PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Modelling eye movements in a categorical search task

Gregory J. Zelinsky, Hossein Adeli, Yifan Peng and Dimitris Samaras

Phil. Trans. R. Soc. B 2013 **368**, 20130058, published 9 September 2013

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Research

Cite this article: Zelinsky GJ, Adeli H, Peng Y, Samaras D. 2013 Modelling eye movements in a categorical search task. Phil Trans R Soc B 368: 20130058.
http://dx.doi.org/10.1098/rstb.2013.0058

One contribution of 17 to a Theme Issue 'Attentional selection in visual perception, memory and action'.

Subject Areas:
behaviour, cognition, computational biology

Keywords:
visual search, eye movement guidance, object detection, realistic objects, classification, computational models

Author for correspondence:
Gregory J. Zelinsky
e-mail: gregory.zelinsky@stonybrook.edu

Modelling eye movements in a categorical search task

Gregory J. Zelinsky[1,2,3], Hossein Adeli[1], Yifan Peng[2] and Dimitris Samaras[2]

[1]Department of Psychology, and [2]Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-2500, USA
[3]Center for Interdisciplinary Research (ZiF), University of Bielefeld, Bielefeld, Germany

We introduce a model of eye movements during categorical search, the task of finding and recognizing categorically defined targets. It extends a previous model of eye movements during search (target acquisition model, TAM) by using distances from an support vector machine classification boundary to create probability maps indicating pixel-by-pixel evidence for the target category in search images. Other additions include functionality enabling target-absent searches, and a fixation-based blurring of the search images now based on a mapping between visual and collicular space. We tested this model on images from a previously conducted variable set-size (6/13/20) present/absent search experiment where participants searched for categorically defined teddy bear targets among random category distractors. The model not only captured target-present/absent set-size effects, but also accurately predicted for all conditions the numbers of fixations made prior to search judgements. It also predicted the percentages of first eye movements during search landing on targets, a conservative measure of search guidance. Effects of set size on false negative and false positive errors were also captured, but error rates in general were overestimated. We conclude that visual features discriminating a target category from non-targets can be learned and used to guide eye movements during categorical search.

1. Introduction

Decades of research has made visual search one of the best-understood paradigms for studying how information is selected from the world. During this long and fruitful endeavour (for reviews, see [1,2]), the visual search literature has evolved in both its behavioural methodology and its computational sophistication. There have been at least two substantive advances in behavioural methodology. One is in the preferred dependent measure. Whereas early studies of visual search relied almost exclusively on manual reaction time (RT) and accuracy-dependent measures [2], in the mid-1990s, the mainstream search community started to embrace eye movement-dependent measures ([3–6] for notable earlier studies, see also [7,8]). This trend has accelerated over the years, with oculomotor measures such as initial saccade direction now considered to be a golden standard of search guidance [9,10]. Another advance in behavioural methodology involves the choice of search stimuli. Whereas very simple patterns had been the norm in search studies, this too started to rapidly change in the mid-1990s with the adoption of more visually and semantically complex stimuli, first photo-realistic images of objects [11], then line drawings of scenes [12] and eventually fully realistic scenes [13,14].

Paralleling this evolution in behavioural methods was an evolution in the computational techniques used to study visual search. Modern models of visual search are designed to work with arbitrarily complex stimuli, typically arrays of real-world objects or scenes depicting complex environments. These models differ from their simpler counterparts in many respects but fundamentally in terms of the types of input they accept. Whereas earlier models might have input a list of item locations in a search array and the feature values associated with each of these items, newer models accept only an image—an array of

Royal Society Publishing
Informing the science of the future

pixels that has not been preprocessed into objects or even features. An important contribution of these *image-based models* is that they therefore make explicit how information is extracted from the world and used by the operations controlling search. This information is commonly represented as a pixel-by-pixel map of prioritized activity indicating where in a scene the search process should next be directed. Whether it is called an activation map [15], a saliency map [16], a target map [17] or a priority map [18], the importance of this stage of processing should not be underestimated. To derive these maps, decisions must be made about the types of feature that are used to represent information in images, how to build these features using techniques true to human neurobiology, how to combine and potentially weight these features in order to capture priority and how these prioritized signals change over time owing to lateral interactions or top-down inputs. Once answers are found to all of these questions, there is arguably little left to explain about visual search.

Combining these modern behavioural and computational methodologies, several models have attempted to predict the objects or patterns in an image that are fixated during search [19–22]. Most of these models, however, share a common limitation. Information on the map used to direct search is essentially rank ordered, with gaze sent to the location on this map having the highest ranking following the outcome of an explicit or assumed winner-take-all (WTA) process. Should the corresponding location in the search display not be the target, this location on the priority map is inhibited and the process repeats with gaze sent to the next lower-ranked pattern until the target is eventually fixated or some termination criterion is achieved. While implementing a reasonable scanning heuristic, this reliance on a WTA process means that a single pattern is selected for each fixation and that gaze is sent directly to that pattern. Such a peak-picking dynamic, however, limits the usefulness of these models as realistic accounts of eye movements during search because not all eye movements land on objects. The clearest example of this is the global effect [23], where gaze is directed to the centre of mass of two or more objects. Other examples are off-object fixations, cases in which gaze lands near but not on objects, and background fixations, cases in which relatively unstructured regions of a scene are fixated [11]. Because these models pick peaks of activity to fixate, and because peaks almost always correspond to salient patterns or objects, these off-object fixation behaviours cannot be explained and are therefore dismissed as errata by image-based models of search.[1] This is unfortunate as these ‘errors’, by some estimates, can account for up to 28% of the total fixations made during search [24].

The one image-based model of overt visual search that does not have this limitation is the target acquisition model, (TAM; [17]). TAM works by taking a filter-based decomposition of a target image, presumed to be maintained in visual working memory (VWM; [25]), then correlating these target features with a search image that has been blurred to reflect the retinal acuity limitations existing at each fixation. This operation produces a map of target-distractor similarity referred to as a *target map*, with the point of maximum correlation on this map referred to as the *hotspot*—TAM’s best guess as to the likely location of the target. However, and unlike other map-based methods of prioritizing search behaviour that assume a WTA operation, TAM does not send its simulated fovea directly to the hotspot location. Instead, gaze is sent to the weighted average of the target map activation, similar to the population coding of eye movements that takes place in the superior colliculus [26]. After this eye movement, a new search image, one blurred to reflect acuity limitations at the new fixation position, is again compared with the target features in VWM to obtain a new target map, and this process repeats. To generate sequences of eye movements that are guided to the target, TAM prunes from the target map over time the activity that is least correlated with the target, a dynamic that results in an initially strong expression of averaging behaviour that gradually lessens with each eye movement until the hotspot is selected and the suspected target is acquired.

This relatively simple model was found to capture several aspects of search performance, including: set-size effects, target guidance effects, target-distractor similarity effects, eccentricity effects and even a search asymmetry effect (see [17], for additional details about the model’s methods and how it compared with human behaviour). More recently, TAM was also shown to predict off-object fixations and centre-looking fixations—cases in which the initial saccade during search is directed to the centre of a scene regardless of its starting position [27]. Centre-looking fixations result from averaging over a target map that is still highly populated with activity, as the centroid of this activation will typically be at the scene’s centre. Off-object fixations and demonstrations of a global effect are typically observed later during search, when the target map is sparser and averaging is done over smaller, and often more local, pockets of activity (consult [24] for additional details). These behaviours make TAM the only image-based model of search that predicts not only the core eye movement patterns accompanying search, but also the oculomotor errata that have been neglected by other image-based models. Of course these other image-based models could also adopt a population code for programming eye movements and potentially explain some of these behaviours, but currently they do not.

However, this version of TAM also has serious weaknesses (detailed in [17]), with perhaps the biggest being that the filter-based features and the correlation method that it uses to generate its target map are both overly dependent on knowledge of the target’s exact appearance. This is a problem because this situation almost never exists in the real world. Even when searching for a very familiar target that has been seen hundreds of times before (e.g. car keys), on any given search variability in perspective, scale, lighting, etc., make it impossible to know exactly how this object will appear. This variability, by weakening the match between the target representation and its appearance in the search image, creates a problem for TAM and every other image-based model of visual search. A related but even more serious problem is that searchers often do not even know the specific type of object that they are searching for. If you are in an unfamiliar building and you want to throw away a piece of trash, you need to find *any* trash bin, not a particular one. But these things come in many different shapes and sizes and colours. This is the problem of categorical search—searching for a target that can be any member of an object class—and this is an extremely difficult problem as it requires somehow representing a search target that can literally be very different objects.

Of course our everyday experience tells us that a solution to this problem exists and that it is possible to search for categorically defined targets, but at issue is whether this search is guided. Early work on this topic showed that search is more efficient for a target designated by a picture preview compared with a categorical word cue [28,29], leading some researchers to conclude that it was not possible to guide search to a target category based on a preattentive analysis of visual features [30]. However, an inefficient search is not necessarily unguided, a point made by more recent studies using eye movement-dependent measures. Yang & Zelinsky [31] had observers search for a teddy bear target that was designated either by picture preview (specific) or by instruction (categorical). Although replicating the previously observed advantage for specific targets, they also found that categorical targets were fixated first during search more often than would be expected by chance. Concurrently, Schmidt & Zelinsky [32] found a preferential direction of initial saccades to a wide range of target categories, with this level of guidance increasing with the amount of information about the target provided in the word cue. Subsequent work has extended these findings, showing that categorical target guidance is sensitive to visual similarity relationships (gaze is preferentially directed to distractors that were rated as visually similar to the target category [19]), and that search guidance is also affected by the level that a target is designated in the categorical hierarchy (stronger guidance for subordinate targets, but faster verification for basic-level targets [33]).

Image-based theories have not kept pace with this new research into categorical search. This is because the task of categorical search requires the use of tools that are not yet familiar to the behavioural visual search community. Although the use of scale and orientation-selective filters (e.g. Gabors) have long been used by vision scientists to extract and represent the visual features of objects [34,35], simple applications of these tools to categorical search are doomed to failure—obviously, the features extracted from the word ‘teddy bear’ would bear no resemblance to those of an actual teddy bear, leading to no guidance. However, the computer vision community has been working on the problem of object class detection for decades and has made reasonable progress, now able to reliably detect many dozens of object categories [36–38]. Central to these techniques is the development of robust visual features to deal with the variability among members of an object class, and the use of machine learning methods to learn these features from training sets—exemplars of target images not used during testing. From these categorical features, object detectors can be built and used to derive a probability that a member of the target category exists in an image, with the totality of these probabilities for each location in an image creating again a sort of priority map that can be used to guide search. This conceptual similarity means that it may be possible to borrow these techniques from computer vision and use them to build behavioural models of categorical visual search.

There have been notable previous attempts to apply methods from computer vision to categorical search. The contextual guidance model [39] combines a bottom-up saliency map with a scene gist descriptor to approximate guidance by top-down knowledge about where categories of targets are likely to appear in a scene. This model was shown to predict the distributions of fixations made by observers searching for three target categories, people, paintings and cups/mugs, demonstrating that knowledge of a specific scene type (e.g. a city street) can be used to constrain the space over which a search is made for a particular class of target (e.g. people). The SUN model [40] differs from the contextual guidance model in predicting distributions of fixations using a probabilistic framework that learns the appearance-based features that are discriminative of a target category. It does this by combining simple difference-of-Gaussian filters (similar to TAM) with a probabilistic support vector machine (SVM) to train three classifiers: people/background, paintings/background and mugs/background. Applying this model to the search data collected by Torralba *et al*. [39], they found that target appearance was as useful as contextual guidance in predicting where people look for targets. In [41], the contextual guidance model was combined with information about the target category’s appearance to again predict distributions of fixations, this time in the context of a pedestrian search task. Target appearance was quantified using the pedestrian detector from [42]. Perhaps unsurprisingly, they found that a model combining both forms of top-down information better predicted fixations than either model separately.

By integrating object class detection techniques with TAM [17], the present work also strengthens the bridge between the behavioural and computer vision communities, while differing from previous work in several respects. Most fundamentally, whereas this previous work focused on *where* people look for targets, we focus on *how* people look for targets (see also [43]). Consequently, our model attempts to predict not distributions of fixations but rather properties of the eye movements that people make as they search. Our analyses consider measures of the number of eye movements made during search, the amplitudes of these eye movements, and the distances between the landing positions of eye movements and the nearest objects (indicating whether fixations were on or off of objects). Crucially, we also report the percentage of trials in which the very first eye movement landed on the target. We consider these immediate target fixations to be a true measure of a person’s ability to guide their search using a preattentive analysis of appearance-based features. Analyses of distributions or clusters of fixations, while also measuring guidance in a sense, are less able to distinguish contributions of preattentive guidance from post-attentive factors associated with object recognition. Our work also differs from previous work in its focus on visual search and factors known to affect search behaviour. We therefore report the effects of a set-size manipulation in the context of a target-present/absent search task. Although [41] also used a present/absence design, [39] and [40] did not. It is perhaps also worth noting that [39] and [40] used a task in which participants had to count the instances of a target class in a scene. While this is undoubtedly related to search, it is also likely different in that a counting task may encourage a more systematic inspection of a scene at the expense of a search urgency that may be needed to fully engage guidance. In this sense, our work is more closely related to [41], but adopts the more biologically plausible features and emphasis on target appearance found in [40].

2. Material and methods

We evaluated this categorical version of TAM against a previously collected behavioural dataset ([31], experiment 1). Observers in this study performed a response-terminated present/absent search for a teddy bear target in 6, 13 or 20 object displays (approx. 26° × 20°). The target was designated by instruction and not by a specific picture preview, making the search categorical. Distractors were from random object categories, and neither teddy bear targets nor distractors were repeated throughout the experiment. Each of the 12 observers participated in 180 trials, which were evenly divided into the present/absent and set-size conditions. Eye movement data were collected using an EyeLink II eye tracker (SR Research) with default saccade detection settings.

The computational methods followed those previously described for TAM [17], with three exceptions: (i) the use of a categorical target map, replacing the target map used by TAM, (ii) the inclusion of routines allowing for target-absent search and (iii) the use of a new method to transform search images to approximate the retinal acuity limitations existing at each fixation, replacing the method used by TAM. Each change is discussed in detail below, but the original sources should be consulted for additional details regarding the computational and behavioural methods.

(a) Categorical target map

The selection of features and methods used to create the model's new categorical target map was informed by recent work comparing and evaluating the potential for several existing computational models to predict categorical guidance and recognition in the context of a search task [44]. Out of the nine approaches tested in this study, the one that best predicted both search components was a model that combined a simple colour feature with a biologically plausible model of object recognition known as HMAX [45]. Based on this previous analysis, we therefore adopted these HMAX + COLOUR features for use in this study. Our goal in this initial effort was to simply replace the appearance-based target map used by TAM with a categorical target map derived from these features, keeping most of TAM's other eye movement-specific routines unchanged. This was done not only to preserve TAM's demonstrated ability to predict several aspects of overt search behaviour, but also to better isolate the effect of this extension to categorical stimuli on its behaviour.

The HMAX model attempts to describe human object recognition performance using only the initial feed-forward visual processing known to be performed by simple and complex cells in primary visual cortex [45]. In the basic four-layer version of the model, the responses of simple cells (S1), approximated by a bank of Gabor filters applied to an image, are pooled by complex cells (C1) using a local maximum operation, thereby producing limited invariance to changes in position and scale. Category learning is accomplished by randomly selecting C1 layer patches from training images of the category. These C1 layer maps are then filtered by simple cells (S2) to obtain feature maps for each of the sampled training patches. The final C2 features used for object detection are obtained by taking the maximum response within the S2 maps for each patch, forcing the number of features to equal the number of patches. Our implementation used a bank of Gabor filters with 16 scales and eight orientations and extracted 1000 C1 patches from positive training samples for use as prototypes for classification.

The HMAX model accepts only greyscale images, but colour is known to be an important feature for guiding search [46–48]. We therefore combined the HMAX model with a simple colour histogram feature [49]. Colour was defined in DKL space [50]. This colour space approximates the sensitivity of short-, medium- and long-wavelength cone receptors, and captures the luminance and colour opponent properties of double opponent cells. The colour histogram feature used 10 evenly spaced bins over three channels, luminance, red–green and blue–yellow, each normalized to the [0, 1] range. To compute this feature, we first converted images from RGB space to DKL space using the conversion described in [20], then built an image pyramid using three scales per image. From each layer of this pyramid, we sampled 24 × 24 pixel image patches, where each patch was separated by 12 pixels, and computed a colour histogram for each sampled patch. Prototypes were obtained by randomly selecting 250 patches across the three pyramid layers from positive training samples, with the maximum response to each prototype over a window becoming the colour feature for that window. Note that this produces a C2-like feature for colour similar to the Gabor-based features used by the HMAX model. Concatenating this colour feature with HMAX produced a 1250-dimensional HMAX+COLOUR feature.

Using these HMAX+COLOUR features, we trained a linear SVM classifier [51] to discriminate teddy bears from non-bears. The positive training samples were 136 images of teddy bears [31]; the negative samples were 500 images of non-bears randomly selected from the Hemera object collection. None of these training images appeared as targets or distractors in the search displays. For each search display used in the behavioural experiment, the 1280 × 960 pixel image was first blurred (see the following section on retina transformation) and then processed by sliding an HMAX + COLOUR object detector over the image. The resulting map of detector responses was then converted to probabilities based on their distances from the linear SVM classification boundary [52] and finally smoothed by replacing each point on the map with an average computed over a 20 × 20 pixel window centred on that point. This categorical target map, a pixel-by-pixel estimation of evidence for the target category, was used by our model to generate sequences of eye movements following the method described by Zelinsky [17]. Figure 1 shows two examples of categorical target maps (figure 1*c*,*d*) as they existed at the start of each trial, with the scan paths that ultimately culminated in correct target-present (figure 1*e*) and target-absent (figure 1*f*) search decisions.

(b) Collicular retina transformation

For each fixation that the model made as it searched, the search image was blurred to approximate the retinal acuity limitations that would exist when viewing the scene from that fixation. This *retina transformation* was done to better equate the visual information available to human searchers with the information used by the model. In order to make the model's target map more like the one known to exist in the superior colliculus [53], we replaced the pyramid-based approach used by TAM [54,55] with the one based on the projection of retinal ganglion cells to the superficial layers of the superior colliculus.

Visual information on the retina is coded densely at the fovea, but this density decreases with increasing distance from the fovea [56]. However, the distribution of retinal afferents on the collicular surface is roughly uniform with equal density, resulting in an overrepresentation of information from central vision relative to peripheral vision [57]. Using the method, parameters and local connectivity assumptions described in [58], we established a mapping between visual space and collicular space, then estimated the receptive field for each neuron in the superficial colliculus. This was done by assuming a fixed-size and Gaussian-distributed pattern of collicular activation (with sigma of 0.015 mm in the current implementation, based on a 4 mm colliculus), then finding the region of visual space corresponding to this activation. This allows for a distorted region of visual space to be approximated by a roughly circular region of activity on the colliculus. For each neuron, we then averaged the region in the image corresponding to its receptive field, weighted by the Gaussian function in collicular space. Applying this transformation produces an image that becomes increasingly blurred with distance from the fovea, thereby capturing the progressive loss of acuity as objects are viewed further in the visual periphery (figure 1*a*,*b*). TAM therefore not only 'saw' the same objects shown to the behavioural participants, but these objects were also blurred on a fixation-by-fixation basis to approximate

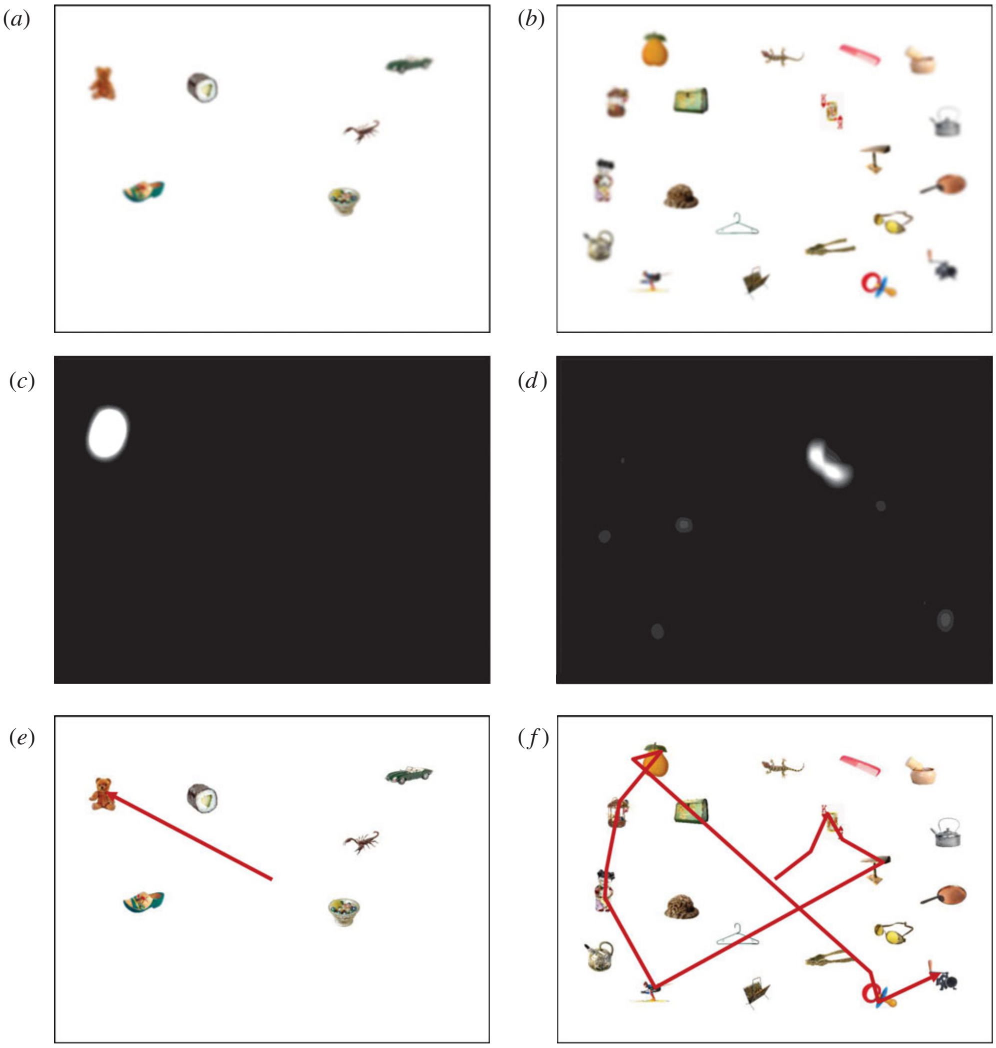

Figure 1. Representative search displays and behaviour from the model. (*a,c,e*) and (*b,d,f*) Images from two different trials. (*a*) Set size 6 search display, blurred to reflect acuity limitations as viewed from the centre. (*b*) Set size 20 search display, also blurred to reflect acuity limitations as viewed from the centre. (*c*) Target map generated from (*a*), reflecting activation before the initial eye movement. (*d*) Target map generated from (*b*), reflecting activation before the initial eye movement. (*e*) Eye movements preceding a correct target-present judgement. (*f*) Eye movements preceding a correct target-absent judgement. Note that each fixation would have associated with it a new blurred search display (as shown in *a,b*) and a new target map (as shown in *c,d*). The degree of blur can be seen by comparing (*a,b*) with (*e,f*). (Online version in colour.)

the information on the collicular map that is available to guide eye movements during search.

(c) Target-absent search

The previous implementation of TAM was purely a model of target acquisition—how gaze is directed to a target when it is present in a search display; functionality did not exist to terminate search in the absence of a target. To extend TAM to target-absent search, we therefore added a termination criterion in the form of a target-absent threshold. The target-absent threshold works much like the target-present threshold in TAM. Just as TAM pre-attentively monitors its target map for activity exceeding a high target-present threshold—which would allow its search to terminate with a target-present response—it now also monitors its target map for activity exceeding a target-absent threshold. Like the target-present threshold, this target-absent threshold is also a probability [0, 1]. Search continues so long as any activity on the target map exceeds this threshold, with the model making eye movements to suspected targets and ultimately rejecting these patterns as distractors in the case of a target-absent trial. Because each of these distractor rejections is accompanied by an injection of spatially localized inhibition on the target map [17] that implements a form of inhibition of return [59],[2] the net effect of this behaviour is the progressive removal of the activation peaks from the target map—the activity arising from the most target-like patterns. With the removal of these hotspots of peak activity, the maximum probability on the target map steadily decreases, eventually dipping below the target-absent threshold and causing the model to terminate with a target-absent response. In the present implementation, the target-absent threshold was set at 0.008 and not varied across search conditions. Inhibition was administered over a 150 × 150 pixel window. This window size was chosen to cover the objects appearing in the search display and was also fixed across conditions.

3. Results and discussion

All behavioural data were from Yang & Zelinsky ([31]; experiment 1). The search displays from this experiment were input to the model, which generated simulated eye movement and search behaviour for comparison to the behavioural participants. The parameters used for these simulations are described in detail in [17], although the model's three main user-supplied parameter settings were changed to: *target-present threshold* = 0.9; *target map increment threshold* (+TMT) = 0.2, and *target map decrement threshold* (−TMT) = 0.0001. These changes were made to better capture the different levels of confidence that likely exist between target-specific and categorical search. When TAM is searching for a specific target, knowledge of the target's appearance means that a very high target-present threshold can be used to minimize false positive errors. However, in the case of a categorical target, greater uncertainty in appearance requires a lower target-present threshold. Had the threshold from the study of Zelinsky [17] been used, the model would have terminated each trial with a target-absent response—not because targets would be missed, but because they would not be recognized. A similar logic applies to the target-absent threshold. In piloting, we explored a range of target and distractor values on the categorical target map in order to gain a sense of the signal-to-noise ratio in this task, then set the target-present and target-absent thresholds to obtain reasonable error rates. We assume that human participants engage in a similar learning process when settling on their decision criteria and perhaps in setting the size of their rejection window. The +TMT parameter affects the thoroughness in which the search display is inspected by gaze, whether it is conservative with many small saccades or more liberal with larger saccades. A smaller value was used in [17] to accommodate the scene stimuli used in that study (experiments 1 and 2), and the relatively small inter-item spacing (2.1°) among the object array stimuli (experiment 3*a*–*c*). The minimum inter-item spacing in the present experiment was 4°, making a highly conservative search unnecessary and the adoption of a larger +TMT reasonable. This, too, is something that human participants might realistically learn during practice. The range of +TMT values was explored only coarsely, and no attempt was made to find the optimal setting for this parameter. As for the –TMT parameter, this value was changed to 0.0001 because the value used in [17] was found to be unnecessarily small and needlessly increased model run-time.

(a) Search accuracy

Both the model and the behavioural participants made a target-present or target-absent decision in response to each search display. Although not the principle focus of this study, these decisions can be evaluated as a measure of task accuracy. Figure 2*a* plots false negative and false positive rates from human searchers and the model. Two patterns are worth noting. First, behavioural accuracy was superior to the model in each of the set-size conditions. This reflects the fact that object recognition is still an open problem in computer vision; even state-of-the-art methods cannot recognize a member of an object class as well as humans. Second, behavioural false negative rates increased with set size, $F_{2,33} = 137.66$, $p < 0.001$, but false positive rates decreased slightly as the number of distractors in the display increased, $F_{2,33} = 8.56$, $p = 0.001$. The model captured the former effect of set size on false negative errors (its slope was within the 95% CI of the behavioural slope) but not the latter effect of set size on false positive errors—as objects were added to a display it was more likely to mistakenly recognize a distractor as a target.

This departure from the behavioural data with respect to the false positive set-size effect might again be explained by the model's generally weaker object detection ability—each

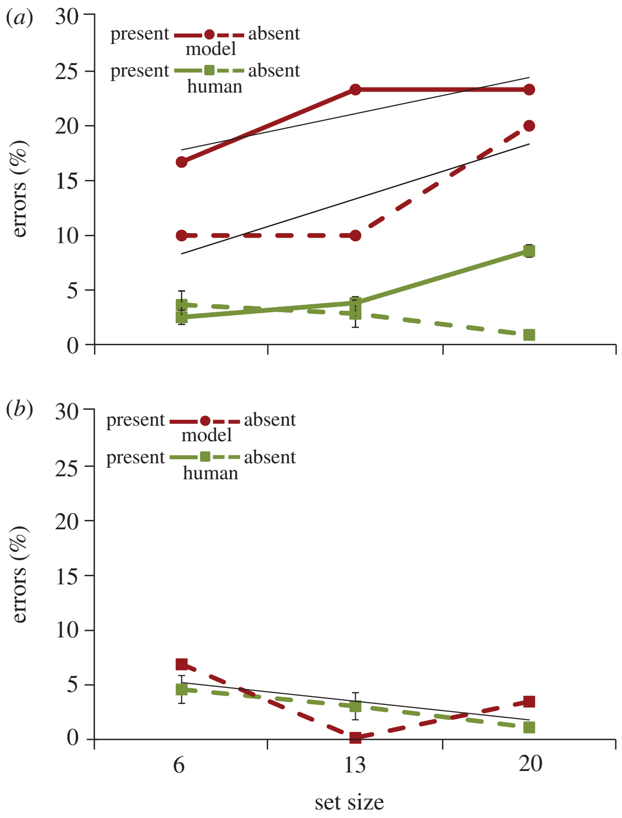

Figure 2. Percentages of false negative (solid lines) and false positive (dashed lines) errors for the human searchers (green (light) lines) and the model (red (dark) lines) as a function of set size. Error bars attached to the behavioural means indicate a 95% CI and trend lines are fit to the model data. (*a*) Data from all trials, regardless of eye position relative to an object. (*b*) Data from target-absent trials in which gaze was within 1° of a distractor at the time of the target-present search decision. Note that the behavioural data are essentially unchanged from (*a*), reflecting the fact that human observers almost always looked at an object prior to making their search decision. (Online version in colour.)

additional object created another opportunity to confuse a distractor with a teddy bear. But there is another factor potentially contributing to this discrepancy. Human searchers invariably chose to fixate an object before using it in a target-present search decision. This may be owing to either a strategic desire to accumulate higher resolution information about the object to inform the decision or simply because the slower manual response created the opportunity for the faster eye movement to reach the object before the decision was made. The model produced neither of these behaviours; it made a target-present decision as soon as its evidence that an object was a target exceeded a target-present threshold, even when this evidence was based on peripherally degraded information. This means that the model sometimes made its target-present search decision before fixating an object, where it might have found contra-indicating target information that may have prevented a false positive error. To evaluate this possibility, we analysed only those target-absent trials in which the model and searchers responded target present while fixating within 1° of the distractor's centre (figure 2*b*). Consistent with our prediction, the model made fewer false positive errors when it was fixating on a distractor immediately prior to its search decision. Moreover, this post-fixation false positive rate no longer increased with set size, and in fact now decreased with set size—bringing its behaviour more in line with human searchers. This suggests that if human searchers were to also maximize the speed of their search decisions, they too might show a positive relationship between false positive errors and set size.

(b) Set-size effects

Although subtle set-size effects were found in the error data, the effect of number of distractors is typically more pronounced in measures of RT for response-terminated search tasks. Our model does not currently make search RT predictions, but it does predict that the number of eye movements made before a present or absent search decision should increase with set size. Because number of fixations correlates highly with search RT [6], this measure can therefore be used to estimate search efficiency. Previous work has shown that TAM can produce a range of fixation-based set-size effects [17], but this was demonstrated for previewed search targets and very simple O/Q stimuli. The present work asks whether set-size effects can also be modelled for real-world objects and in the context of a categorical search task.

Figure 3 plots the average number of fixations leading up to a correct search decision, as a function of set size. Included in this measure is the initial fixation at the display's centre at the start of each trial. Human searchers showed clear set-size effects in both the target-present and target-absent conditions—the more objects in the display, the more fixations were needed to find the target or to conclude that the target was absent (target present, $F_{2,33} = 20.53$, $p < 0.001$; target absent, $F_{2,33} = 200.67$, $p < 0.001$). We also found the commonly observed interaction between set size and target presence—the slope of the fixations × set-size function was steeper for target-absent search compared with target-present search, $t_{11} = 14.02$, $p < 0.001$. The model captured all of these seminal search patterns. Moreover, and unlike the error analyses, its numbers of fixations were within the 95% CIs surrounding each of the six behavioural means. This is an impressive level of prediction; not only was the model successful in predicting the effect of set size on search efficiency, but it also successfully predicted the actual numbers of fixations made by human observers as they searched. This finding also highlights a significant generalization from TAM, showing that the processes used by the model to produce set-size effects are robust to changes in stimulus complexity and knowledge about target appearance.

(c) Search guidance

Figure 3 showed that for the 13 and 20 target-present set-size conditions, there were fewer eye movements made during search than half the number of objects in the display, and that this was true for both human searchers and the model. This pattern can be interpreted as evidence for guidance—had search been unguided a random direction of gaze to objects would have resulted in half of the objects being fixated, on average. However, a more compelling measure of search guidance is the percentage of trials in which the first eye movement made during search landed on the target, what we refer to as an *immediate fixation*. Yang & Zelinsky [31] conducted a related analysis for their observers and found significantly above-chance probabilities of targets being the first-fixated objects in each of their set-size conditions. Figure 4 plots a re-analysis of the Yang & Zelinsky [31] data to show the more conservative immediate fixation rate, together with the first-fixation-on-target percentages from the model. Trials in which no eye movement was made during search were excluded.[3] Increasing set size resulted in fewer initial eye movements from searchers landing on the target, $F_{2,33} = 182.19$, $p < 0.001$. The model's search also became less guided as objects were added to the search displays. More significantly, except for the set size 13 condition where the immediate fixation rate inched above the 95% CI surrounding the behavioural mean, the model was able to accurately predict the percentages of trials in which the first eye movement during categorical search landed on the target. This suggests that the guidance signal obtained from the categorical target map, and ultimately from an SVM-based model using HMAX+COLOUR features, is comparable to the guidance signal used by human observers as they searched for a categorically defined target.

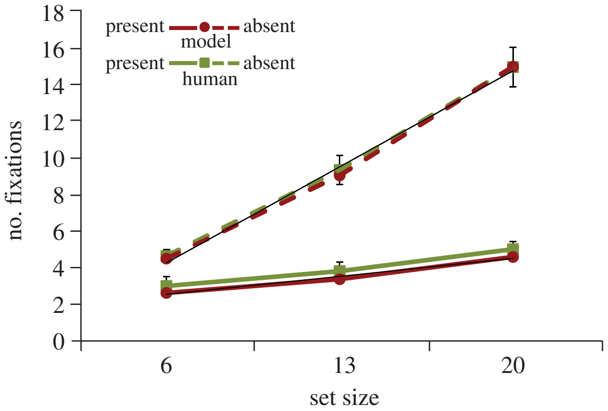

Figure 3. The average number of fixations prior to the target-present (solid lines) and target-absent (dashed lines) search decisions for human searchers (green (light) lines) and the model (red (dark) lines) as a function of set size. Data are from correct trials only, and the initial fixation during search was included in the measure for both the model and the behavioural participants. Error bars attached to the behavioural means show 95% CIs and trend lines are fit to the model data. (Online version in colour.)

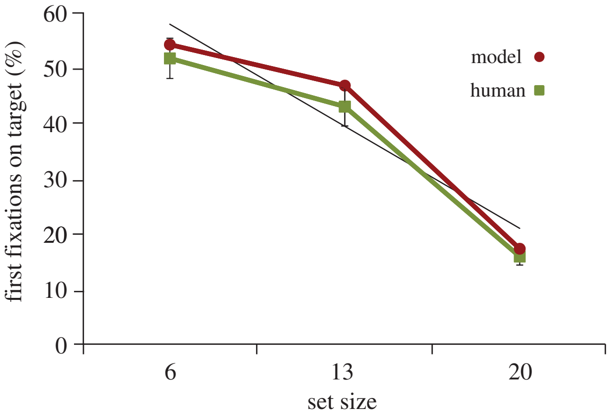

Figure 4. The percentage of target-present trials in which the first eye movement made during search landed within 1° of the target's centre, as a function of set size. The green (light) line indicates data from human searchers; the red (dark) line indicates data from the model. Error bars attached to the behavioural means show 95% CIs, and a trendline is fit to the model data. (Online version in colour.)

(d) On-object versus off-object fixations and saccade amplitudes

Further analyses of saccades and fixations were conducted to obtain a more complete picture of how well the model's

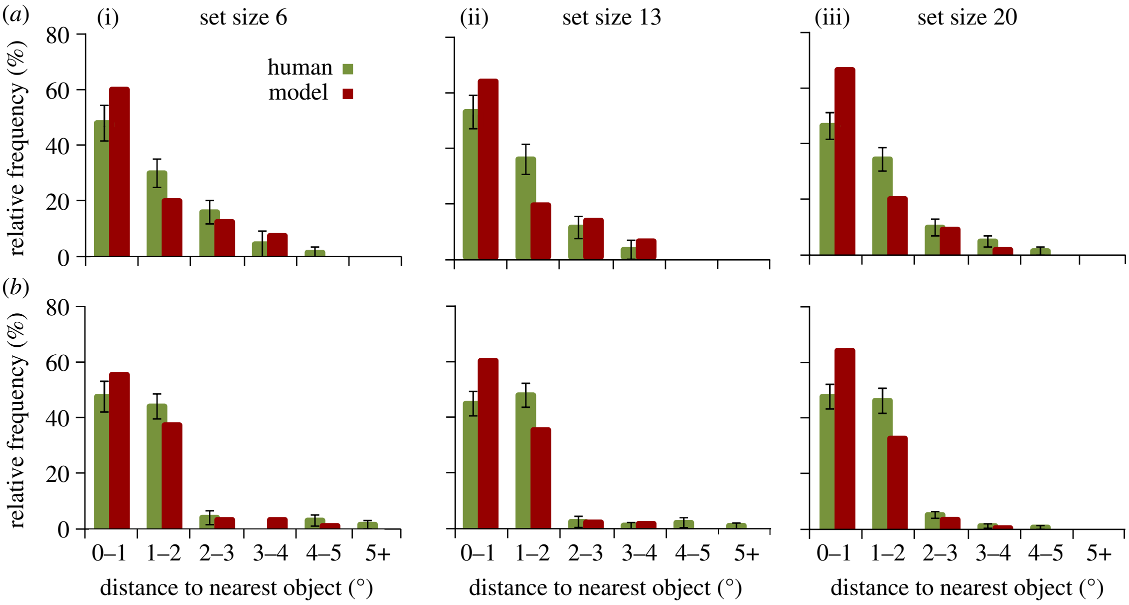

Figure 5. Relative frequency histograms of mean Euclidean distance between gaze and the centre of the nearest object for all non-initial fixations made during search for the model (red/dark) and human searchers (green/light). Analyses of the target-present data are shown in row (*a*); target-absent analyses are shown in row (*b*). The 6, 13 and 20 set-size conditions are shown in (i, ii and iii), respectively. Error bars indicate 95% CIs, and error trials were excluded from all analyses. (Online version in colour.)

behaviour captured that of human searchers. The first of these analyses looked at on-object versus off-object fixations. As described in [17,24], TAM is unique among image-based models in that it does not always make eye movements that land on objects—will this new, categorical version of TAM preserve this signature behaviour of its predecessor? Figure 5 shows histograms of mean distance between gaze and the centre of the nearest object for all non-initial fixations made during search. Initial fixations were excluded as these would conflate this distance measure with the minimum 4° object eccentricity used in the experiment [31]. There are two patterns of note. First, off-object fixations appeared quite prominently in the target-present data (row (*a*)). Given that objects in this task ranged in size from 1° to 4° [31], the two leftmost bars of each histogram indicate cases in which objects were fixated accurately. Excluding these on-object fixations, and averaging over the three set sizes, leaves 17% of the model's fixations falling off of objects. This is nearly identical to the 18% off-object fixation rate from the human searchers. Second, there were fewer off-objects fixations in the target-absent data (row (*b*)) compared with the target-present data. This, however, reflects the fact that many more fixations were made on target-absent trials. Because off-object fixations are most prominent early in search [24], their relative number decreases as the total number of fixations increases. Note also that the model's on-object fixations tended to be more accurate than those from human searchers. As also discussed in [24], this is an unsurprising observation given that the model's behaviour is perfectly constrained to the average of the object's spatial extent. Humans, although also showing this tendency to fixate the centres of objects [62], will certainly be more variable in their behaviour.

A second analysis compared saccade amplitudes between the behavioural participants and the model. The model's behaviour is initially biased towards the centre of the display configuration, meaning that its initial eye movements may be relatively short, shorter than the minimum 4° eccentricity to the nearest object. To evaluate this detailed prediction of the model, figure 6 shows histograms of saccade amplitudes for the first and second saccades (row (*a*) and row (*b*), respectively), as well as the average amplitudes for all subsequent saccades (row (*c*)). This was done only for the target-absent data, as saccade amplitudes on target-present trials were highly influenced by the presence of the target. As predicted, the model's initial averaging behaviour was expressed in saccade amplitude, but this was modulated by set size. For the 6 and 13 object displays, there was a clear rightward skew in the saccade amplitude distributions, whereas for 20 object displays, this distribution flattened with many more small amplitude saccades. This general pattern appeared for both the model and human searchers, and in some sense was to be expected—larger set sizes would, on average, result in an object appearing at the minimum 4° eccentricity. More interesting is the fact that saccade amplitudes smaller than this minimum distance increased in frequency with set size, so much so that they accounted for the majority of cases in the set size 20 condition for both human searchers and the model. We interpret this pattern as evidence for the centre biasing behaviour discussed in [24]. Given that gaze started at the display's centre, the centroid of the dense set size 20 displays would likely be near this starting gaze position, resulting in the observed small amplitude saccades. However, for the set size 6 condition, the sparser display configurations meant that the centroid was often farther from the display's centre, resulting in larger amplitude saccades. For the second and subsequent saccades, this pattern disappeared, and, indeed, patterns overall became less distinct. However, there was a relatively consistent bimodality in the amplitude distributions, suggesting a preference for short (1°–3°) and long (5°+) saccades. This pattern, appearing most prominently in the third (and subsequent) saccade amplitude distributions, suggests the use of two functionally different types of saccade during search, one that brings gaze to neighbouring objects in an item-by-item fashion and another that re-orients gaze to new regions of the display.

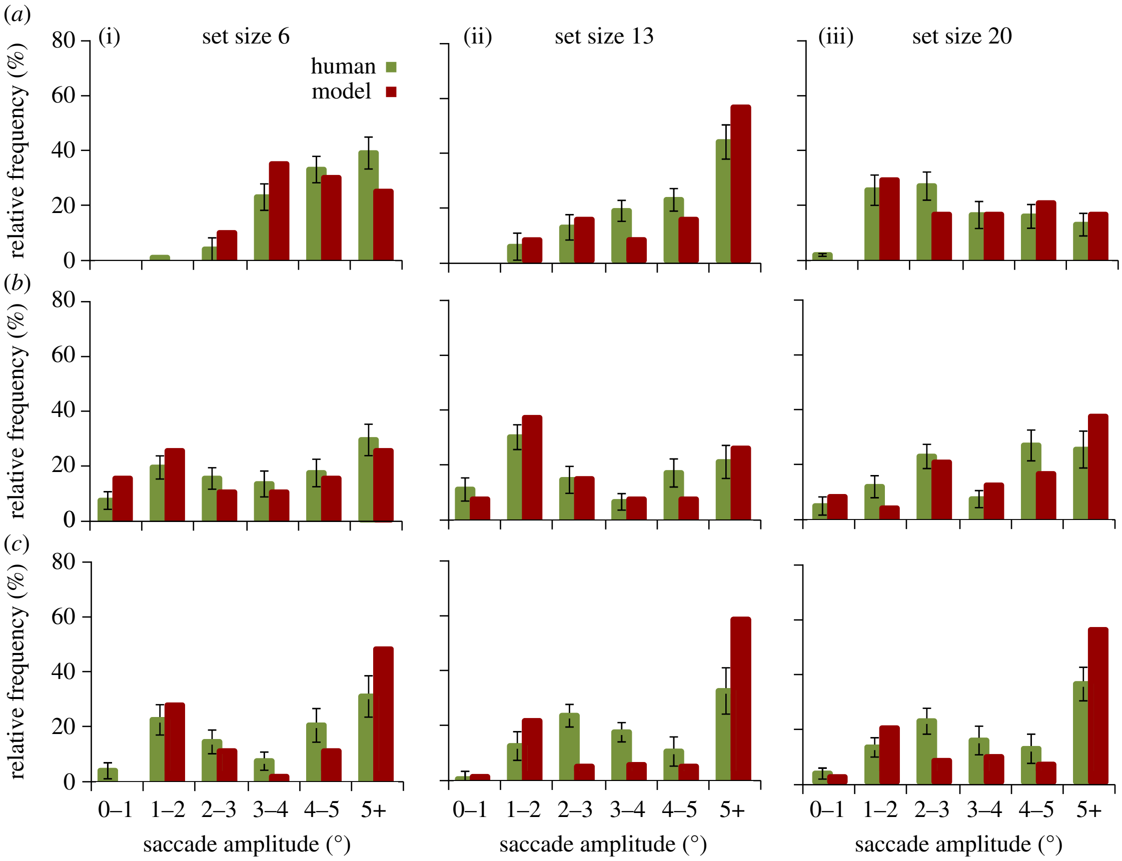

Figure 6. Relative frequency histograms of mean saccade amplitudes on correct target-present trials for the first and second saccades made during search (row (*a*) and row (*b*), respectively), as well as the average amplitudes for all subsequent saccades (third and greater, row (*c*)). Analyses for the model are in red (dark), and human searchers in green (light). The 6, 13 and 20 set-size conditions are shown in (i, ii and iii), respectively. Error bars indicate 95% CIs. (Online version in colour.)

This latter behaviour might serve the purpose of inspecting remote objects or pockets of items missed during search, or perhaps to revisit previously inspected objects so as to increase search confidence (see [63] for a related suggestion).

4. General discussion

A new evolution is underway in the search literature. Search tasks do not always come with specific targets; very often we need to search for dogs, or chairs, or pens, without any clear idea of the visual features comprising these objects. Despite the prevalence of these tasks, the question of how categories of realistic targets are acquired by gaze during search has attracted surprisingly little research. The present work adds to our understanding of this important topic by extending an existing model of eye movements during visual search, TAM [17], to the problem of categorical search. The TAM model was severely limited in that its representation of the target, and consequently the target map that it used to guide its search, required knowledge of the target's exact appearance in the form of a picture preview. In the present model, we replaced this target-specific target map with one capable of capturing information about the target category. Using this new categorical target map, we showed that multiple aspects of search efficiency and guidance could be modelled. It was able to predict accurately the numbers of fixations that were made before both target-present and target-absent search judgements, as well as how this behaviour changed as a function of the number of objects in the search display. Similarly, it predicted the percentages of trials in which the first eye movement during search landed on the target—a very conservative measure of search guidance—and once again captured how this behaviour changed with set size. Finally, the model captured the behavioural expression of off-object fixations and saccade amplitudes suggestive of saccade averaging. Collectively, these findings demonstrate that the changes made to TAM in order to accommodate categorical targets did not sacrifice the model's ability to broadly capture the patterns of eye movements that people make as they search.

The fact that this model was able to describe these different aspects of search is encouraging. Because it differed from its predecessor only in its target map and not in any of the processes used to generate eye movements, this suggests that these processes are generalizable and robust to changes in search condition—they work for target-specific search as well as categorical search. Relatedly, although the features used by this model differed from those used by TAM, these HMAX+COLOUR features were still based on responses from simple Gabor-like filters. Methods from computer vision for representing categories of objects typically use far more complex features, raising the concern that the successful modelling of categorical search would require sacrificing assumptions of biological plausibility. This appears not to be the case. Although the present demonstrations were limited to only one target class, teddy bears, the favourable comparisons between model and human search behaviour in this limited context shows that human-like patterns of search guidance and efficiency are possible using relatively simple features. An important direction for future work will be to demonstrate that these same features can be used to model the search for other target categories.

We also compared the model's target detection abilities to human target detection performance, and in this regard the model fell short. Object detection is a subtask of visual search, one that is often treated separately from the questions of search guidance and efficiency that dominate this literature. Indeed, many theories of visual search treat the problem of target detection as a sort of 'black box'; once attention is directed to an object, that object is assumed to be detected as a target or rejected as a distractor by some largely unspecified process [15,16]. Although the present model predicted very human-like effects of set size on false negative and false positive detections, the error rates were higher overall compared with human searchers. This result seemingly contradicts recent work that used the same HMAX+COLOUR features and found largely comparable target detection rates between model and humans [44]. However, that study used only a set size of four and the results were obtained outside the context of a model making multiple eye movements en route to an object. This would seem to suggest that the increased opportunities for false positives and false negatives accompanying larger set sizes during free viewing search cannot yet be modelling at a human level of accuracy, at least not using the current features and methods. Whether this means that modelling object detection performance in the context of categorical search will require different and more powerful features than those needed to model search guidance is a question that must be informed by future work integrating object recognition with visual search.

Finally, our work has theoretical implications for categorical search. Empirical work has made clear that gaze can be guided to categorically defined targets [19,31–33], but these studies only speculated as to how such guidance was possible. This categorical version of TAM makes explicit this speculation in the form of a working computational model.

One method of implementing categorical guidance would be to assume the existence and use of preattentively available features coding semantic attributes of the target category. Upon presentation of a search display, these attributes could be evaluated in parallel for each object, with gaze sent to the one having the greatest semantic similarity to the target (for related suggestions, see [64–66]). Although the present model does not rule out the possibility that preattentive semantic features are used for search guidance, it shows that these features are unnecessary and that a far simpler possibility exits. According to this model, the features used for categorical guidance are purely visual, as has been claimed for target-specific search [17], with the only difference between the two being how these features are derived and where they reside. In the case of target-specific TAM, these target features were presumed to be extracted by perceptual processes directly from the target image shown at preview and maintained in VWM for use as a guiding template. In the case of categorical TAM, discriminative features are believed to be learned from previous exposures to targets and non-targets. These features would likely reside in visual long-term memory, where they can be retrieved and used as a sort of categorical target template. In this sense, visual search may have available two paths by which a guidance signal might originate, one from information about a specific target entering the eyes and another from information about a categorical target that has been learned and coded into visual long-term memory. Important directions for future work will be to determine the conditions under which one form of target representation is used over the other, and how these two sources of guidance information might be combined. It is also important to determine the resolution of these categorical target representations. Work by Maxfield & Zelinsky [33] suggests that templates for categorical targets are available even for subordinate and superordinate categories, begging the question of what the limits are to these categorical representations. Can even a single previously viewed exemplar be used as a guiding target template, and are there some categories for which no target representations are available (and if so, why)? These, too, will be directions for future work.

Acknowledgements. Some of the work described in this study was conducted while G.J.Z. was a Fellow at the ZiF Center for Interdisciplinary Research.

Funding statement. This work was supported by NSF grant no. 1111047 to G.J.Z. and D.S., and NIMH grant no. R01-MH063748 to G.J.Z.

Endnotes

[1] Note that this limitation is in some sense ironic. Because image-based models are by definition not object-based, they are in principle able to explain off-object fixations. The fact that they do not results from their use of a WTA code to programme eye movements rather than one using a larger population of activity on the priority map.

[2] Recent work questions the existence of spatial inhibition of return [60–61], and to the extent that this is true it would violate this assumption of TAM—and every other image-based model of search that relies on inhibition of return to avoid becoming trapped in local minima on a priority map. However, our current position is that the literature has not yet settled on the role of spatial inhibition of return in search and scene viewing, and for this reason, we believe that it is premature for models to excise this mechanism from their function.

[3] Note that these trials were not excluded in the data plotted in fig. 3 of Yang & Zelinsky [31].

References

1. Eckstein MP. 2011 Visual search: a retrospective. *J. Vis.* **11**(5), 1–36. (doi:10.1167/11.5.14)
2. Wolfe JM. 1998 Visual search. In *Attention* (ed. H Pashler), pp. 13–71. London, UK: University College London Press.
3. Findlay JM. 1997 Saccade target selection during visual search. *Vis. Res.* **37**, 617–631. (doi:10.1016/S0042-6989(96)00218-0)
4. Williams DE, Reingold EM, Moscovitch M, Behrmann M. 1997 Patterns of eye movements during parallel and serial visual search tasks. *Can. J. Exp. Psychol.* **51**, 151–164. (doi:10.1037/1196-1961.51.2.151)
5. Zelinsky GJ. 1996 Using eye saccades to assess the selectivity of search movements. *Vis. Res.* **36**, 2177–2187. (doi:10.1016/0042-6989(95)00300-2)
6. Zelinsky GJ, Sheinberg D. 1997 Eye movements during parallel–serial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 244–262. (doi:10.1037/0096-1523.23.1.244)
7. Engel F. 1977 Visual conspicuity, visual search and fixation tendencies of the eye. *Vis. Res.* **17**, 95–108. (doi:10.1016/0042-6989(77)90207-3)
8. Williams LG. 1967 The effects of target specification on objects fixated during visual search. *Acta Psychol.* **27**, 355–360. (doi:10.1016/0001-6918(67)90080-7)

9. Chen X, Zelinsky GJ. 2006 Real-world visual search is dominated by top-down guidance. *Vis. Res.* **46**, 4118–4133. (doi:10.1016/j.visres.2006.08.008)

10. Schmidt J, Zelinsky GJ. 2011 Visual search guidance is best after a short delay. *Vis. Res.* **51**, 535–545. (doi:10.1016/j.visres.2011.01.013)

11. Zelinsky GJ, Rao RPN, Hayhoe MM, Ballard DH. 1997 Eye movements reveal the spatio-temporal dynamics of visual search. *Psychol. Sci.* **8**, 448–453. (doi:10.1111/j.1467-9280.1997.tb00459.x)

12. Henderson JM, Weeks P, Hollingworth A. 1999 The effects of semantic consistency on eye movements during scene viewing. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 210–228. (doi:10.1037/0096-1523.25.1.210)

13. Eckstein MP, Drescher B, Shimozaki SS. 2006 Attentional cues in real scenes, saccadic targeting and Bayesian priors. *Psychol. Sci.* **17**, 973–980. (doi:10.1111/j.1467-9280.2006.01815.x)

14. Zelinsky GJ, Schmidt J. 2009 An effect of referential scene constraint on search implies scene segmentation. *Vis. Cogn.* **17**, 1004–1028. (doi:10.1080/13506280902764315)

15. Wolfe JM. 1994 Guided search 2.0: a revised model of visual search. *Psychon. Bull. Rev.* **1**, 202–238. (doi:10.3758/BF03200774)

16. Itti L, Koch C. 2000 A saliency-based search mechanism for overt and covert shifts of visual attention. *Vis. Res.* **40**, 1489–1506. (doi:10.1016/S0042-6989(99)00163-7)

17. Zelinsky GJ. 2008 A theory of eye movements during target acquisition. *Psychol. Rev.* **115**, 787–835. (doi:10.1037/a0013118)

18. Bisley JW, Goldberg ME. 2010 Attention, intention, and priority in the parietal lobe. *Annu. Rev. Neurosci.* **33**, 1–21. (doi:10.1146/annurev-neuro-060909-152823)

19. Alexander RG, Zelinsky GJ. 2011 Visual similarity effects in categorical search. *J. Vis.* **11**(8), 1–15. (doi:10.1167/11.8.9)

20. Hwang AD, Higgins EC, Pomplun M. 2009 A model of top-down attentional control during visual search in complex scenes. *J. Vis.* **9**(5), 1–18. (doi:10.1167/9.5.25)

21. Peters RJ, Iyer A, Itti L, Koch C. 2005 Components of bottom-up gaze allocation in natural images. *Vis. Res.* **45**, 2397–2416. (doi:10.1016/j.visres.2005.03.019)

22. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW. 2008 SUN: a Bayesian framework for saliency using natural statistics. *J. Vis.* **8**(7), 1–20. (doi:10.1167/8.7.32)

23. Findlay JM. 1982 Global visual processing for saccadic eye movements. *Vis. Res.* **22**, 1033–1045. (doi:10.1016/0042-6989(82)90040-2)

24. Zelinsky GJ. 2012 TAM: Explaining off-object fixations and central fixation biases as effects of population averaging during search. *Vis. Cogn.* **20**, 515–545. (Special Issue on Behavioral and Computational Approaches to Reading and Scene Perception). (doi:10.1080/13506285.2012.666577)

25. Olivers C, Peters J, Roos H, Roelfsema P. 2011 Different states in visual working memory: when it guides attention and when it does not. *Trends Cogn. Sci.* **15**, 327–334.

26. McIlwain JT. 1991 Distributed spatial coding in the superior colliculus: a review. *Vis. Neurosci.* **6**, 3–13. (doi:10.1017/S0952523800000857)

27. Tatler BW. 2007 The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* **7**(14), 1–17. (doi:10.1167/7.14.4)

28. Vickory TJ, King L-W, Jiang Y. 2005 Setting up the target template in visual search. *J. Vis.* **5**, 81–92. (doi:10.1167/5.1.8)

29. Wolfe JM, Horowitz TS, Kenner N, Hyle M, Vasan N. 2004 How fast can you change your mind? The speed of top-down guidance in visual search. *Vis. Res.* **44**, 1411–1426. (doi:10.1016/j.visres.2003.11.024)

30. Castelhano MS, Pollatsek A, Cave KR. 2008 Typicality aids search for an unspecified target, but only in identification and not in attentional guidance. *Psychon. Bull. Rev.* **15**, 795–801. (doi:10.3758/PBR.15.4.795)

31. Yang H, Zelinsky GJ. 2009 Visual search is guided to categorically-defined targets. *Vis. Res.* **49**, 2095–2103. (doi:10.1016/j.visres.2009.05.017)

32. Schmidt J, Zelinsky GJ. 2009 Search guidance is proportional to the categorical specificity of a target cue. *Q. J. Exp. Psychol.* **62**, 1904–1914. (doi:10.1080/17470210902853530)

33. Maxfield JT, Zelinsky GJ. 2012 Searching through the hierarchy: how level of target categorization affects visual search. *Vis. Cogn.* **20**, 1153–1163. (doi:10.1080/13506285.2012.735718)

34. Daugman J. 1980 Two-dimensional spectral analysis of cortical receptive field profiles. *Vis. Res.* **20**, 847–856. (doi:10.1016/0042-6989(80)90065-6)

35. Olshausen B, Field D. 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609. (doi:10.1038/381607a0)

36. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. 2010 Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1627–1645. (doi:10.1109/TPAMI.2009.167)

37. Russakovsky O, Lin Y, Yu K, Fei-Fei L. 2012 Object-centric spatial pooling for image classification. In *Lecture Notes in Computer Science*, (*Proceedings of the 12th European Conference on Computer Vision; ECCV, 8–11 October, Firenze, Italy*), 1–15.

38. Everingham M, van Gool L, Williams CKI, Winn J, Zisserman A. 2012 *The PASCAL visual object classes challenge 2012 results*. See http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

39. Torralba A, Oliva A, Castelhano M, Henderson JM. 2006 Contextual guidance of attention in natural scenes: the role of global features on object search. *Psychol. Rev.* **113**, 766–786. (doi:10.1037/0033-295X.113.4.766)

40. Kanan C, Tong M, Zhang L, Cottrell G. 2009 SUN: top-down saliency using natural statistics. *Vis. Cogn.* **17**, 979–1003. (doi:10.1080/13506280902771138)

41. Ehinger KA, Hidalgo-Sotelo B, Torralba A, Oliva A. 2009 Modelling search for people in 900 scenes: a combined source model of eye guidance. *Vis. Cogn.* **17**, 945–978. (doi:10.1080/13506280902834720)

42. Dalal N, Triggs B. 2005 Histograms of oriented gradients for human detection. *IEEE Conf. on Computer Vision and Pattern Recognition* **1**, 886–893.

43. Zhang W, Yang H, Samaras D, Zelinsky GJ. 2006 A computational model of eye movements during object class detection. In *Advances in neural information processing systems 18* (eds Y Weiss, B Scholkopf, J Platt), pp. 1609–1616. Cambridge, MA: MIT Press.

44. Zelinsky GJ, Peng Y, Berg AC, Samaras D. In press. Modeling guidance and recognition in categorical search: bridging human and computer object detection. *J. Vis.* **12**, 957. (doi:10.1167/12.9.957)

45. Serre T, Wolf L, Poggio T. 2006 Object recognition with features inspired by visual cortex. *IEEE Conf. on Computer Vision and Pattern Recognition* **2**, 994–1000.

46. Motter BC, Belky EJ. 1998 The guidance of eye movements during active visual search. *Vis. Res.* **38**, 1805–1815. (doi:10.1016/S0042-6989(97)00349-0)

47. Rutishauser U, Koch C. 2007 Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *J. Vis.* **7**(6), 1–20. (doi:10.1167/7.6.5)

48. Williams DE, Reingold EM. 2001 Preattentive guidance of eye movements during triple conjunction search tasks. *Psychon. Bull. Rev.* **8**, 476–488. (doi:10.3758/BF03196182)

49. Swain M, Ballard D. 1991 Color indexing. *Int. J. Comp. Vis.* **7**, 11–32. (doi:10.1007/BF00130487)

50. Derrington AM, Krauskopf J, Lennie P. 1984 Chromatic mechanisms in lateral geniculate nucleus of macaque. *J. Physiol.* **357**, 241–265.

51. Chang CC, Lin CJ. 2001 *LIBSVM: a library for support vector machines*. See http://www.csie.ntu.edu.tw/cjlin/libsvm.

52. Platt JC. 2000 Probabilities for SV machines. In *Advances in large margin classifiers* (ed. AJ Smola), pp. 61–74. Cambridge, MA: MIT Press.

53. McPeek RM, Keller EL. 2002 Saccade target selection in the superior colliculus during a visual search task. *J. Neurophysiol.* **88**, 2019–2034.

54. Geisler WS, Perry JS. 2002 A real-time foveated multiresolution system for low-bandwidth video communication. *SPIE Human Vision and Electronic Imaging* **3299**, 294–305.

55. Geisler WS, Perry JS. 2002 Real-time simulation of arbitrary visual fields. In *Proc. of the Eye Tracking Research & Applications Symp.* (*ACM*), pp. 83–87. New York, NY: ACM.

56. Fischer B. 1973 Overlap of receptive field centers and representation of the visual field in the cat's optic tract. *Vis. Res.* **13**, 2113–2120. (doi:10.1016/0042-6989(73)90188-0)

57. McIlwain JT. 1986 Point images in the visual system: new interest in an old idea. *Trends Neurosci.* **9**, 354–358. (doi:10.1016/0166-2236(86)90113-X)

58. Ottes F, Van Gisbergen J, Eggermont J. 1986 Visuomotor fields of the superior colliculus: a

quantitative model. *Vis. Res.* **26**, 857–873. (doi:10.1016/0042-6989(86)90144-6)

59. Posner MI, Cohen Y. 1984 Components of visual orienting. In *Attention and performance X: control of language processes* (eds H Bouma, DG Bouwhuis), pp. 531–556. Hillsdale, NJ: Erlbaum.

60. Smith TJ, Henderson JM. 2011 Looking back at Waldo: oculomotor inhibition of return does not prevent return fixations. *J. Vis.* **11**(1), 1–11. (doi:10.1167/11.1.3)

61. Wilming N, Harst S, Schmidt N, König P. 2013 Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. *PLoS Comput. Biol.* **9**, 1–13. (doi:10.1371/journal.pcbi.1002871)

62. Nuthmann A, Henderson JM. 2010 Object-based attentional selection in scene viewing. *J. Vis.* **10**(8), 1–19. (doi:10.1167/10.8.20)

63. Tatler BW, Hayhoe MM, Land MF, Ballard DH. 2011 Eye guidance in natural vision: reinterpreting salience. *J. Vis.* **11**(5), 1–23. (doi:10.1167/11.5.5)

64. Becker MW, Pashler H, Lubin J. 2007 Object-intrinsic oddities draw early saccades. *J. Exp. Psychol. Hum. Percept. Perform.* **33**, 20–30. (doi:10.1037/0096-1523.33.1.20)

65. Bonitz VS, Gordon RD. 2008 Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychol.* **129**, 255–263. (doi:10.1016/j.actpsy.2008.08.006)

66. Underwood G, Foulsham T. 2006 Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Q. J. Exp. Psychol.* **59**, 1931–1949. (doi:10.1080/17470210500416342)