Knowledge-driven perceptual organization reshapes information sampling via eye movements

Marek A. Pedziwiatr [1,2], Elisabeth von dem Hagen [3], and Christoph Teufel [1]

[1] Biological and Computational Vision Lab, Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University

[2] Department of Biological and Experimental Psychology, Queen Mary University of London

[3] Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University

Word count: 13,728

Author Note

Marek A. Pedziwiatr https://orcid.org/0000-0002-3959-8666

Elisabeth von dem Hagen https://orcid.org/0000-0003-1056-8196

Christoph Teufel https://orcid.org/0000-0003-3915-9716

We have no conflicts of interest to disclose.

Data from this study is openly available under the following link: [link to be provided upon publication]

Correspondence concerning this article should be addressed to Marek A. Pedziwiatr and Christoph Teufel, Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff University, Maindy Road, Cardiff, Wales, UK, CF24 4HQ. Emails: m.pedziwiatr@qmul.ac.uk and teufelc@cardiff.ac.uk

Abstract

Humans constantly move their eyes to explore the environment. However, how image-computable features and object representations contribute to eye-movement control is an ongoing debate. Recent developments in object perception indicate a complex relationship between features and object representations, where image-independent object-knowledge generates objecthood by reconfiguring how feature space is carved up. Here, we adopt this emerging perspective, asking whether object-oriented eye-movements result from gaze being guided by image-computable features, or by the fact that these features are bound into an object representation. We recorded eye movements in response to stimuli that initially appear as meaningless patches but are experienced as coherent objects once relevant object-knowledge has been acquired. We demonstrate that fixations on identical images are more object-centred, less dispersed, and more consistent across observers once these images are organised into objects. Gaze guidance also showed a shift from exploratory information-sampling to exploitation of object-related image-areas. These effects were evident from the first fixations onwards. Importantly, eye-movements were not fully determined by knowledge-dependent object representations but were best explained by the integration of these representations with image-computable features. Overall, the results show how information sampling via eye-movements is guided by a dynamic interaction between image-computable features and knowledge-driven perceptual organization.

Keywords: eye movements; perceptual organization; prior knowledge; object perception; natural scenes

Public Significance Statement: To explore and make sense of the world around us, we have to move our eyes. This study shows how our brain combines simple image-features such as edges and contrast with knowledge about objects to guide our eyes through a visual scene.

Knowledge-driven perceptual organization reshapes information sampling via eye movements

Human visual experience carves up the world into objects (Feldman, 2003; Wagemans et al., 2012), distinct entities that are critical in structuring our interaction with the environment. When searching for a specific item in a scene or when exploring the world with no purpose other than to obtain information, humans tend to look at the centre of objects (e.g., Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013; Stoll, Thrun, Nuthmann, & Einhäuser, 2015). While these object-oriented effects of information sampling are well established, the current literature provides little consensus about which specific aspects of objects influence programming of eye movements (Borji & Tanner, 2016; Federico & Brandimonte, 2019; Hayes & Henderson, 2021; Henderson, Malcolm, & Schandl, 2009; Kilpelaïnen & Georgeson, 2018; Nuthmann, Schütz, & Einhäuser, 2020; Van der Linden, Mathôt, & Vitu, 2015). This issue is complicated by the fact that it is often not clear exactly what constitutes an ‘object’ or how objects relate to image-computable features: except for special cases such as hallucinations (Horga & Abi-Dargham, 2019; Powers, Mathys, & Corlett, 2017; Teufel et al., 2015), features are necessary for visual object representations to arise but they are often not sufficient. Indeed, a growing number of studies using human psychophysics (Christensen, Bex, & Fiser, 2015; Lengyel, Nagy, & Fiser, 2021; Lengyel et al., 2019; Neri, 2017; Ongchoco & Scholl, 2019; Teufel, Dakin, & Fletcher, 2018) neuroimaging (Flounders, González-García, Hardstone, & He, 2019; Hsieh, Vul, & Kanwisher, 2010) , and animal electrophysiology (Gilbert & Li, 2013; Liang et al., 2017; Self et al., 2019; Self, van Kerkoerle, Supèr, & Roelfsema, 2013; Walsh, McGovern, Clark, & O’Connell, 2020) suggest that in order for object representations to emerge, prior object-knowledge has to interact with sensory processing. By contrast to conventional models of object recognition (DiCarlo, Zoccolan, & Rust, 2012; Kourtzi & Connor, 2011; Kriegeskorte, 2015; Marr & Nishihara, 1978), these studies demonstrate that prior object-knowledge effectively generates objecthood by reconfiguring sensory mechanisms that process visual inputs, thereby changing how feature space is carved up into meaningful units (Teufel & Fletcher, 2020). In other words, a given cluster of features is an object not by virtue of the features themselves but because these features are *represented as an object*. In the current study, we demonstrate that this objecthood, i.e., the fact that certain features are bound into an object representation, affects eye movements. Specifically, we show that the dynamic re-shaping of feature space by knowledge-driven perceptual organization that underlies the emergence of objecthood has a substantial influence on information sampling via eye movements in human observers.

The most influential early saliency models – that is, computational methods used to predict human eye-movements – largely disregarded objects, arguing that programming of eye-movements is determined by an analysis of low-level features such as luminance, colour, and orientation (Harel, Koch, & Perona, 2007; Itti & Koch, 2001). According to these early accounts, the visual system computes feature maps, which highlight areas in the image that attract fixations (Zelinsky & Bisley, 2015). Over the past 15 years, however, several studies have emphasised the importance of objects in guiding information sampling (Einhäuser, Spain, & Perona, 2008; Hayes & Henderson, 2021; Hwang et al., 2011; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013; Pilarczyk & Kuniecki, 2014; Stoll et al., 2015). For instance, in one of the early studies, Einhäuser and colleagues (2008) found that maps of object locations outperform maps derived from a low-level feature model in predicting human fixations. Moreover, human observers show a tendency to look at the centre of objects rather than their edges, contrasting with predictions from early low-level feature models (Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013; Stoll et al., 2015; Borji & Tanner, 2016; see also Vincent, Baddeley, Correani, Troscianko, & Leonards, 2009). These effects have been interpreted as demonstrations of the importance of objects in oculomotor control.

Other lines of evidence suggest that the fact that human observers primarily fixate at object locations can be explained by low-level mechanisms (Borji et al., 2013; Elazary & Itti, 2008; Kilpelaïnen & Georgeson, 2018; Masciocchi et al., 2009). For instance, a recent attempt to assess the unique contribution of features vs. objects to oculomotor control suggests that object-centred effects are, at least partly, driven by low-level features that correlate with objects (Nuthmann et al., 2020). This conclusion is in line with a careful psychophysical study, suggesting that the tendency of human observers to focus on the centre of objects might be controlled by a relatively simple process that programs eye-movements towards homogeneous luminance surfaces on the basis of luminance-defined edges (Kilpelaïnen & Georgeson, 2018). This result provides a potential mechanism for the finding that fixations that occur shortly after image onset tend to be located close to the stimulus centre not only for objects but also for non-objects if low-level properties are matched (Van der Linden et al., 2015). Together, these results suggest that the tendency to fixate on the centre of objects might not be related to objecthood itself but is controlled by mechanisms that respond to relatively low-level features in the input. Note, however, that the study by van der Linden and colleagues (2015) also suggests that guidance of eye-movements that are generated later after image onset might be affected by semantic aspects of object. This finding potentially indicates a time course according to which locations of early fixations are mainly determined by low-level, image-computable features while locations of later fixations might be determined by high-level object representations (see also Anderson, Ort, Kruijne, Meeter, & Donk, 2015 and Wolf & Lappe, 2021).

Many previous studies that aim to show the contribution of objects to oculomotor control relied on a comparison of eye movements to saliency models that calculate image-computable feature maps as their null hypothesis (for example, Einhäuser et al., 2008; Pilarczyk & Kuniecki, 2014; Stoll et al., 2015). This approach has led to important insights regarding oculomotor control but is hampered by the fact that the specific methodological choices regarding the type of saliency model and object map are critical in determining the interpretation. In fact, in the previous literature, the use of different models has led to categorically different conclusions, even if they have been applied to identical or very similar data sets (Borji et al., 2013; Einhauser, 2013; Einhäuser et al., 2008; Henderson et al., 2021; Henderson & Hayes, 2017; Pedziwiatr et al., 2021a, 2021b; Stoll et al., 2015). Importantly, independently of the favoured interpretation of these findings, there is a more fundamental aspect that is easily overlooked. Specifically, contrasting outputs of low-level feature models with 'objects', and the tendency to conceptualise these as categorically different – although possible to reconcile (Borji & Tanner, 2016; Nuthmann et al., 2020; Stoll et al., 2015) – interpretations**,** has concealed a fundamental similarity between these explanations. Namely, comparable to how low-level models deal with simple features, most studies implicitly treat 'objects' as image-computable properties. This notion is also the basis for state-of-the-art computer vision models that aim to predict human fixations (e.g., Kroner, Senden, Driessens, & Goebel, 2020; Kümmerer et al., 2017a): these models use deep convolutional neural networks trained on object recognition to extract high-level features that are directly computed from the image. In other words, the different approaches in the current eye-movement literature can be understood as lying on a continuum, with their position being defined by the type of features they emphasise. This notion is made explicit in a recent study by Schütt and colleagues (Schütt, Rothkegel, Trukenbrod, Engbert, & Wichmann, 2019): the authors explicitly conceptualised objects as high-level features that are computed in a bottom-up fashion, and contrasted their contribution to the guidance of eye-movements with the contribution of low-level features.

While the theoretical precision of the study by Schütt and colleagues is exceedingly helpful in clarifying the different positions, conceptualising objects as high-level features directly conflicts with current developments in object perception. Two aspects of the complex relationship between features and objects are particularly relevant: first, several recent studies demonstrate that features are not always sufficient for object representations to arise (Flounders et al., 2019; Hsieh et al., 2010; Lengyel et al., 2019, 2021; Ongchoco & Scholl, 2019; Teufel et al., 2018). Rather, objecthood emerges as a consequence of the interaction between current visual input and perceptual organization processes that are based on prior object-knowledge. Second, once object representations have been generated, top-down influences reconfigure the way in which even some of the earliest cortical mechanisms process low-level visual features (Christensen et al., 2015; Flounders et al., 2019; Hsieh et al., 2010; Lengyel et al., 2021, 2019; Neri, 2014, 2017; Ongchoco & Scholl, 2019; Teufel et al., 2018). For instance, psychophysical studies show that early feature-detector units are sharpened for currently relevant input based on top-down influences from object representations (Teufel et al., 2018). This reconfiguration of information processing is detectable in early retinotopic cortices (Flounders et al., 2019; Hsieh et al., 2010). Overall, these findings thus cast serious doubt on the notion that the human visual system computes image features independently of the inferred object structure of the environment (Neri, 2017).

This novel perspective of object perception has fundamental implications for our understanding of information sampling via eye movement. First, if objecthood emerges from the interaction between features and prior knowledge, then the question of whether objects guide eye movements cannot be answered by an approach that exclusively focuses on how image-computable feature space is carved up by the visual system, regardless of whether the considered features are low- or high-level. Second, the novel perspective of object perception means that a full understanding of the role of objects in eye-movement control has to move away from regarding feature space as static, instead taking into account the plasticity of low-level sensory processing introduced by dynamic interactions with object representations. Here we address both of these issues. We analysed gaze data from human observers viewing stimuli, which, on initial viewing, are experienced as a collection of meaningless black and white patches. After gaining relevant object knowledge, however, the observers' visual system organizes the sensory input into meaningful object representations. These stimuli allow us to test the hypothesis that eye-movements are guided by objecthood per se – i.e., the fact that certain features are *represented as an object* – rather than by the high-level features associated with objects. Across three experiments (see Fig. 3 for a roadmap through them), we demonstrate that, consistent with our hypothesis, the knowledge-driven perceptual organization of identical inputs substantially re-shapes eye-movement patterns, with the selection of fixation locations being driven by a combination of image-computable features and the knowledge-dependent object representations. Moreover, these effects are already present at the first fixation. In summary, we show that a fundamental human visual behaviour – information sampling via eye movements – is guided by a dynamic interaction between image-computable features and object representations that emerge when prior object-knowledge restructures sensory input.

Experiment 1 – Methods

Overview

In Experiment 1, observers viewed black and white two-tone images while their eye movements were recorded. Two-tone images are derived from photographs of natural scenes ('templates'). Each two-tone appears as meaningless patches on initial viewing. Once an observer has acquired relevant prior object-knowledge by viewing the corresponding template, however, processes of perceptual organization in the visual system bind the patches of the two-tone image into a coherent percept of an object (see caption of Figure 1 for instructions of how to experience the effect).

Figure 1

Example of a two-tone image

Note. On initial viewing, this image appears as meaningless black and white patches. To be able to perceptually organize it into a meaningful percept, the reader is advised to first carefully look at the template image from which this two-tone was derived, presented on Fig. 2. An animated version of the blending between this two-tone and its template is provided in Supplementary Materials. Note that the example two-tone image is for illustration only, it was not used in the study. Image copyrights owner: author C. T.

Two-tone images provide a tool to manipulate object perception without changing the visual features of the stimulus. They are therefore ideally suited to test the hypothesis that human oculomotor control is determined by object representations that are not constituted by image-computable features but emerge via an interaction between image-computable features and prior object-knowledge. According to this idea, eye movements in response to two-tone images should be influenced by whether the observer experiences the input as an object percept. Specifically, patterns of fixations on identical two-tone images should be more similar to the ones from the corresponding template when an observer experiences the two-tone image as a meaningful object percept compared to when they experience it as meaningless patches.

To test these predictions, we recorded eye-movements of 36 human observers who viewed two-tone images before and after being exposed to the relevant templates (Before, After, and Template conditions, respectively; see Fig. 2). In the Before condition, observers perceive two-tone images as meaningless black and white patches. In the After condition, prior object-knowledge allows them to bind patches into meaningful object percepts. Crucially, any potential differences in eye movements between the Before and the After conditions cannot be explained by image-computable features because these are identical across these conditions; the only aspect that has changed is the prior object-knowledge that observers have access to. Experiment 1 established the key effects; to exclude alternative explanations, we conducted Experiments 2 and 3 (see Fig. 3 for design details). The experiments were not preregistered. Experimental data is openly available under the following link: [link to be provided upon publication]

Observers

The primary units of analysis were not individual observers, but the distribution of fixations from all observers on individual images. Therefore, we selected the number of observers based on the estimation of how well our empirical fixation distributions approximate the theoretical distributions which would be obtained from the population of infinitely many observers. Previous work has shown that fixations from 18 observers provide a sufficiently good approximation for natural scenes viewed for three seconds (as in our experiment), and that further increasing the number of observers results only in marginal improvements (Judd et al., 2012). However, one of our analyses – reported in the Supplement – required splitting our sample into two groups and we therefore recruited 36 observers in total (mean age 20.06 years, 7 men), ensuring sufficient amounts of data in each group after the split. All participants were Cardiff University students, had normal or corrected-to-normal vision, participated in the study voluntarily, and received either money or study-credits as a reimbursement. All experiments reported in this article were approved by the Cardiff University School of Psychology Research Ethics Committee.

Stimuli

We used 30 pairs of images, where each pair consisted of a two-tone image and its template in greyscale. These stimuli were a subset of stimuli used in a previous study (Teufel et al., 2015), where details of template selection and two-tone image generation can be found. In brief, template images were taken from the Corel Photo library. The main objects depicted in the images were either animals (25 images), humans (three images), or animals and humans (two images). Twenty-five images depicted one main object, five images depicted two main objects. Regarding specific object-parts, seven images depicted mainly one head, two mainly two heads, 18 depicted a head with a full body, and three images depicted two full bodies with heads. Two-tones were generated by smoothing and binarising template images. A good two-tone image should be perceived as a collection of meaningless patches prior to seeing its template but observers should be able to easily bind the stimulus into a coherent percept of an object after they see the template. Extensive tests on naïve observers were conducted to select both the template images, and the parameters of smoothing and binarisation that guarantee that the created two-tones have these desired properties. Note that two-tone images are different from Mooney stimuli (Mooney 1957). By contrast to two-tone images, Mooney stimuli can be, and are designed to be, recognized spontaneously (without need for prior knowledge).

Experimental setup

The experiment was conducted in a dark testing room. Participants sat 56 cm from the monitor, with their head supported by a chin and forehead rest. Their eye-movements were recorded using an EyeLink 1000+ eye-tracker (with 500 Hz sampling rate) placed on a tower mount. The experiment was controlled by in-house developed code written in Matlab R2016b (Mathworks, Natick, MA) and using the Psychophysics Toolbox Version 3 (Brainard, 1997; Kleiner et al., 2007). Images were presented centrally on the screen, against a mid-grey background. Images measured 21.9 degrees of visual angle (788 pixels) horizontally and 14.6 degrees (526 pixels) vertically.

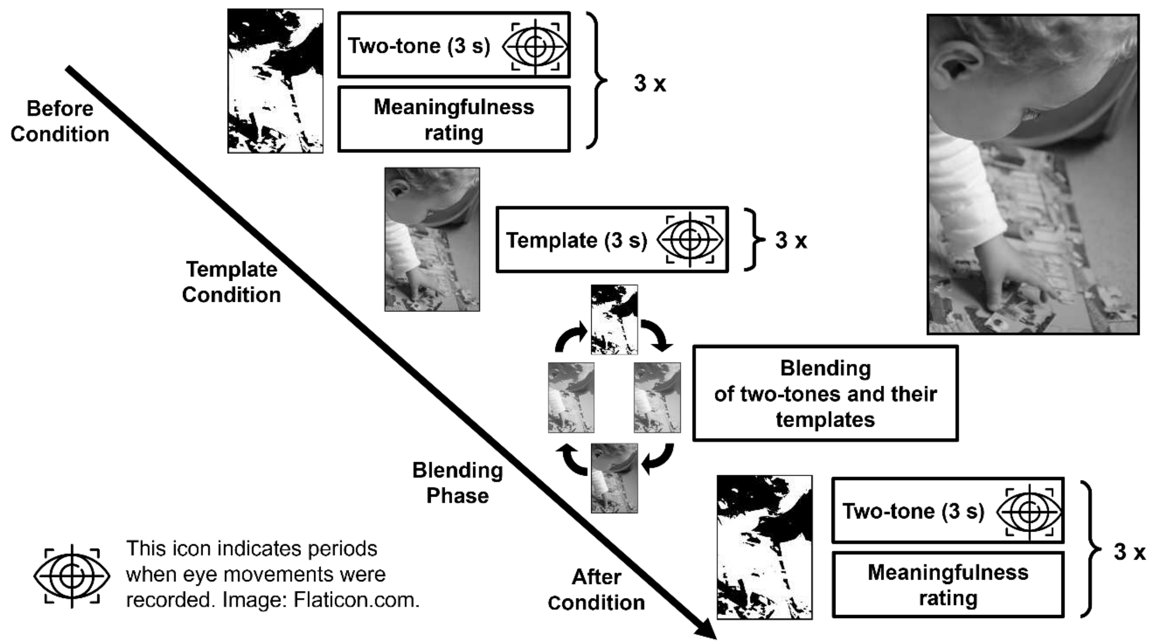
Procedure

The experiment consisted of ten blocks; a single block is schematically illustrated in Fig. 2. Before the start of the procedure, a 13-point eye-tracker calibration and validation was conducted. Each block started with the Before condition, in which three two-tones were presented in a sequence, each for 3 seconds. Observers were instructed to carefully look at these images, but they were not specifically told to search for objects. Two-tone images were preceded by a centrally-located fixation-dot displayed for 1 second. They were followed by a visual analogue scale, which observers adjusted by pressing ‘z’ and ‘m’ buttons on a keyboard to indicate how meaningful they experienced the two-tone image to be. The instruction given to the observers prior to the experiment was also displayed above the scale, saying: ‘Please indicate how clearly the scene or object in the image appeared to be.’ The scale was continuous, with the following labels placed at five linearly spaced points above the scale: 'Very unclear', 'Unclear', 'Neither clear nor unclear', 'Clear', 'Very clear'. Meaningfulness ratings were used as a manipulation check. After each rating, a blank screen was displayed for 500 ms. The Before condition was followed by the Template condition, in which template images were displayed while eye-movements were recorded – again, each for 3 seconds, preceded by a fixation dot. After the Template condition, we ensured that observers had enough object-knowledge to bind two-tone images into meaningful object percepts by presenting six cycles of dynamic blending between two-tones and their templates (Blending Phase). Each cycle began with the presentation of a template image for two seconds. This was then linearly blended into the corresponding two-tone image, with the full transition from template to two-tone taking 4 seconds. The two-tone image remained on the screen for 2 seconds and then was blended back into the template, remaining on the screen for another 2 seconds. Each of the three image-pairs used in a block was presented in a full blending procedure twice with the order pseudorandomised such that the same pair was never used twice in a row. The subsequent cycles of blending were separated with a blank screen presented for 500 ms. After the Blending Phase, the After condition was presented, which was identical to the Before condition except that images were presented in a newly randomized order. There was a break every two blocks, and the eye-tracker was re-calibrated. For each observer, images were assigned to blocks randomly and were presented in a pseudo-random order within each block. The pseudo-randomization ensured that the image shown last in the Blending Phase was never presented at the beginning of the After condition. Total experiment time was ~50 minutes.

Instructions were delivered verbally and on-screen. Key elements of the procedure were illustrated visually: observers were shown a single two-tone image (which was not used in the actual experiment), rated its meaningfulness, viewed the blending procedure with the template and, finally, viewed the same two-tone again and were asked to provide a meaningfulness rating.

Figure 2

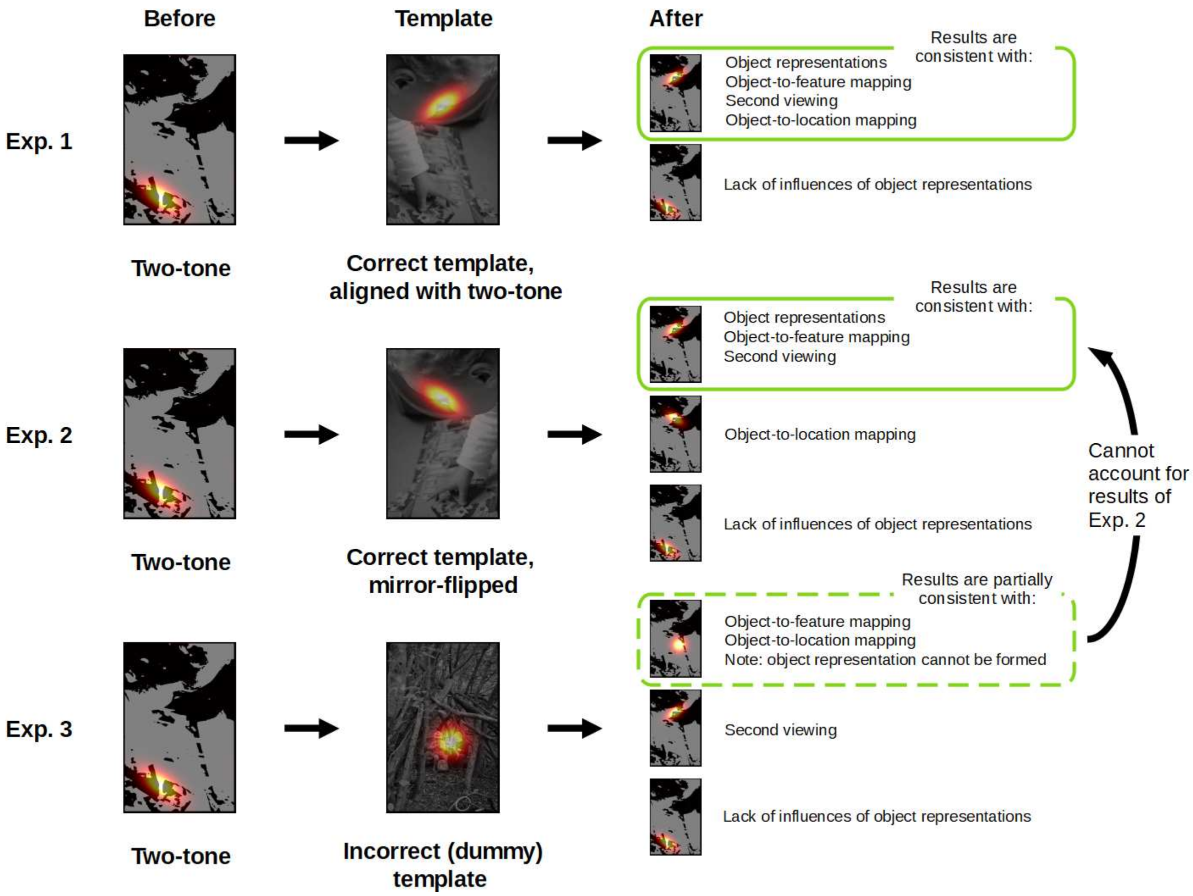
Experiment 1 – Outline of a single experimental block



Note. In each block, observers first free-viewed three two-tone images (Before condition). After presentation of each image, they were asked to rate its perceived meaningfulness. Then, the grayscale templates of these three two-tones were presented (Template condition). In the next part of the block, observers viewed the two-tones gradually blended with their templates six times (Blending Phase). The After condition was identical to the Before condition in all aspects except for the order of presentation of the two-tone images. In the upper right corner, the template of the two-tone image from Fig. 1 is presented (copyrights owner: author C. T.).

Figure 3

Summary of key experimental manipulations, predictions, and findings of Experiment 1, 2, and 3



Note. The heatmaps superimposed over the example stimuli illustrate hypotheses we test in our experiments. The After column illustrates potential experimental outcomes, with green rectangles indicating interpretations consistent with the results for each experiment. All three experiments had identical designs except for the type of image shown in the Template condition and in the Blending Phase. In Experiment 1, the original grayscale photograph used to generate the two-tone image provided observers with the prior object-knowledge required to organise the two-tone image into a coherent object percept in the After condition. We found that gaze guidance in the After condition was similar to that in the Template condition (first row, right top panel), suggesting that knowledge-driven perceptual organization is an important driver of oculomotor control. In Experiments 2 and 3, we excluded potential alternative explanations. In Experiment 2, we presented mirror-flipped template images. This manipulation allowed us to exclude the possibility that when viewing the templates, observers learned the position of objects in the images, and re-visited these locations in the After condition. In Experiment 3, ‘dummy templates’ unrelated to the two-tone images were presented, which allowed us to exclude the possibility that second-viewing of the two-tone images could explain the results. Moreover, this design allowed us to test whether observers had learned to map the features of a two-tone image to locations of objects in the template images. We found a small effect consistent with this idea, but it was too small to fully account for the main findings.

Data pre-processing and analysis methods

The default EyeLink algorithm was used to extract fixation-locations from the eye-movement recordings. Further data pre-processing was done in Matlab. For each image, we discarded the initial fixation that was directed at the fixation-dot presented before image onset. We also discarded fixations not landing within the image-boundaries. Further details regarding data exclusions can be found in the Data exclusion section of the Supplement. For each image in each condition, we generated heatmaps (see examples on Fig. 5E) by smoothing the discrete distribution of fixations with a Gaussian filter, cutoff frequency of –6dB (implementation provided by Bylinskii and colleagues; Kümmerer et al., 2020), and then normalizing the smoothed distribution to the zero-one range.

The majority of our analyses focused on the similarity between two heatmaps. As a similarity index, we calculated Pearson’s linear correlation coefficient using Matlab implementation (Kümmerer et al., 2020). This measure is intuitive, commonly used in the literature (Wilming, Betz, Kietzmann, & König, 2011), and its values have a straightforward interpretation. In the current study, values ranged between zero and one, with one indicating that two heatmaps are identical and zero indicating a maximal dissimilarity. In the Supplement, we provide the results of key analyses using a different metric to quantify the similarity between two heatmaps, the histogram intersection (SIM; Bylinskii, Judd, Oliva, Torralba, & Durand, 2016), showing a similar pattern of results. For statistical comparisons, we primarily relied on standard null-hypothesis-significance-testing techniques implemented in R (R Core Team, 2020) and Matlab. Unless otherwise stated, the t-tests reported throughout the text are paired-sample t-tests. In order to assess the amount of evidence for a lack of a difference between groups of measurements, we used Bayes factors (BFs) calculated using bayesFactor R package (Morey & Rouder, 2018).

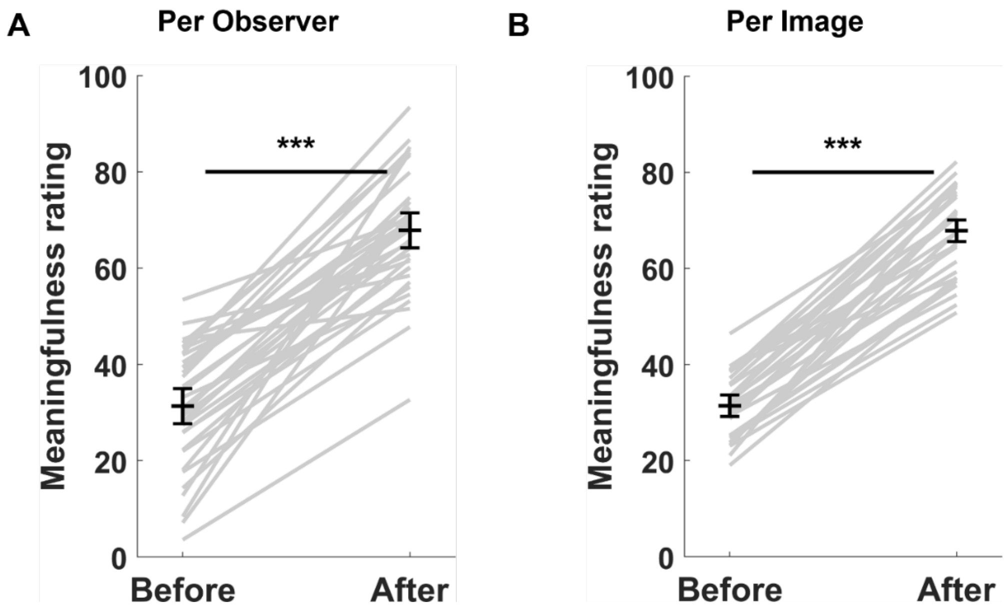
Experiment 1 – Results

Manipulation check: analysis of meaningfulness ratings

In the Before and After conditions, observers rated the perceived meaningfulness of two-tone images. Averaging these ratings per image showed that the two-tones were perceived as more meaningful in the After compared to the Before condition (Fig. 4A and B; $t(29) = 23.84$, $p < 0.001$; mean difference Mdiff = 0.36, 95% confidence interval CI = [0.33, 0.4]). The same pattern of results held when the ratings were averaged per observer ($t(35) = 14.42$, $p < 0.001$; Mdiff = 0.37, 95% CI = [0.31, 0.42]). These results provide a manipulation check, suggesting that observers are able to organize two-tone images into meaningful object representations after but not before acquiring relevant prior object-knowledge.

Figure 4

Meaningfulness ratings for two-tone images in the Before and After conditions averaged per observer (A) and per image (B)



Note. The following conventions are used in this and all remaining figures: asterisks on plots indicate p-values: *** indicates $p \leq 0.001$, ** indicates $p \leq 0.01$, * indicates $p \leq 0.05$, and 'n.s.' indicates the lack of statistical significance. Grey lines indicate values for individual observers (panel A) and images (panel B). Black horizontal bars indicate means. They are surrounded with 95% confidence intervals for within-subjects designs, calculated using Cousineau-Morey method (Cousineau, 2005; Morey, 2008).

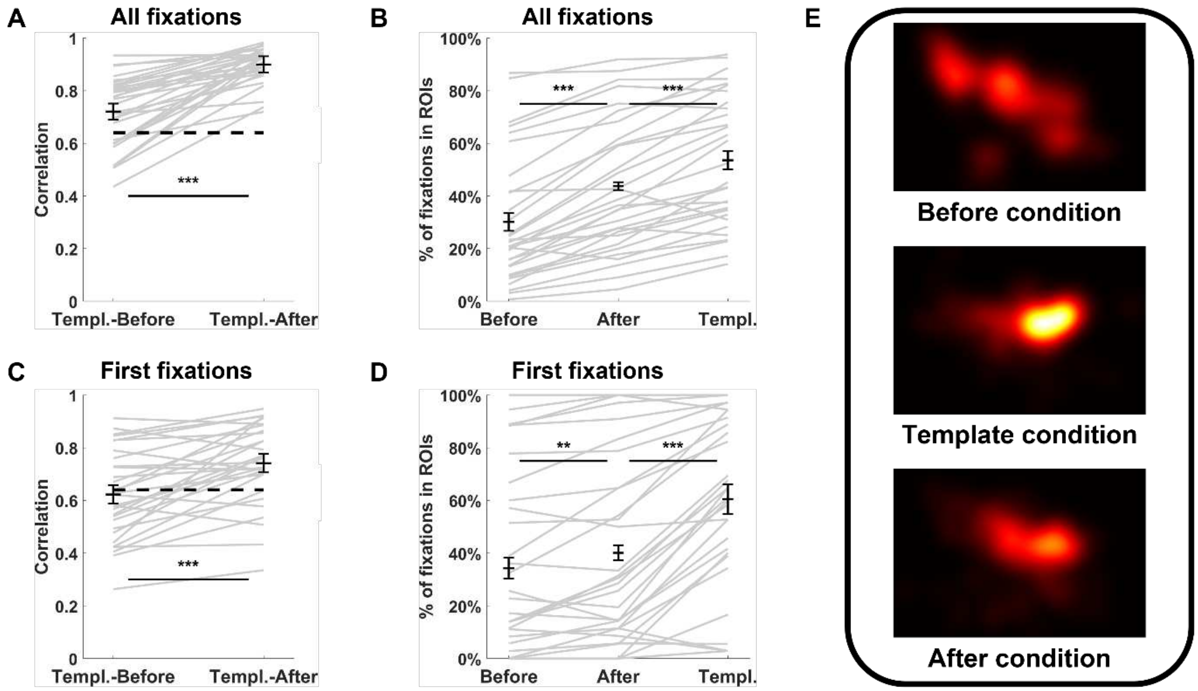
Analysis of similarity between heatmaps

If knowledge-dependent object representations drive eye movements, the spatial distribution of fixations recorded in response to two-tone and template images should be more similar when two-tone images elicit object representations (After condition) compared to when they do not (Before condition). To test this hypothesis, we compared the similarities of heatmaps across pairs of conditions (Fig. 5A). As predicted, we found a higher similarity between the Template-After pair ($M = 0.90$, $SD = 0.07$) compared to the Template-Before pair ($M = 0.72$, $SD = 0.13$; $t(29) = 8.39$, $p < 0.001$; mean difference $M_{diff} = 0.18$, 95% CI = [0.14, 0.22]). This result suggests that gaze patterns in response to two-tone images more closely resemble eye movements from the templates when the two-tones were perceived as containing meaningful objects, as compared to when they were perceived as meaningless patches.

While there was a clear difference in similarity between the two pairs, at first glance the Template-Before similarity might seem unexpectedly high. Importantly, however, the distribution of fixations on images is not only determined by the characteristics of the visual input but also by general factors that are independent of the image (Tatler & Vincent, 2009). One key general factor is the centre bias, a tendency of humans to look at the centre of an image rather than regions closer to the edges (Tatler, 2007). A meaningful evaluation of the difference in similarities between Template-Before and Template-After pairs therefore requires a baseline that accounts for this bias. Given that there is no consensus on exactly how to model centre bias (Hayes & Henderson, 2020), and that systematic studies of centre bias only exist for a limited number of combinations of image sizes and aspect ratios (Clarke & Tatler, 2014), we adopted a data-driven approach to derive a centre bias. Specifically, we modelled a centre bias for our data by creating a single heatmap (labelled ‘Centre’) from all fixations registered throughout the experiment. The rationale for this approach is that by averaging across all images and all observers, the remaining heatmap should include only those factors that are general to all images and observers (i.e., centre bias) in our dataset. We found a statistically robust difference in similarity scores between the Template-Centre and Template-Before pairs (Template-Centre: $M = 0.64$, $SD = 0.16$; Template-Before: $M = 0.72$, $SD = 0.13$; $t(29) = 2.40$, $p = 0.023$; $M_{diff} = 0.08$, 95% CI = [0.01, 0.14]). Importantly, however, this difference was small, suggesting that a centre bias explained most, but not all, of the Template-Before similarity.

Figure 5

Results of Experiment 1



Note. A) Similarities between heatmaps from template and two-tone images, where the two-tones were viewed either in the Before or in the After condition. The dashed horizontal line illustrates the baseline, i.e., the expected similarity with the Template condition based purely on centre bias. B) Proportion of fixations landing within the regions-of-interest (ROIs) in each condition. ROIs included important object parts (e.g., the heads of depicted animals). C, D) The same analyses as on panels A and B but conducted including only first fixations from the Before and After conditions. E) Sample heatmaps illustrating the distributions of fixations in all three conditions of Experiment 1 for one two-tone/template pair. These maps were created from all fixations registered on the images. Pixel values of all three maps were jointly normalised to zero-one range, so colour values (indicating fixation densities) are comparable across panels.

We ran a further analysis (full details in Supplement) to address the influence of knowledge-dependent object representations by comparing heatmaps from identical visual inputs only. In other words, instead of analysing the similarity between heatmaps from a two-tone image and its template image (different visual inputs), we evaluated the similarities in heatmaps when the same two-tone image was viewed in the Before and the After conditions (identical visual inputs). The findings provide further support for the influence of object-knowledge on gaze guidance (see supplement for details).

Regions-of-interest (ROI) analysis

The analyses of heatmap similarities suggests that prior object-knowledge contributes to eye-movement control. We used a region-of-interest (ROI) analysis to assess in a more fine-grained manner the extent to which changes in fixation patterns related directly to object representations. We exploited the fact that animal and human heads are known to attract fixations in natural scenes (Cerf, Paxon Frady, & Koch, 2009; Drewes, Trommershäuser, & Gegenfurtner, 2011). On each template, we manually labelled each pixel associated with a head (recall that all templates depicted animals and/or humans). The resulting masks, which covered 9% of the image area on average (SD = 12%, median = 3%), served as the ROIs for the template and its associated two-tone image. The average distance of the centre of gravity of each masks (as determined by Matlab function *regionprops*) to the image centre was 3.83 degrees of visual angle (SD = 2.91) and the distance to the central vertical image axis was 2.19 degrees (SD = 2.23). For each image and condition, we calculated the proportion of fixations landing within the ROIs (Fig. 5B). This metric increased in the After compared to the Before condition, indicating that changes in fixations were object-specific (Before: M = 30%, SD = 24; After: M = 44%, SD = 25; $t(29) = 8.64$, $p < 0.001$; Mdiff = 0.14, 95% CI = [0.1, 0.17]). Furthermore, there were more fixations within the ROIs in the Template compared to the After condition (Template: M = 54%, SD = 25; $t(29) = 6.02$, $p < 0.001$; Mdiff = 0.1, 95% CI = [0.06, 0.13]). Overall, the ROI analysis provides clear evidence to suggest that the influence of knowledge-dependent object representations on fixation patterns is object-specific.

Analysis of first fixations

In order to assess the time-course of the influence of knowledge-dependent object representations on oculomotor control, we repeated our previous analyses exclusively for first fixations. This restriction did not change the overall pattern of the results (see Fig. 5C and D), suggesting that even first fixations were influenced by object representations that emerged as a consequence of the observer's prior knowledge. Specifically, the statistical analysis showed that for first fixations, the similarity between Template and After was higher than for Template and Before (Template-After: $M = 0.74$, $SD = 0.15$; Template-Before: $M = 0.62$, $SD = 0.17$; $t(29) = 4.91$, $p < .001$; $M_{diff} = 0.12$, 95% CI = [0.07, 0.17]). This finding was corroborated by an ROI analysis of first fixations: the proportion of first fixations landing on ROIs was higher in the After than in the Before condition, and also higher in Template than in After (Before: $M = 34\%$, $SD = 34$; After: $M = 40\%$, $SD = 35$; Template: $M = 60\%$, $SD = 32$; Before-After: $t(29) = 3.61$, $p = 0.001$; $M_{diff} = 0.06$, 95% CI = [0.03, 0.09]; Template-After: $t(29) = 6.41$, $p < 0.001$; $M_{diff} = 0.2$, 95% CI = [0.14, 0.27]). Taken together, these results suggest that knowledge-dependent object representations emerge fast enough to influence even the first eye-movements after stimulus onset.

Analysis of combined effects of image-computable features and prior knowledge

Our analyses so far indicate that knowledge-dependent object representations play a role in gaze guidance, beginning with the first fixation after image onset. However, these analyses do not assess the role of the interaction between image-computable features and object representations. In order to address this point, we capitalized on common and distinct characteristics shared between the After condition and each of the remaining conditions (Before and Template). In particular, image-computable features of Before and After conditions are identical, but they differ in the extent to which observers experienced object representations. Specific similarities in fixation patterns between Before and the After conditions, which go beyond general factors such as centre bias, can therefore be attributed to the image-computable features of two-tone images. Conversely, the After and the Template conditions have the reverse relationship: they lead to similar object representations but differ in image-computable features. We exploited this situation to characterize the contribution of these gaze-guidance factors in the After condition.

For this purpose, we created linear combinations of heatmaps from the Before and Template conditions to compare with the heatmaps of the After condition (Fig. 6). Each new linear-combination heatmap was calculated from the Before and the Template conditions' heatmaps, using the formula:

$$w_{Template} * heatmap_{Template} + w_{Before} * heatmap_{Before} \qquad (1)$$

where w is a weight for the heatmap indicated by the subscript. Incorporating the normalization assumption ($w_{\mathrm{Template}} + w_{\mathrm{Before}} = 1$), we created a continuum of heatmaps spanning the range between being fully determined by the Template heatmap to being fully determined by the Before heatmap. This continuum was uniformly sampled with a step-size of 0.05. This procedure led to a set of heatmaps, which capture factors driving eye movements in the Before and the Template conditions to varying degrees. Evaluating the similarity of these new heatmaps with those from the After condition allowed us to determine the relative contribution of image-computable features and object representations to gaze guidance in the After condition. To focus on the time course, we conducted this analysis separately for first fixations and for all the remaining fixations.

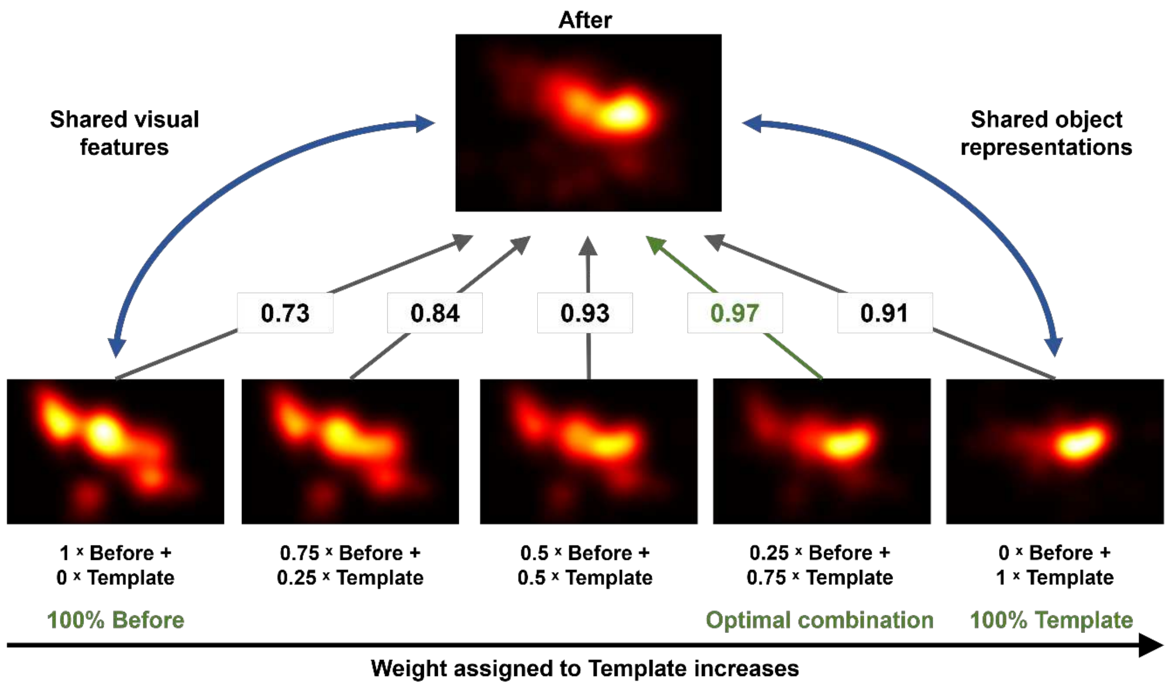
The results of this similarity analysis suggest that both first and all remaining fixations in the After condition were guided synergistically by image-computable features and object representations (Fig. 7). The linear-combination heatmaps that had the highest similarity with the first fixations in the After condition showed an influence from the Template heatmap but also had a substantial contribution from the Before heatmap (wTemplate = 0.4, wBefore = 0.6; mean correlation $M = 0.85$, $SD = 0.06$; see Fig. 7A). Statistical analyses indicated that the heatmaps from the After condition were more similar to this optimal linear-combination heatmap than to either the Before or the Template conditions alone (Optimal-After vs. Before-After: $t(29) = -2.67$, $p = 0.012$; Mdiff = 0.03, 95% CI = [0.01, 0.04]; Optimal-After vs. Template-After: $t(29) = 5.70$, $p < 0.001$; Mdiff = 0.11, 95% CI = [0.07, 0.15]).

The findings for all remaining fixations from the After condition were similar (Fig. 7B). However, the linear combinations that were optimal for these fixations were more strongly influenced by the Template heatmap (wTemplate = 0.65, wBefore = 35; mean correlation $M = 0.95$, $SD = 0.03$). Yet, even for these later fixations, there was a substantial influence of image-computable factors as captured by the Before heatmaps. This idea is supported by the statistical analysis, which indicates that the heatmaps from the After condition were more similar to the optimally combined heatmaps compared to both the Before and the Template condition alone (Optimal-After vs. Before-After: $t(29) = 6.49$, $p < 0.001$; Mdiff = 0.09, 95% CI = [0.06, 0.12]; Optimal-After vs. Template-After: $t(29) = 5.48$, $p < 0.001$; Mdiff = 0.05, 95% CI = [0.03, 0.06]).

Overall, the analysis suggests that image-computable features and object representations guide eye movements in a synergistic manner (see also Borji & Tanner, 2016). The contribution of these two factors vary over time, with object representations playing a less important role in first fixations than in later fixations. Yet, both factors already influence first fixations.

Figure 6

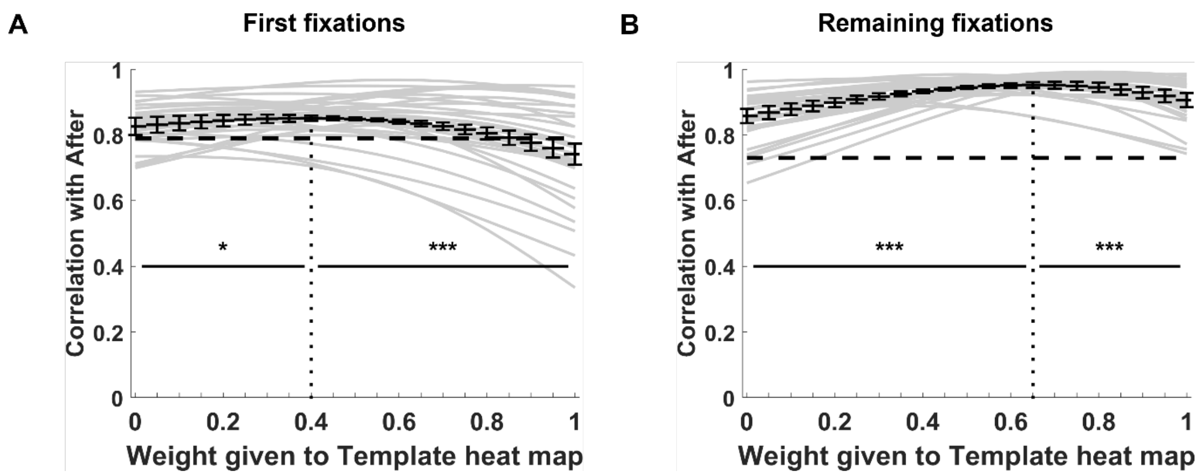
Linear combination analysis – illustration for a single two-tone image



Note. The bottom row shows heatmaps that have been created by linearly combining the heatmaps from the Before and the Template conditions, as indicated by the text below each image. These linear-combination heatmaps were compared to the heatmap of the After condition as indicated by arrows. Numbers on the arrows indicate correlation values. The blue, double-pointed arrows illustrate the fact that the After condition shares image-computable features and object representations with the Before and the Template condition, respectively. To enable visually comparing all heatmaps shown on the figure, their pixel values were jointly normalised to zero-one range.

Figure 7

Similarities of heatmaps from the After condition to different linear combinations of heatmaps from the Template and Before conditions



Note. A) Similarities obtained when only first fixations from the After condition are considered. B) The same analysis but for all the remaining fixations (i.e., without first) from the After condition. The weights of the linear combinations for which the similarity is maximal are indicated by the dotted vertical lines. Dashed vertical lines on both panels indicate the baseline, i.e., the average similarities of the respective After heatmaps to centre bias model (M = 0.79, SD = 0.09 for first fixations; M = 0.73, SD = 0.14 for the remaining ones).

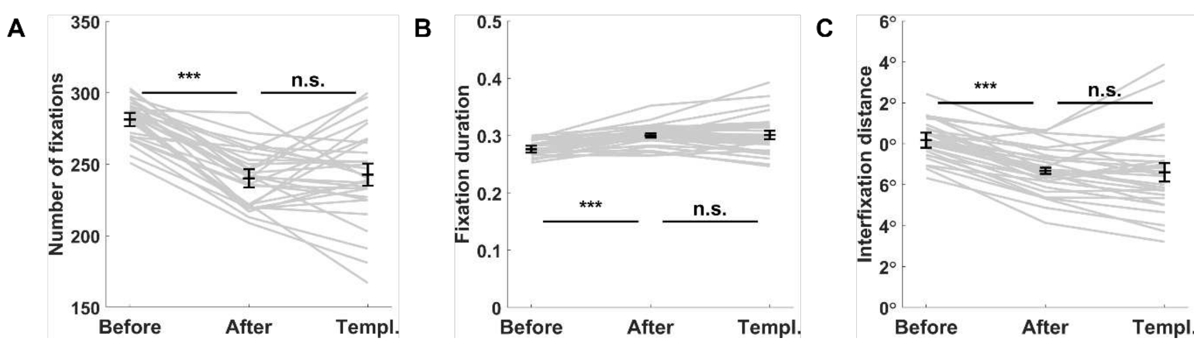
Analysis of other characteristics of oculomotor behaviour

In our final analyses of Experiment 1, we assessed the extent to which knowledge-dependent object representations affect characteristics of eye movements that might be indicative of a more fundamental change in the observers' information-sampling strategy. First, we calculated the mean number of fixations, average fixation duration (in seconds), and average Euclidean distance between consecutive fixations (interfixation distance, in degrees of visual angle) per image, and compared them across conditions (Fig. 8). Compared to the Before condition, the After condition showed a decrease in the number of fixations (values summed across observers separately for each image; Before: $M = 281.37$, $SD = 13.22$; After: $M = 240.10$, $SD = 19.32$; $t(29) = 12.76$, $p < 0.001$; $M_{diff} = 41.27$, 95% CI = [34.65, 47.88]), an increase in the fixation duration (Before: $M = 0.28$, $SD = 0.01$; After: $M = 0.30$, $SD = 0.02$; $t(29) = -8.22$, $p <$

0.001; Mdiff = -0.02, 95% CI = [0.02, 0.03]), and a decrease in interfixation distance (Before: M = 4.09, SD = 0.45; After: M = 3.34, SD = 0.55; t(29) = 11.24, p < 0.001; Mdiff = 0.75, 95% CI = [0.61, 0.89]). We did not find statistically significant differences between the Template and the After conditions for any of these metrics (number of fixations: t(29) = -0.50, p = 0.621; Mdiff = -2.67, 95% CI = [-13.58, 8.25]; fixation duration: t(29) = -0.24, p = 0.816; Mdiff = 0 , 95% CI = [-0.01, 0.01]; interfixation distance: t(29) = 0.32, p = 0.755; Mdiff = 0.04, 95% CI = [-0.19, 0.27]; descriptive statistics for these three respective characteristics for Template condition: M = 242.77, SD = 31.76; M = 0.30, SD = 0.03; M = 3.3, SD = 0.96).

Figure 8

Number of fixations (A), fixation duration (B), and interfixation distance measured in degrees of a visual angle (C)



Note. All three were calculated per image and compared between conditions.

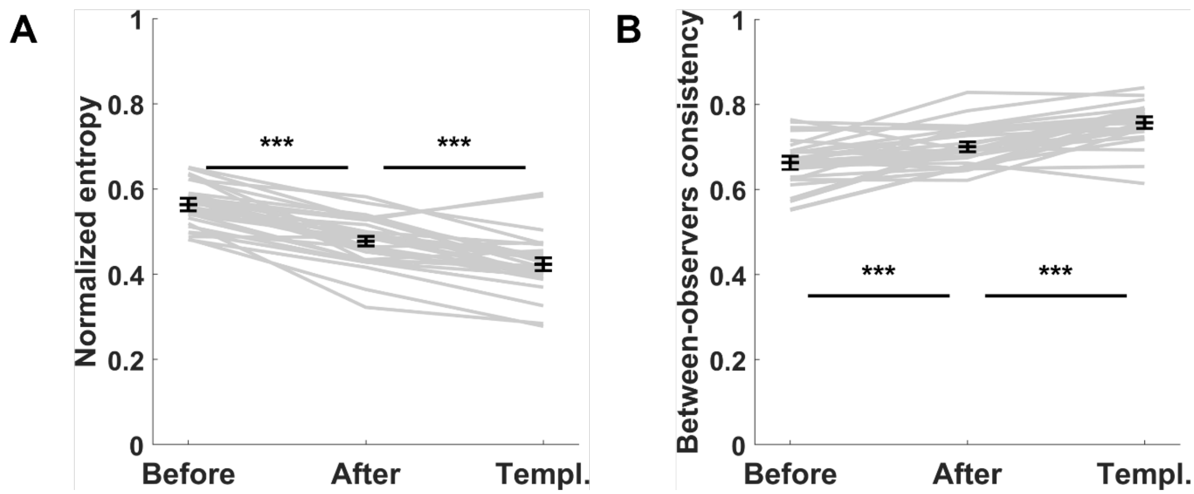
These findings are consistent with the idea that observers shift from exploring the whole stimulus in the Before condition towards extracting information only from selected parts in the After and Template conditions. To further substantiate this interpretation, we calculated the normalized entropy for the heatmaps in the different conditions (Fig. 9A). This measure is thought to index the extent to which an observer's behaviour is exploratory (Gameiro, Kaspar, König, Nordholt, & König, 2017; Kaspar et al., 2013). Normalized entropy was lowest in the

Template condition, increased in the After condition, and was highest in the Before condition (Before: $M = 0.56$, $SD = 0.05$; After: $M = 0.48$, $SD = 0.06$; Template: $M = 0.42$, $SD = 0.07$; Before-After: $t(29) = 9.92$, $p < 0.001$; Mdiff = 0.09, 95% CI = [0.07, 0.10]; After-Template: $t(29) = 6.28$, $p < 0.001$; Mdiff = 0.05, 95% CI = [0.04, 0.07]). In other words, observers showed the highest exploratory behaviour in the Before condition, followed by the After and the Template condition.

In our final analysis, we wanted to know if object representations would result in more homogenous gaze behaviour across observers (Fig. 9B). We quantified between-observers consistency by averaging the similarity between each observer's individual heatmap to the heatmaps of all remaining observers (Lyu et al., 2020). This metric increased both between the Before and After conditions and between the After and Template conditions (Before: $M = 0.66$, $SD = 0.05$; After: $M = 0.7$, $SD = 0.05$; Template: $M = 0.76$, $SD = 0.05$; Before-After: $t(29) = 3.96$, $p < 0.001$; Mdiff = 0.04, 95% CI = [0.02, 0.06]; After-Template $t(29) = 6.96$, $p < 0.001$; Mdiff = 0.06, 95% CI = [0.04, 0.07]), suggesting that object representations increase consistency in information-sampling behaviour across observers.

Figure 9

Normalized entropy and between-observers consistency



Note. A) Normalized entropy of fixation distributions (in arbitrary units) as a measure of their spread. Higher values indicate more exploratory behaviour of observers. B) Between-observers consistency in selecting fixation targets measured by how similar (on average) fixations of a single observer were to fixations of all the remaining observers pooled together.

Experiment 1 – Discussion

In Experiment 1, we measured eye-movements in response to grayscale images of scenes containing objects and two-tone images derived from these templates. On initial viewing, two-tone images are experienced as meaningless black-and-white patches. Once an observer has acquired relevant prior object-knowledge, however, the visual system organizes the patches into a coherent percept of an object. We demonstrate that, when a two-tone image is perceived as showing a coherent object rather than meaningless patches, gaze guidance changes in several ways. First, and most importantly, fixation patterns on two-tone images become more similar to those measured in response to the template when two-tones lead to object representations vs. when they are experienced as meaningless patches. Moreover, fixation locations become more object-specific. Importantly, however, we also demonstrate that object representations do not fully dominate gaze guidance, but that image-computable feature space and object representations interact in determining where people look. While the data suggest a specific temporal development of this interaction, we also observe that the influence of knowledge-dependent object representations is already present in the first eye-movement after image onset, suggesting that the emergence of knowledge-driven object representations precedes the first eye-movement. Object representations also lead to fewer fixations, longer fixation durations, shorter interfixation distances as well as a less exploratory pattern of eye movements and more consistency across observers. Overall, these results suggest that object representations, which are not fully determined by image-computable features but depend on an observer's prior object-knowledge have a substantial influence on eye movements. Note that the images were presented in batches of three (see the *Procedure* section), ensuring that they were not fully predictable. These results are therefore unlikely to be explained by planning of eye movements done before the onset of the image in the After condition.

It is, however, possible that the change in fixation patterns observed in Experiment 1 were caused by a memory process unrelated to knowledge-driven perceptual organization. Specifically, it has been suggested that eye movements performed during memory retrieval of an image resemble the eye movements performed when seeing this stimulus for the first time (Noton & Stark, 1971; see Wynn, Shen, & Ryan, 2019 for a recent review and Foulsham & Kingstone, 2013 for criticism). According to this alternative explanation, two-tone images in the After condition might have acted as cues that triggered the retrieval of the corresponding template, and this retrieval might have been accompanied by the re-enactment of gaze behaviour from the Template condition. A simpler but overall similar alternative explanation of the results from Experiment 1 might suggest that memory-retrieval of template images resulted in the observers voluntarily directing their gaze towards locations in the two-tone images, which they remembered to be occupied by objects. According to both explanations, the factor driving changes in eye movements in the After condition is the mapping of objects to locations that the observers remember from the Template condition, rather than perceptual organization induced by prior object-knowledge. To exclude these alternative explanations, which we label the 'object-to-location mapping' interpretation, we conducted Experiment 2.

Experiment 2

Overview

Experiment 2 was identical to Experiment 1 in all aspect except that the template images were flipped along the vertical axis ('mirror-flipped') from left to right. Consequently, the screen locations occupied by objects differed between the Template condition and the remaining conditions. This simple manipulation allowed us to adjudicate between the different alternative interpretations mentioned in the previous section: according to the object-to-location mapping hypothesis, which suggests that observers merely revisited the parts of the display, which contained objects during the presentation of template images, we would expect a high similarity between heatmaps from the After and Template conditions, despite the lack of overlap in spatial location of objects in these two conditions. If, however, the effects observed in Experiment 1 were attributable to knowledge-dependent object representations, we would expect the similarity between the After and Template conditions to be low (see Fig. 3 for illustration). Moreover, by mirror-flipping the heatmaps obtained from the mirror-flipped templates, we would expect an increase in similarity to levels seen in Experiment 1 (because this leads to a re-alignment of heatmaps from templates and two-tones).

Experiment 2 – Method

A separate set of 18 Cardiff University students (mean age 19.5 years, 5 men), who did not participate in Experiment 1, served as observers. The design of Experiment 2 was identical to that of Experiment 1 except that the template images were flipped along the vertical axis from left to right for all parts of the experiment. Additionally, during the Blending Phase, the two-tones were flipped such that two-tones and templates were aligned. This condition is labelled FlippedTemplate. Observers were not explicitly informed about the flipping; the instructions were identical to those in Experiment 1.

Experiment 2 – Results

Controlling for the effects of object-to-location mapping

Similar to Experiment 1, the meaningfulness ratings provided by the observers after viewing each two-tone were higher in the After condition than the Before condition both when we averaged them per observer ($t(17) = 6.62$, $p < 0.001$; Mdiff = 0.24, 95% CI = [0.16, 0.31]) and per image ($t(29) = 16.74$, $p < 0.001$; Mdiff = 0.24, 95% CI = [0.21, 0.27]). This result indicates that observers were able to bind the two-tone images into meaningful percepts despite viewing templates, which were presented in a mirror-flipped manner.

The results of the eye-movements data analysis were inconsistent with the object-to-location hypothesis but provided support for the idea that knowledge-dependent object representations influence eye movements (see Fig. 10). In particular, by contrast to the analogous analysis in Experiment 1, heatmap similarities did not differ when comparing the FlippedTemplate-Before pair vs. the FlippedTemplate-After pair (FlippedTemplate-Before: M = 0.46, SD = 0.22; FlippedTemplate-After: M = 0.48, SD = 0.22; $t(29) = 1.45$, $p = 0.158$; Mdiff = 0.03, 95% CI = [-0.01, 0.06]). A BF of 0.50 suggested that the data provided evidence in favour of there being no difference between conditions, but that this evidence was weak. Importantly, once the heatmaps from the template and two-tone images were re-aligned, by flipping the heatmaps of the FlippedTemplate condition, the similarity between the RealignedTemplate and the After condition was higher than the similarity between RealignedTemplate and Before (RealignedTemplate-Before: M = 0.68, SD = 0.15; RealignedTemplate-After M = 0.8, SD = 0.11; $t(29) = 7.77$, $p < 0.001$; Mdiff = 0.13, 95% CI = [0.09, 0.16]). Moreover, the differences between Template-Before and Template-After were more than four times larger in the Realigned heatmaps than in the Flipped ones (FlippedTemplate-After minus FlippedTemplate-Before: M = 0.03, SD = 0.10; RealignedTemplate-After minus RealignedTemplate-Before: M =

0.13, SD = 0.09), and the difference between these differences was statistically significant ($t(29) = 3.81$, $p < 0.001$; Mdiff = 0.10, 95% CI = [0.05, 0.15]).

Similar to Experiment 1, we conducted an analysis of the proportion of fixations landing within flipped and re-aligned ROIs on the two-tone images to assess in more detail whether fixations are specifically object-oriented. Note, however, that to the extent to which ROIs cross the central vertical axis of an image, flipped ROIs overlap with re-aligned ROIs (this happened in 16 images, with an average overlap of 49.59 % (SD = 29.16 %) of pixels). To ensure that ROIs are unique, in this analysis, we used flipped and re-aligned ROIs from which the overlap between the two had been removed. The proportion of fixations landing in the flipped ROIs did not differ between the After and the Before conditions (Before: M = 7%, SD = 7; After: M = 7%, SD = 7; $t(29) = 0.14$, $p = 0.888$; Mdiff = 0, 95% CI = [-1, 2]). The same metric for the realigned ROIs indicated a clear difference between the two conditions, with more fixations landing in the realigned ROI in the After than the Before condition, indicating that changes in fixations were object-specific (Before: M = 16%, SD = 11; After: M = 19%, SD = 11; $t(29) = 3.55$, $p < 0.01$; Mdiff = 4%, 95% CI = [2, 6]).

Similar to the findings for all fixations, heatmap similarities did not differ when comparing the FlippedTemplate-Before pair vs. the FlippedTemplate-After pair for first fixations (FlippedTemplate-Before: M = 0.44, SD = 0.25; FlippedTemplate-After: M = 0.45, SD = 0.27; $t(29) = 0.47$, $p = 0.645$; Mdiff = 0.01, 95% CI = [-0.03, 0.05]). By contrast to all fixations, however, the equivalent comparison for the realigned pairs did also not show a significant difference, albeit with a numerically larger effect in the direction expected from Experiment 1 (RealignedTemplate-Before: M = 0.57, SD = 0.21; RealignedTemplate-After M = 0.62, SD = 0.20; $t(29) = 1.87$, $p = 0.072$; Mdiff = 0.05, 95% CI = [0, 0.09]).

The ROI analyses for first fixations corroborated this pattern of results. We found no significant differences between the After and the Before conditions in the proportion of fixations landing in the flipped ROI (Before: M = 7%, SD = 10; After: M = 6%, SD = 8; Before-After: t(29)

= -0.87, $p = 0.391$; Mdiff = -1, 95% CI = [-4, 2]) and the realigned ROI (Before: M = 13%, SD = 18; After: M = 16%, SD = 17; Before-After: $t(29) = 1.66$, $p = 0.107$; Mdiff = 3, 95% CI = [-1, 6]), albeit with a numerical pattern in line with that of all fixations.

Comparison between Experiments 1 and 2

The spatial misalignment of template and two-tone images had an influence on how well observers were able to disambiguate the two-tones, as indicated by the finding that the (per-image) average increase in the meaningfulness ratings in Experiment 2 were smaller than in Experiment 1 ($t(29) = 8.63$, $p < 0.001$; Mdiff = 0.12, 95% CI = [0.09, 0.15]). In order to contrast the effects on gaze guidance across experiments, we directly compared the increase in similarity between the Template-Before vs. Template-After pairs across Experiments 1 and 2. Given that both experiments differed with respect to the number of observers who contributed to the heatmaps of each image, we included fixations only from 18 observers from Experiment 1 (drawn randomly). We found that the increase in similarity between the Template-Before vs. Template-After pairs were larger in Experiment 1 than Experiment 2 (Experiment 1: M = 0.17, SD = 0.13; Experiment 2: M = 0.13, SD = 0.09; $p = 0.0174$; Mdiff = 0.05, 95% CI = [0.01, 0.08]). To ensure that the outcome did not depend on the specific set of observers from Experiment 1, we repeated this analysis for 20 different, randomly drawn sets and obtained the same pattern of outcomes for 19 of them.

Experiment 2 – Discussion

In sum, despite the spatial misalignment of objects in template and two-tone images, fixations were strongly influenced by object locations in Experiment 2. There was no evidence to suggest that mapping objects to locations played a role in gaze guidance. It is noteworthy, however, that the spatial misalignment between template and two-tone images in Experiment 2 had an attenuating effect on the influence of objects on eye movements compared to

Experiment 1. Interestingly, this attenuation in gaze guidance data was mirrored by an attenuation in the meaningfulness ratings, reflecting the ability of observers to use prior knowledge to organise two-tone images into meaningful object percepts (which was, nevertheless, robust). This finding is consistent with our overall interpretation that knowledge-driven object representations are important in eye-movement control.

While the analysis of first fixations showed a pattern that was numerically similar to that of all fixations, none of the analyses reached significance. In other words, by contrast to Experiment 1, first fixations in Experiment 2 did not show significant object-oriented effects, probably because the spatial misalignment between template and two-tone images resulted in less efficient perceptual organization of the latter into a meaningful percept (as suggested by the comparison of the meaningfulness ratings between Experiments 1 and 2). Importantly, analyses of first fixations also provided no evidence to suggest that a process of object-to-location mapping played any role in guiding first fixations during viewing of the two-tone images. Taken together, the results from Experiment 2 exclude the possibility that gaze guidance in the After condition is based on a mapping of objects to locations via retrieval of this information from the Template condition.

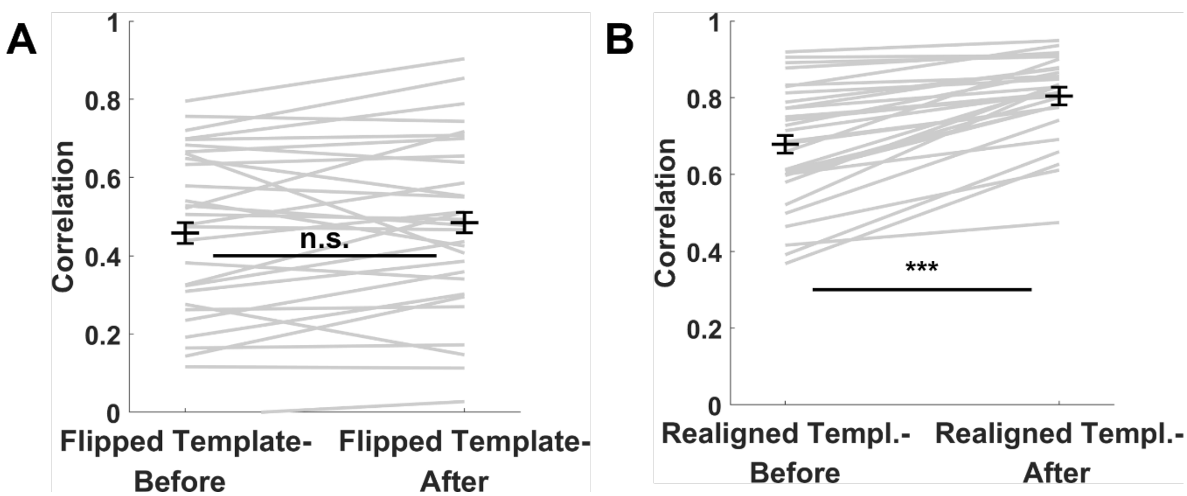
In a third experiment, we addressed two further alternative explanations of the results from Experiment 1. First, it is possible that during the phase when two-tone images are blended with templates, observers learn to associate specific image-features in the two-tone images with object locations in the templates. When viewing two-tone images in the After condition, these feature-object associations might guide fixations towards these specific visual patterns, irrespective of transformations such as those introduced by the mirror-flipping. While this possibility might seem implausible, there is evidence to suggest that such learning processes are an important factor in oculomotor control (Alfandari, Belopolsky, & Olivers, 2019).

A final alternative explanation of our results from both Experiment 1 and 2 relates to potential order effects. It is possible that the changes in fixation patterns between Before and

After conditions resulted from viewing two-tones for a second time, rather than from knowledge-dependent perceptual organization. In other words, observers might sample information from different image regions on second compared to first viewing, irrespective of the kind of information they acquire in the meantime. We conducted Experiment 3 to exclude the possibility that (i) feature-object associations, or (ii) any order effects could explain the effects of Experiments 1 and 2.

Figure 10

Results of Experiment 2



Note. A) Similarities between heatmaps from two-tone images and mirror-flipped templates, where the two-tones were viewed either in the Before or in the After condition. The heatmaps derived from the mirror-flipped template images were used either before (A) or after (B) the mirror-flipping was reverted by ‘flipping back’ these heatmaps and realigning them with the heatmaps from two-tone images.

Experiment 3

Overview

Experiment 3 adopted the same procedure as the previous experiments except that the templates from Experiment 1 (‘real templates’) were replaced with different images that were unrelated to the two-tones (‘dummy templates’). This experimental design allowed us to test whether feature-object associations provide a plausible explanation for the findings of Experiment 1 and 2. Specifically, observers might associate certain features in the two-tone images with objects in the templates during the Blending Phase. When viewing two-tone images in the After condition, these feature-object associations could drive fixations towards image locations in the two-tones that overlap with objects in the respective (dummy) templates. These effects should be observable despite observers not having acquired the prior object-knowledge required to organize the two-tone images into coherent percepts. Moreover, the design also allowed us to assess whether order effects could explain the findings from Experiments 1 and 2.

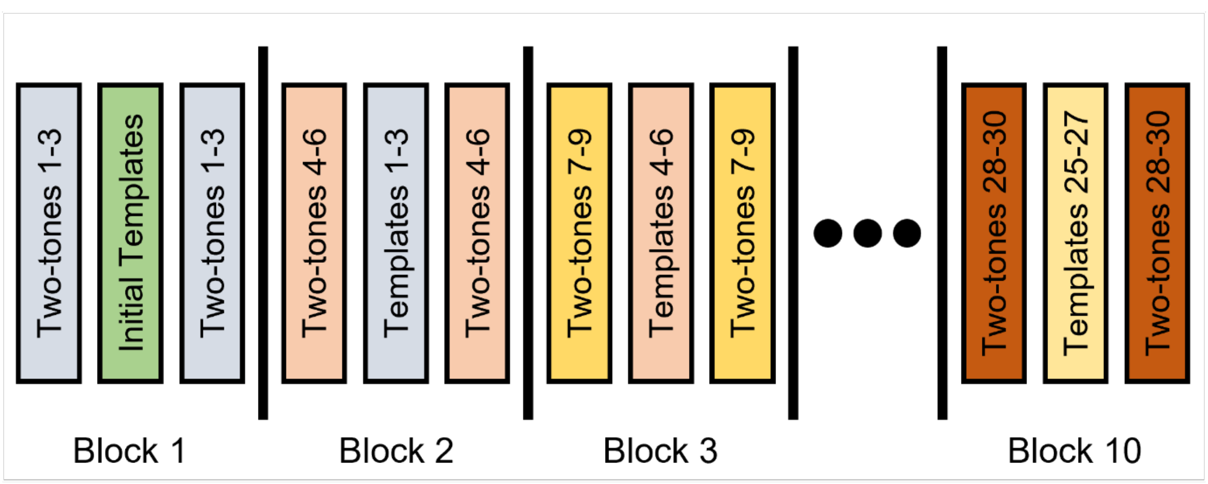
Experiment 3 – Method

Experiment 3 was completed by 20 observers (mean age 19.55, 5 men) who did not participate in the previous two experiments. All were Cardiff University students. The procedure was identical to the previous experiments except that in each block, the templates used in the Template condition and in the Blending Phase were unrelated to the two-tones presented in this block (‘dummy templates’). Each two-tone had a unique dummy template paired with it and this pairing was fixed for all observers. Importantly, each dummy template was a ‘real template’ of a different two-tone presented in the preceding block during the experiment (see Fig. 11). While templates in this experiment could thus not provide object knowledge that would help organize the two-tone image into an object percept in the After condition, we were nevertheless able to register eye movements on the real templates. Measuring fixations on real templates was necessary to assess whether simply viewing a two-tone for a second time, without prior object-knowledge, would lead to increased similarity between heatmaps of two-tone images in the After and their real templates, as seen in the previous experiments.

In the first block, the same dummy templates – greyscale images not related to any of the two-tones – were always presented. In all other blocks, the assignment of stimuli to experimental blocks was pseudo-randomized for each observer individually in a way which guaranteed that dummy templates presented in any given block were the real templates of two-tones presented in the preceding block (see Fig. 11). To ensure that we included data from the same number of observers for each two-tone and template, we had to discard fixations registered on the two-tones presented in the final experimental block and fixations from the dummy templates from the first blocks ('initial templates'). Note that – because we pseudo-randomized the order of stimulus presentation for each observer individually – for different images we had to discard data from different observers. Importantly, however, for each image set consisting of a two-tone (viewed in Before and After condition), its dummy template, and its real template, we retained data from a homogenous group of 18 observers (out of 20 who completed the experiment), but the composition of these groups was different for different image sets.

Figure 11

Randomization schema used in Experiment 3



Note. Within each block, stimuli were presented in a randomised order (as in Experiment 1 and

2). The presentation of images was arranged in such a way that templates in, e.g., Block 2, were the real templates of the two-tone images in Block 1. This order allowed us to register fixations for the real templates (for comparison with fixation on two-tone images) while omitting the opportunity for the observer to acquire the relevant prior object-knowledge that would allow them to disambiguate the two-tone images.

Experiment 3 – Results

Analysis of meaningfulness ratings

The analysis of meaningfulness ratings demonstrated that, as expected, observers were not able to bind the two-tone images into coherent object percepts even in the After condition (Fig. 12A and B). In particular, the differences in ratings between Before and After conditions were not statistically significant, both when the data were averaged per observer ($t(19) = 1.49$, $p = 0.152$; Mdiff = 0.02, 95% CI = [-0.01, 0.06]) or per image ($t(29) = 1.97$, $p = 0.058$; Mdiff = 0.02, 95% CI = [0, 0.05]). In the former case, Bayes factor analysis suggested weak evidence for the lack of differences (BF = 0.60), while in the latter no clear conclusions could be drawn (BF = 1.07).

Controlling for the effects of object-to-feature mapping

Experiment 3 tested the hypothesis that the effects observed in the two previous experiments might be explainable by a learned association between feature clusters in two-tones and object locations on templates. Specifically, it is possible that during blending of two-tone images and templates, observers learn to associate specific features of the two-tones with object locations in the templates and then re-visit these features when viewing the two-tone images in the After condition. Our analysis indicated that the similarity in heatmaps in the DummyTemplate-After pair was higher compared to the DummyTemplate-Before pair (Fig. 12C). This increase in similarity, although significant in a statistical sense, was small

(DummyTemplate-Before: M = 0.46, SD = 0.21; DummyTemplate-After: M = 0.52, SD = 0.22; $t(29) = 4.70$, $p < 0.001$; Mdiff = 0.06, 95% CI = [0.03, 0.08]). Nevertheless, the analysis provided evidence to suggest that feature-object associations might guide oculomotor control to a limited extent. Alternatively, these results could be driven by memory retrieval of object-locations in the templates: while Experiment 2 showed that memory retrieval does not play a role when perceptual organization takes place, this process may become important when the stimulus remains unorganized with no object representations to guide eye movements. In either case, it is interesting that the analysis of first fixations did not indicate a difference between DummyTemplate-Before and DummyTemplate-After (DummyTemplate-Before: M = 0.41, SD = 0.21; DummyTemplate-After: M = 0.45, SD = 0.24; $t(29) = 1.38$, $p = 0.179$; Mdiff = 0.04, 95% CI = [-0.02, 0.01])). This finding suggests a different temporal development of the influence on gaze guidance by object representations vs. by object-to-location or object-to-feature mappings: while the former is present from the first fixations, the latter kick in only after the first fixation (and potentially only if no object representations are available to provide guidance).

The ROI analyses corroborated the findings for heatmaps: for all fixations, we found a significant difference between the After and the Before conditions in the proportion of fixations landing in the ROIs of DummyTemplates (Before: M = 0.21, SD = 24; After: M = 0.23, SD = 0.27; Before-After: $t(29) = 2.64$, $p = 0.013$; Mdiff = 0.02, CI = [0.01, 0.04]). Note that this difference was similar in magnitude to the equivalent difference regarding the ROIs of real Templates (see the *Controlling for order effects* section; difference between the differences: $t(29) = -1.27$, $p = 0.212$; Mdiff = -0.02, 95% CI = [-0.05, 0.01]). Finally, first fixations showed no difference in the proportion of fixations landing in the ROIs of the DummyTemplates (Before: M = 0.26, SD = 0.34; After: M = 0.27, SD = 0.35; Before-After: $t(29) = 0.79$, $p = 0.435$; Mdiff = 0.01, 95% CI = [-0.02, 0.05]).

Object-to-location mapping: comparison between Experiments 1 and 3

While the results reported in the previous section suggest that object-to-location or object-to-feature mapping might influence gaze guidance in the After condition (after the first fixation), the key question is whether these effects can explain the results found in Experiment 1. To address this issue, we directly compared the increase in similarity between the Template-Before vs. Template-After pairs across Experiments 1 and 3. Given that both experiments differed with respect to the number of observers who contributed to the heatmaps of each image, we adopted a similar approach for that used to compare Experiments 1 and 2 (i.e., we randomly drew 18 observers from Experiment 1 and repeated this analysis for 20 different, randomly drawn sets). This analysis indicates that the change in similarity between the Template-Before vs. Template-After pairs was larger in Experiment 1 than in Experiment 3 (Experiment 1: $M = 0.17$, $SD = 0.13$; Experiment 3: $M = 0.06$, $SD = 0.07$; $t(29) = 4.15$, $p < 0.001$; $M_{diff} = 0.11$, 95% CI = [0.06, 0.17]; results for one of the 20 sets).

Our results (for all fixations) thus demonstrate that the processes responsible for changing gaze-patterns between Before and After conditions in Experiment 3 cannot fully explain the analogous changes in Experiment 1. One possible explanation for this finding is that it might be more difficult to learn object-to-feature mappings in Experiment 3 than Experiment 1 (during viewing of the template images and the blending phase). If we assume that gaze is guided by this mapping process, then less robust learning might explain the differences in effect size for all fixations in Experiments 1 and 3. Importantly, however, the differences in temporal trajectories found in the two experiments might be difficult to reconcile with this idea: by contrast to Experiment 1, we found no evidence for a change between Before and After in Experiment 3 for first fixations. This pattern of results suggests that (partly) different processes that are characterised by different temporal trajectories are at work in the two experiments. Specifically, we argue that the influence of object representations is present from the first fixations onwards (as seen in Experiment 1), while object-to-feature or object-to-location mapping kicks in later (as seen in Experiment 3), and potentially only if no object representations are available to provide guidance. Overall, the pattern of results in Experiments 1 and 3 suggest that the findings for first fixations cannot be explained by either object-to-feature or object-to-location mapping, even if these processes might contribute to, but not fully explain, the effect seen in all fixations.

Controlling for order effects

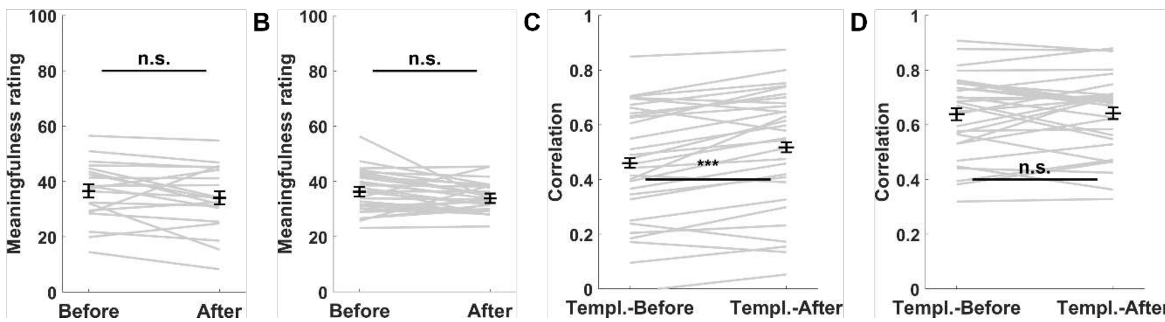
In the final analysis, we considered the possibility that order effects explain the key findings of Experiment 1 and 2. Specifically, we asked whether viewing the same two-tones for a second time without receiving prior object-knowledge could change fixation patterns such that they would resemble the patterns from the (real) templates. Recall that the design of Experiment 3 ensured that observers saw each two-tone image twice, each time without prior object-knowledge (Before and After conditions, respectively) and they also saw the real template for these two-tones in the following block. If the findings in Experiments 1 and 2 resulted, at least partly, from an order effect, we would expect that the similarity in fixation patterns in the (real) Template-After pair would be higher than in the (real) Template-Before pair in the current experiment.

The results were inconsistent with this 'second-viewing' hypothesis (Fig. 12D). The heatmap similarities between the real templates and the corresponding two-tones viewed in the Before and After conditions were not statistically different (Template-Before $M = 0.64$, $SD = 0.15$; Template-After $M = 0.64$, $SD = 0.14$; $t(29) = 0.22$, $p = 0.830$; Mdiff = 0, 95% CI = [-0.03, 0.03]). Moreover, a Bayes factor analysis provided evidence to support a lack of a difference ($BF = 0.20$). We found a similar result for first fixations (Template-Before $M = 0.52$, $SD = 0.21$; Template-After $M = 0.53$, $SD = 0.18$; $t(29) = 1.01$, $p = 0.323$; Mdiff = 0.02, 95% CI = [-0.02, 0.06]). Finally, the ROI analyses for both all fixations (Before: $M = 0.27$, $SD = 0.24$; After: $M = 0.27$, $SD = 0.33$; Before-After: $t(29) = 0.32$, $p = 0.212$; Mdiff = -0.02, 95% CI = [-0.05, 0.01]) and first fixations corroborated these findings (Before: $M = 0.31$, $SD = 0.34$; After: $M = 0.31$, $SD =$

0.33; Before-After: t(29) = 0.44, p = 0.666; Mdiff = 0.01, 95% CI = [-0.03, 0.05]).

Figure 12

Results of Experiment 3



Note. Meaningfulness ratings averaged per observer (A) and per image (B). C) Comparison of heatmap similarities between two-tones (viewed in the Before and After conditions) and their dummy templates (i.e., unrelated images). D) Comparison of heatmap similarities between two-tones (viewed in Before and After conditions) and their real templates.

Discussion

When an observer explores the environment with no specific task other than to obtain information, eye movements are typically directed towards object locations. Here, we consider this effect in light of emerging evidence highlighting the complex and intricate relationship between image-computable features and high-level object representations in visual perception. Specifically, we ask whether object-oriented eye-movements result from gaze being guided by high-level features or by objecthood, i.e., the fact that these features are bound into an object representation. We recorded eye movements in response to two-tone images, stimuli that appear as meaningless patches on initial viewing but, once relevant object-knowledge has been acquired, are organized into coherent and meaningful percepts of objects. In the current study, prior object-knowledge was provided in the form of template images, i.e., the unambiguous photographs from which the two-tone images had been generated. Across three experiments, fixation patterns on the same two-tone images differed substantially depending on whether observers experienced them as meaningless patches or organized them into object representations. In particular, when organized into object representations, we found that fixation patterns on two-tone images were more similar to those on templates, more focused on object-specific, pre-defined regions-of-interest, less dispersed, and more consistent across observers. These effects were evident from the first fixations on an image. Importantly, eye-movements on two-tone images were best explained by a simple model that takes into account both low-level features and high-level, knowledge-dependent object representations. Together, these findings highlight the importance of dynamic interactions between image-computable features and knowledge-driven perceptual organization in guiding information sampling via eye-movements in humans.

The idea that knowledge-driven object representations restructure human eye-movements is supported by both our general assessment of fixation distributions between two-tone images and template, and also by a more specific analysis focusing on fixations within regions-of-interest. These findings provide strong support for the hypothesis that objecthood per se contributes to the process of selecting fixation targets in images. In our experimental design, image-computable visual features are insufficient for object representations to emerge, their formation is dependent on prior object-knowledge. This characteristic of two-tone images is an important experimental tool: it allows us to decisively rule out the possibility that human oculomotor control during free viewing relies solely on image-computable features, regardless of whether these features are low- or high-level (Zelinsky & Bisley, 2015). The simple but critical result in this regard is the finding that eye-movement patterns differed dependent on whether observers had formed object representations despite the fact that the features in the stimuli remained identical. Of course, despite being highly impoverished, two-tone images might still contain some of the features that give rise to object representations in the Template images.

Note, however, that Before and After conditions have identical featural overlap with the Template condition, and differences in eye-movements between Before and After can therefore not be explained by this factor.

In addition to its use as an experimental tool, however, the dependence of object representations on prior knowledge is also important from a conceptual perspective. Specifically, the finding that fixations were guided by knowledge-dependent representations demonstrates that for the oculomotor system, objects cannot be conceptualised (exclusively) as image-computable, high-level features (Schütt et al., 2019). As highlighted in the introduction, Schütt and colleagues' (2019) study is one of the few that is explicit about this conceptualisation. While other studies have been less clear about exactly what constitutes an object, many treat them in a manner that (implicitly) equates object representations to complex high-level features (Borji & Tanner, 2016; Einhäuser et al., 2008; Nuthmann et al., 2020; Pajak & Nuthmann, 2013; Stoll et al., 2015). While these studies contribute to our understanding of the role of low- vs. high-level features in gaze control, they are not able to (and did not intend to) dissociate the influence of image-computable features from the influence of objecthood per se. Here, we show that objecthood that is relevant for guiding eye-movements is a characteristic that is distinct from the collection of any low- or high-level features. In our study, objecthood emerges in the interaction between prior object-knowledge and the visual input. Whether object representations that are relevant for oculomotor control are always distinct from the featural input is a difficult question that we cannot answer with our data. However, the size, the speed, and the incidental nature of these effects suggests that they might be characteristic of eye-movement control in everyday visual behaviour.

Our findings contrast in interesting ways with previous work that studied the relationship between eye-movements and object representations using ambiguous, bi-stable object stimuli (Kietzmann et al. 2011, 2015). These studies demonstrate that fixation patterns typical for one of the two interpretations of these stimuli often precede the emergence of the first percept corresponding to that interpretation. Thus, eye movements might play a role in the accumulation of image-computable evidence for competing stimulus interpretations, potentially suggesting that specific fixation patterns facilitate selection of one of two possible interpretations. In contrast to this finding, our results suggest that the influence of object representations precedes the first saccade. While our data provide no means to reconcile these contrasting findings, one possibility is a bi-directional relationship, where object representations guide eye-movements (as shown here) and eye-movements also support the generation of object representations (as shown in the studies by Kietzmann and colleagues). The use of a design that focusses on the role of eye movements in the accumulation of image-computable evidence for competing stimulus interpretations might be the reason why Kietzmann and colleagues mainly picked up on the latter component.

Manipulating low-level features is another approach aiming at dissociating feature-based and object-based effects. It was adopted by Stoll and colleagues (2015), who reduced contrast – a low-level feature contributing to saliency – in image areas containing objects. Given that in this study, objects are defined by high-level features, this approach provides a useful tool to assess the influence of low- vs. high-level features. It does not, however, allow for distinguishing between high-level features and objecthood per se as we do in the current study.

Equally important as the finding that knowledge-driven object representations guide human gaze is the fact that they do not fully determine the selection of fixation locations. While eye-movements on two-tone images changed once they elicited object representations such that fixation distributions became more similar to fixations on template images, substantial differences in eye-movements remained between these two conditions. Our linear combination analysis suggests that this disparity is systematic and can be explained by the differences in the features in two-tone vs. template images. In this analysis, we generated linear combinations with varying proportions of the heatmaps from the Template and Before conditions. We then assessed the similarities between these combined heatmaps and the heatmaps from the After condition. These similarities peaked for combined heatmaps that were determined by the fixation distributions from both the Template and the Before conditions (and not just one of them). The finding thus demonstrates that when observers experienced the percept of an object in the two-tone images (After condition), fixations were best explained by a combination of the factors guiding eye movements in the Before and the Template conditions. Specifically, even when observers perceived an object in the two-tone images, their eye movements were only partly determined by the factors that guide eye movements in response to the template image. The image-computable features that drive eye movements in response to two-tone images when no object is perceived (Before condition) still made a substantial contribution to gaze guidance. Note that the linear combination analysis was conducted on a per-image basis. The finding that both features and objecthood contribute to eye-movement control can therefore not be explained by averaging across different images, with some leading to purely feature-driven and other to purely representation-driven eye-movement control.

The finding that features remain important for eye-movement control even after having been bound into a high-level object representation potentially challenges some of the strong claims regarding the role of features vs. objects in gaze guidance. For instance, the cognitive relevance theory (Henderson et al., 2009) proposes that visual features do not contribute to oculomotor control directly but provide the means to generate a representation of potential fixation locations that have not yet been ranked for priority. High-level factors operate on this ‘flat landscape’ to determine the ultimate fixation locations. In other words, features are important only as potential carriers of higher-level representations and do not contribute to eye-movement control by themselves. According to this idea, as long as visual features give rise to similar object representation, these representations should guide eye movements towards similar locations, independently of the specific characteristics of features. Therefore, to the extent to which two-tones and templates lead to similar object representations, both image types should result in similar eye-movement patterns independent of their featural differences.

Contrasting with this notion, in the analysis of linear combinations, we found that the specific features that support these high-level representations continue to exert a sizeable influence on eye-movements. Specifically, we demonstrate that the same features that guided eye-movements when no object representation was present (Before condition) still had an influence on gaze guidance when an object representation had been generated (After). Therefore, to the extent to which two-tones and templates lead to similar object representations, we would have expected both image types to result in similar eye-movement patterns independent of their featural differences. Contrasting with this notion, we found that, while features can be flexible carriers of object representations that guide eye-movements as predicted by the cognitive relevance theory, the specific features that support these high-level representations persist to exert a sizeable influence.

In terms of the time-course of eye-movements, we provide clear evidence that already the first fixations after image onset are affected by objecthood. Interestingly, however, the linear combination analysis indicates that for first fixations the relative influence of features is stronger – and, therefore, the relative influence of objecthood weaker – compared to later fixations. Thus, while the influence of knowledge-dependent object representations emerges quickly, the linear combination analysis suggests that the effects of knowledge-driven perceptual organization continue to build beyond the first fixation, by contrast to the effects of features. Nevertheless, our data suggest that the influence of knowledge-dependent object representations emerges quickly and exerts an influence from the earliest fixations.

At image onset, when the eyes are stationary prior to the first saccade, most of the image is viewed via peripheral vision with only a small part being inspected with high-resolution foveal vision. The analysis of the first fixations therefore suggest that the visual system is able to generate knowledge-dependent object representations quickly and largely based on information from peripheral vision. Due to the optical, anatomical, and neurophysiological characteristics of the primate visual system, peripheral vision is limited in various respects (Rosenholtz, 2016), but there is good evidence that it provides enough information to generate a gist representation of a visual scene that can guide subsequent eye movements (Anderson, Donk, & Meeter, 2016; Castelhano & Henderson, 2007; Melissa L.H. Võ & Schneider, 2010). Exactly how detailed this gist representation is, which features it contains, and whether objects are represented varies depending on a number of different factors (Malcolm, Groen, & Baker, 2016; Wallis, Bethge, & Wichmann, 2016). Note, however, that this question is of limited relevance in the current context because features in two-tone images – independently of whether they are viewed by foveal or peripheral vision – are necessary but, by themselves, not sufficient to determine the high-level object representations we study here. However, one notion that might help in explaining the rapid influence of knowledge-dependent object representations on eye movements is provided by the suggestion that object recognition involves a predictive process that is triggered by low spatial-frequencies in the input (Bar et al., 2006; Bar, 2003, 2004, 2021; Bullier 2001). Specifically, low spatial-frequency information is thought to be fed forward by fast projections to high-level brain systems that connects this rudimentary input to prior object-knowledge. This process narrows down the search space of possible hypotheses about object identities in the input, thereby scaffolding and shaping a more precise perceptual experience of the input. It is therefore tempting to speculate that, in our experiment, first fixations were guided by object representations that are based on the process that links impoverished low spatial-frequency image content to prior knowledge, while later fixations might be based on fuller object representations. This idea rests on the assumption that two-tone images provide low spatial-frequency information to peripheral vision that allows the linking of two-tone images to memory representations of template images. Given that the image-processing operations required to generate two-tone images mainly affect high spatial-frequency components and have less impact on low spatial frequencies, this assumption seems plausible.

While our analyses mainly focused on locations of fixations, other aspects of oculomotor control are also influenced by knowledge-dependent perceptual organization. Specifically, we observed a decrease in saccade length and an increase in fixation duration when two-tone images were organized into object representations (After condition) compared to when they were not (Before condition). Both changes are indicative of a shift from image exploration to image exploitation (Gameiro et al., 2017; Kaspar et al., 2013), an interpretation that was also supported by the decrease in entropy across the two conditions. The oculomotor system constantly has to decide whether to keep the eyes still in order to be able to further inspect the currently fixated scene region – a process referred to as exploitation –, or to perform a saccade to explore another part of the image. Interestingly, in our study, the shift from exploration to exploitation went along with an increase in the amount of fixations landing on objects. This finding suggests that the visual system prioritizes objects in a specific way: it exploits object locations for further information while abandoning exploration of the remaining parts of the image. In other words, our data demonstrate that clusters of features that are bound into, and provide support for, object representations become interesting for the visual system over non-object related feature clusters (for a similar finding, see Król & Król, 2019). The shift from exploitation to exploration once objecthood is established also leads to higher consistency across observers. This finding suggests that guidance of exploration is either more idiosyncratic or that image-computable features that are not bound into object representations do not provide strong constraints for oculomotor control. Conversely, object representations, even when supported by exactly the same features, have a structuring or normative effect on information sampling. In other words, while observers explore features in different ways, they exploit objects in similar ways.

In summary, we demonstrate that gaze guidance is best understood by dynamic interactions between image-computable features and knowledge-dependent perceptual organization. Specifically, our findings demonstrate the importance of objecthood per se – i.e., representations that are not reducible to image-computable features – in oculomotor control but also indicate a persistent contribution of object-independent features. We demonstrate that when visual input remains identical, the emergence of knowledge-dependent object representations substantially restructures information sampling via eye-movements. However, we also show that even when image-computable features are bound into object representations, they still retain some influence on eye movements, challenging the idea that the role of features is limited to being carriers for high-level representation without direct influence on eye-movements. Finally, we also show that the emergence of object representations results in an overall change of the information-sampling strategy of the visual system, leading to the prioritization of information extraction from features that are bound into object representations, at the expense of exploration of the entire image.

CRediT authorship statement:

M.P.: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing - Original Draft, Writing - Review & Editing

E. v.d. H.: Methodology, Writing - Review & Editing

C.T.: Conceptualisation, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision, Resources

References

Alfandari, D., Belopolsky, A. V., & Olivers, C. N. L. (2019). Eye movements reveal learning and information-seeking in attentional template acquisition. *Visual Cognition*, *27*(5–8), 467–486. https://doi.org/10.1080/13506285.2019.1636918

Anderson, N. C., Donk, M., & Meeter, M. (2016). The influence of a scene preview on eye movement behavior in natural scenes. *Psychonomic Bulletin & Review*, 1–8. https://doi.org/10.3758/s13423-016-1035-4

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., & Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision*, *15*(5), 9. https://doi.org/10.1167/15.5.9.doi

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609. https://doi.org/10.1162/089892903321662976

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454. https://doi.org/10.1073/pnas.0507062103

Bar, M. (2021). From Objects to Unified Minds. *Current Directions in Psychological Science*, 30(2), 129–137. https://doi.org/10.1177/0963721420984403

Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of einḧauser et al.’s data. *Journal of Vision*, *13*(10), 1–4. https://doi.org/10.1167/13.10.18

Borji, A., & Tanner, J. (2016). Reconciling Saliency and Object Center-Bias Hypotheses in Explaining Free-Viewing Fixations. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(6), 1214–1226. https://doi.org/10.1109/TNNLS.2015.2480683

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4).

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, 36(2-3), 96–107. http://doi.org/10.1016/S0165-0173(01)00085-6

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 740–757. https://doi.org/10.1109/TPAMI.2018.2815601

Castelhano, M. S., & Henderson, J. M. (2007). Initial Scene Representations Facilitate Eye Movement Guidance in Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753–763. https://doi.org/10.1037/0096-1523.33.4.753

Cerf, M., Paxon Frady, E., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12), 1–15. https://doi.org/10.1167/9.12.1

Christensen, J. H., Bex, P. J., & Fiser, J. (2015). Prior implicit knowledge shapes human threshold for orientation noise. *Journal of Vision*, *15*(9), 1–15. https://doi.org/10.1167/15.9.24

Clarke, A. D. F., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41–51. https://doi.org/10.1016/j.visres.2014.06.016

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Drewes, J., Trommershäuser, J., & Gegenfurtner, K. R. (2011). Parallel visual search and rapid animal detection in natural scenes. *Journal of Vision*, *1120*(2), 1–21. https://doi.org/10.1167/11.2.20

Einhauser, W. (2013). Objects and saliency: Reply to Borji et al. *Journal of Vision*, *13*(10), 20–20. https://doi.org/10.1167/13.10.20

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18.1-26. https://doi.org/10.1167/8.14.18

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3). https://doi.org/10.1167/8.3.3

Federico, G., & Brandimonte, M. A. (2019). Tool and object affordances: An ecological eye-tracking study. *Brain and Cognition*, *135*(May), 103582. https://doi.org/10.1016/j.bandc.2019.103582

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences*, *7*(6), 252–256. https://doi.org/10.1016/S1364-6613(03)00111-6

Flounders, M. W., González-García, C., Hardstone, R., & He, B. J. (2019). Neural dynamics of visual ambiguity resolution by perceptual prior. *ELife*, *8*, 1–25. https://doi.org/10.7554/eLife.41861

Foulsham, T., & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology: General*, *142*(1), 41–56. https://doi.org/10.1037/a0028227

Gameiro, R. R., Kaspar, K., König, S. U., Nordholt, S., & König, P. (2017). Exploration and Exploitation in Natural Viewing Behavior. *Scientific Reports*, *7*(1), 1–23. https://doi.org/10.1038/s41598-017-02526-1

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363. https://doi.org/10.1038/nrn3476

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., & Scholte, H. S. (2013). From Image Statistics to Scene Gist: Evoked Neural Activity Reveals Transition from Low-Level Natural Image Structure to Scene Category. *Journal of Neuroscience*, 33(48), 18814–18824. http://doi.org/10.1523/JNEUROSCI.3128-13.2013

Harel, J., Koch, C., & Perona, P. (2007). Graph-Based Visual Saliency. In *Advances in Neural Information Processing Systems 19* (Vol. 19, pp. 545–552). The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0073

Hayes, T. R., & Henderson, J. M. (2020). Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, and Psychophysics*, *82*(3), 985–994. https://doi.org/10.3758/s13414-019-01849-7

Hayes, T. R., & Henderson, J. M. (2021). Looking for Semantic Similarity: What a Vector-Space Model of Semantics Can Tell Us About Attention in Real-World Scenes. *Psychological Science*, *32*(8), 1262–1270. https://doi.org/10.1177/0956797621994768

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(October). https://doi.org/10.1038/s41562-017-0208-0

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2021). Meaning maps capture the density of local semantic features in scenes: A reply to Pedziwiatr, Kümmerer, Wallis, Bethge & Teufel (2021). *Cognition*, (January), 104742. https://doi.org/10.1016/j.cognition.2021.104742

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856. https://doi.org/10.3758/PBR.16.5.850

Horga, G., & Abi-Dargham, A. (2019). An integrative framework for perceptual disturbances in psychosis. *Nature Reviews Neuroscience*, *20*(12), 763–778. https://doi.org/10.1038/s41583-019-0234-1

Hsieh, P.-J. J., Vul, E., & Kanwisher, N. (2010). Recognition alters the spatial pattern of fMRI activation in early retinotopic cortex. *Journal of Neurophysiology*, *103*(3), 1501–1507. https://doi.org/10.1152/jn.00812.2009

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192–1205. https://doi.org/10.1016/j.visres.2011.03.010

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Judd, T., Durand, F., Torralba, A., Azam, S., Gilani, S. O., Jeon, M., … Torralba, A. (2012). A Benchmark of Computational Models of Saliency to Predict Human Fixations. In *MIT Technical Report* https://doi.org/10.5220/0005678701340142

Kaspar, K., Hloucal, T. M., Kriz, J., Canzler, S., Gameiro, R. R., Krapp, V., & König, P. (2013). Emotions' Impact on Viewing Behavior under Natural Conditions. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0052737

Kietzmann, T. C., & König, P. (2015). Effects of contextual information and stimulus ambiguity on overt visual sampling behavior. *Vision Research*, *110*(Part A), 76–86. https://doi.org/10.1016/j.visres.2015.02.023

Kietzmann, Tim C., Geuter, S., & König, P. (2011). Overt Visual Attention as a Causal Factor of Perceptual Awareness. *PLoS ONE*, *6*(7), e22614. https://doi.org/10.1371/journal.pone.0022614

Kilpelaïnen, M., & Georgeson, M. A. (2018). Luminance gradient at object borders communicates object location to the human oculomotor system. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-19464-1

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(14).

Kourtzi, Z., & Connor, C. E. (2011). Neural representations for object perception: Structure, category, and adaptive coding. *Annual Review of Neuroscience*, *34*, 45–67.

https://doi.org/10.1146/annurev-neuro-060909-153218

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Król, M., & Król, M. (2019). The world as we know it and the world as it is: Eye-movement patterns reveal decreased use of prior knowledge in individuals with autism. *Autism Research*, *12*(9), 1386–1398. https://doi.org/10.1002/aur.2133

Kroner, A., Senden, M., Driessens, K., & Goebel, R. (2020). Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, *129*, 261–270. https://doi.org/10.1016/j.neunet.2020.05.004

Kümmerer, M., Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., … Torrabla, A. (2020). MIT/Tübingen Saliency Benchmark. Retrieved from https://saliency.tuebingen.ai/

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding Low-and High-Level Contributions to Fixation Prediction. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4799–4808. https://doi.org/10.1109/ICCV.2017.513

Wolf, C., & Lappe, M. (2021). Salient objects dominate the central fixation bias when orienting toward images. *Journal of Vision*, *21*(8), 1–21. https://doi.org/10.1167/jov.21.8.23

Lengyel, G., Nagy, M., & Fiser, J. (2021). Statistically defined visual chunks engage object-based attention. *Nature Communications*, *12*(1), 1–12. https://doi.org/10.1038/s41467-020-20589-z

Lengyel, G., Žalalytė, G., Pantelides, A., Ingram, J. N., Fiser, J., Lengyel, M., & Wolpert, D. M. (2019). Unimodal statistical learning produces multimodal object-like representations. *ELife*, *8*, 1–21. https://doi.org/10.7554/eLife.43942

Liang, H., Gong, X., Chen, M., Yan, Y., Li, W., & Gilbert, C. D. (2017). Interactions between feedback and lateral connections in the primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(32), 8637–8642. https://doi.org/10.1073/pnas.1706183114

Lyu, M., Choe, K. W., Kardan, O., Kotabe, H. P., Henderson, J. M., & Berman, M. G. (2020). Overt attentional correlates of memorability of scene images and their relationships to scene semantics. *Journal of Vision*, *20*(9), 2. https://doi.org/10.1167/jov.20.9.2

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making Sense of Real-World Scenes. *Trends in Cognitive Sciences*, 20(11), 843–856. http://doi.org/10.1016/j.tics.2016.09.003

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain)*, *200*(1140), 269–294. https://doi.org/10.1098/rspb.1978.0020

Masciocchi, C. M., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, 9(11). https://doi.org/10.1167/9.11.1

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology*, *11*(4), 219–226. https://doi.org/10.1037/h0083717

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*. Retrieved from https://cran.r-project.org/package=BayesFactor

Neri, P. (2014). Semantic control of feature extraction from natural scenes. *Journal of Neuroscience*, 34(6), 2374–2388. http://doi.org/10.1523/JNEUROSCI.1755-13.2014

Neri, P. (2017). Object segmentation controls image reconstruction from natural scenes. In *PLoS Biology* (Vol. 15). https://doi.org/10.1371/journal.pbio.1002611

Noton, D., & Stark, L. (1971). Scanpaths in Eye Movements during Pattern Perception. *Science*, *171*(3968), 308–311. https://doi.org/10.1126/science.171.3968.308

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 20. https://doi.org/10.1167/10.8.20

Nuthmann, A., Schütz, I., & Einhäuser, W. (2020). Salience-based object prioritization during active viewing of naturalistic scenes in young and older adults. *Scientific Reports*, *10*(1), 22057. https://doi.org/10.1038/s41598-020-78203-7

Ongchoco, J. D. K., & Scholl, B. J. (2019). How to Create Objects With Your Mind: From Object-Based Attention to Attention-Based Objects. *Psychological Science*, *30*(11), 1648–1655. https://doi.org/10.1177/0956797619863072

Pajak, M., & Nuthmann, a. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision*, *13*(2013). https://doi.org/10.1167/13.5.2

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021a). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition*, *206*(10), 104465. https://doi.org/10.1016/j.cognition.2020.104465

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021b). There is no evidence that meaning maps capture semantic information relevant to gaze guidance: Reply to Henderson, Hayes, Peacock, and Rehrig (2021). *Cognition*, (April), 104741. https://doi.org/10.1016/j.cognition.2021.104741

Pilarczyk, J., & Kuniecki, M. J. (2014). Emotional content of an image attracts attention more than visually salient features in various signal-to-noise ratio conditions. *Journal of Vision*, *14*(12), 4–4. https://doi.org/10.1167/14.12.4

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600.

https://doi.org/10.1126/science.aan3458

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rosenholtz, R. (2016). Capabilities and Limitations of Peripheral Vision. *Annual Review of Vision Science*, *2*, 437–457. https://doi.org/10.1146/annurev-vision-082114-035733

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R., & Wichmann, F. A. (2019). Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision*, *19*(3). https://doi.org/10.1167/19.3.1

Self, M. W., Jeurissen, D., van Ham, A. F., van Vugt, B., Poort, J., & Roelfsema, P. R. (2019). The Segmentation of Proto-Objects in the Monkey Primary Visual Cortex. *Current Biology*, *29*(6), 1019-1029.e4. https://doi.org/10.1016/j.cub.2019.02.016

Self, M. W., van Kerkoerle, T., Supèr, H., & Roelfsema, P. R. (2013). Distinct Roles of the Cortical Layers of Area V1 in Figure-Ground Segregation. *Current Biology*, 2121–2129. https://doi.org/10.1016/j.cub.2013.09.013

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research*, *107*, 36–48. https://doi.org/10.1016/j.visres.2014.11.006

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14). https://doi.org/10.1167/7.14.4

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6–7), 1029–1054. https://doi.org/10.1080/13506280902764539

Teufel, C., Dakin, S. C., & Fletcher, P. C. (2018). Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-28845-5

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews*

Neuroscience, *21*(4), 231–242. https://doi.org/10.1038/s41583-020-0275-5

Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P. R., … Fletcher, P. C. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences*, *112*(43), 13401–13406. https://doi.org/10.1073/pnas.1503916112

Van der Linden, L., Mathôt, S., & Vitu, F. (2015). The role of object affordances and center of gravity in eye movements toward isolated daily-life objects. *Journal of Vision*, *15*(5), 1–18. https://doi.org/10.1167/15.5.8

Vincent, B. T., Baddeley, R., Correani, A., Troscianko, T., & Leonards, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition*, *17*(6–7), 856–879. https://doi.org/10.1080/13506280902916691

Võ, M. L. H., & Schneider, W. X. (2010). A glimpse is not a glimpse: Differential processing of flashed scene previews leads to differential target search benefits. *Visual Cognition*, *18*(2), 171–200. https://doi.org/10.1080/13506280802547901

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin*, *138*(6), 1172–1217. https://doi.org/10.1037/a0029333

Wallis, T. S. A., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision*, 16(2). http://doi.org/10.1167/16.2.4

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268. https://doi.org/10.1111/nyas.14321

Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and Limits of Models of

Fixation Selection. *PLoS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0024038

Wynn, J. S., Shen, K., & Ryan, J. D. (2019). Eye movements actively reinstate spatiotemporal mnemonic content. *Vision*, *3*(2). https://doi.org/10.3390/vision3020021

Zelinsky, G. J., & Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 154–164. https://doi.org/10.1111/nyas.12606