Perception, 2007, volume 36, pages 1123 – 1138

doi:10.1068/p5659

How does the purpose of inspection influence the potency of visual salience in scene perception?

Tom Foulsham, Geoffrey Underwood
School of Psychology, University of Nottingham, Nottingham NG7 2RD, UK;
e-mail: lpxtf@psychology.nottingham.ac.uk
Received 14 August 2006, in revised form 22 January 2007

Abstract. Salience-map models have been taken to suggest that the locations of eye fixations are determined by the extent of the low-level discontinuities in an image. While such models have found some support, an increasing emphasis on the task viewers are performing implies that these models must combine with cognitive demands to describe how the eyes are guided efficiently. An experiment is reported in which eye movements to objects in photographs were examined while viewers performed a memory-encoding task or one of two search tasks. The objects depicted in the scenes had known salience ranks according to a popular model. Participants fixated higher-salience objects sooner and more often than lower-salience objects, but only when memorising scenes. This difference shows that salience-map models provide useful predictions even in complex scenes and late in viewing. However, salience had no effects when searching for a target defined by category or exemplar. The results suggest that salience maps are not used to guide the eyes in these tasks, that cognitive override by task demands can be total, and that modelling top – down search is important but may not be easily accomplished within a salience-map framework.

1 Introduction

The concept of a salience map is often used as a model of eye guidance and attentional capture (Hugli et al 2005; Itti 2005; Itti and Koch 2000, 2001; Koch and Ullman 1985). This is theorised to be an explicit, preattentive representation that allows shifts in attention to be directed to those regions of an image that are deemed salient. The concept of a salience map has the advantage of being neurally plausible (Treue 2003). Often this representation is computed purely on the basis of visual factors that highlight discontinuities in various low-level characteristics such as intensity or colour. Thus attention is guided towards the most visually salient location—that which stands out from its background to the greatest degree. This accords well with the concept of 'pop-out' effects in visual search where singletons which differ from the other items in the display capture attention (eg Baldassi and Burr 2004; Nothdurft 2002). In addition, measurement of image statistics suggests that local measurements, such as contrast, are higher at fixated regions than at unfixated or randomly selected regions (Mannan et al 1996; Zetzsche 2005), suggesting that such dimensions are indeed predictive of fixation location.

The most detailed model of visual salience is that proposed by Itti and Koch (2000, 2001). In this model centre – surround mechanisms enhance the regional differences in features at various spatial scales, and these are normalised to give a set of feature maps quantifying the discontinuities in each feature. The feature dimensions are intensity, colour, and orientation, although for dynamic situations motion and flicker channels can also be computed. The values for each scale are combined, and the feature maps linearly summed to give an overall salience map. Attention is hypothesised to move to the most salient point in the map, which is then transiently inhibited, allowing attention to disengage and move to the next most salient point, and so on. The model is advantageous in that it provides a straightforward account of dynamic attentional selection alongside specific predictions of which regions should be attended in which order.

Itti and Koch (2000) demonstrated reliable correlations between model performance and that of human participants in feature and conjunctive search (Treisman and Gelade 1980); though there are effects which they cannot model, for example that conjunction search slopes depend on the feature combinations used (Wolfe 1998b) and in a more naturalistic task searching for military vehicles in photographs. Using a different method, Parkhurst and colleagues compared the model-predicted salience at the fixated locations of four subjects with that expected by a chance fixation distribution (Parkhurst et al 2002). The resulting 'chance-adjusted salience' was significantly positive, indicating a greater than chance probability of fixating regions with high salience. This was particularly true for early fixations (though this was convincingly argued to be artifactual by Tatler et al 2005) and when the images used were artificial fractals rather than natural home interiors, perhaps emphasising the contribution of scene meaning. Such studies have emphasised the importance of visual salience, even in natural settings, although their correlational basis makes firm conclusions difficult. For example, in landscape photographs, edges and high-contrast information are often concentrated around the horizon. If this area is more likely to be fixated, it might be because of its higher visual salience, or it might be that gist information ("it's a landscape ...") and past experience ("people, houses and other interesting objects are often found on the meridian of landscapes ...") combine to determine gaze. Correlational studies can often not distinguish between these interpretations, so experimental manipulation is required.

While salience-map models emphasise bottom – up processes in scene perception, top – down processes determined by purpose of inspection can also have an influence. Scene semantics may also attract attention if, for example, an unexpected object is present. Bottom – up visual factors might then combine with scene-specific knowledge to direct the eyes to the most informative parts of the scene. The question whether a semantically incongruent object can be fixated immediately (Henderson et al 1999; Loftus and Mackworth 1978; Underwood and Foulsham 2006), coupled with the finding that scene schema and layout information may become available very early in viewing (Biederman et al 1982; Potter et al 2002), have led to discussion of the role of such top – down knowledge in guiding early attention. A 'salience-map hypothesis' (Henderson et al 1999) suggests that initial fixation placement in scene viewing is always based on bottom – up information, and it is only after a region is selected that scene semantics can begin to influence eye-guidance (by encouraging refixation of semantically informative objects, for example). A direct test of this hypothesis has suggested that a weaker version is more appropriate where visual salience does not determine early fixation placement regardless of cognitive demands (Underwood et al 2005). Torralba (2003) has shown how contextual priors regarding the likely layout of features, and not just salience, also effect early attention.

An important component of top – down influence is provided by the task being undertaken (Hayhoe and Ballard 2005). Of course, when changes in task lead to different scanning patterns over the same stimulus (Yarbus 1967), bottom – up salience models are fundamentally unable to provide a full explanation. In such a case the bottom – up features, and any salience derived only from them, remain the same, and so changes in scanning behaviour must be due to changes in top – down control. Such a change in task is brought about when the target of a search varies. So, for example, the eyes will move differently when searching a car park for one's car than when searching for somewhere to park, despite the scene remaining constant. Whilst some of the earliest research into eye movements and scene perception used multiple tasks (Yarbus 1967), other investigators (eg Parkhurst et al 2002) have attempted to limit the impact of task by simply instructing participants to view the stimuli. However, fixation placement does tend to vary depending on the task (Triesch et al 2003; Welchman and Harris 2003), and this is true whether trying to remember scenes, searching for something (Henderson et al 1999; Underwood et al 2005), or verifying concurrently presented sentences (Underwood et al 2004). The top–down guidance in these cases can explain the rather weak relationship between salience and fixation in natural situations.

How does top–down control of attention and eye movements function? Many models of attention have used visual search as an example of top–down guidance whereby knowledge of the target affects search [eg guided search, see Wolfe (1998a) for a review]. Search tasks provide a key paradigm for testing the interplay between bottom–up features or salience and top–down information. Pomplun (2006) has quantified task-driven control within complex search tasks using natural stimuli. He investigated 'saccadic selectivity' and low-level target features present elsewhere in the scene that attracted fixation. Recently, Tatler et al (2005) have further investigated exactly how cognitive and visual factors interact, concluding that a 'strategic divergence' model best accounts for the data. This model suggests that the salience map has a constant influence on eye guidance but that cognitive influences change over time. Modelling the effect of task within a salience-map framework continues to be a focus for some research (Lee et al 2005; Navalpakkam and Itti 2005). We have recently suggested that task demands allow *cognitive override* of the salience map (Underwood et al 2005), and visual salience may interact with semantic incongruity (or *cognitive* salience) in a similar way (Underwood and Foulsham 2006). Cognitive override signifies not just the top–down control inherent in search tasks but task-driven processes operating to an extent that bottom–up influences are negligible. Of course, the low-level features of a scene are necessary for preattentive parsing into relevant search regions and objects but the extent of salience-map computations in search is unclear.

How might salience-based search work? Target knowledge could potentially restrict or weight differently the features contributing to the salience map to those possessed by the target, and could also enhance regions of the map corresponding to the likely locations of this object (in cooperation with scene layout or gist). These influences of task are similar to those included in Findlay and Walker's (1999) model of saccade generation as processes of spatial and search selection, where regions or features are boosted on the basis of the target. Rao et al (2002) produced impressive fits to human data using a top–down iconic search model. In this model, salience is computed by correlating filter responses to a target with filter responses at each location in a to-be-searched scene. Zelinsky et al (2005) recently argued that this top–down model was actually hindered by the addition of bottom–up information based purely on feature contrast. Similarly, in a real-world walking task Turano et al (2003) found that a top–down model using coarse target features and location information outperformed a salience model. Navalpakkam and Itti's (2005) model precisely specifies how task can influence their bottom–up model of attention. Learned visual representations of targets bias the low-level visual processing which generates the salience map. So, if searching for something with prominent vertical edges, the feature map indicating the presence of 90° orientations, along with the whole of the orientation channel, will be more strongly weighted relative to features and channels not present in the target (eg red–green contrast and the colour channel). Likely target locations are primed with a 'task relevance map' which is generated on the basis of gist and layout information and task knowledge and which combines with the salience map to give overall attentional guidance.

Despite the obvious inadequacy of bottom–up salience models to explain all eye movements in natural scenes (Turano et al 2003; Underwood et al 2005; Yarbus 1967; Zelinsky et al 2005) this aspect of eye guidance continues to be proposed as important (Parkhurst et al 2002). Few researchers have demonstrated convincing effects of salience on scene viewing using an experimental approach where objects or regions in natural scenes are actually manipulated. By using such stimuli, we can look for a causal role for salience on eye movements in a natural task. Moreover, rather than looking just at simple displays or the most striking scene regions, we look at colour photographs and objects which should be fixated later in viewing. A previous study used similar scenes with distracting objects that were the most salient region (Underwood et al 2005). Natural search is often performed in cases where the target is not the most salient object in the scene and an ideal model would still be able to predict relative potency later in viewing. As a result, this experiment uses objects that are not identified as the most salient part of the scene. Is visual salience still capable of attracting attention in such conditions?

Although in this paper we use the Itti and Koch (2000) implementation as a measure of salience, it should not be taken as ignoring other possibilities. There are several good reasons for using this model: it is precise and well specified and can be implemented with relative ease and with natural images. There are some problems with the model, however. For example it does not, by default, consider the eccentricity of image regions, despite the fact that visual information from peripheral regions is of a necessarily lower spatial resolution (Carrasco et al 1995). The exact features that go into salience calculations may also be different in other bottom – up models. This study is largely a test of the Itti and Koch (2000) model, although we believe that the definition of salience in this model is at the very least highly correlated with that proposed in other models.

In this study we look at the potency of objects in natural scenes to attract overt attention as a function of their visual salience, as coded by a salience map, and also as a function of the task being undertaken. There are two main reasons why the interaction of task and salience is of interest. First, consideration of task may resolve questions about how far overt attentional selection is bottom – up. At least part of the variation in estimating the stimulus dependence of attention may be due to failing to take (implicit or explicit) task demands into account. Vision is an active process, and even when 'free viewing' scenes it is likely that participants' knowledge and presumptions of what behaviour is expected of them will influence their performance.

Second, it may allow conclusions about exactly how top – down and bottom – up processes interact. We compare here three tasks: a 'memory-encoding' task which requires participants to inspect photographs in preparation for a recognition task, encouraging them to look at the details, and as in Henderson et al (1999) and Underwood et al (2005) the effect of search instructions is also investigated. Two search variants are used, one in which the target is defined by a broad category ('category search'), and one in which it is a specific instance of the same category ('instance search'). In all tasks the attention given to objects with known salience is measured. If task demands influence a visual salience map by selectively weighting those features present in preconceived targets, a number of predictions can be made. Fixations in a memory task should be guided more on the basis of visual salience than those in a search task, as in the former there is no specific target. As a result, there will be little or no prior knowledge about specific informative features to weight a salience map, and fixations should be determined by default, bottom – up control. Top – down guidance in this task should be less pronounced than in search, where a clear target could bias the salience map. In category and instance search, the same objects, with constant bottom – up salience, will function as targets. Previous studies of search have assumed that detailed iconic representations of the target are available, although in the real world this may not be the case. Kenner and Wolfe (2003) suggest that visual search is faster when an exact picture of the target is seen. In instance search more feature information about the target is available than in category search. Does such information make search more efficient and less likely to be distracted by other salient areas?

If top – down instructions can be input into the same salience map used in the memory conditions, in the form of filtering or weighting by expected features, then this process might well be enhanced in instance search, owing to more specific target information. This prediction relies on some assumptions and so it is a tentative one. It is assumed that a target indicated by a verbal label can be efficiently translated into a set of features, that this set will be more restricted in instance search, and that this difference will be exploited by the eye-guidance system. Similar assumptions are present in recent models of search (Navalpakkam and Itti 2005; Rao et al 2002). It should also be noted that if the search process is based on a bottom – up salience map then some features not present in the target might still be salient enough to attract fixations, even after target-based salience reduction, unless all other features are reduced to zero. Does raw visual salience have less impact in instance search that in category search (as there is more specific target information to bias the model)? In other words, is cognitive override by task more frequent in instance search?

2 Method

2.1 *Participants*

Three groups, each containing fifteen student volunteers, participated in this experiment. All had normal or corrected-to-normal vision and were naive to the purpose of the experiment. Inclusion in the study was contingent on reliable eye-tracking calibration and, in particular, on maintaining a central fixation at the beginning of the majority of trials. Three participants were replaced, as they did not meet these criteria.

2.2 *Materials*

The same set of 48 digital photographs was used for stimuli in each condition, and these were taken with a 5 MP digital camera. They were presented on a colour computer monitor at a resolution of 1024 by 768 pixels. A fixed viewing distance of 60 cm gave an image that subtended 31 deg by 25 deg visual angle.

All photographs showed office scenes. Many different instances were used and all the scenes contained standard office furniture (desk, chairs, computer) along with a selection of smaller office objects such as books and stationery. Pictures contained similar amounts of office clutter, and no scenes were used more than once. There were 24 experimental stimuli, all of which contained a principal object (a piece of fruit) alongside the standard office clutter. Four types of fruit were used (apple, lemon, orange, and pear), with six pictures containing each type. These objects were chosen as they are of a similar size and have smooth contours and constant colouring, factors important in determining visual salience in the Itti and Koch salience model. In the interests of clarity, the fruit will be referred to as targets in the remainder of this report, although they were only highlighted as such to participants in two of the three task conditions. In each picture, the target could be located anywhere in the scene (within physical scene constraints) but was positioned 12° from the centre. The visual salience of the target was manipulated as described below.

A salience map was computed for each picture that allowed the relative salience of different regions to be measured and used as a selection criterion. The salience map was generated by a software implementation of a model of visual salience, developed by Itti and Koch (2000) and described in detail there. This model extracts variations in orientation, intensity, and colour at various spatial scales and combines them to produce a map showing the most visually salient locations. Following selection, the most salient 'peak' is inhibited, allowing the next most salient region to evolve, thereby simulating dynamic attentional selection. In this case standard settings were used, as detailed and justified in Itti and Koch (2000), and the salience-map computation consisted of filtering at 8 spatial scales, resulting in 6 centre – surround feature maps per feature. The salience map has a resolution of 1/16 of the original image, and an enlarged example is included in figure 1 (bottom panels). The present experiment is principally concerned with the order of fixation and so the rank of the regions selected by the winner-takes-all network was used as selection criterion rather than specific values. The choice of stimuli was made on the basis that none of the target objects was highly salient enough to feature in the first 5 peaks predicted by the model. Lower-salience objects were chosen here in order to extend scrutiny of the salience-map model to longer and more natural viewing periods, as opposed to using the number one most salient object in the scene, as was the case in Underwood et al (2005). Targets were further classified as *medium* salience (featuring between the 5th and the 10th peak) or *low* salience (featuring after the 10th peak), allowing the effect of salience to be explored. An equal number of medium- and low-salience pictures were included, with each target fruit equally represented in both. In practice, the salience

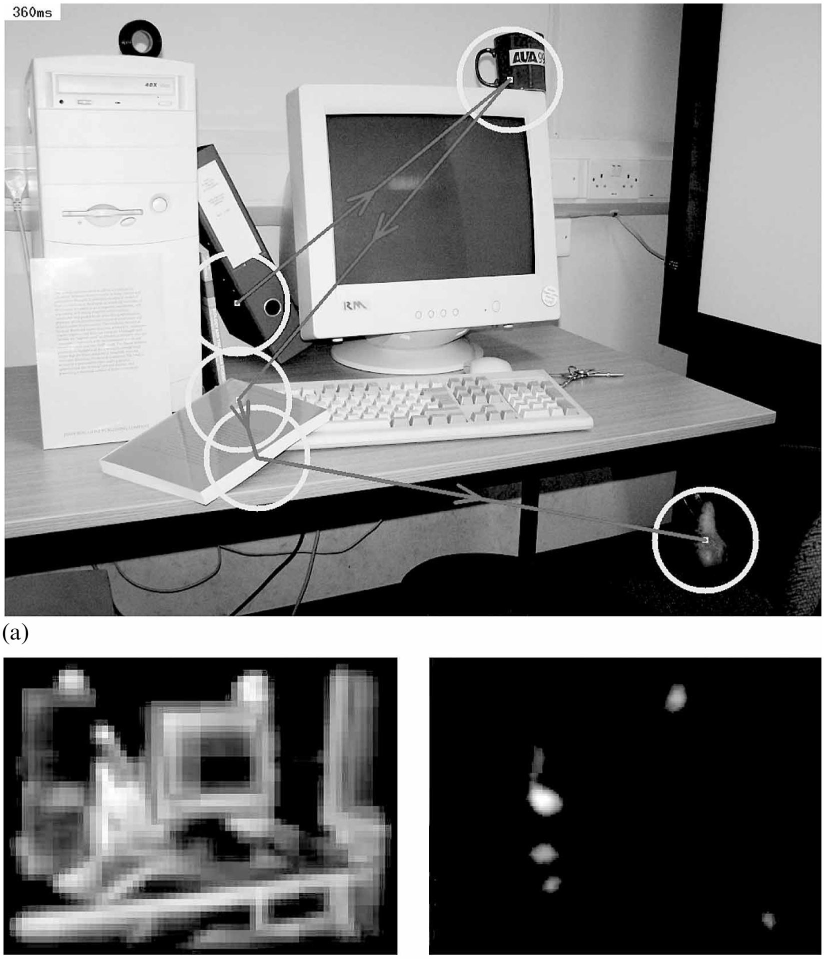


Figure 1. An example stimulus from the medium-salience condition, with ranks from the salience model (a). The area inside the circles indicates the focus of attention which, among other things, determines the region of inhibition following a fixation. A non-normalised salience map (b), which is formed from the combination of intensity, colour, and orientation conspicuity maps, is included with the final salience map (c), formed after normalisation and lateral competition processes. In both cases brighter areas indicate higher salience. Note that while the corner of a folder and the mug feature early in the salience map, the target pear is ranked 5th.

measure was an indicator of how much the target stood out from its background compared to the other distinctive objects in the scene. Thus the same object could be made less salient by placing it on a background of similar brightness and colour. Target objects were un-occluded. As a further control, the most salient region in the picture (the first predicted peak) was always on the opposite side of the picture from the target. Figure 1 shows an example stimulus with graphical output from the salience software indicating the salient regions in terms of their predicted ordinal salience.

A further 24 pictures did not contain a target and were used as controls. In addition, three sets of 8 practice pictures were prepared to familiarise subjects with the tasks.

2.3 *Design*

Visual salience of the target was a within-subjects manipulation with two levels (low versus medium). In addition, subjects were randomly allocated to one of three task conditions. These were a task simulating encoding for a memory test ('memory encoding') and two search tasks ('category search' and 'instance search'). Thus the between-groups factor of task had three independent levels.

2.4 *Apparatus*

An SR Research EyeLink I system was used to monitor eye movements. The system was head-mounted and sampled pupil position from the right eye every 4 ms. Calibration was repeated to ensure a spatial accuracy better than 0.5°. Fixation and saccade events were detected on the basis of velocity, acceleration, and saccade motion thresholds which were $30^\circ\ s^{-1}$, $4000^\circ\ s^{-2}$, and 0.15°, respectively. A chin-rest was used to minimise head movement and ensure a constant viewing position, and head position was recorded remotely.

2.5 *Procedure*

A preliminary calibration phase ensured that the apparatus was recording correctly. Participants were then shown written instructions on the screen. Prior to each picture, a drift correction marker and then a fixation cross, both in the centre of the screen, were presented, which confirmed that initial fixation was in the centre.

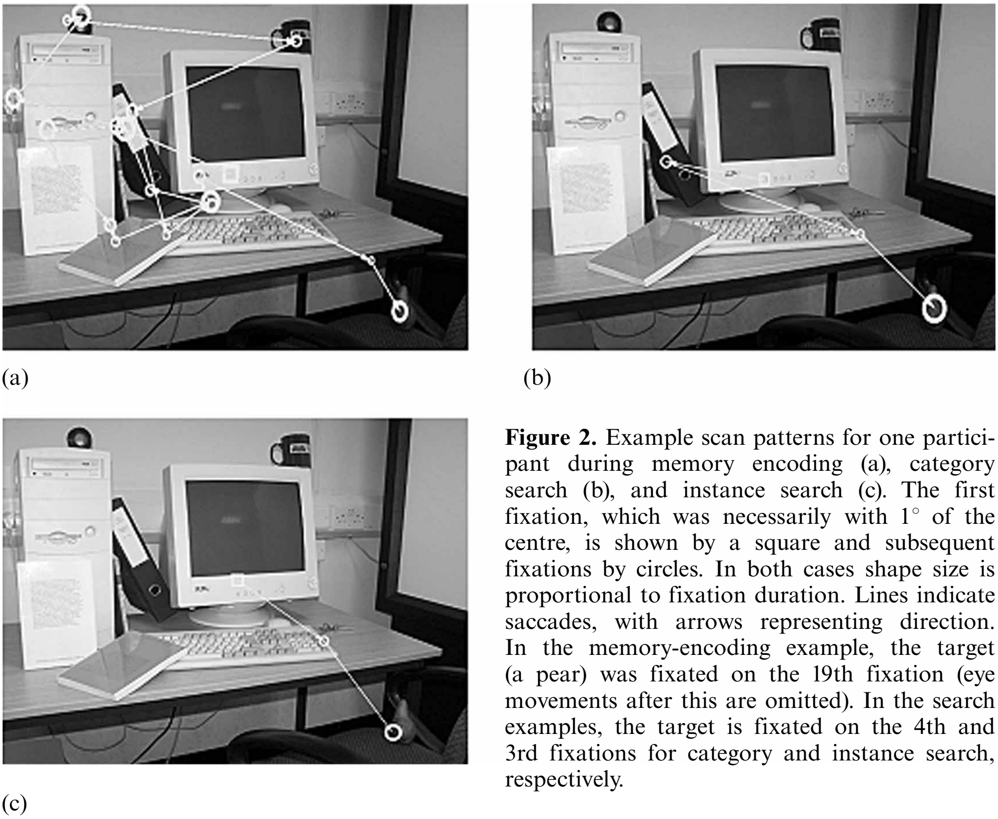
In the memory-encoding condition, instructions told the participant to view the scenes "in preparation for a memory test". Viewing was self-paced, and subjects were told to press a key to see the next picture. In a short training phase, subjects were shown some practice pictures followed by a two-alternative forced-choice recognition test featuring one picture they had seen previously and one that differed in the location or presence of an object (neither contained a target from the main experiment). This memory test was not given other than in the practice session, although most subjects expected it, as our concern was with attention and eye guidance during scene perception. This task was designed to simulate viewing of the whole scene with no particular preference for any one object, and has been used previously by Henderson et al (1999, experiment 1) and Underwood et al (2005). Following the practice session, all 48 pictures were shown in a randomised order with each one being terminated by the participant's key-press. The target will be labelled as such in order to conform to the terminology of the other conditions, although in this condition there was no reason to look at the object and participants had no knowledge of its significance. Few subjects identified any significance of the targets in this condition, and, owing to the large number of control pictures without fruit, we are confident that they were not deemed important.

In the first search condition, instructions informed the participant that the task was to search for the target, a piece of fruit, in each picture. This task was therefore a category search, looking for members of the category 'fruit'. If the target was present, participants had to press the 'Y' key, and otherwise the 'N' key, as quickly and accurately as possible. Following a practice phase, all 48 pictures (half of which contained the target) were presented in a randomised order, with stimulus offset triggered by the response.

The second search task was similar, but here the target was a particular instance of the category 'fruit' (apple, lemon, orange, or pear). This target was indicated by written instructions at the beginning of each of 4 blocks (one for each type of fruit, eg "the target for this block is an APPLE"). In each case the subjects had to respond by pressing the 'Y' or 'N' key to indicate if the target was present. Each block consisted of 12 pictures in a randomised order, 6 of which contained the target (3 of low and 3 of medium salience) and 6 of which contained no target. All participants viewed all 4 blocks in a random order. This condition will be referred to as 'instance search'.

3 Results

A range of eye-movement measures was computed from the raw data that showed fixation coordinates for each time sample. Although targets varied slightly in size (with mean dimensions of 1.9 deg by 2.1 deg), target fixations were identified where fixation coordinates were within a standard 100 × 100 pixels (3.1 deg × 3.1 deg) square that was centred on the target and which fitted all instances. Fixations were excluded if less than 100 ms in duration. In addition, fixation location at picture onset had to lie within 1° of the centre of the screen for that trial to be included. This was encouraged by the central fixation cross prior to each picture and was used as a strict way of ensuring the eccentricity of the target. This condition was not met for 14% of all trials and in these cases no further measures for that trial were included. Trials in the search tasks that led to an incorrect response were also removed. Figure 2 depicts an example of the general scan patterns made by observers on the 3 tasks whilst viewing the same stimulus as that in figure 1.



(a)

(b)

(c)

Figure 2. Example scan patterns for one participant during memory encoding (a), category search (b), and instance search (c). The first fixation, which was necessarily with 1° of the centre, is shown by a square and subsequent fixations by circles. In both cases shape size is proportional to fixation duration. Lines indicate saccades, with arrows representing direction. In the memory-encoding example, the target (a pear) was fixated on the 19th fixation (eye movements after this are omitted). In the search examples, the target is fixated on the 4th and 3rd fixations for category and instance search, respectively.

The measures taken reflect the hypothesis that visual salience will affect attentional selection, and therefore how soon and how often an object is fixated, and maybe other cognitive processing, which might be indicated by how long objects and scenes are inspected (Rayner 1998). Finally, for the search tasks, the proportion of correct responses to target pictures (that is, responding 'Y') was analysed. In each case, mean values were calculated across participants for each salience level and task condition. A summary of the measures taken is provided in table 1. A series of analysis of variance (ANOVA) tests was performed to determine statistical reliability. All pairwise comparisons were a posteriori Scheffé tests. Each measure will be discussed in turn.

Table 1. Means (and standard deviations in parentheses) for all the measures taken, organised by task condition and the visual salience (low or medium) of the target.

	Memory encoding		Category search		Instance search	
	low	medium	low	medium	low	medium
Ordinal fixation on target or end of trial	16.51 (6.57)	13.59 (4.15)	4.26 (0.89)	3.61 (0.55)	4.19 (1.02)	3.76 (0.80)
Probability of target being fixated	0.74 (0.20)	0.83 (0.16)	0.89 (0.21)	0.91 (0.24)	0.91 (0.12)	0.88 (0.15)
First-gaze duration/μs	581 (254)	658 (223)	432 (89)	427 (98)	524 (250)	567 (273)
Total inspection duration/μs	9210 (4604)	9044 (4665)	1275 (249)	1188 (204)	1457 (413)	1355 (385)
Number of discrete target fixations	2.20 (1.21)	2.56 (1.19)	1.46 (0.38)	1.46 (0.47)	1.30 (0.38)	1.63 (0.64)
Proportion of correct responses			0.84 (0.18)	0.87 (0.12)	0.85 (0.23)	0.91 (0.11)

3.1 *Ordinal fixation on target or end of trial*

The number of fixations on the picture leading up to a fixation of a target is an indicator of how quickly that object attracts attention. Targets that are potent in attracting attention will be fixated after fewer fixations than other objects. Targets that are less potent will be fixated after more fixations elsewhere in the scene, or the trial will be terminated before target fixation. This measure was therefore analysed to explore whether medium-salience targets attracted attention earlier than low-salience targets, irrespective of task. The earliest a target could be fixated was on the second fixation, as the first was necessarily in the centre of the display. The highest value this measure could have was the total number of fixations on the picture which varied (viewing was self-paced) but had a mean of 30.5, 4.6, and 4.7 in the memory, category-search, and instance-search, conditions (see analysis of total inspection duration which reflects this measure). Figure 3 displays the cumulative probability of fixation for each fixation since picture onset, and for each separate task condition.

A two-way ANOVA with the within-subjects factor of visual salience and the between-groups factor of task was carried out on the participant means. The results showed a highly significant effect of salience ($F_{1,42} = 13.56$, $\mathrm{MSE} = 2.97$, $p = 0.001$), with medium-salience targets fixated before low-salience targets (mean ordinal fixations of 6.99 and 8.32, respectively). As would be expected, task had a very reliable effect ($F_{2,42} = 67.94$, $\mathrm{MSE} = 18.12$, $p < 0.001$) and pairwise comparisons revealed that targets were fixated later when participants viewed pictures for a memory test (where there was no task requirement to look at the target, mean ordinal fixations of 15.05) than in either category search (mean = 3.93 fixations) or instance search (mean = 3.98 fixations) (both ps < 0.001). The two search conditions were not significantly different.

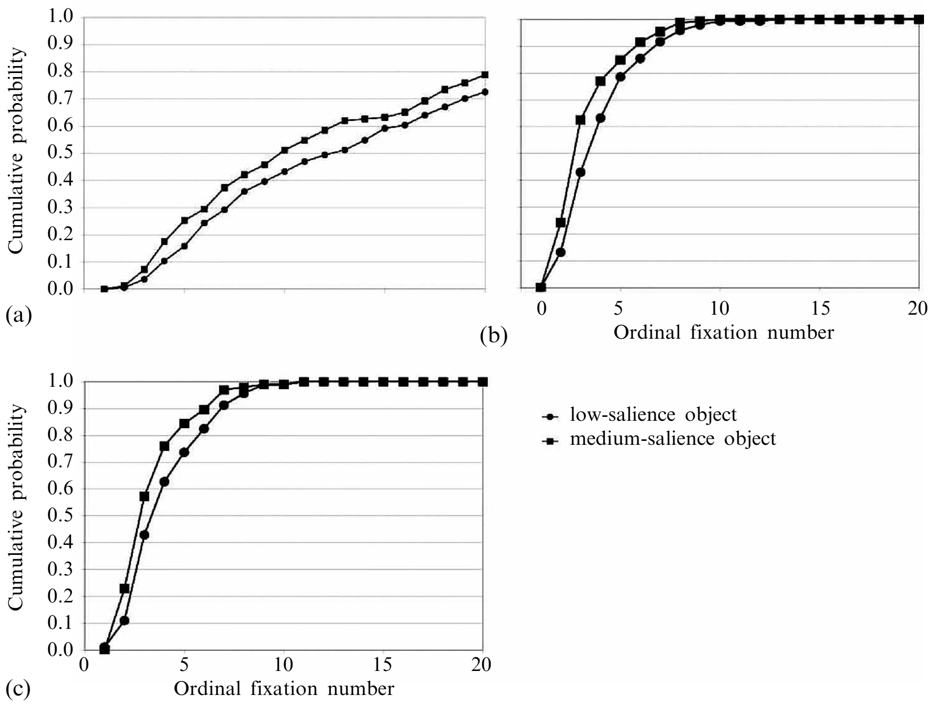


Figure 3. The cumulative probability of a target being fixated at least once as a function of ordinal fixation number since display onset for each task condition: (a) memory encoding, (b) category search, and (c) instance search. Targets were much more likely to be fixated earlier in the search tasks than in the memory task. Note that values may differ from those elsewhere in the report as they do not include those trials where the target was not fixated. Also note that the *x*-axis is shown up until 20 fixations, though some trials would have gone on longer.

There was an interesting interaction between task and salience ($F_{2,42} = 4.81$, $\mathrm{MSE} = 2.97$, $p = 0.013$). Simple main-effects analysis revealed that, while salience had an effect in memory encoding ($F_{1,42} = 21.63$, $\mathrm{MSE} = 2.965$, $p < 0.001$), in all other cases it was not significant. Task had a reliable effect on both levels of salience (low: $F_{2,84} = 71.0$, $\mathrm{MSE} = 10.54$, $p < 0.0001$; medium: $F_{2,84} = 46.0$, $\mathrm{MSE} = 10.54$, $p < 0.0001$).

3.2 *Probability of target fixation*

This measure was taken as a second indicator of the potency of the target in capturing attention. It was calculated from the proportion of trials where fixation lay within the target region at least once during stimulus presentation. If salience is important in all tasks, then fixations will be more likely to lie on medium targets than low targets.

As with the previous measure, a two-way ANOVA was performed on the participant means and, while the main effect of salience approached significance ($F_{1,42} = 3.78$, $\mathrm{MSE} = 0.0041$, $p = 0.059$), the effect of task was not significant ($F_{2,42} = 2.07$, $\mathrm{MSE} = 0.0629$, $p = 0.139$). There was, however, a significant interaction between the two ($F_{1,42} = 7.32$, $\mathrm{MSE} = 0.0041$, $p = 0.002$). Analysis of simple main effects showed that, while task had a significant effect on the probability of fixating low-salience targets ($F_{2,84} = 4.047$, $\mathrm{MSE} = 0.034$, $p = 0.021$), this was not significant with medium-salience targets. In addition, there was a simple main effect of salience only when encoding for a memory test ($F_{1,42} = 15.99$, $\mathrm{MSE} = 0.004$, $p = 0.0003$). This indicated that, in this condition, medium-salience targets were more likely to be fixated than were low-salience targets.

3.3 *First-gaze duration on target*

The duration of the first gaze on an object is an index of how difficult processing that object is (Rayner 1998). Gaze is the sum duration of all consecutive target fixations before fixating outside the region, including the first fixation duration. As the meaning or task demands related to medium- and low-salience targets did not differ, salience should not affect gaze duration. This measure also served as a control that targets did not differ in other ways, such as ease of processing once fixated.

The same ANOVA test as that performed on previous measures was applied here and indicated no significant main effect of salience ($F_{1,42} = 2.50$, $\mathrm{MSE} = 12766$, $p = 0.122$). The first-gaze duration was not different for low- and medium-salience objects. There was a main effect of task on gaze, however ($F_{2,42} = 3.58$, $\mathrm{MSE} = 76789$, $p = 0.037$). A posteriori comparisons showed that gazes in the memory-encoding condition were significantly longer than those in the category-search condition (619 ms and 430 ms, respectively; $p < 0.05$). There was no significant interaction of salience and task on first-gaze durations ($F_{2,42} = 0.99$, $\mathrm{MSE} = 12766$, $p = 0.38$).

3.4 *Total picture-inspection duration*

This measure was taken as the interval between picture onset and the terminating key-press response. As such, it was an indicator of the time required to perform the task before moving on. As above, a two-way mixed ANOVA was computed for the participant means. While there was a highly significant effect of task ($F_{2,42} = 43.14$, $\mathrm{MSE} = 14138314$, $p < 0.001$), neither the within-subjects factor of salience ($F_{1,42} = 0.99$, $\mathrm{MSE} = 319939$, $p = 0.326$) nor the interaction ($F_{2,42} = 0.040$, $\mathrm{MSE} = 319939$, $p = 0.96$) reached significance. As might be expected, comparisons between the different task conditions revealed that pictures were inspected for much longer in memory encoding (mean picture-inspection duration 9127 ms), where the task was more challenging and there was no target end-point, than in either category search (mean 1231 ms) or instance search (mean 1406 ms; both ps < 0.001). There was no significant difference between the total picture-inspection duration in the two search conditions.

3.5 *Number of target fixations per trial*

This measure, the number of times a target was separately fixated on any one trial, was taken to investigate whether certain targets were often refixated. The score for trials where the target was not fixated was zero. The ANOVA test gave a significant effect of salience ($F_{1,42} = 8.25$, $\mathrm{MSE} = 0.146$, $p = 0.006$), indicating that medium-salience targets were fixated more times per trial on average (1.88) than low-salience targets (1.65). There was also a significant effect of task ($F_{2,42} = 7.54$, $\mathrm{MSE} = 1.12$, $p = 0.002$), and a posteriori comparisons indicated that targets were fixated more times per trial (2.38 on average) in the memory-encoding condition than in either search condition (both ps < 0.01). No other differences were significant. The two factors of task and salience did not interact ($F_{2,42} = 2.05$, $\mathrm{MSE} = 0.146$, $p = 0.141$).

3.6 *Proportion of correct responses*

For the two search tasks, the proportion of correct responses to those pictures with each type of target (responding 'Y' to target pictures or the 'hit rate') was analysed with a 2×2 ANOVA with salience (low versus medium) and task (instance versus category search). The hit rate was high in all cases and there were no significant effects (salience: $F_{1,28} = 1.16$, $\mathrm{MSE} = 0.024$, $p = 0.291$; task: $F_{1,28} = 0.293$, $\mathrm{MSE} = 0.0315$, $p = 0.593$; salience by task interaction: $F_{1,28} = 0.084$, $\mathrm{MSE} = 0.024$, $p = 0.773$). Several participants were 100% accurate and false alarms in these tasks were relatively rare, occurring 8.1% of the time across conditions.

3.7 *Potency of most salient region*

There was a smaller, non-significant effect of target salience on ordinal-target fixation and target-fixation probability in the search tasks than in the memory task. A possible objection to these results is that by the 5th or 10th predicted fixation, salience values are lower. This forces the model to rank regions that may be only marginally different, and so may lead to an unfair evaluation of its performance. Although we believe that a complete model should account for the time span of a whole trial, a further test of salience is to look at fixations on the most salient region in the picture. This region was defined as that ranked first by the salience model (it is the corner of the folder in figure 1) and was bounded by a rectangle of the same size as the targets. Fixations on this region were recorded and the proportion of trials where the most salient region was fixated was analysed with the same two-way ANOVA as previously. Task was highly significant ($F_{2,42} = 191.2$, $\text{MSE} = 0.0215$, $p < 0.001$), with the most salient region capturing attention much more often in a memory task (mean 0.84) than in either search task (category mean 0.19; instance mean 0.21; $p < 0.001$). Interestingly, there was also an effect of salience ($F_{2,42} = 6.31$, $\text{MSE} = 0.0091$, $p < 0.05$) such that the most salient region was fixated more often when the target was medium salience (0.44) than when it was low salience (0.39). There was no interaction ($F_{2,42} = 1.02$, $\text{MSE} = 0.0091$, $p = 0.369$).

A valid objection to this analysis is that memory trials contained more fixations so that, even if fixations were allocated randomly, the memory task would be expected to contain more fixations on the most salient region. To resolve this, the above analysis was repeated with only the first five fixations from the memory task. Search trials contained five fixations on average, making the two comparable. Task remained significant ($F_{2,42} = 4.00$, $\text{MSE} = 0.019$, $p < 0.05$) indicating that even in the first five fixations the most salient region was more potent at attracting attention in the memory task (0.29) than in the search tasks (means as above). The effect of target salience remained ($F_{2,42} = 5.02$, $\text{MSE} = 0.013$, $p < 0.05$) and the interaction was not significant ($F_{2,42} < 1$).

3.8 *Summary of results*

As would be expected, the task instructions had a large effect on viewing behaviour with people making more fixations, longer picture inspections, and longer first gazes when encoding for a memory test than when searching for something. Search was efficient so that, of course, targets were fixated much earlier when subjects were actively looking for them. Salience had an effect on the ordinal fixation of targets such that medium-salience targets were fixated earlier than low-salience ones, and this was the case even late in the trial.

There were a number of particularly interesting results. Salience had a significant effect on fixation probability (how often) and ordinal fixation number (how early objects were fixated) only when pictures were viewed for memory encoding, and not during search. In any one trial, targets were refixated more often if they were more salient. There were no effects of salience in the two search tasks, and there were no significant differences between category and instance search. Accuracy in the search tasks did not differ with target salience or search variant. The most salient region in the scene was more likely to be fixated in the memory task than in the search task, even when differing numbers of fixations were controlled. There was also evidence that the most salient region was more potent when the target was of medium salience.

4 Discussion

The experimental manipulation of salience had a significant effect. Differences in visual salience caused the objects of interest to be fixated earlier when they were ranked higher according to Itti and Koch's (2000) salience algorithm. However, this was only the case when viewing in preparation for a memory test, and not when searching for the objects. Higher-salience objects were also more likely to be fixated in the memory task. This suggests that bottom – up selection is important when scanning photographs, but that these effects are not independent of task. The salience-map model correctly predicted which object would be fixated first, as medium-salience targets were on average fixated earlier (than low-salience targets) and were by definition predicted to be fixated earlier. There was a general trend for targets to be fixated later than the ranks generated by the model (for example, medium-salience objects were by definition ranked between 5th and 10th by the model but were fixated after 13.6 fixations on average). This suggests that the fit between the model and real data was not perfect. Other models, either modifications of a salience map taking into account different features or other accounts of bottom – up selection, might do better. In the memory task the objects did not differ semantically in importance with regard to scene or task context but only in low-level discontinuities.

The fact that an effect of salience can be found even fairly late in scene viewing (after more than 10 fixations on average) is particularly interesting. Parkhurst et al (2002) suggested that the effect of salience decreases over viewing time. If this is the case, the present results indicate that, even late in viewing, salience is still a significant factor. Henderson et al's (1999) general framework suggests that items are always initially fixated on the basis of salience, but once acquired can be evaluated in terms of cognitive demands which determine later processing. Henderson et al, along with many other researchers, used line drawings where the salience discussed here is effectively meaningless, so it is important that in the present research we used photographs (as did Underwood et al 2005) allowing the salience hypothesis to be tested fully. In the present study, salience had no effect on an index of processing (first-gaze duration) suggesting that, while bottom – up processes were important for attentional engagement, disengagement was not dependent on this. However, there was a tendency for higher-salience objects to be refixated, despite receiving presumably equal processing on the first gaze. This is shown by the fact that there were on average around two discrete fixations on these targets per trial, and this suggests that bottom – up selection continued to be important and may have triggered reflexive shifts when it probably would have been more efficient (in terms of the memorising task) to fixate elsewhere. The dynamics of the salience-map model are not incompatible with this finding as, although fixated regions are inhibited, this inhibition is transient and may not lead to complete suppression. How strong and for how long is this suppression? These control processes are beyond the scope of the current article but the importance of salience following the first acquisition of a region is worthy of further study.

The scale of the effects of task makes it clear how important this factor is in eye guidance. When searching for a target, as opposed to viewing scenes for a later memory test, scenes were inspected for much less time and targets were fixated much earlier, gazed at for less time, and refixated less often. It is problematic to assume that there is any such thing as 'free' viewing, and as models become more sophisticated it is important that experiments describe the control task conditions carefully, so as to both anchor results in the real world and enhance model predictiveness. Changes in fixation behaviour with search tasks similar to those found here were reported by Henderson et al (1999) and Underwood et al (2005). Eye-movement strategies in more complex everyday tasks have also been studied (Hayhoe and Ballard 2005), and Ballard and Sprague (2005) have recently argued that fully embodied models which take into account the role of attention in oculomotor routines are more useful. The results of the present study emphasise that a strong version of the salience-map hypothesis must be rejected on the grounds that cognitive demands can influence eye guidance and that this can happen before the object is first acquired. Comparison between the memory-encoding task and the search tasks here shows that search instructions can override visual salience, allowing earlier target fixation. This behaviour might be modelled with modified feature weights based on the target, as in the approach of Navalpakkam and Itti (2005). In addition, the difference between medium- and low-salience targets was reduced to the point where it was not reliable under search instructions, suggesting that salience is not important in search. Given the argument that top – down influences take longer to influence viewing, the analysis of fixations on the most salient region is interesting. Despite limiting this to early viewing, this region was more often fixated in the memory task than in the search task. The fact that this region was also more likely to be fixated when the target was more salient is hard to explain and may merit further study.

How does cognitive override of attention work? Several researchers have identified the types of top – down knowledge that may be available in a task and the way in which this might interact with bottom – up salience. In Findlay and Walker's (1999) model the 'where' pathway, which determines which regions are fixated, can be affected top – down in three ways. Spatial selection occurs when areas of the salience map are inhibited or potentiated on the basis of knowledge of target locations. Similarly, Torralba (2003) includes location probability as one of several contextual priors which influence attention. Object-location predictability was minimised here, and targets were not strongly cued by the gist of the scene, but location bias may have encouraged saccades to some areas of the display (as targets were always the same distance from the centre). Findlay and Walker's (1999) search selection, which has a parallel in Torralba's (2003) target-driven control parameter, enhances the salience of features present in the target. In a similar way, Navalpakkam and Itti's (2005) model weights the salience map on the basis of learned features of the target. Findlay and Walker's (1999) final process is the slightly underspecified concept of intrinsic salience, which allows some features (such as contours) to be intrinsically potent at capturing attention and some to develop this salience following medium-term learning. The present results might be explained by any one of these processes in that overt attention was presumably drawn to the targets in the search task on the basis of features. Some more detailed conclusions are possible, however. If search selection proceeds by potentiating a salience map, then salience should still have an effect on the time to fixate the target (both medium- and low-salience target regions will be potentiated as they contain target features, but medium-salience targets will still produce a higher peak). There was no significant effect of salience in the search conditions here, suggesting that this conceptualisation of the effect of task may be incorrect in this case. It was not the case that there was a difference in the influence of visual salience between the two variants of the search task. This might be predicted as in instance search more specific information about target features is available to bias the map. For example, in the category search, while likely shape and scale features might be primed (fruit targets are round and of a similar size), in the instance search colour was also identified by the target description (lemon targets are yellow, not green or orange like pears and oranges). In this case, there would be fewer possible saccade targets so search might be expected to be more efficient and less susceptible to interference from bottom – up visual salience of other regions. Kenner and Wolfe (2003) reported that visual search is quicker when an exact representation of the target is known than a general category, but the total inspection durations of the two search tasks here (which was also the time to respond) does not show this trend. There were no significant differences in the measures obtained between the two search variants, so if more information about the target was available, it does not appear to have been used in moving the eyes and responding more efficiently.

Williams et al (2005) report that memory for (and attention to) distractors in visual search was greater for objects which shared target features. While no distractors were specified here, predictions can be made for eye movements to distractors of various types in an instance search of natural scenes. Semantic category distractors (for example a banana while searching for an apple) and featural distractors (for example a green ball while searching for an apple) should attract attention more than unrelated objects in the scene, and might do so to different degrees. If salience is unimportant in search, as the results presented here suggest, the salience of such distractors will not affect their ability to draw attention.

The results reported here confirm that low-level salience is important in determining fixation location and order, but only in certain tasks. There was little evidence that visual salience was important in eye guidance in a search situation. Instead, they suggest that cognitive override in search may be an all-or-nothing process that does not rely on the same salience map as less targeted viewing. Although previous research has tended to focus on what might be referred to as 'default' or 'intrinsic' salience, the present results cannot fully reject a version where this property is weighted by target characteristics. There is a problem here as the process of feature weighting is often underspecified and what is meant by 'salience', be it intrinsic or modified by task, becomes unclear. Perhaps a better framework, and one which is now commonly being adopted, is the Bayesian one of Torralba (2003). In this approach local, hard-wired salience is one of several priors and is kept separate from a 'target-driven control' parameter. It is clear that a purely salience-based model is incapable of explaining the natural complexities of eye-movement behaviour and the findings reported here emphasise some of the difficulties of addressing top–down control within such a framework.

Acknowledgments. We thank Laurent Itti for the use of his visual salience software, and Editha van Loon for installing it. We are also grateful for the helpful comments of Ben Tatler and an anonymous reviewer on a previous version of this manuscript.

References

Baldassi S, Burr D C, 2004 "'Pop-out' of targets modulated in luminance or colour: the effect of intrinsic and extrinsic uncertainty" *Vision Research* **44** 1227 – 1233

Ballard D, Sprague N, 2005 "Modeling the brain's operating system", in *Brain, Vision, and Artificial Intelligence, Conference Proceedings* pp 347 – 366

Biederman I, Mezzanotte R J, Rabinowitz J C, 1982 "Scene perception—detecting and judging objects undergoing relational violations" *Cognitive Psychology* **14** 143 – 177

Carrasco M, Evert D L, Chang I, Katz S M, 1995 "The eccentricity effect: Target eccentricity affects performance on conjunction searches" *Perception & Psychophysics* **57** 1241 – 1261

Findlay J M, Walker R, 1999 "A model of saccade generation based on parallel processing and competitive inhibition" *Behavioural and Brain Sciences* **22** 661 – 721

Hayhoe M, Ballard D, 2005 "Eye movements in natural behavior" *Trends in Cognitive Sciences* **9** 188 – 194

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

Hugli H, Jost T, Ouerhani N, 2005 "Model performance for visual attention in real 3D color scenes", in *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach, Conference Proceedings* Part 2, pp 469 – 478 (*Lecture Notes in Computer Science* 3562)

Itti L, 2005 "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes" *Visual Cognition* **12** 1093 – 1123

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

Itti L, Koch C, 2001 "Computational modelling of visual attention" *Nature Reviews Neuroscience* **2** 194 – 203

Kenner N, Wolfe J M, 2003 "An exact picture of your target guides visual search better than any other representation" *Journal of Vision* **3** 230 (abstract)

Koch C, Ullman S, 1985 "Shifts in selective visual attention: towards the underlying neural circuitry" *Human Neurobiology* **4** 219 – 227

Lee K, Buxton H, Feng H F, 2005 "Cue-guided search: A computational model of selective attention" *IEEE Transactions on Neural Networks* **16** 910 – 924

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

Mannan S, Ruddock K, Wooding D, 1996 "The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images" *Spatial Vision* **10** 165 – 188

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

Nothdurft H C, 2002 "Attention shifts in salient targets" *Vision Research* **42** 1287 – 1306

Parkhurst D, Law K, Niebur E, 2002 "Modeling the role of salience in the allocation of overt visual attention" *Vision Research* **42** 107 – 123

Pomplun M, 2006 "Saccadic selectivity in complex visual search displays" *Vision Research* **46** 1886 – 2000

Potter M C, Staub A, Rado J, O'Connor D H, 2002 "Recognition memory for briefly presented pictures: The time course of rapid forgetting" *Journal of Experimental Psychology: Human Perception and Performance* **28** 1163 – 1175

Rao R P N, Zelinsky G J, Hayhoe M, Ballard D, 2002 "Eye movements in iconic visual search" *Vision Research* **42** 1447 – 1463

Rayner K, 1998 "Eye movements in reading and information processing: 20 years of research" *Psychological Bulletin* **124** 372 – 422

Tatler B W, Baddeley R J, Gilchrist I D, 2005 "Visual correlates of fixation selection: effects of scale and time" *Vision Research* **45** 643 – 659

Torralba A, 2003 "Modeling global scene factors in attention" *Journal of the Optical Society of America A* **20** 1407 – 1418

Treisman A, Gelade G, 1980 "A feature-integration theory of attention" *Cognitive Psychology* **12** 97 – 136

Treue S, 2003 "Visual attention: the where, what, how and why of saliency" *Current Opinion in Neurobiology* **13** 428 – 432

Triesch J, Ballard D H, Hayhoe M M, Sullivan B T, 2003 "What you see is what you need" *Journal of Vision* **3** 86 – 94

Turano K A, Geruschat D R, Baker F H, 2003 "Oculomotor strategies for the direction of gaze tested with a real-world activity" *Vision Research* **43** 333 – 346

Underwood G, Foulsham T, 2006 "Visual saliency and semantic incongruency influence eye movements when inspecting pictures" *Quarterly Journal of Experimental Psychology A* **59** 1931 – 1950

Underwood G, Foulsham T, Loon E van, Underwood J, 2005 "Visual attention, visual saliency, and eye movements during the inspection of natural scenes", in *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach, Conference Proceedings* Part 2, pp 459 – 468 (*Lecture Notes in Computer Science* 3562)

Underwood G, Jebbett L, Roberts K, 2004 "Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search" *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* **57** 165 – 182

Welchman A E, Harris J M, 2003 "Task demands and binocular eye movements" *Journal of Vision* **3** 817 – 830

Williams C C, Henderson J M, Zacks R T, 2005 "Incidental visual memory for targets and distractors in visual search" *Perception & Psychophysics* **67** 816 – 827

Wolfe J M, 1998a "Visual search", in *Attention* Ed. H Pashler (Hove, Sussex: Psychology Press) pp 13 – 71

Wolfe J M, 1998b "What can 1,000,000 trials tell us about visual search?" *Psychological Science* **9** 33 – 39

Yarbus A L, 1967 *Eye Movements and Vision* (New York: Plenum)

Zelinsky G J, Zhang W, Yu B, Chen X, Samaras D, 2005 "The role of top – down and bottom – up processes in guiding eye movements in visual search", in *Neural Information Processing Systems* Eds Y Weiss, B Scholkopf, J Platt (Cambridge, MA: MIT Press) pp 1569 – 1576

Zetzsche C, 2005 "Natural scene statistics and salient visual features", in *Neurobiology of Attention* Eds L Itti, G Rees, J K Tsotsos (London: Elsevier) pp 226 – 233

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION

VOLUME 36 2007

www.perceptionweb.com