complex everyday environments, H&O's model would benefit from a thorough consideration of the *functional* characteristics of visual search.

If visual search functions to gather information for guiding our interactions with the environment, where the information resides and what action the information controls is of key concern. Gibson (1979) asserted that information resides in structured energy arrays that surround us. For example, the optic array is the surrounding light patterned by reflection against the surfaces, objects and persons (including the observer!) of the environment. Hence, the structured light patterns are specific to, and inform about, the environment. Accordingly, perception is the pick-up of this information in the optic array. Thus, it is crucial to recognize that an observer always moves, even if these movements would be restricted to the eyes! The information the observer exploits is a continuous flux, but within these unceasing transformations some patterns remain unchanged or invariant. Broadly speaking, Gibson proposed that invariances specify the (unchanging) environment, while the changes specify how the observer relates to the environment; they inform the observer about the actions that the environment affords. This implies that the observer's body, head, and eye movements co-structure information for guiding the observer's interaction with the environment, and thus must be part of any unified account of visual search.

To illustrate let us return to the soccer goalkeeper trying to stop a penalty kick. Typically, the ball moves at a speed that leaves a goalkeeper insufficient time to decide which side to dive on the basis of ball flight information. Therefore the goalkeeper must anticipate the direction of the dive based on information that resides in the penalty taker's movements. Expert goalkeepers distinguish themselves from their less successful counterparts in how they visually search the penalty taker's body for gathering this information. They make a small number of fixations of longer duration to fewer locations (Savelsbergh et al. 2002). Intriguingly, they particularly make long fixations on the empty space *in between* the non-kicking leg and the ball instead of making a sequence of fixations between different locations (Piras & Vickers 2011). This finding concurs with analyses that the most reliable information is distributed across different body locations rather than being located at one joint or body part (Diaz et al. 2012). H&O's FVF model can easily accommodate these observations. Within the model the FVF is defined as "the area of the visual field around fixation from which a signal can be expected to be detected" (sect. 5.1, para. 2). Importantly, the field is not fixed but varies in size. The smaller the field, the more fixations are needed and the more time the observer needs to search. Accordingly, expert goalkeepers may have a larger FVF than less skilled goalkeepers, allowing them to exploit the distributed information with less extensive visual search. This skilled search behaviour potentially provides experts with more reliable and timely information for ball interception.

H&O's perspective is largely limited to (typical) seated-monitor paradigms, which address how eye movements are used to search the environment. Yet in complex everyday environments, eye movements are but one means of gathering information, as a person's search also relies heavily on head and (whole) body movements. The over-reliance on seated-monitor experimental tasks leads to a limited view of visual search and may especially obscure its functional aspects. Perception and action, even in an *identical* environment, exploit different information and induce different patterns of visual search (Van Doorn et al. 2009). In this respect, we have previously reported that a soccer goalkeeper's visual search is fundamentally different when watching a penalty taker on a screen and verbally predicting kick direction compared to when actually facing a penalty taker in real-time and attempting to intercept the ball (Dicks et al. 2010). Perhaps surprisingly, on the pitch less time appears to be spent searching the body, while fixation towards the ball increases. It is likely that the different functional requirements affect the spatio-temporal structure of visual search in many more ways than a change in the magnitude of the FVF. A crucial challenge for any account of visual search, including the FVF model, is to spell out in more detail how functional requirements systematically affect the eye, head, and body movements for gathering information in complex everyday environments. To this end, looking further than monitors is a necessity!

Don't admit defeat: A new dawn for the item in visual search

doi:10.1017/S0140525X16000285, e159

Stefan Van der Stigchel[a] and Sebastiaan Mathôt[b]

[a]*Department of Experimental Psychology, Helmholtz Institute, Utrecht University, 3584 CS Utrecht, The Netherlands;* [b]*Aix-Marseille University, CNRS, LPC UMR 7290, Marseille, 13331 Cedex 1, France.*

s.vanderstigchel@uu.nl
http://www.attentionlab.nl
s.mathot@cogsci.nl
http://www.cogsci.nl/smathot

Abstract: Even though we lack a precise definition of "item," it is clear that people do parse their visual environment into objects (the real-world equivalent of items). We will review evidence that items are essential in visual search, and argue that computer vision – especially deep learning – may offer a solution for the lack of a solid definition of "item."

To say that items do not play a role in visual search is to admit defeat. Even though we lack a precise definition of "item," it is clear that people do parse their visual environment into objects (the real-world equivalent of items in visual search). In this commentary, we will review evidence that items are essential in visual search; furthermore, we will argue that computer vision – especially deep learning – may offer a solution for the lack of a solid definition of "item."

In the model of Hulleman & Olivers (H&O), search proceeds on the basis of fixations that are used to scan a visual scene for a target. Although we appreciate its parsimony, the model lacks a crucial aspect of visual search: the decision where to look next. The model simply assumes that an arbitrary new location is selected. Yet there is abundant evidence that fixation selection is not random but rather results from integration of top-down and bottom-up influences in a common saccade map (Meeter et al. 2010; Trappenberg et al. 2001). That is, we look mostly at things that are salient or behaviorally relevant (Theeuwes et al. 1998) and, crucially, behavioral relevance is related to how we parse visual input into items (e.g., Einhäuser et al. 2008). Consider repetition priming: People preferentially look at distractor items that resemble previous target items (Becker et al. 2009; Meeter & Van der Stigchel 2013); or its complement, negative priming: People avoid distractor items that resemble previous distractor items (Kristjánsson & Driver 2008). In addition, there are many object-based attention effects in visual search. For example, we tend to shift our attention and gaze within, compared to between, objects (Egly et al. 1994; Theeuwes et al. 2010); and, if an attended object moves, the focus of attention follows (Theeuwes et al. 2013). We could list even more object-based effects, but our main point is: Items matter, whether we know how to define them or not. Therefore, by denying a role for items in visual search, H&O ignore, or at least downplay the importance of, a substantial part of the visual-search literature.

But how can there ever be a role for items in models of visual search if we do not even know what "item" means? Possibly, our language simply lacks the vocabulary to define "item" or "object." Many researchers, such as David Marr, have speculated that it is impossible to define "object" (Marr 1982) – and we agree. But rather than abandon items altogether (and admit defeat!) we

should adopt recent computational approaches to object recognition as an alternative to formal definitions.

Consider a modern deep-learning network: an artificial neural network that consists of many nodes across many layers. (We will not discuss one specific network, but focus on the general architecture that is shared by most networks.) Such models are inspired by the architecture of our visual system by implementing a complex arrangement of nodes, each of which only looks at small portions of the input image (Krizhevsky et al. 2012). First, this network is trained on a large set of example images, which can be either labeled (e.g., Krizhevsky et al. 2012), unlabeled (e.g., Le et al. 2012), or a mix (LeCun et al. 2010). Crucially, in all cases training occurs by example, without explicit definitions. Next, when the trained network is presented with an image, nodes in the lowest layers respond to simple features, such as edges and specific orientations (Lee et al. 2009), reminiscent of neurons in lower layers of the visual cortex (Hubel & Wiesel 1959). Nodes in higher layers of the network respond to progressively more complex features, until, near the top layers of the network, nodes have become highly selective object detectors; for example, a node may respond selectively to faces, cats, human body parts, cars, and so forth. (Le et al. 2012). These nodes are reminiscent of neurons in the temporal cortex, which also respond selectively to object categories such as faces or hands (Desimone et al. 1984), 194). Importantly, deep-learning networks detect objects in those real-world scenes that H&O consider problematic (He et al. 2015; Krizhevsky et al. 2012); and they do so without explicit definitions, seemingly like humans do.

Combining deep-learning networks with traditional visual search models could explain how people explore their environment, item by item. As a starting point, we could take the model of H&O, and replace their bag of items with active nodes in high layers of a deep-learning network – that is, nodes that respond selectively to high-level features of the input (for example, cats), and for which the activation exceeds a certain threshold (Le et al. 2012). This would provide H&O's model with a bag of items to search through, without being fed any definition of "item." Of course, in its simplest form, this combined model is far from perfect. First, it does not explain object-based effects of the kind that we discussed above. Second, it assumes that the entire visual field is parsed at once, and does not take into account eye movements – the very idea that H&O rightfully want to get away from. But this simple combined model would be a good starting point that combines cognitive psychology with computer vision. And when combining principles from both disciplines, improvements readily come to mind. For example, a deep-learning network could be fed with eye-centered visual input that takes into account the functional viewing window.

In conclusion, we feel that H&O have been too quick to admit defeat. They have constructed a parsimonious model that explains visual-search behavior well without requiring items. Now all we need to do is put the item back in.

Where the item still rules supreme: Time-based selection, enumeration, pre-attentive processing and the target template?

doi:10.1017/S0140525X16000297, e160

Derrick G. Watson

Department of Psychology, University of Warwick, Coventry, CV4 7AL, United Kingdom.

d.g.watson@warwick.ac.uk

Abstract: I propose that there remains a central role for the item (or its equivalent) in a wider range of search and search-related tasks/functions than might be conveyed by the article. I consider the functional relationship between the framework and some aspects of previous theories, and suggest some challenges that the new framework might encounter.

Hulleman & Olivers (H&O) make a convincing case that researchers have tended to study and model search either solely from a covert attention or solely from an eye movement (EM) perspective and that if the field is to move forward there needs to be a concerted effort to combine the two – a sentiment with which I agree fully. The message is that we should replace the idea of the item with a combination of EMs and the extraction of information from fixations via a Functional Viewing Field (FVF) mechanism/perspective. EMs guide the FVF sequentially to regions from which information is extracted in parallel until the target is found. Because the size of the FVF changes as a function of target discriminability there is no role for the "item" within this framework. H&O argue that even when the task is to locate a target, the search process itself need not be item-based. Nonetheless, this of course still leaves (some) room for the item in visual search (it is the product of the search, and the target "template" will likely always be item-based).

In response, I will argue that item representations do play a central role in at least some search tasks. The "preview benefit" (Watson & Humphreys 1997) is just one finding that supports this view. In preview search, one set of distractors is presented (previewed) before a second set that contains the target. We find that people can ignore the previewed items and restrict their search to the second set of stimuli. According to the inhibitory visual marking account, this is achieved with stationary stimuli by developing a template of the locations of the old items and applying inhibition to those locations. This biases attention (and eye movements) away from those items, creating a search advantage for newly arriving stimuli. Granted, the localization of the initial items might not need to proceed via an item-by-item process (see above). However, because the inhibitory template is item- (location) based, and influences the subsequent search process, I would suggest that here "the item" (and its location) continues to play a crucial role in the subsequent search process itself. Indeed, if the locations of the old items change when the new items arrive, the preview benefit disappears (e.g., Zupan et al. 2015). In contrast, when preview items move, inhibition is applied mostly to feature maps (Andrews et al. 2011; Watson & Humphreys 1998), removing the need to track, localize, or process individual items (an example of part of a search theory in which the item is explicitly not important).

A second example in which the item probably remains salient can be found in enumeration tasks. Here people do not search for a single target but have to search for all targets (with or without the annoyance of distractors; Trick & Pylyshyn 1994) and report how many are present. In contrast to absent/present search, it is essential that items are not revisited because re-counting an item will lead to an error. With relatively coarse FVFs and an overlapping sequence of FVFs, ensuring that items are not recounted could be difficult. Perhaps here FVFs would be so small that search would effectively be item-by-item. Indeed, beyond four items enumeration appears to be especially reliant on EMs (Simon & Vaishnavi 1996; Watson et al. 2007).

Selection in time and counting things are two conditions in which the item might remain central to the task, but there are others. I wonder, for example, how contextual cuing (Chun 2000) will work without the spatial configuration of "items."

Moving on, does the FVF implicitly maintain the notion of an item? H&O argue that theories such as Attentional Engagement Theory (AET) are item-based because individual stimuli are grouped and rejected until the target is found. However, the FVF argument proposes that a stimulus emerges from the FVF which presumably is the result of some kind of competition between visual entities within the FVF. Is it possible that one episode of FVF processing equates to an entire search process in AET? So have we simply replaced the "item" from AET with more abstract visual entities within the FVF? Presumably there needs to be some individuation of "things" within the FVF for a target to emerge – aren't these "things" still just items? Notably, even though just a proof of concept, the entities fed into H&O's