 patches across the three pyramid layers and used these as prototypes. The maximum response to each prototype over a window was used as the color feature for that window, producing a C2-like feature for color similar to the Gabor-based features used by the HMAX model.

+ COLOR

As part of a limited evaluation of how combinations of features might affect predictions of search guidance and recognition, we also concatenated our 250-dimensional color histogram model (COLOR) with the above described SIFT-BOW, SIFT-SPM, V1, and HMAX models, creating color versions of each and giving us the nine models considered in this study. These combined models will be referred to as: SIFT-BOW+COLOR, SIFT-SPM+COLOR, V1+COLOR, and HMAX+COLOR.

For each object detector model, we trained a linear SVM classifier (Chang & Lin, 2001) to discriminate teddy bears from nonbears. All nine classifiers were trained using the same set of positive samples, 136 images of teddy bears, and negative samples, 500 images of nonbears randomly selected from the Hemera object collection. Following standard practice (see Perronnin, Akata, Harchaoui, & Schmid, 2012; Zhu, Vondrick, Ramanan, & Fowlkes, 2012), we used a per-class weighting factor to increase the weight for each positive training sample by a factor of sqrt(# negatives / # positives). Testing occurred on a trial-by-trial and object-by-object basis. Training and testing images were completely disjoint sets.

Critically, test images were either blurred or not depending on whether predictions were obtained for the guidance or recognition components of the search task. To model guidance, each image of a test object was blurred to reflect the retinal acuity limitations that existed for the human observer viewing that object at 8.9° eccentricity and in the position that it appeared in the search display (as blurring would change slightly depending on radial position). Blurring was accomplished using the target acquisition model (TAM; Zelinsky, 2008), and that work should be consulted for additional details. For every trial from the behavioral experiment we applied a sliding window detector, one for each of the nine models, over the four blurred objects from the corresponding search display, and obtained the maximum detector response for each blurred object. These responses were then converted to probabilities based on their distances from the linear SVM classification boundary using a method of probability estimation (Platt, 2000).[2] Search guidance was estimated directly from these probabilities, with our prediction of the object that should be fixated first on a given search trial being the one having the highest teddy bear probability among the four. To model the recognition decision following an eye movement to the first fixated object we then classified an unblurred version of that object as either a teddy bear or a nonbear. Our modeling of guidance and recognition therefore approximated, for each search trial, the visual conditions existing during the first pass through the guidance-recognition cycle; the models not only "saw" the same objects as the behavioral participants, but these objects were either blurred or not to reflect the differing information available to the guidance and recognition decisions.

	Display type		
	TP	TA-HML	TA-HM
Reaction time (ms)	650 (33.2)	793 (54.3)	857 (64.8)
Errors (%)	4.8 (1.6)	3.4 (1.3)	4.3 (1.0)
Trials with no fixated object (%)	0.6 (0.6)	4.6 (2.5)	4.5 (2.5)

Table 1. Summary search performance by display type. *Note:* Values in parentheses indicate one standard error of the mean.

Results and discussion

Manual button press search performance is summarized in Table 1. Response times differed by display type, $F(2, 21) = 4.98$, $p < 0.05$. Search judgments were faster in the target-present condition compared to the two target-absent conditions, $t(7) \geq 3.55$, $p \leq 0.009$ (paired-group)[3], and there was a reliable advantage in TA-HML trials over TA-HM trials, $t(7) = 3.02$, $p < 0.05$. Also reported in Table 1 are the relatively rare cases in which observers failed to look at any object during search. Although these no-fixation trials did not differ by condition, $F(2, 21) = 1.21$, $p > 0.1$, there was a trend toward more of these in the target-absent data, probably due to observers desiring to make a confirmatory eye movement to the target when it was present. An analysis of button press errors revealed no differences between conditions, $F(2, 21) = 0.29$, $p > 0.1$. Our failure to find the typical pattern of increased misses relative to false positives is likely due to the fact that an explicitly target-similar item was embedded in each target-absent display. A more detailed analysis of errors by object type will be reported in a following section.

Search guidance

Search guidance was quantified as the probability that an object was first fixated during search, what we refer to as an *immediate fixation*. This oculomotor measure of guidance is more conservative than the time taken to first fixate an object (Alexander & Zelinsky, 2011), as this latter time-to-target measure allows for fixations on distractors that may conflate guidance with distractor rejection processes.

Figure 3A shows the probabilities of immediate fixations on targets and similarity-rated nontargets for each of the three display conditions (TP, TA-HML, TA-HM) and each of the object types within these conditions (targets and high-similarity, medium-similarity, and low-similarity distractors). Note that these probabilities were adjusted to correct for multiple instances of medium-similarity objects appearing in a display. This adjustment allows for a meaningful chance baseline of 0.25 against which all conditions can be compared, although it resulted in probabilities within a display condition not summing to one. Turning first to the target-present data (leftmost bar), the probability of immediately fixating a teddy bear target was ~ 0.8, far above chance. This replicates previous work showing strong search guidance to categorically defined targets (Alexander & Zelinsky, 2011; Maxfield & Zelinsky, 2012; Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009).

More interesting are data from the target-absent conditions, where we found a similar pattern for the high-similarity bearlike distractors. In the TA-HM condition (rightmost two bars) the percentage of immediate fixations on bearlike distractors was again well above both chance and the immediate fixation rate for medium-similarity distractors, $t(7) \geq 11.53$, $p \leq 0.001$, but significantly less than immediate fixations on the actual targets, $t(7) = 4.75$, $p = 0.002$. In the TA-HML condition we again found strong guidance to high-similarity distractors, which was also well above chance and stronger than guidance to either the medium or low-similarity objects, $t(7) \geq 13.59$, $p < 0.001$. Most notably, this level of guidance did not significantly differ from guidance to the actual targets, $t(7) = 1.93$, $p = 0.096$. What makes this finding remarkable is the fact that these objects were not teddy bears, but rather things like pumpkins and backpacks and pieces of furniture. Not only is this additional and converging evidence for categorical guidance, but it demonstrates that this guidance is very sensitive to the arguably quite subtle visual similarity relationships between nontarget categories of objects and a target class (see also Alexander & Zelinsky, 2011), at least for the teddy bear targets used in this study. As for whether search was guided *away* from objects that did not look like bears, we found below chance rates of immediate fixations on both medium-similarity and low-similarity distractors, $t(7) \geq 8.85$, $p < 0.001$, but only marginally weaker guidance to the low-similarity distractors compared to the medium-similarity distractors, $t(7) = 2.26$, $p = 0.06$. However, there was an approximately 10% increase in immediate looks to the high-similarity distractor in the TA-HML condition compared to the TA-HM condition, $t(7) = 11.25$, $p < 0.001$. This result is consistent with the difference observed in manual reaction times and suggests that including a low-similarity distractor in the display did affect search

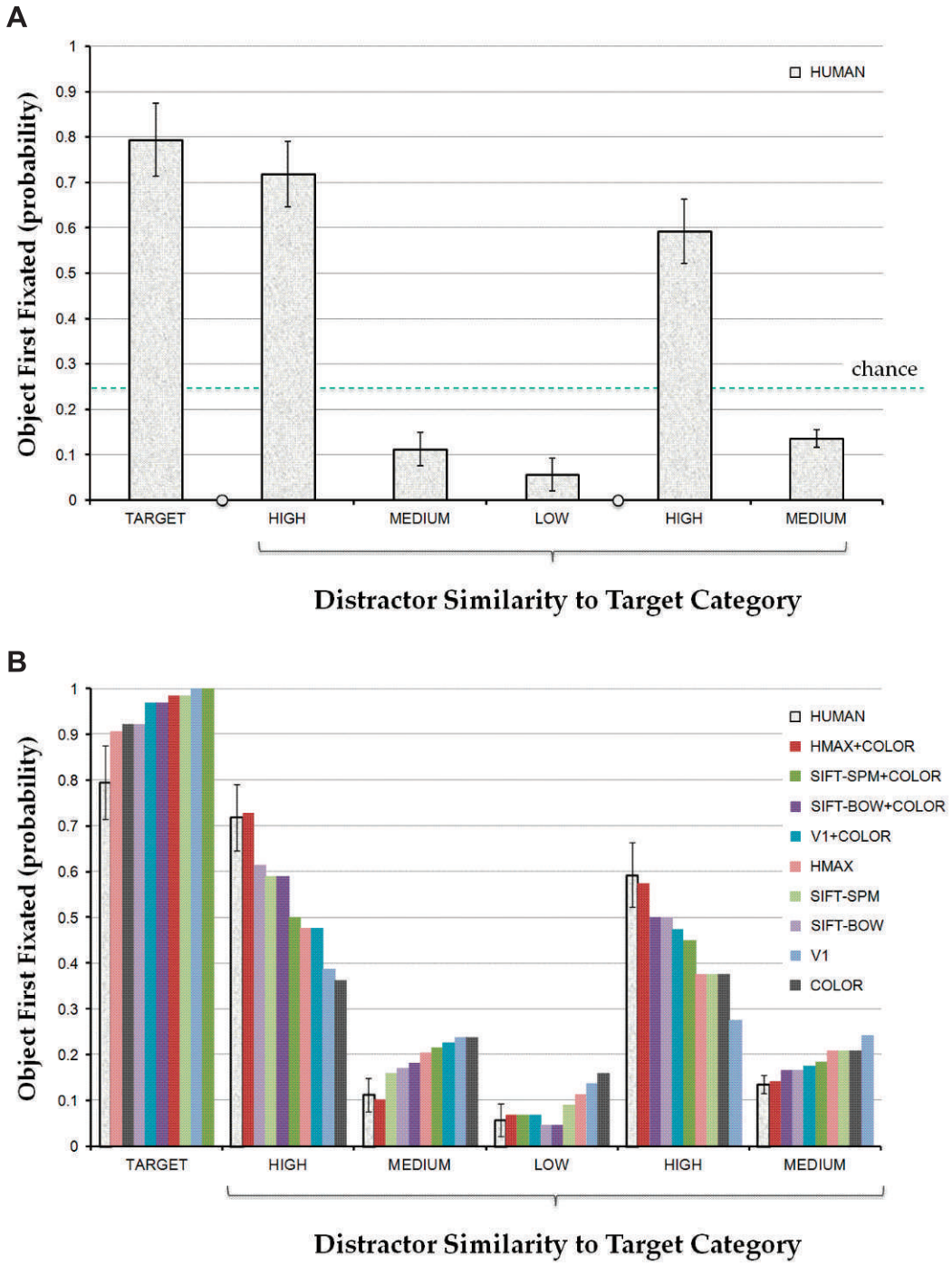

Figure 3. Probabilities of immediate fixations on objects, grouped by target present (leftmost bar), TA-HML (next three bars), and TA-HM (rightmost two bars) display conditions and object type (TARGET, HIGH-similarity, MEDIUM-similarity, and LOW-similarity). (A) Behavioral data. (B) Behavioral data and data from the nine models tested, with model performance plotted in order of decreasing match to the behavioral data. The error bars attached to the behavioral means show 95% confidence intervals. The dashed line in (A) indicates chance. See text for additional details.

behavior, probably as a result of these distractors competing less than medium-similarity distractors for the attraction of gaze.

Figure 3B replots the behavioral data with the corresponding data for the nine models tested, each color coded and plotted in order of decreasing match to the human behavior. This ordering of the data means that the best matching model is indicated by the leftmost colored bar for each object type. As in the case of the behavioral data, each colored bar indicates the probability that a given object detector would have selected a particular type of object for immediate

fixation. This selection was made on a trial-by-trial basis, as described in the computational methods section; the likelihood of an object being a bear was obtained for each of the four objects in a search display, with our prediction of the first-fixated object on that trial being the one with the highest likelihood estimate. This again was done on blurred versions of each object, so as to approximate the visual conditions existing at the time of search guidance.

The leftmost group of bars shows the probabilities that an immediate fixation to the teddy bear target would have occurred on target present trials. What is clear from these data is that all of the models overestimated the ~0.8 behavioral probability of first fixating a teddy bear, forming a cluster in the [0.9, 1.0] range. This high level of guidance was expected, and indicates that the models were generally successful in capturing the fact that teddy bears, more so than nonbear objects, tend to look like other teddy bears. However, this resulted in some of the worst performing models, those that gave a relatively low teddy bear likelihood estimate to an actual teddy bear, best matching human behavior—although it should be noted that all of the models fell outside the 95% confidence interval surrounding the behavioral mean. But interpreting this disagreement between model and behavioral guidance is complicated by the very real possibility of a ceiling effect in the behavioral guidance measure. Unlike computers, humans are subject to motivational lapses, momentary distractions, and a host of idiosyncratic biases that combine to create a ceiling on how strong target-related guidance might become.

More informative are the target-absent data, where guidance was well below ceiling. There are two noteworthy patterns to extract from the morass of bars. First, each of the nine models captured the general effect of target-distractor similarity appearing in the behavioral categorical guidance data—immediate fixations were most likely for high-similarity distractors, followed by medium-similarity and then low-similarity distractors. This encouraging level of agreement, although restricted to only a single target class, gives reason for optimism in using features and methods from computer vision to capture the perceptual confusions underlying behavioral guidance differences during categorical search. Second, some of these models did much better than others. Out of the nine models tested, the one that best predicted behavioral guidance by far was the HMAX+COLOR model, as indicated by the same red color appearing next to the behavioral data in each of the target-absent conditions. Not only did this model capture the pattern of target-distractor similarity effects on categorical guidance, it also captured the magnitude of these effects. Guidance to each of the distractor types was within the 95% confidence interval surrounding the behavioral mean, and it was the only model of the nine for which this happened consistently. Also interesting is the role that color played in these predictions. Whereas performance of the HMAX and COLOR models alone was mediocre, with each appearing in the middle or end of each group, combining these two models dramatically improved predicted guidance. This impressive level of agreement is surprising given that all guidance estimates were based on the raw probabilities outputted by the models—no parameters were used to fit the behavioral data. The base rates of the models could therefore have been anywhere—they just happened to align almost perfectly with the behavioral guidance data in the case of the HMAX+COLOR model.

Recognition of the first-fixated objects

How successful were participants in deciding whether the previously blurred object used to guide their search was a target or a distractor? Figure 4A plots the percentages of behavioral recognition errors by object type and display condition, where recognition is defined in terms of the oculomotor response to the first fixated object. The leftmost bar indicates cases in which observers incorrectly made a target-absent response while looking at the target or at any point after gaze left the target—suggesting that the target was not recognized as a member of the teddy bear category. Figure 5A shows some teddy bear exemplars for which such misses occurred. By this measure, cases in which an observer would make a target-present response after leaving the target and fixating some other object might also be scored as target recognition errors, although this never happened. The bars to the right show cases in which the first-fixated object was mistakenly recognized as a teddy bear, as indicated by the observer making a target-present response while looking at that distractor or within 500 ms after gaze left the distractor. Figure 5C shows some distractors for which this occurred. Cases in which a target-present response was made after shifting gaze away from the first-fixated object (for longer than 500 ms) were not scored as errors under this measure, but this happened on only one target-absent trial over all participants.

A clear message from these data is that observers were quite good at recognizing the object first fixated during search, with very few false negative (3.02%) or false positive (3.91%, averaged over conditions) errors. However, a closer look at the false positives revealed an interesting trend. There were more errors to the high-similarity bearlike distractors compared to the medium-similarity distractors. This was true for both the TA-HML and TA-HM conditions, $t(7) \geq 2.44$, $p < 0.05$. There were also more errors to medium-similarity

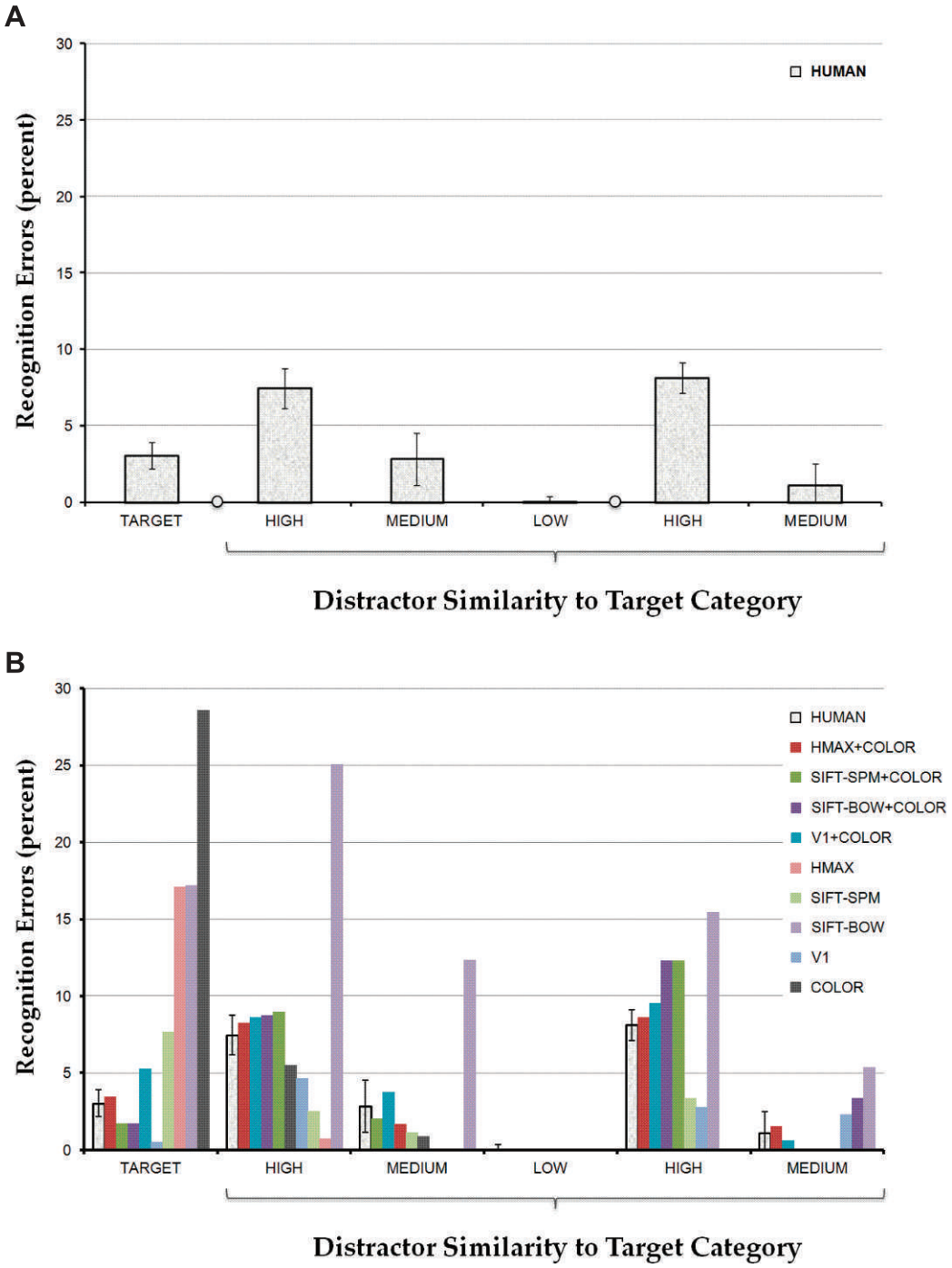

Figure 4. Percentages of recognition errors made to the first-fixated objects, grouped by target present (leftmost bar), TA-HML (next three bars), and TA-HM (rightmost two bars) display conditions and object type (TARGET, HIGH-similarity, MEDIUM-similarity, and LOW-similarity). (A) Behavioral data. (B) Behavioral data and data from the nine models tested, with model performance plotted in order of decreasing match to the behavioral data. The error bars attached to the behavioral means show 95% confidence intervals. See text for additional details.

distractors compared to low-similarity distractors, $t(7) = 2.83$, $p < 0.05$, where accuracy was essentially perfect. These patterns suggest that the same target-distractor similarity relationships used to guide search to bearlike distractors might also increase the probability of recognizing these objects as actual bears.

Figure 4B replots the behavioral recognition errors with the error rates from the nine models, each again color coded and plotted in order of decreasing match to the behavioral data. There are two patterns of note. First, and unlike the categorical guidance data, there were extreme differences in recognition performance

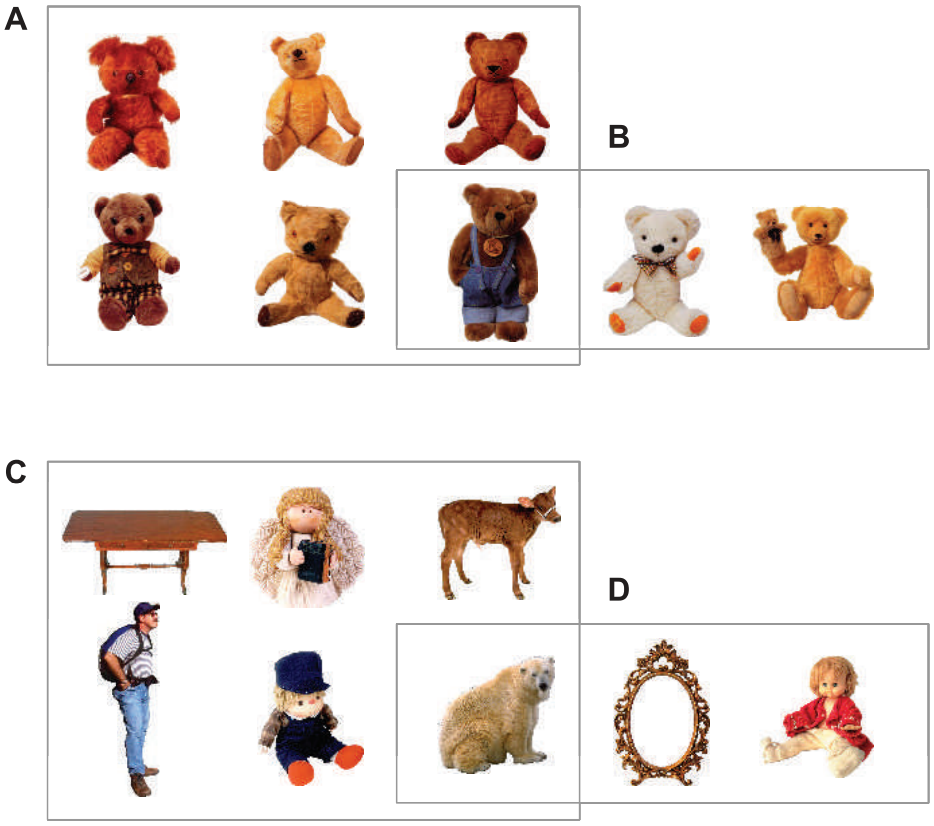

Figure 5. Examples of objects resulting in a recognition error, where the error was defined in terms of the oculomotor response to the object (see text for details). (A) Objects not recognized as a teddy bear by at least one observer. (B) Objects not recognized as a teddy bears by the HMAX+COLOR model. Note that the teddy bear in blue jeans was missed by both the model and an observer. (C) Objects mistakenly recognized as a teddy bear by at least one observer. (D) Objects mistakenly recognized as a teddy bear by the HMAX+COLOR model. Note that the polar bear elicited a false positive response by the model and two observers.

among the models tested. This was particularly true in the case of the target-present data, where the miss rates for some models (COLOR, SIFT-BOW, and HMAX) indicate great difficulty in recognizing exemplars from the teddy bear category. For other models recognition accuracy was quite high, on par or even better than that from human observers. Several models even captured the relatively subtle relationship between recognition accuracy and target-distractor similarity reported in the behavioral data; false positives rates were highest for high-similar distractors and lowest for low-similarity distractors. Second, of the nine models that we tested in this study the HMAX+COLOR model again gave the best overall match to human behavior. This was true for both the false negative and false positive errors, with the model's performance again falling within the 95% confidence intervals surrounding all of the behavioral means. Moreover, this exceptional agreement was not due to a floor effect, as this level of agreement extended to the high-similarity distractors that were mistakenly recognized as a target at a rate clearly above floor. Figure 5B shows examples of teddy bears that were not recognized by the HMAX+COLOR model, and Figure 5D shows distractors that this model mistakenly recognized as teddy bears.

Also interesting is the role that color played in recognition, which can be seen by looking at only the four best matching models in Figure 4B. Except for the SIFT-SPM model, which was the fourth best matching model for medium-similarity distractors in the TA-HML condition, all of the other models included a color histogram feature. Although color was expected to be helpful in modeling guidance, this suggests that it helped in modeling recognition as well.

The previous analyses used unblurred objects for both training and testing, under the assumption that foveal views of objects, obtained after these objects had been fixated, would best serve recognition. However, it may be that people learn representations of blurred objects so as to mediate their recognition in the visual periphery, given that this is how objects are typically first seen. To test this possibility we retrained each of the models on blurred versions of the objects used previously for training, then tested these models on the same blurred objects that were used to evaluate search guidance. The training conditions (blurred) therefore

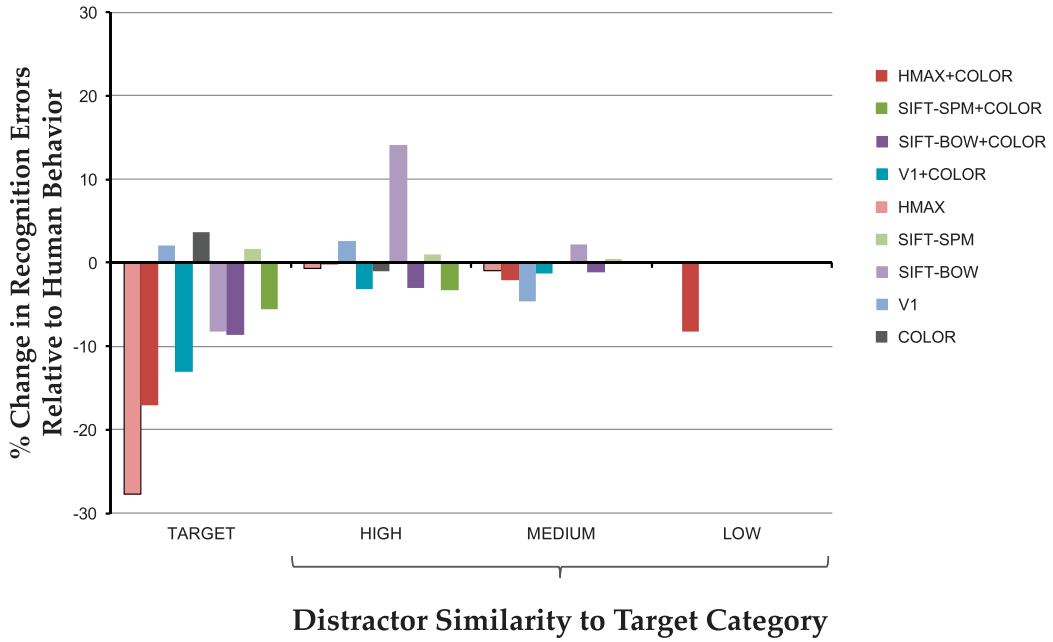

Figure 6. Percent change in fit to the behavioral recognition error rates as a result of adding blur to the sets of training and testing images for each of the nine models tested. Positive values indicate better fits to behavioral recognition following the addition of blur, negative values indicate worse fits. Data are grouped by TARGET, HIGH-similarity, MEDIUM-similarity, and LOW-similarity object types.

matched the testing conditions (also blurred). The results from this analysis are shown in Figure 6, plotted as the percent change in fit to the behavioral recognition rates as a result of adding blur. Positive values indicate better fits to behavioral recognition; negative values indicate worse fits. Turning first to the target-absent data we found very little effect of blur on the false positive error rates; nontargets were recognized as such regardless of whether they were blurred or not.[4] But the target-present data revealed large costs; the recognition of blurred bears was much poorer than the recognition of unblurred bears, despite the models being trained and tested on blurred objects. Moreover, this was particularly true for some of the more biologically plausible models, including HMAX+COLOR. To the extent that blurred target representations are learned and compared to patterns viewed in the visual periphery, this finding suggests that this variety of recognition would be less accurate than recognition performed centrally. More generally, we interpret this as support for the suggestion that eye movements are made during search for the purpose of offsetting retinal acuity limitations and making better and more confident target decisions (Zelinsky, 2008, 2012). Given the high (and presumably unacceptable) probability of failing to recognize a teddy bear seen peripherally (Figure 6), people choose to improve the probability of recognition success (Figure 4B) by shifting their gaze to that object and obtaining a foveal view.

General discussion

In the present study we asked whether search guidance and object recognition are meaningfully different, or whether they should be treated as essentially the same process (see Eckstein, Beutter, Pham, Shimozaki, & Stone, 2007, for a related question). The search literature has historically separated these behaviors, treating recognition as a sort of black box in a repeating guidance-recognition cycle. Fueling this cycle is what might be called an "it can't be so simple" fallacy. There is good reason to believe that relatively simple visual features are used to guide search (Wolfe, 1998; Wolfe & Horowitz, 2004), but object recognition is a *really* hard problem and requires features that can't possibly be this simple, leading to the conclusion that guidance and recognition must be qualitatively different things. Under this view, recognition is therefore an integral part of the search *task*, but it is distinct from the actual search *process* that underlies guidance.

To explicitly evaluate the assumption that search guidance and recognition are separate processes we explored two classes of models, one approximating the conditions existing during guidance and the other approximating the conditions existing during recognition. The guidance models were given sets of four blurred objects, each corresponding to the objects in a search display, and predicted the object that would be fixated first. The recognition models were given

unblurred versions of these same first-fixated objects, and classified these objects as teddy bear targets or nonbear distractors. We also manipulated the visual similarity between these objects and the target class, as well as the types of features and/or methods that were used by the models. In the context of this categorical bear search task, we found that guidance and recognition could be well described by several relatively simple computational models, all without the use of any explicit fit parameters. However, of the nine models that we tested the one that best predicted both categorical guidance and recognition was not an implausibly complex model from computer vision, but rather a basic version of an HMAX model, one that included a color feature. This HMAX+COLOR model not only captured the finding that gaze was guided to nontarget objects in proportion to their similarity to the target class, it also captured the magnitude of these guidance effects for each of the similarity conditions that we tested. This same model also captured the behavioral false negative and false positive rates, as well as the effect of target-distractor similarity on these recognition errors. In summary, under conditions that closely approximate the information available to observers, namely whether the objects were blurred or not depending on viewing from pre- or post-fixation, we found that the HMAX+COLOR model was able to predict behavioral guidance and recognition during a categorical teddy bear search with impressive accuracy.

The fact that a single model could be trained to predict search guidance and recognition is informative, and suggests that the historical belief that these search components reflect different processes may have been misguided. Although currently limited to just a single target category, our data demonstrates that the recognition of a visually complex class of objects can be accomplished at human levels using the same simple visual features believed to underlie search guidance. In this limited teddy bear context, the "it can't be that simple" fallacy is exactly that—a fallacy; the same probability estimates used to predict eye movements to blurred objects could also be used to classify unblurred versions of these objects as targets or distractors. These commonalities raise the intriguing possibility that search guidance and recognition may not only be more similar than what has been believed in the literature, but that they may be one and the same process. Under this view, target guidance can be conceptualized as object recognition performed on blurred patterns viewed in the visual periphery. Perhaps the guidance-recognition cycle is not so much a cycle as it is a target recognition decision that is distributed across a shift of attention or gaze.

The suggestion that guidance and recognition are essentially the same process, if true, would have consequences for search theory. Perhaps the biggest of these is its implication for the age-old early versus late selection debate (see Pashler, 1998). If search guidance is preattentive, which is a truism, and recognition and guidance are the same thing, which is our claim, then it follows that recognition should also be preattentive—the definition of late selection. Why then is search not guided based on semantic information, as would be expected if all of the objects in a search display were preattentively recognized and processed to the level of semantic meaning?

One obvious answer to this question is that perhaps semantic guidance *is* possible. There are now several lines of evidence arguing that scenes are preattentively analyzed into meaningful semantic objects. One of these is based on the preferential guidance of gaze to objects that are semantically (Bonitz & Gordon, 2008; Loftus & Mackworth, 1978; Underwood & Foulsham, 2006; Underwood, Templeman, Lamming, & Foulsham, 2008) or syntactically (Becker, Pashler, & Lubin, 2007) inconsistent with a scene's context. This suggests, not only the preattentive recognition of the objects in a scene, but the parallel evaluation of these objects with respect to scene consistency for the purpose of detecting violations and generating a guidance signal (but see Henderson, Weeks, & Hollingworth, 1999; Võ & Henderson, 2009, 2011, for counterarguments). More recently, Hwang, Wang, and Pomplun (2011) extended this idea beyond violation-based guidance, showing that gaze in a free scene viewing task was also more likely to be directed to an object that was semantically related to the one that was previously fixated. Similar evidence for semantic guidance was reported in the context of a search task. Still another line of evidence comes from the emotion and attention literature, where it is commonly believed that threatening objects are preferentially fixated during search (e.g., LoBue & DeLoache, 2008; Öhman, Flykt, & Esteves, 2001). Given that no discriminative visual features are likely to exist for the category of "threatening things," to the extent that this is true it would seem to require a form of late selection.

However, our suggestion that guidance is a form of preliminary recognition performed on blurred objects does not require the full preattentive recognition of every object in a search display. The reason for this stems from the distinction between object recognition and object detection. Recognition is the attachment of meaning to a pattern via comparison to patterns that have been learned and committed to memory. This is true for both biological and computer systems. Preattentive recognition would mean that this process occurs automatically and in parallel for every pattern appearing in a scene. Walking into an opening reception of a conference would therefore cause names to be attached to all familiar attendees, and every other

object in the scene, regardless of where or how attention was allocated. Detection is the determination of whether a scene contains a particular pattern, with preattentive detection being the simultaneous comparison of this target pattern to all the patterns appearing in a scene. The analogous preattentive detection task would be walking into an opening reception with the goal of finding a particular colleague, and having all the patterns in a scene automatically evaluated and prioritized with respect to this goal. The product of a preattentive detection process is therefore a target map (Zelinsky, 2008) or a priority map (Bisley & Goldberg, 2010) that can be used to guide overt or covert attention; the product of a preattentive recognition process would be a map of meaningful objects, with the semantic properties of each being available to guide attention. Although building a target map is not trivial, especially when targets can be entire object classes, this task is vastly simpler than the task of building a map of recognized objects. This latter possibility has even been criticized as being biologically implausible on the basis of computational complexity, with the comparison of every pattern in a scene to every pattern in memory potentially resulting in a combinatorial explosion (Tsotsos, 1990).

Our contention is that a map of target detection probabilities is preattentively constructed from blurred information obtained from the visual periphery and that this map is used by the search process to guide gaze to the most likely target candidates. This puts us on the "early" side of the early versus late selection debate, as this target map is currently derived from purely visual analyses. Furthermore, we contend that these same target detection probabilities are used to make preattentive target/nontarget classification decisions for patterns appearing throughout the visual field. This is essentially the approach adopted in computer vision, where classifiers are routinely built on top of object detectors. Because classification decisions and the priority decisions used for guidance are both based on the same target detection probabilities, which are themselves derived from the same visual features, guidance and recognition under this framework become a distinction without a difference.

It is our belief that throughout the course of visual search the visual system is guiding gaze preattentively to the most likely targets while simultaneously attempting to preattentively classify objects in this visual information as targets or distractors. As for how the brain might construct a preattentive *categorical target map* to serve these dual functions, this question is beyond the scope of the present study. However, there seem at least two possibilities: either detectors for specific objects are duplicated at each region of the visual field, or a single, nonretinotopically organized mechanism compares in parallel a target representation to a search scene.[5] We prefer the latter explanation, which is most consistent with evidence from neuroimaging suggesting the representation of multiple object categories across the human ventral stream (e.g., Grill-Spector, 2009; Haxby et al., 2001), perhaps topologically organized into a semantic space (Huth et al., 2012). Given a single visual representation of a specific or categorical target, the brain might implement the parallel comparison of a target and a scene by biasing the low-level features of the target pattern, a mechanism also consistent with current theories of attention (e.g., Bundesen, Habekost, & Kyllingsbaek, 2005; Desimone & Duncan, 1995).

If guidance and classification decisions are as similar as we suggest and indeed are performed concurrently, why then is guidance necessary at all—why not base the final search decision on the outputs of the preattentive classifiers? There are again at least two possible answers. One is that parallel classification may simply not be possible in biological systems. It may be that only one pattern at a time can be routed to a single central classifier, making guidance important in its role of prioritizing these serial classification decisions (Broadbent, 1957; see Pashler, 1998, for a review). If true, this path would lead us back to a curious form of guidance-recognition cycle, one in which guidance is parallel and recognition serial despite both using basically the same information. A second possibility, and the one that we favor, is that classification *can* be performed on multiple patterns in parallel, but that the target detection probabilities underlying these classifications are often too low to make confident search decisions. In the context of the present study, a teddy bear in the search display would almost always have the highest probability of being the target, but peripheral blurring may prevent this object from reaching a threshold needed for classification. Following an eye movement to this object, the teddy bear would no longer be blurred and a confident classification decision could be made. Under this framework, target detection is parallel and without a capacity limit (see also Zelinsky, 2008), with guidance being a sort of movement to a more confident search decision. Often this movement is literal. We believe that the role of eye movements during search is to improve the quality of the visual information so as to make better and more confident classification decisions, and not to turn the wheel of the guidance-recognition cycle. Although additional constraints may one day be needed to satisfy new data, certainly this second path should be followed to its end before venturing down the path of serial search.

The bridge between the behavioral and computational approaches outlined here also has implications for the future direction of search. Visual search is one of the great success stories in the study of human

perception (Eckstein, 2011; Nakayama & Martini, 2011; Wolfe, 1998). Through the hard work of many researchers over many decades, this community now has an impressive understanding of the search process under conditions in which people have very precise knowledge of how a target will appear in a scene, as in the case of a picture preview. However, when precise information about a target's appearance is not known, as is the case for nearly every search task that we perform in our day-to-day lives, most computational models of search break down, returning the literature to a sort of theoretical infancy (but see, Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Judd, Ehinger, Durand, & Torralba, 2009; Zhang, Yang, Samaras, & Zelinsky, 2006). In this sense the literature has always had in front of it a theoretical wall—categorical search—with enormous progress made up to this wall, but very little beyond. This wall was built to prevent the search community from venturing into domains for which it didn't yet have the proper tools to find satisfying answers to questions—domains such as object recognition and categorization. But with recent advances in computer vision and machine learning, these tools are becoming available and the wall is starting to crumble—first under the weight of new behavioral findings (Maxfield & Zelinsky, 2012; Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009), and increasingly from the force of a new theoretical perspective (Alexander & Zelinsky, 2011; Chikkerur, Serre, Tan, & Poggio, 2010; Nakayama & Martini, 2011; Walther & Koch, 2007). The same factors and concerns that raised the wall are now causing its demolition.

Using the class of techniques and features described in this study, the search community can now move forward to explore the many difficult questions surrounding categorical search, and do so without sacrificing assumptions of biological plausibility. An essential step in doing this will be to reconceptualize guidance and recognition behavior as part of a single integrated search process, one in which object detectors are actively used in the service of classification decisions. It may be that the single template used to guide search (Olivers, Peters, Roos, & Roelfsema, 2011) may not be a template of the target at all, but rather a signature set of features that discriminate the target or target class from distractors. It may also be that target representations are limited, not in the number or resolution of their features, as capacity-limited thinking would suggest, but rather in the types of classifiers that can be trained and used to perform different search tasks. Support for these bold proposals will require evidence accumulated over many studies, not just one. The present work, while suggestive of this framework, is limited in that the proposed relationship between guidance and recognition was demonstrated for only a single category of targets. Future work will certainly need to address this generalization concern by considering a wider range of target categories. We will also replace the target templates used by a modal model of visual search (Zelinsky, 2008) with object detectors from computer vision, thereby developing this framework into a new and computationally explicit theory of attention and eye movements during categorical search.

Keywords: categorical guidance, object detection, visual similarity, classifiers, eye movements

Acknowledgments

This work was supported by NSF grant 1111047 to GJZ and DS, and NIMH grant R01-MH063748 to GJZ.

Commercial relationships: none.
Corresponding author: Gregory J. Zelinsky.
Email: Gregory.Zelinsky@stonybrook.edu.
Address: Department of Psychology, Stony Brook University, Stony Brook, NY, USA.

Footnotes

[1]One might usefully think of a detection window as being a sort of fovea, one that is passed in raster-like fashion over an image in a sequence of tiny movements with essentially zero delay.

[2]The capacity and effective VC dimension (Vapnik-Chervonenkis; Blumer, Ehrenfeucht, Haussler, & Warmuth, 1989) of the SVM models were controlled using the loss-regularization trade-off (C) chosen via cross validation.

[3]Paired-group *t* tests were used for all comparisons between conditions and between object types.

[4]When interpreting the improvement observed for the SIFT-BOW model it is important to consider this model's extremely poor fit to recognition behavior reported in Figure 4B. Rather than concluding that the addition of blur resulted in a good fit to behavioral recognition, it is probably fairer to conclude that adding blur to a SIFT-BOW model resulted in a somewhat less poor fit.

[5]Note that not considered here is the mechanism actually used in the present study—a moving detector window. We adopted this mechanism as a computational convenience and do not consider it to be a realistic alternative for how a categorical target map might be constructed by the brain.

References

Alexander, R. G., & Zelinsky, G. J. (2011). Visual similarity effects in categorical search. *Journal of Vision*, *11*(8):9, 1–15, http://www.journalofvision.org/content/11/8/9, doi:10.1167/11.8.9. [PubMed] [Article]

Alexander, R. G., & Zelinsky, G. J. (2012). Effects of part-based similarity on visual search: The Frankenbear experiment. *Vision Research, 54,* 20–30.

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 20–30.

Becker, S. I. (2011). Determinants of dwell time in visual search: Similarity or perceptual difficulty. *PLoS ONE, 6*(3), e17740. doi:10.1371/journal.pone.0017740.

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

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19*(6), 1162–1182.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology, 20*(1), 38–64.

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience, 33,* 1–21.

Blumer, A., Ehrenfeucht, D., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM, 36*(4), 929–965.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica, 129,* 255–263.

Broadbent, D. (1957). A mechanical model for human attention and immediate memory. *Psychological Review, 64,* 205–215.

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

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science, USA, 89,* 60–64.

Bundesen, C., Habekost, T., & Kyllingsbaek, S. (2005). A neural theory of visual attention: Bridging cognition and neurophysiology. *Psychological Review, 112*(2), 291–328.

Chang, C. C., & Lin, C. J. (2001). *LIBSVM: A Library for Support Vector Machines*. Internet site: http://www.csie.ntu.edu.tw/cjlin/libsvm. Accessed April, 2011.

Chelazzi, L., Duncan, J., Miller, E., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology, 80,* 2918–2940.

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research, 50,* 2233–2247.

Cockrill, P. (2001). *The teddy bear encyclopedia*. New York, NY: DK Publishing Inc.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, 1,* 22.

Deng, J., Berg, A., Li, K., & Fei-Fei, L. (2010). What does classifying more than 10,000 image categories tell us? *Proceedings of the 11th European Conference on Computer Vision (ECCV), 5,* 71–84.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, *357*(1), 241–265.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193–222.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science, 293,* 2470–2473.

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5):14, 1–36, http://www.journalofvision.org/content/11/5/14, doi:10.1167/11.5.14. [PubMed] [Article]

Eckstein, M. P., Beutter, B. R., Pham, B. T., Shimozaki, S. S., & Stone, L. S. (2007). Similar neural representations of the target for saccades and perception during search. *Neuron, 27,* 1266–1270.

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

Edelman, S., & Duvdevani-Bar, S. (1997). A model of

visual recognition and categorization. *Philosophical Transactions of the Royal Society B Biological Sciences, 352,* 1191–1202.

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*(6), 945–978.

Elazary, L., & Itti, L. (2010). A Bayesian model for efficient visual search and recognition. *Vision Research, 50*(14), 1338–1352.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392,* 598–601.

Everingham, M., van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2012*). The PASCAL Visual Object Classes Challenge 2012 Results.* http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1627–1645.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*(2), 191–197.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience, 2*(6), 568–573.

Grill-Spector, K. (2009). What has fMRI taught us about object recognition?. In S. Dickinson, M. Tarr, A. Leonardis, & B. Schiele (Eds.), *Object categorization: Computer and human vision perspectives* (pp. 102–128). Cambridge, UK: Cambridge University Press.

Gross, C. G., Bender, D. B., & Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of monkey. *Science, 166,* 1303–1306.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293,* 2425–2430.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210–228.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*(3), 480–517.

Huth, A. G., Nishimoto, S., Vu, A., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron, 76,* 1210–1224.

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2007). How chromaticity guides visual search in real-world scenes. *Proceedings of the 29th annual Cognitive Science Society* (pp. 371–378). Austin, TX: Cognitive Science Society.

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, *9*(5): 25, 1–18, http://www.journalofvision.org/content/9/5/25, doi:10.1167/9.5.25. [PubMed] [Article]

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research, 51,* 1192–1205.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *IEEE 12th International Conference on Computer Vision (ICCV),* 2106–2113. doi:10.1109/ICCV.2009.5459462.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3,* 759–763.

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

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial Pyramid matching for recognizing natural scene categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2,* 2169–2178.

LoBue, V., & DeLoache, J. S. (2008). Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychological Science, 19,* 284–289.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 565–572.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B, Biological Sciences, 200,* 269–294.

Maxfield, J. T., & Zelinsky, G. J. (2012). Searching through the hierarchy: How level of target categorization affects visual search. *Visual Cognition*, *20*(10), 1153–1163.

Motter, B. C., & Belky, E. J. (1998). The guidance of eye movements during active visual search. *Vision Research, 38,* 1805–1815.

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research, 51,* 1526–1537.

Nascimento, J., & Marques, J. (2006). Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, *8*(4), 761–774.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45,* 205–231.

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

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130,* 466–478.

Olivers, C., Peters, J., Houtkamp, R., & Roelfsema, P. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences, 15,* 327–334.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40,* 1227–1268.

Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1984). Neurons responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception. *Human Neurobiology, 3,* 197–208.

Perronnin, F., Akata, Z., Harchaoui, Z., & Schmid, C. (2012). Towards good practice in large-scale learning for image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 1–8. doi:10.1109/CVPR.2012.6248090.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology, 4*(1), e27. doi:10.1371/journal.pcbi.0040027

Platt, J. C. (2000). Probabilities for SV machines. In A. J. Smola (Ed.), *Advances in large margin classifiers* (pp. 61–74). Cambridge, MA: MIT Press.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343,* 263–266.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. New York, NY: Oxford University Press.

Russakovsky, O., Lin, Y., Yu, K., & Fei-Fei, L. (2012). Object-centric spatial pooling for image classification. *Proceedings of the European Conference on Computer Vision (ECCV), 2,* 1–15.

Rutishauser, U., & Koch, C. (2007). Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *Journal of Vision*, *7*(6):5, 1–20, http://www.journalofvision.org/content/7/6/5, doi:10.1167/7.6.5. [PubMed] [Article]

Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, *62*(10), 1904–1914.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2,* 994–1000.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415,* 318–320.

Swain, M., & Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Tanaka, J. W. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19,* 109–139.

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

Torralba, A., Murphy, K., Freeman, W., & Rubin, M. (2003). Context-based vision system for place and object recognition. *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 273–280.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review, 113,* 766–786.

Tsotsos, J. K. (1990). Analyzing vision at the com-

plexity level. *Behavioral and Brain Sciences, 13*(3), 423–445.

Ullman, S. Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology, 59,* 1931–1949.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17,* 159–170.

van de Sande, K., Uijlings, J., Gevers, T., & Smeulders, A. (2011). Segmentation as selective search for object recognition. *IEEE International Conference on Computer Vision (ICCV),* 1879–1886.

Vedaldi, A., Gulshan, V., Varma, M., & Zisserman, A. (2009). Multiple kernels for object detection. *IEEE International Conference on Computer Vision (ICCV),* 606–613.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1,* 511–518.

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, http://www.journalofvision.org/content/9/3/24, doi:10.1167/9.3.24. [PubMed] [Article]

Võ, M. L.-H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving window paradigm. *Attention, Perception, & Psychophysics, 73*(6), 1742–1753.

Walther, D. B., & Koch, C. (2007). Attention in hierarchical models of object recognition. In P. Cisek, T. Drew, & J. F. Kalaska (Eds.), *Computational Neuroscience: Theoretical Insights into Brain Function, Progress in Brain Research, 165,* 57–78.

Wang, H.-C., Hwang, A. D., & Pomplun, M. (2009). Object frequency and predictability effects on eye fixation durations in real-world scene viewing. *Journal of Eye Movement Research, 3*(3), *3*, 1–10.

Williams, D. E., & Reingold, E. M. (2001). Preattentive guidance of eye movements during triple conjunction search tasks. *Psychonomic Bulletin and Review, 8,* 476–488.

Williams, L. G. (1967). The effects of target specification on objects fixated during visual search. *Acta Psychologica, 27,* 355–360.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–71). London, UK: University College London Press.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5,* 1–7.

Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. *Cerebral Cortex, 15,* 1234–1242.

Yang, H., & Zelinsky, G. J. (2009). Visual search is guided to categorically-defined targets. *Vision Research, 49,* 2095–2103.

Zelinsky, G. J. (2012). TAM: Explaining off-object fixations and central fixation biases as effects of population averaging during search [Special issue]. *Visual Cognition, 20,* 4–5, 515–545.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787–835.

Zelinsky, G. J., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The role of top-down and bottom-up processes in guiding eye movements during visual search. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems* (Vol. 18, pp. 1569–1576). Cambridge, MA: MIT Press.

Zhang, L., & Cottrell, G. W. (2005). Holistic processing develops because it is good. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Annual Cognitive Science Conference* (pp. 2428–2433). Mahwah, NJ: Erlbaum.

Zhang, W., Yang, H., Samaras, D., and Zelinsky, G. J. (2006). A computational model of eye movements during object class detection. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems* (Vol. 18, pp. 1609–1616). Cambridge, MA: MIT Press.

Zhu, L., Chen, Y., Yuille, A., & Freeman, W. (2010). Latent hierarchical structural learning for object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 1062–1069.

Zhu, X., Vondrick, C., Ramanan, D., & Fowlkes, C. (2012). Do we need more training data or better models for object detection? *British Machine Vision Conference,* 1–11.