Looking as if you know: Systematic object inspection precedes object recognition

Linus Holm Department of Psychology, Umeå University, Umeå, Sweden

Johan Eriksson Department of Psychology, Umeå University, Umeå, Sweden

Linus Andersson Department of Psychology, Umeå University, Umeå, Sweden

Sometimes we seem to look at the very object we are searching for, without consciously seeing it. How do we select object relevant information before we become aware of the object? We addressed this question in two recognition experiments involving pictures of fragmented objects. In Experiment 1, participants preferred to look at the target object rather than a control region 25 fixations prior to explicit recognition. Furthermore, participants inspected the target as if they had identified it around 9 fixations prior to explicit recognition. In Experiment 2, we investigated the influence of semantic knowledge in guiding object inspection prior to explicit recognition. Consistently, more specific knowledge about target identity made participants scan the fragmented stimulus more efficiently. For instance, non-target regions were rejected faster when participants knew the target object's name. Both experiments showed that participants were looking at the objects as if they knew them before they became aware of their identity.

Keywords: object recognition, eye movement, visual awareness, attention

Citation: Holm, L., Eriksson, J., & Andersson, L. (2008). Looking as if you know: Systematic object inspection precedes object recognition. *Journal of Vision, 8*(4):14, 1–7, http://journalofvision.org/8/4/14/, doi:10.1167/8.4.14.

Introduction

At one instance, there is just a bunch of objects in the refrigerator, in the next, you clearly discern the package of milk. Do we systematically inspect the object before we become aware of its presence, or do we see it the moment we lay eyes on it? Several object recognition studies suggest there is little need for systematic inspection in object recognition. For instance, Potter and Faulconer (1975) showed that humans can identify objects presented for only 50 ms in rapid serial visual streams (i.e., one stimulus picture immediately follows the next). Consistently, eye movement studies of ambiguous figures indicate that attention and perception are tightly linked to the extent that it remains unclear which is cause and effect (Ellis & Stark, 1978; Ito et al., 2003; Pomplun, Ritter, & Velichovsky, 1996). Furthermore, in visual search studies, participants generally fixate the target at the instant they overtly indicate its presence with a button press. It has even been suggested that detection and recognition are parts of the same process (Grill-Spector & Kanwisher, 2005).

Although detection and explicit recognition can be closely related in some instances (Grill-Spector & Kanwisher 2005), more difficult situations might tease them apart. Indeed, a few studies using more complex stimulus (McCarley, Kramer, Wickens, Vidoni, & Boot, 2004; Pomplun et al., 1996) indicated that explicit recognition and deployment of attentional resources can be dissociated to some extent. If the information gained from each fixation is poor, it seems reasonable to aggregate it across fixations to make a more accurate recognition decision.

The primary objective of this study was to test the hypothesis that information is gradually accumulated before explicit object recognition. For this purpose, we registered eye movements in a fragmented object recognition task. Fragmented objects are well suited for this purpose because there is no immediate pop-out of the target following presentation and there is only one single transition in perceptual awareness of the target object (i.e., explicit recognition). In addition, the change in perceptual awareness seems to occur very fast (see Figure 1), and similar perceptual transitions in ambiguous figures have been shown to occur within a few 100 ms. (Kornmeier & Bach, 2004). Therefore, systematic inspection of the object preceding explicit object recognition should reflect selection processes at a stage where the observer is yet agnostic about the object's identity. That is, systematic eye movements on the object prior to recognition should not reflect a gradual build up of perceptual awareness because the transition in awareness is expected to be instantaneous (see also Eriksson, Larsson, Riklund-Åhlström, & Nyberg, 2004, regarding neural correlates of the transition in visual awareness of fragmented objects).

doi: 10.1167/8.4.14 Received June 8, 2007; published April 18, 2008 ISSN 1534-7362 © ARVO

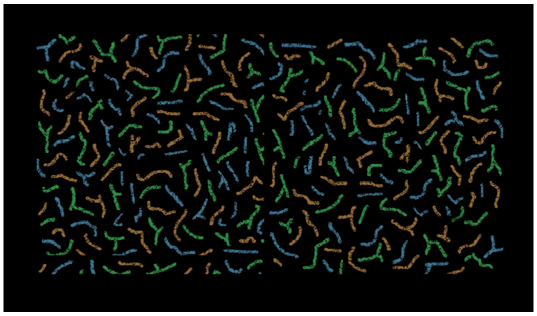

Figure 1. An example of an average difficult stimulus picture (55% of participants identified it in Experiment 1). Overall average performance for the stimuli was 51%. Can you find the blue horse to the left?

When objects are not immediately discernible, the observer has to decide which information to select for further processing in order to recognize the content. Under such uncertain conditions, the observer might draw on prior knowledge to guide information selection. Recognition would then resemble hypothesis testing (Bar et al. 2006; Deco & Shürmann, 2000; Kersten, Mamassian, & Yuille, 2004; Torralba, Oliva, Castelhano, & Henderson, 2006) with the eyes. Provided the observer has the right hypothesis, he would structure the visual input by moving his eyes to diagnostic regions for confirmation. The structured information hence accumulated would then be adequate for determining the objects' identity.

A secondary objective of this study was to test whether eye movements are guided by prior knowledge in object recognition. Decisions about where to look next could be influenced by perceptual information such as gestalt properties. Alternatively, eye movements might be guided by semantic knowledge (e.g., looking for horse-like shapes in the picture). For instance, when the target's identity is uncertain, one might expect extended periods of viewing in limited regions due to increased iteration between perception and memory (e.g., testing different semantic interpretations of the object). Instead, when one knows the identity of the object beforehand, a more sweeping pattern of eye movements would be expected, reflecting search in the scene rather than in memory (see also Oliva, Wolfe, & Arsenio, 2004).

In Experiment 1, we tested the hypothesis that information is systematically sampled prior to recognition in a fragmented object recognition test. If object recognition is a function of progressive object information accumulation, then eye movements should increasingly fixate the object prior to the recognition decision. Alternatively, recognition decisions might be independent of preceding information, and there is no increase in object fixations prior to recognition.

In Experiment 2, we tested the influence of prior knowledge on information acquisition by manipulating the accuracy of prior target information. If object recognition is characterized by hypothesis testing, and semantic information affects perception (see e.g., Dolan et al., 1997; Vecera & Farah, 1997) and attention deployment (Moores, Laiti, & Chelazzi, 2003), then accurate prior knowledge should facilitate object recognition by guiding the eyes to diagnostic object regions.

Experiment 1

In Experiment 1, we tested the hypothesis that explicit recognition (i.e., the point in time when participants become aware of the objects presence) is preceded by systematic object information selection in a fragmented object recognition test. Explicit recognition should be preceded by eye movements on the object resembling the scan pattern exhibited after explicit recognition. In other words, participants should look at the object as if they had recognized it before they explicitly recognize it. Systematic information selection would be expressed as consecutive eye movements on the object.

Method

Participants

Twenty-one Umeå University students participated in the experiment (13 females). They were between 19 and 32 years old (M = 24.8). Participants had normal or corrected-to-normal visual acuity (with contact lenses) and color vision (assessed by the Ishihara color test).

Stimulus

A set of 20 fragmented animal pictures were selected from the Eriksson et al. (2004) study. The pictures consisted of four footed animals (16), birds (3), and other (snail, 1). In order to avoid confounding target fixations with general tendencies to favor central regions in scenes (Mannan, Ruddock, & Wooding, 1997), the fragmented animals were positioned to the right or the left in the picture (see Figure 1). In the other half of the picture, a 180° or 90° reoriented copy of the first half was inserted, but with the animal-related fragments altered to make recognition impossible (see Figure 1). This way, each half of a stimulus contained similar basic level visual parameters. Each animal was defined by one of three colors (blue, brown, or green). The animals covered on average 9% (SD = 2%) of the screen area. Five different lists of presentation order were created.

Apparatus

An EyeLink I eye-tracker (SR Research Ltd.) with a sampling rate of 250 Hz and a spatial precision of approximately 0.5 degrees was used to monitor the subjects' eye movements. An eye movement was defined as a saccade when its distance exceeded .15° or its velocity reached 35°/s or when its acceleration reached $9500°/s^2$. The pictures were presented in 800 × 600 pixel resolution on a 19-in. CRT screen with an 85-Hz refresh rate.

Procedure

The participants were seated approximately 50 cm away from the computer screen with a chin support which provided constant viewing distance and reduced head movement. They were instructed to identify fragmented line drawings of animals. Before test, participants were shown an example of the stimuli and had to search for the target animal. All participants managed to identify the animal after the experimenter had disclosed its color and location.

At test, each of the twenty stimulus pictures was presented for 30 s. Each stimulus picture was only presented once to every participant. Participants were told to identify the animal as quickly and accurately as possible. Participants indicated their recognition by pressing a button on a response box. After recognition, participants were instructed to keep looking at the target until the picture disappeared. After picture removal, participants reported the animal.

Results and discussion

One participant was excluded from analysis, having misinterpreted the instructions. The average performance of the remaining 20 participants was 10.2 items recognized (*SD* = 3.62) of 20 possible. Participants recognized the animals on average after 14.4 s (*SD* = 2.5) corresponding to 44.1 fixations (*SD* = 11.3). All reported items were correctly identified, except for two responses (one participant identified a snail as an elephant and a gorilla as a dog), corresponding to 1% of responses.

To analyze eye movement data, the target region of each stimulus was first defined by a polygon circumscribing the outlines of the animal with an eccentricity corresponding to 0.5° (see yellow outline in Figure 2). For comparison, we created a control region of exactly the

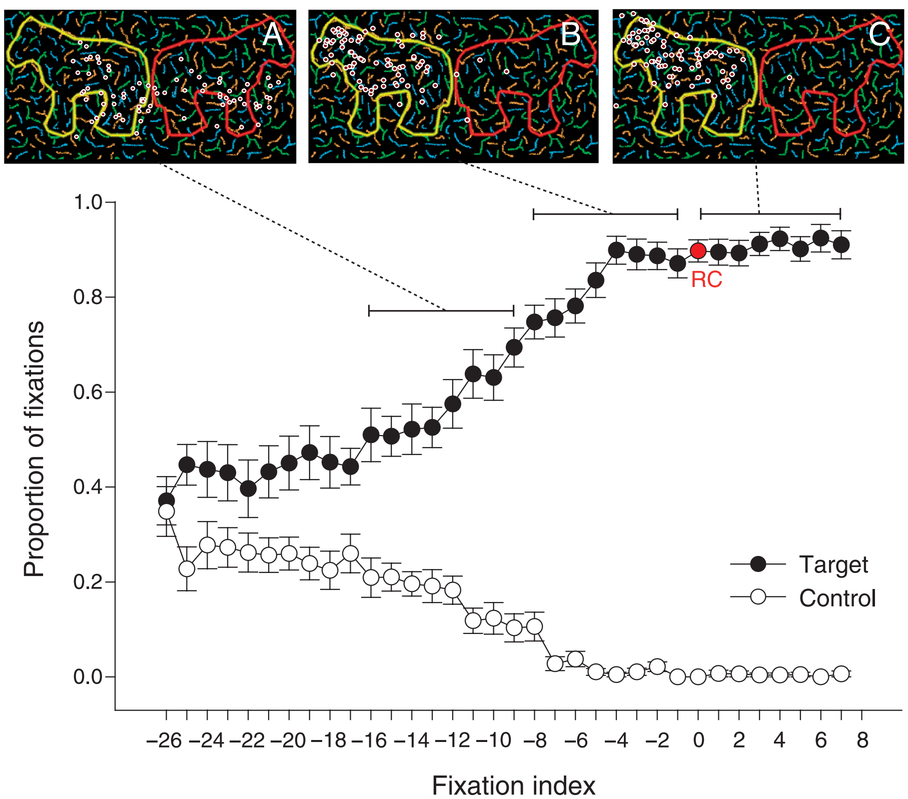

Figure 2. The three stimulus pictures on top are overlaid with aggregated fixations from the ten participants who identified this particular item. Fixations in pictures A–C are reflected as dots, with an equal number of fixations in each picture. The yellow and red outlines define the target and control regions, respectively. Picture A illustrates fixations 16 to 9 before target recognition. Picture B illustrates the 8 to 1 fixations preceding target recognition, and picture C illustrates the 0 to 7 fixations following target recognition. The graph illustrates target and control region fixation probability as a function of fixation index. Data are synchronized around recognition (RC, fixation index = 0). Error bars indicate *SEM*.

same size as the target polygon but mirror-reversed to the opposite half of the picture (see red outline in Figure 2). The mirror control region hence constituted a spatial control region, although it did not contain any structured object information. First fixation in the target region after stimulus onset occurred on average 5.65 fixations after stimulus onset. The corresponding average for the control region was 6.82. The difference was not statistically reliable, $t(19) = 1.23$, $p = .23$, indicating that the target region did not stand out as perceptually salient over the control region.

As suggested by Figure 2, the target region was increasingly fixated in favor of the control region, as participants approached explicit recognition. Data presented in Figure 2 were limited to the last 26 fixations prior to recognition because earlier data points would have been based on few observations (due to fast responses) and were hence less reliable. However, it seems like the proportion of target and control fixations would have converged on a similar proportion, had earlier fixation data been available. This conclusion is also justified by the finding that there was no significant difference in number of fixations before first entry into the target and control regions, respectively. The probability of fixating the target region rather than the control region was statistically significant at 25 fixations prior to recognition, $t(19) = 2.77$, $p < .05$.

As indicated by Figure 2, participants reached ceiling for proportion of target fixations several fixations prior to explicit recognition. This suggests that participants had detected and systematically inspected the target region before they became aware of its semantic content. Consistently, the average number of consecutive fixations on target (termed "gaze" henceforth) preceding recognition was 9.2 ($SD = 3.5$).

Experiment 2

Experiment 1 showed that object information was systematically selected prior to explicit object recognition. Participants inspected the objects so closely before their overt recognition decisions, that it seems possible they were accurately hypothesis testing with their eyes, guided by visual characteristics such as gestalt properties or even by semantic interpretations. Ryan, Althoff, Whitlow, and Cohen (2000) found that non-amnesic participants revisited changed scene regions more often than amnesic patients did, even when neither participant category was able to explicitly identify the change. Therefore, it seems like memory might guide the eyes to diagnostic regions even in the absence of target awareness. Similarly, Hollingworth, Williams, and Henderson (2001) found prolonged gaze on changed target regions in a change blindness paradigm, even when participants were unaware of the change.

To strongly test the hypothesis testing account of information selection, we manipulated participants' prior object knowledge. If information selection is guided by prior knowledge, more accurate knowledge should produce more efficient information selection. Specifically, when the observer has the right hypothesis regarding the objects' identity, there is less need to recruit long-term memory in recognition because the right representation is already retrieved. Therefore, visual features would be more efficiently interpreted with accurate object knowledge. This might be translated into shorter gazes, as the observer is able to reject or to confirm target presence in a region based on less exchange between perception and memory. Instead, under less accurate knowledge conditions, the observer has to compare the visual information to a large set of identities in semantic memory. This should produce more iteration between visual information selection and memory, as alternative object interpretations need to be ruled out. Hence, longer gazes to both control and target regions would be expected under less specific target knowledge.

If only gestalt properties influenced eye movements, there should be no difference in gaze between conditions.

Methods

Participants

Sixteen Umeå University undergraduates (7 females) participated in the experiment for payment (approximately USD 6). They were between 21 and 32 years ($M = 24.3$) and had no prior experience of similar experiments. Participants had self reported normal or corrected-to-normal visual acuity (contact lenses) and color vision (assessed with the Ishihara color test).

Stimulus

The stimulus set of Experiment 1 was extended with 20 fragmented pictures from the pool of pictures used by Eriksson et al. (2004). These additional items depicted non-living, man-made objects. The objects (number of items in parenthesis) consisted of furniture (3), vehicles (5), weapon (1), musical instruments (3) articles of clothing (4), tools (3), and other (anchor, 1). The new items were prepared in the same manner as in Experiment 1. The new target objects were of similar size to those of Experiment 1. Half of the items were presented following a text label indicating its superordinate category membership (e.g., animal), whereas the other half was preceded by its basic level name (e.g., monkey). Across participants, each item was presented equally often following the superordinate as the basic level labels. In addition to the items from Eriksson et al. (2004), a set of ten practice stimulus items were created from Snodgrass and Vanderwart (1980) along the same principles mentioned above.

Procedure

The procedure of Experiment 1 was used with the following exceptions. To maximize the number of recognized items and hence obtain more data, participants were given training in identifying fragmented objects on the practice set before test. At test, each fragmented picture was preceded by a label appearing for 3 s. Fragmented pictures were presented for 60 s, or until recognition as indicated with a button press. The pictures were presented for an additional 3 s after button press in order to sample data post recognition. After each trial, participants were instructed to report the identity, the position, and the color of the target item.

Results and discussion

One participant abandoned the experiment due to a headache. Recognition scores for the remaining 15 participants were based on correct reports of name, color, and position of the target item. On average, participants correctly recognized 18.9 and 15.8 out of 20 items from the basic and the superordinate conditions, respectively. The difference in performance between the conditions was statistically reliable, $t(14) = 5.43$, $p < .001$.

The same overall pattern of results from Experiment 1 was replicated in that participants increasingly favored the target region rather than the control region prior to explicit recognition (see Figure 3). Furthermore, there was a steeper increase in proportion of target fixations in the basic condition compared to the superordinate condition, suggesting that specificity of prior target knowledge affected how participants scanned the stimulus. Specifically, it seems like recognition in the basic condition required less target information sampling than the superordinate condition.

The first gaze on the target regions averaged 3.70 and 5.66 fixations in the basic and the superordinate conditions respectively, $t(14) = 3.56$, $p < .01$. First gaze towards the control region was 2.46 and 2.91 fixations for the basic and the superordinate conditions, respectively, $t(14) = 2.88$, $p < .05$. That is, the control region was faster rejected when the observer had more specific target knowledge. Finally, the last gaze prior to recognition was 5.49 and 9.12 for the basic and the superordinate conditions, respectively, $t(14) = 4.08$, $p < .001$.

General discussion

This study investigated how visual information is selected prior to explicit object recognition. Both experiments showed that the fragmented objects were inspected several fixations before explicit object recognition. In Experiment 1, the target region was favored over the similarly sized control region even 25 fixations before target recognition. Interestingly, participants often made saccades to the center of the fragmented target (see picture B in Figure 2), which suggests that they had inferred the

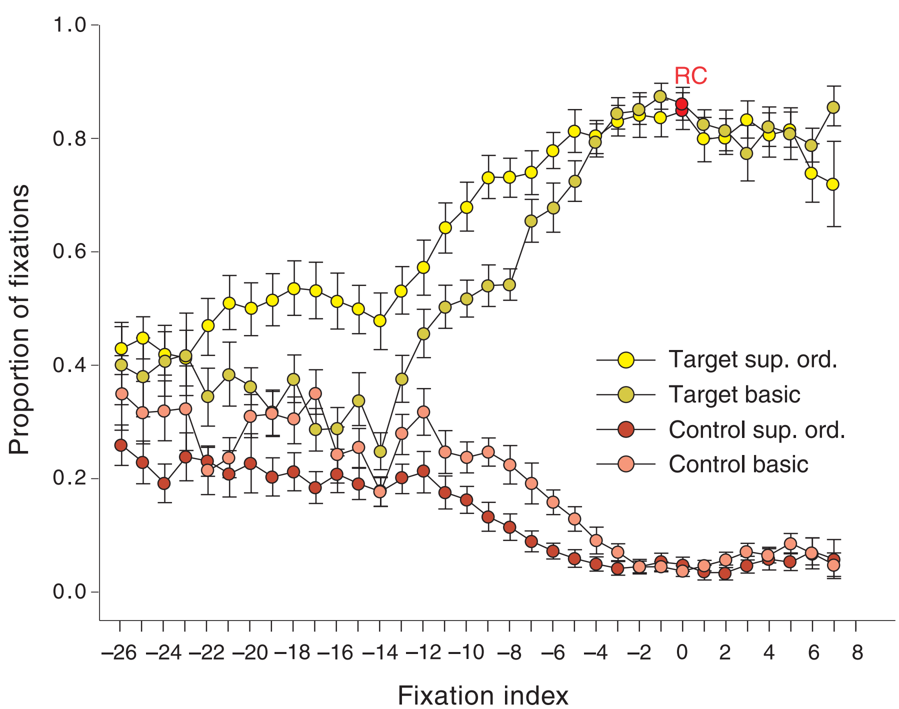

Figure 3. The graph illustrates target and control region fixation probability as a function of fixation index for the basic and superordinate condition, respectively. Data are synchronized around recognition (RC, fixation index = 0). Error bars indicate *SEM*.

target prior to explicit recognition (Melcher & Kowler, 1999). Furthermore, both experiments showed that the target region was inspected very systematically before explicit recognition because target recognition was preceded by several consecutive target fixations. Specifically, objects were fixated as if the observer had recognized the objects around nine fixations (i.e., last gaze) prior to explicit object recognition in Experiment 1. An alternative account would be that participants were just slow to overtly respond following identification and continued to look at the object while preparing their response. However, that interpretation seems unlikely, considering that the time required for nine fixations is much longer than what even a generous estimate of reaction time would suggest.

There was no immediate preference for the target region as compared to the similarly sized control region in Experiment 1; participants' fixated target and control regions equally fast after stimulus onset. This result is inconsistent with a perceptual saliency account of the systematic target inspection prior to recognition. Instead, the results indicate that the processes preceding explicit recognition are influenced by more complex visual information (e.g., gestalts) or hypotheses based on prior knowledge (e.g., semantic memory).

Experiment 2 showed that prior target knowledge affected information selection as indicated by eye movements. With basic level knowledge about object identity, participants employed a more sweeping pattern of fixations across the stimulus pictures compared to superordinate object knowledge. Specifically, participants would both reject and confirm target presence more readily in the basic condition because gazes on both control and target regions were shorter compared to the superordinate condition. This finding is also consistent with knowledge facilitating the ability to group visual contour information (for a discussion, see Geisler, Perry, Super, & Gallogly, 2001).

The difference in number of object fixations prior to recognition might reflect a response bias between the basic and the superordinate conditions (Hollingworth & Henderson, 1998). Specifically, with only superordinate information, participants could have been more conservative in responding and therefore made more fixations on the object before responding. However, this interpretation is not consistent with the finding that participants inspected the control region more meticulously in the superordinate condition.

Because the present study used highly artificial stimulus, its generalization to object recognition in more natural conditions is limited. However, even under more natural conditions, object recognition might involve several fixations and iterative information sampling. For instance, recognizing an animal in a tree or making fine within-category judgments such as finding the right utensil in the drawer seem to resemble the situation tested in the present experiments. Future experiments might investigate difficult recognition tasks in natural conditions.

Conclusions

Collectively, our findings suggest that information can be accumulated across fixations prior to explicit object recognition. Furthermore, the findings support the view that information selection in object recognition is carried out in a hypothesis testing manner (Deco & Schürman, 2000; Friston, 2003). Our results support recent findings on predictive coding (Friston, 2003) in object recognition (Summerfield, Egner, et al. 2006; Summerfield, Lepsien, Gitelman, Mesulam, & Nobre, 2006), suggesting that expectations guide attention in object recognition and visual search (see also Holm & Mäntylä 2007; Moores et al., 2003). Importantly, predictive information sampling seems to precede explicit object recognition, suggesting that object recognition is a function of implicit inference (Helmholtz, 1910). Therefore, we look as if we know before we know.

Acknowledgments

We are thankful for the valuable comments by Timo Mäntylä, Bruno Laeng, and Stephen Engel in the preparation of this manuscript. The findings described have been reported in a doctoral thesis (Holm, 2007).

Commercial relationships: none.
Corresponding author: Linus Holm.
Email: linus.holm@psy.umu.se.
Address: Department of Psychology, Umeå University, S-901 87 Umeå, Sweden.

References

Bar, M., Kassam, K. S., Ghuman, A. S., Boschyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Science of the United States of America, 103,* 449–454. [PubMed] [Article]

Deco, G., & Schürmann, B. (2000). A neuro-cognitive visual system for object recognition based on testing of interactive attentional top-down hypotheses. *Perception, 29,* 1249–1264. [PubMed]

Dolan, R. J., Fink, G. R., Rolls, E., Booth, M., Holmes, A., Franckowiak, R. S., et al. (1997). How the brain learns to see objects and faces in an impoverished context. *Nature, 389,* 596–599. [PubMed]

Ellis, S. R., & Stark, L. (1978). Eye movements during the viewing of Necker cubes. *Perception, 7,* 575–581. [PubMed]

Eriksson, J., Larsson, A., Riklund-Åhlström, K., & Nyberg, L. (2004). Visual consciousness: Dissociating the neural correlates of perceptual transitions from sustained perception with fMRI. *Consciousness and Cognition, 13,* 61–72. [PubMed]

Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16,* 1325–1352. [PubMed]

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41,* 711–724. [PubMed]

Grill-Spector, K., & Kanwisher, N. (2005). As soon as you know it is there, you know what it is. *Psychological Science, 16,* 152–160. [PubMed]

Helmholtz, H. von (1910). *Handbuch der physiologischen optik: Vol. 3* (3rd ed.). Leipzig: Leopold Voss.

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

Hollingworth, A., Williams, C. C., & Henderson, J. M. (2001). To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review, 8,* 761–768. [PubMed]

Holm, L. (2007). *Predictive eyes precede retrieval: Visual recognition as hypothesis testing*. Unpublished doctoral dissertation, Umeå University, Sweden.

Holm, L., & Mäntylä, T. (2007). Memory for scenes: Refixations reflect retrieval. *Memory & Cognition, 35,* 1664–1674. [PubMed]

Ito, J., Nikolaev, A. R., Luman, M., Aukes, M. F., Nakatani, C., & van Leeuwen, C. (2003). Perceptual switching, eye movements, and the bus paradox. *Perception, 32,* 681–698. [PubMed]

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55,* 271–304. [PubMed]

Kornmeier, J., & Bach, M. (2004). Early neural activity in Necker-cube reversal: Evidence for low-level processing of a gestalt phenomenon. *Psychophysiology, 41,* 1–8. [PubMed]

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception, 26,* 1059–1072. [PubMed]

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15,* 302–306. [PubMed]

Melcher, D., & Kowler, E. (1999). Shapes, surfaces and saccades. *Vision Research, 39,* 2929–2946. [PubMed]

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual attention. *Nature Neuroscience, 6,* 182–189. [PubMed]

Oliva, A., Wolfe, J. M., & Arsenio, H. C. (2004). Panoramic search: The interaction of memory and vision in search through a familiar scene. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 1132–1146. [PubMed]

Pomplun, M., Ritter, H., & Velichovsky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception, 25,* 941–948. [PubMed]

Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature, 253,* 437–438. [PubMed]

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science, 11,* 454–461. [PubMed]

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174–215. [PubMed]

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science, 314,* 1311–1314. [PubMed]

Summerfield, J. J., Lepsien, J., Gitelman, D. R., Mesulam, M. M., & Nobre, A. C. (2006). Orienting attention based on long-term memory experience. *Neuron, 49,* 905–916. [PubMed] [Article]

Torralba, A., Oliva, A., Castelhano, S. M., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113,* 766–786. [PubMed]

Vecera, S. P., & Farah, M. J. (1997). Is visual image segmentation a bottom-up or an interactive process?. *Perception & Psychophysics, 59,* 1280–1296. [PubMed]