Research Article

Psychological Science
1–14

DOI: 10.1177/09567976241279198
www.psychologicalscience.org/PS

Gaze Behavior Reveals Expectations of Potential Scene Changes

Nicolas Roth[1,2], Jasper McLaughlin[1], Klaus Obermayer[1,2,3], and Martin Rolfs[1,3,4]

[1]Cluster of Excellence Science of Intelligence, Technische Universität Berlin; [2]Institute of Software Engineering and Theoretical Computer Science, Technische Universität Berlin; [3]Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany; and [4]Department of Psychology, Humboldt-Universität zu Berlin

Abstract

Even if the scene before our eyes remains static for some time, we might explore it differently compared with how we examine static images, which are commonly used in studies on visual attention. Here we show experimentally that the top-down expectation of changes in natural scenes causes clearly distinguishable gaze behavior for visually identical scenes. We present free-viewing eye-tracking data of 20 healthy adults on a new video dataset of natural scenes, each mapped for its potential for change (PfC) in independent ratings. Observers looking at frozen videos looked significantly more often at the parts of the scene with a high PfC compared with static images, with substantially higher interobserver coherence. This viewing difference peaked right before a potential movement onset. Established concepts like object animacy or salience alone could not explain this finding. Images thus conceal experience-based expectations that affect gaze behavior in the potentially dynamic real world.

Keywords

top-down attention, static vs. dynamic, animacy, free-viewing, real-world scenes, spatiotemporal expectations

Received 2/9/24; Revision accepted 7/30/24

Introduction

Humans explore the rich visual features of natural scenes by actively sampling information through eye movements. For a long time, research on where humans decide to move their eyes has focused on gaze behavior in static scenes. In contrast to the real world, however, humans who visually explore images in a controlled setting already know that the presented scene will not change over time. When exploring a dynamic scene, on the other hand, observers may have the top-down expectation that objects could move and the environment could evolve, even if these things appear static at the moment.

Given a static observer or camera, scene changes are characterized by object motion (e.g., a sitting person standing up), or sudden onsets, including changes in color or luminance (e.g., a traffic light changing color). The influence of the expectation of such scene changes on gaze behavior is difficult to assess because these changes by themselves—as bottom-up stimuli—have a strong effect on eye movements (Carmi & Itti, 2006; Itti, 2005). It is no surprise, therefore, that differences in gaze behavior between static and dynamic scenes have been explained with the influence of motion and flicker (Smith & Mital, 2013), which consistently attract attention and lead to a high interobserver coherence (Dorr et al., 2010; Mital et al., 2011). Here, we present an experimental paradigm that isolates the influence of the top-down expectation of scene changes from their bottom-up salience.

Previous studies have demonstrated that anticipation of motion can influence gaze behavior even in static scenes. Açık et al. (2014) compared eye-tracking data

Corresponding Author:
Nicolas Roth, Technical University Berlin, Cluster of Excellence Science of Intelligence
Email: roth@tu-berlin.de

of natural dynamic scenes and static frames taken from the same movies. They showed that implied motion in the static scenes is equally effective at attracting gaze during the first second of viewing as the real motion in the respective dynamic scene is. A similar viewing benefit, particularly for animate objects in static scenes, led to the animate-monitoring hypothesis (New et al., 2007), which suggests that animate objects capture attention because of the evolutionary benefit of their detection. This hypothesis is supported by animals attracting ultrarapid saccades in as little as 120 ms (Kirchner & Thorpe, 2006) and animate objects being more frequently detected in inattentional-blindness tasks (Calvillo & Hawkins, 2016; Calvillo & Jackson, 2014). Hence, animacy and contextual cues are important influences on attention allocation, even in static images. We hypothesized that a more general explanation for why implied motion and animate objects attract attention is that they signal scene locations that may potentially change.

To address this fundamental question, we presented observers with a variety of real-world scenes and assessed how they visually explored images compared with visually identical scenes in which temporal changes can be expected. We hypothesized that we would find a significant difference in gaze behavior between the conditions because of the distinct top-down expectations. Over time, we expected this effect to become stronger the closer the presentation time came to the onset of the expected scene change or motion. Across space, we expected the largest difference in gaze behavior in those parts of the scene to which people assigned the highest *potential for change*, meaning those places where independent raters expected motion or other scene changes to happen. We show that images indeed evoke different experience-based expectations than the potentially dynamic real world and that the newly introduced quantity potential for change explains this effect better than established measures of saliency or animacy.

Statement of Relevance

We use eye movements to actively sample information from an ever-changing environment. When we look at images such as paintings or photographs, however, we already know that nothing in the scene will change. In this work, we explored how the knowledge of when to expect scene changes influences where we look. To isolate the effect of this expectation and avoid conflating it with actual scene changes, we showed the same scenes to observers as static images and initially static but subsequently dynamic videos. Analyzing eye-tracking data for both conditions, we uncovered a substantial difference in exploration behavior for visually identical scenes: If observers expect movement or change, they more thoroughly explore parts of the scene where these are most likely. Our results reveal observers' experience-based expectations, reinforcing the idea that we can learn most about attention in a dynamic world by studying it in dynamic scenes.

Research Transparency Statement

General Disclosures

Conflicts of interest: All authors declare no conflicts of interest. **Funding:** This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2002/1 "Science of Intelligence" - project number 390523135. M.R. was supported by the Heisenberg program of the Deutsche Forschungsgemeinschaft (DFG grants RO3579/8-1 and RO3579/12-1). **Artificial intelligence:** No artificial intelligence assisted technologies were used in this research or the creation of this article. **Ethics:** The ethical review board of the Department of Psychology at Humboldt-Universität zu Berlin approved the experimental procedure.

Study Disclosures

Preregistration: The study was preregistered (https://doi.org/10.17605/OSF.IO/K2JTE) on 2022-09-30. The research hypotheses (https://doi.org/10.17605/OSF.IO/24GSB/), stimuli recording (https://doi.org/10.17605/OSF.IO/FWS93/), data acquisition (https://doi.org/10.17605/OSF.IO/95VXK/), and measured variables (https://doi.org/10.17605/OSF.IO/2M6D7) were *not formally* preregistered, however, the timestamps indicate that the contents have not been changed since 2022-09-30, prior to data collection which began on 2022-11-24. There were minor deviations from the preregistration (for details, see Supplementary File Table S1). The data preprocessing and statistical analysis methods were not preregistered. **Materials:** All study materials are publicly available (https://osf.io/vj5dr/files/osfstorage). **Data:** All primary data are publicly available (https://osf.io/x2gaz). **Analysis scripts:** All analysis scripts are publicly available (https://github.com/rederoth/LPA_experimental_code, registered at https://osf.io/vj5dr/files/osfstorage). **Computational reproducibility:** The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

Method

Participants

We recorded complete eye-tracking datasets of 20 participants with normal or corrected-to-normal vision (10 female; mean age: 27.95 years, range 21–38 years; 1 left-handed, 1 ambidextrous). We determined the sample size by computing 90% power contours as a function of sample size and trial number (Baker et al., 2021) on the basis of our pilot data ($N = 4$; 20 out of 80 scenes used in the pilot were replaced in the main study for wider variety). For a lower boundary, we calculated the effect size as the mean difference in potential for change (PfC; $\mu_\Delta = 0.80$) between conditions in the last 200 ms before the scenes unfroze and the corresponding within-subject (ws) standard deviation in PfC ($\sigma_{ws} = 1.56$) and between-subject (bs) standard deviation ($\sigma_{bs} = 0.37$), resulting in a desired sample size equal to or greater than 6. We decided on a sample size of 20—as preregistered before the pilot data was analyzed—so we would also be able to analyze the effect size depending on the presentation time.

In the main study, three additional participants could not be reliably calibrated because of reflections in their glasses, so they did not complete the session, and we excluded their data. Participants were recruited through word of mouth and campus mailing lists and received €10 per hour or course credits as compensation. The ethical review board of the Department of Psychology at Humboldt-Universität zu Berlin approved the experimental procedure, and we obtained written informed consent from all participants before including them in the study.

Apparatus

We implemented the experimental design in MATLAB (The MathWorks, Natick, MA) using the Psychophysics and EyeLink toolboxes (Cornelissen et al., 2002; Kleiner et al., 2007) on an Ubuntu 20.04 operating system. The experimental code can be accessed at https://github.com/rederoth/LPA_experimental_code/. Binocular eye movements were recorded with an EyeLink 1000 Plus tabletop system (SR Research, Osgoode, Ontario, Canada) operating at a sampling rate of 1000 Hz, and the participants' responses were collected using a standard American-English keyboard. To minimize head movement, participants used a chin rest. Visual stimuli were displayed on a wall-mounted 16:9 video-projection screen 180 cm in front of the study participants, measuring 150×84 cm (Stewart Luxus Series GrayHawk G4, Stewart Filmscreen, Torrance, CA), which approximately corresponds to $45.2° \times 26.4°$ of visual angle. We used a PROPixx projector (VPixx Technologies, Saint-Bruno-de-Montarville, Quebec, Canada) operating at its native vertical refresh rate of 120 Hz and a resolution of 1920 × 1080 pixels. To avoid high eccentricities, we showed images and videos in the central 1536×864 pixels (scaled using bilinear interpolation), corresponding to $38.2° \times 21.5°$. The scenes were presented in color and with a γ value of 2.2 and a luminance ranging from a minimum of 0.08 cd/m^2 to a maximum of 71.68 cd/m^2 on a black background in a dark room.

Stimulus design

The recorded dataset contains 80 different scenes, mostly of everyday situations (e.g., office work, traffic, zoo visit). Forty scenes depict at least one animate object (person or animal), and the other 40 scenes show only inanimate objects. We ensured variation in the number, semantic category, and location of objects, and in whether the scene was indoors or outdoors. A representative sample of scenes is shown in the Supplemental Material available online (Fig. S1). Each raw recording of a scene consists of a 10-s video. The videos were recorded with a Lumix DC GH-5 camera mounted on a tripod in 4K resolution at 25 frames per second. The first 5-s period of each video shows few scene changes or none at all, followed by a movement onset or other sudden onset, such as a change in luminance or color in a specific part of the scene. This resulted in changes to the scene of variable magnitude.

The central frame of the raw videos (right before the scene-change onset) was extracted and used for presentation in the *static-image* condition. In the *frozen-video* condition, we showed the same central frame for the initial 5-s period and then played either the second half of the raw video (high-change condition) or the first half reversed in time (low-change condition), both with 50% probability for each presented scene in the video blocks. Hence, different observers saw different videos in the high- versus low-change condition. If we had shown videos only in the high-change condition, observers could have expected a scene change as soon as the video began to move. We thus introduced the low-change condition to evoke the expectation that one or more objects in the scene might move or otherwise change, while keeping observers uncertain about the expectation of such salient events in any given scene. Reversing the frames of the low-change condition prevented a cut between the central frame and the video and was not noticeable to the observers because there was no directed motion in this part of the video. An example video that illustrates this editing process is available in the Supplemental Material (Video S1).

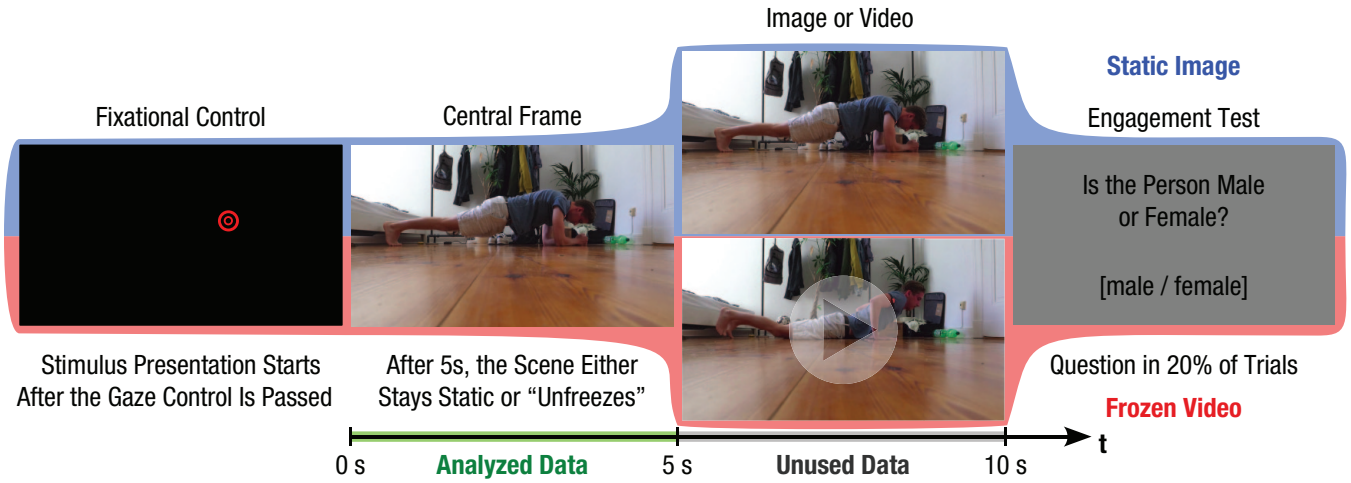

Fig. 1. Experimental design and procedure. At the beginning of each block, observers were informed whether the displayed scenes for free viewing were shown in the static-image condition (blue) or the frozen-video condition. In both conditions, the trial was started after the fixational control was passed. In the first 5 s, the presented scenes were visually identical in both conditions. In the static-image condition, the stimulus remained unchanged, whereas in the frozen-video condition, the scene "unfroze" and played as a dynamic video (randomly, either in the low- or high-change condition) for the remaining 5 s. To keep the observers engaged, we randomly asked a simple question about the scene content after 20% of the trials.

Procedure

The experiment was performed in a single session with a blocked design consisting of four static-image blocks and four frozen-video blocks (Fig. 1), which were randomly interleaved. After we tested their visual acuity, participants were instructed to freely explore the scenes as they wished. They were informed that they would see a total of 80 scenes, each occurring once during a static-image block and once during a frozen-video block; there would be occasional questions to ensure that they were still attending to the scene content. Each block contained 20 trials in which a randomly sampled scene was shown. Text displayed on the screen before each block indicated whether the upcoming block contained images or videos. To ensure that participants had read the information, we asked them to confirm, using the right or left arrow keys respectively, whether the next block would be images or videos. Each trial was preceded by a fixation control, which displayed a red circle of 0.22° radius filled with a red dot on a black background. The position of the fixation target was randomly sampled on the rectangular space on the screen that the video was to occupy. Retrieving samples at a maximum rate of 1000 Hz, fixation control was passed after 400 ms of fixation within a 2° radius around the target red circle. After 2 s without valid fixation or after 50 broken fixations or refixations (set high, to give participants the time to look away before actively starting the trial by fixating correctly), the trial would be aborted and repeated at the end of the respective block, and a new 9-point calibration would be requested. Upon successful fixation control, the scene content was presented.

In a static-image block, the central frame of the raw video scene was presented as a static image for 10 s. In a frozen-video block, we first presented the central frame of the raw video for 5 s, which was then seamlessly followed by 5 s of dynamic video (see Stimulus Design section). We showed videos with both low and high amounts of change to avoid priming participants' expectations too strongly. The video frames played at 24 frames per second (to synchronize with the projector's refresh rate, amounting to frame durations of 41.66 ms rather than the recorded 40 ms per frame). Thus, each block contained 200 s (20 × 10 s) of stimulus presentation time. A black background followed the presentation of each scene. After each trial, there was a 20% chance that a predetermined question with two answer options appeared. We designed the questions to encourage a balanced exploration of the scene by ensuring that in less than half of the scenes the subject of the question was the object with the highest PfC score (cf. Fig. 1, where the question is about the person). For about 60% of the questions, exploring the objects with high PfC was not informative (e.g., "Is the scene outside or inside?" or "Are there any plants in the room?"). To ensure we would not introduce a bias between the static-image and frozen-video conditions, we randomly sampled from the same pool of questions, and all questions could be answered from the static image's visual information. Participants selected their answers to the questions at their own pace by using the arrow keys. Once the trials in a block were completed, participants were informed that

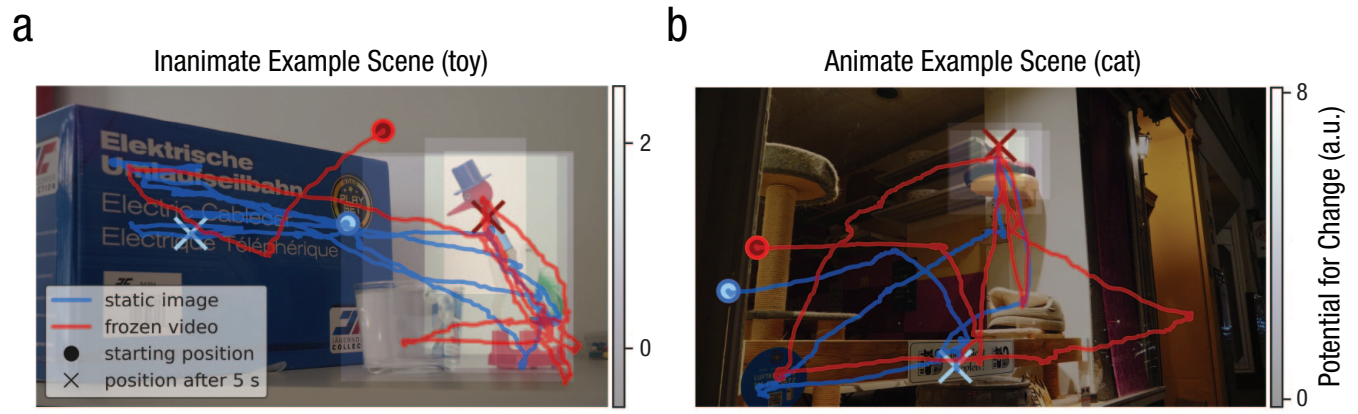

Fig. 2. Direct comparison of the raw-gaze data of one observer when exploring the scene as an image (blue) or as a frozen video (red). The overlaid boxes denote the normalized potential for change (PfC) annotations; bright (dark) indicates high (low) normalized PfC values; a.u. = arbitrary units.

the block was finished. They could take a break after each block, and there was a mandatory short break after Block 4.

Scene annotation

For each of the 80 scenes in our recorded dataset, we annotated the central frame (used in the static-image condition and as the initial frame of the frozen-video condition) with pixelwise maps that quantify the PfC rating and the visual saliency and that segment the scene in semantic objects.

PfC rating. The PfC in each scene was assessed by five independent labelers. Each labeler was asked to draw a single bounding box around the object or area in each scene where they think the PfC is highest—that is, where they expect movement or other changes in the scene to be most likely. The labelers had the option to redraw the box for each scene before they confirmed their selection, and they fulfilled this task without time constraints. Their ratings show a high level of consistency (see Fig. S2 in the Supplemental Material) with a mean pairwise *intersection over union* (defined for two bounding boxes as the area of intersection divided by the area of union) of 0.423. In 74% of scenes, the bounding boxes of all labelers overlapped (i.e., all assessments include the point of maximum PfC). The quantitative pixelwise PfC map for each scene was calculated by adding all bounding boxes with an equal weight of one, dividing this map by its standard deviation, and subtracting the mean value (see Fig. 2 for examples).

Visual saliency. We computed high-level salience maps for each scene with the DeepGaze IIE model, using the provided center bias from the MIT1003 dataset (Linardos et al., 2021). By combining multiple pretrained deep-neural-network backbones, this model generalizes well to unseen datasets and is considered the current state-of-the-art method in predicting fixation probability densities on static images. We also computed low-level salience maps on the basis of color, orientation, and luminance (cf. Itti et al., 1998) using the SaliencyToolbox (Walther & Koch, 2006). To better match the PfC maps, we normalized the saliency maps in the same way (M = 0, SD = 1).

Object segmentation. We segmented object masks using Mask R-CNN (He et al., 2017) as implemented in the Detectron 2 framework (Wu et al., 2019). We manually refined the semantic segmentation masks and controlled for obvious segmentation mistakes (e.g., if prominent objects were not detected). We calculated object-based PfC or saliency scores as the average value of the respective map within the object mask. The object-based metrics are then scaled between 0 and 1, where 1 (0) corresponds to the object with each scene's highest (lowest) score.

Gaze data

We classified eye-movement events using the velocity-based eye-movement-event detection algorithm REMoDNaV (Dar et al., 2021). This algorithm can distinguish between fixation, smooth pursuit, saccades, post-saccadic oscillations, and blinks and is applicable to static and dynamic scenes. To account for the high quality of our data, we adapted the default parameters of the adaptive noise level (5.0 → 3.0, as in Nyström & Holmqvist, 2010), reduced the length of the Savitzky-Golay filter (0.019 → 0.005), and increased the dilatation of missing data (0.01 → 0.025). For the analysis of the gaze data, we

exclusively used the first 5 s of each trial when the stimulus was static in both conditions; therefore, no smooth pursuit occurred. We included the gaze positions during fixation events only in the quantitative evaluation—that is, we excluded the gaze data during saccades, post-saccadic oscillations, and blinks from further analysis. We considered a gaze position (x, y) to be on an object if pixel (x, y) corresponded to an object mask in the semantic segmentation map. To account for potential tracking inaccuracies, we assigned the gaze position to all masks within a tolerance radius of 0.5° of visual angle if (x, y) lay on no object mask (the background).

Coherence measure

We measured the interobserver coherence by calculating the normalized scanpath saliency (NSS) in a "leave-one-out" fashion (Dorr et al., 2010). The time-resolved NSS score for each observer was based on the value of their gaze position on a spatiotemporal map of the smoothed fixation locations of all other observers. For each trial of observer i, we generated a fixation-density map F_t for every point in time t, which contains Gaussians around the gaze positions $x_j(t)$ of other observers viewing the same scene at this time:

$$F_i^t(x) = \sum_{i \neq j} e^{-\frac{(x(t) - x_j(t))^2}{2(\sigma_x^2 + \sigma_y^2)}}. \quad (1)$$

We set the parameters $\sigma_x = \sigma_y = 1.5°$ of visual angle as a tolerance on the basis of the size of the foveal region, while confirming that the results do not qualitatively change for alternative choices ($\sigma_{x,y} = 1$ or 2°).

This map is generated for all $t \in \mathcal{F}_i$, which is the set of time points during which observer i is fixating. The normalized coherence score $C_i(t)$ for observer i over time was then calculated as

$$C_i(t) = \frac{F_i^t(x(t)) - \text{Mean}(F_i^t)}{\text{Std}(F_i^t)}. \quad (2)$$

We confirmed the results of the NSS analysis with a second coherence measure called *attentional synchrony* (Smith & Mital, 2013; see Fig. S4 in the Supplemental Material).

Cluster-based permutation significance testing

We determine significant differences in time-series data through cluster-based permutation tests, as described in detail by Ehinger (2016). For each time point, we computed the difference between the static-image and frozen-video conditions and the corresponding t value on the mean and standard error between participants. We set a critical t value of $t_c = 2.093$, corresponding to a two-sided student t test with a 95% confidence interval (CI) and 19 degrees of freedom ($N = 20$). Consecutive time points exceeding t_c form clusters in time, which are then compared with a random baseline. For this, we randomly permuted the static-image and frozen-video labels for each time point 1,000 times. We identified the largest cluster above t_c in these random permutations and took the 95th percentile of the largest clusters as the cutoff value for significance in the original time series. Because a t value above t_c is required for the whole cluster, this provides a conservative estimate for the point in time when the conditions lead to significantly different behavior.

Results

We focused our analyses on the first 5 s of the gaze data, during which the presented scenes in the static-image and frozen-video conditions were visually identical. Therefore, the observed differences in exploration behavior between the two experimental conditions are exclusively based on the observers' top-down expectations.

Qualitative scanpath comparison

Despite identical visual scenes in the static-image and frozen-video conditions, observers' top-down expectations may differ substantially. For static images, observers know that nothing in the scene can change, whereas frozen videos will eventually unfreeze. We hypothesized that this high-level knowledge brings about systematic differences in exploration behavior. We can qualitatively observe this expected difference in the scanpaths shown in Figure 2, which are representative of the recorded data in the respective condition. Observers initially explore the scenes similarly in both conditions: They direct their gaze toward the center of the screen and quickly scan salient objects, including the objects with high PfC scores (the toy in Fig. 2a and the cat in Fig. 2b). In both conditions, the observers continue exploring the scene further after first looking at the objects with high PfC. In the static-image condition, observers show a tendency to investigate objects with rich visual features, like text. In the frozen-video condition, observers consistently returned to parts of the scene with high PfC toward the end of the initial 5-s period.

Systematic viewing differences in interobserver coherence

We first quantified observers' viewing behaviors by calculating the dynamic changes in interobserver

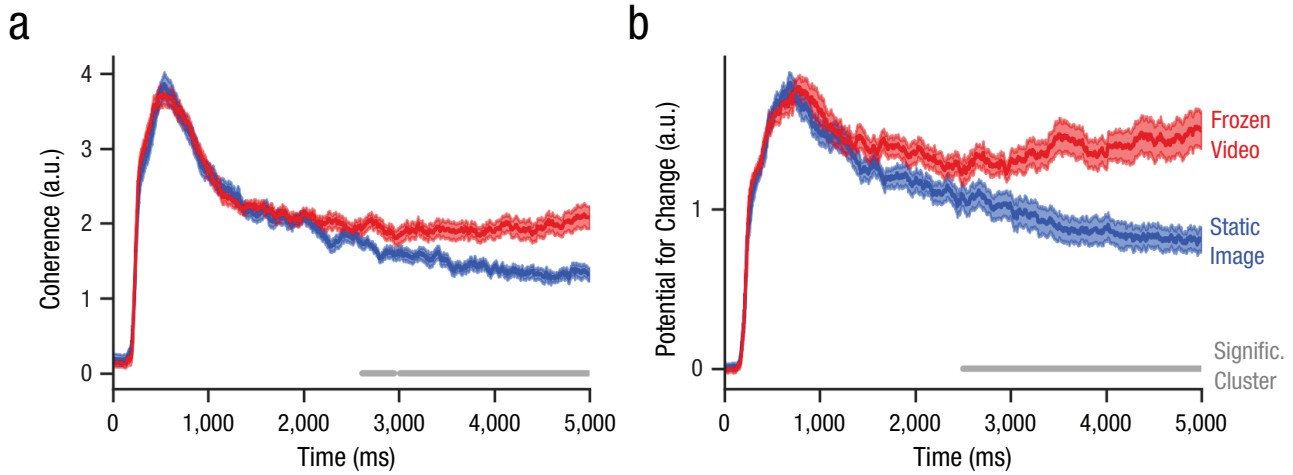

Fig. 3. Quantitative comparison between the static-image condition (blue) and the frozen-video condition (red), averaged across all scenes and observers. The shaded area denotes the standard error of the mean across observers (a.u. = arbitrary units). Lines at the bottom indicate periods in which the two conditions are significantly different (see Method: Cluster-Based Permutation Significance Testing in the Supplemental Material). Shown in (a) is interobserver coherence (normalized scanpath saliency [NSS] score; see Method: Coherence Measure); shown in (b) is average PfC score at the gaze position as a function of stimulus presentation time.

coherence (Fig. 3a), a measure of how consistent each observer's scanpath is with the time-resolved fixation map of all other observers (see Method: Coherence Measure; see Fig. S4 in the Supplemental Material for an alternative coherence measure). Coherence was normalized so that the random starting position in the fixation control for each trial led to an average initial coherence close to zero. Independent of the starting position, observers were highly consistent in exploring a scene's most salient and central objects with their first few eye movements. This is consistent with the literature on the elevated influence of saliency (Donk & Van Zoest, 2008) and center bias (Rothkegel et al., 2017) on the first fixations in a scene, resulting in a prominent peak in coherence between 530 and 540 ms in both conditions. Following this initial peak, coherence decreased again because of observers' individual viewing preferences (cf. de Haas et al., 2019). After 2,613 ms, the average interobserver coherence was significantly higher in the frozen-video condition compared with the static-image condition.

The higher coherence between observers in the frozen-video condition can be explained by observers' consistent exploration of parts of the scene with high PfC (Fig. 3b). We measured the PfC score over time as the normalized value of the PfC map at the current gaze position (see Method: Scene Annotation in the Supplemental Material). As for coherence, the random initial fixation position and the normalization of the PfC maps resulted in an initial average PfC score of around zero. The parts of the scene initially explored with high interobserver coherence correlate with a high PfC rating of the independent group of labelers. The PfC maps also show a center bias and correlate with salience (cf. Fig. S3 in the Supplemental Material). Hence, the peak in PfC roughly coincides with the peak in coherence independent of the viewing condition (692 ms in the static-image condition and 772 ms in the frozen-video condition). In the static-image condition, the average PfC score decreased monotonically afterward. In the frozen-video condition, in contrast, observers consistently showed a stronger tendency to explore parts of the scene with high PfC. As hypothesized, we found a gradual divergence between the two viewing conditions, with the largest difference in gaze behavior in the parts of the scene with high PfC ratings right before the video began to move.

In both the static-image and the frozen-video blocks, scenes were shown in randomized order, allowing us to explore potential effects of their viewing order (see Fig. S5 in the Supplemental Material). We found no connection between the PfC scores and the viewing order of the static images; that is, knowing what happened in the video did not influence the observers' exploration behavior of the image. In the frozen-video condition, the effect was enhanced if the same scene had been seen previously as a static image, suggesting that familiarity with the static scene leads to stronger expectations about potential movement in the unfreezing video.

Systematic object-based viewing differences

We next compared how the observers explored the individual objects in the scenes. For this, we subtracted

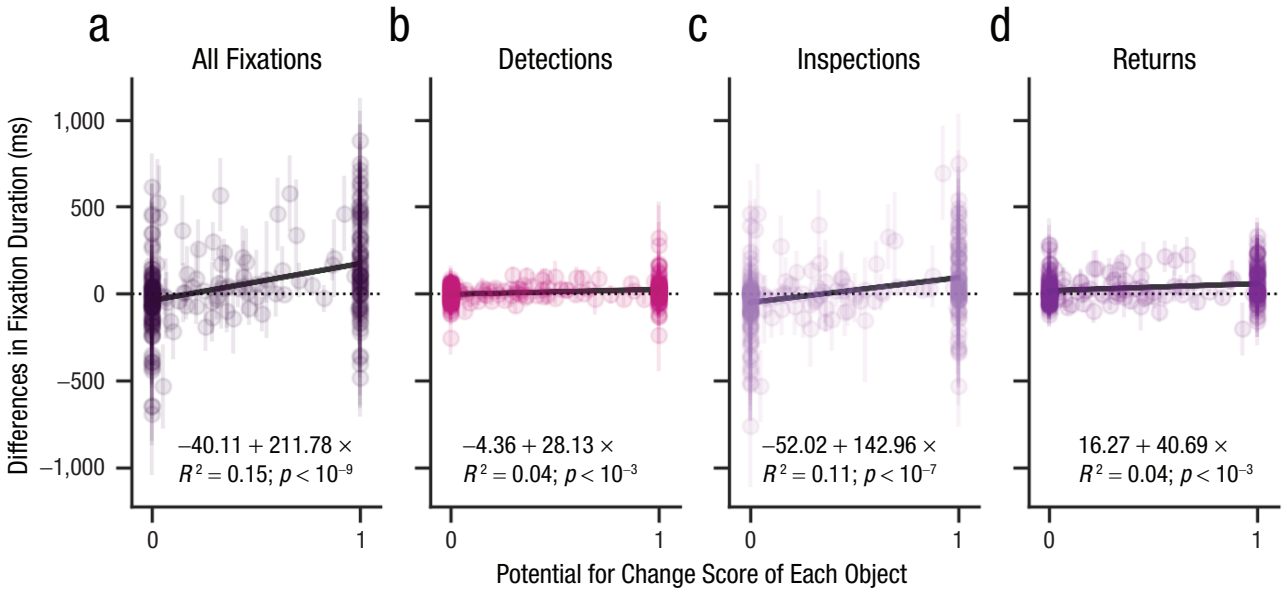

Fig. 4. Differences between the frozen-video condition and the static-image condition in (a) total fixation time, (b) detections, (c) inspections, and (d) returns for each object as a function of its PfC score. Linear regression results (intercept and slope in the equation, p value of the slope, and coefficient of determination R^2) are shown in each panel. Error bars for each object show the standard errors of the mean across participants.

the average total fixation duration per observer in the static-image condition from the average total fixation duration per observer in the frozen-video condition for each object in the dataset. Positive values hence correspond to objects that were fixated longer during the 5-s presentation time in the frozen-video condition compared with the first 5 s of the corresponding static-image trial. Indeed, some objects were consistently fixated for longer durations in one condition than in the other (Fig. 4a). Importantly, the object-based PfC score, calculated on the basis of the PfC ratings across the object-segmentation mask (see Method: Object Segmentation in the Supplemental Material), explained a significant amount of the variance of this difference. A linear regression indicated that objects with the highest PfC score in a scene (normalized to 1; see Method: Object segmentation) had, a larger difference in total fixation duration between the conditions (on average, 212 ms larger) than objects with the lowest PfC score (normalized to 0). Objects with a PfC of 0 were, on average, fixated for 40 ms longer in the static-image condition.

We further distinguished fixation events by dividing them into three distinct categories (Linka & de Haas, 2023; Roth et al., 2023): *Detections* (Fig. 4b) uncover an object for the first time; *inspections* (Fig. 4c) target the same object as the previous fixation and therefore further explore its details; and *return* events come back from elsewhere to revisit a previously fixated object (Fig. 4d). We did not expect a large effect of the condition on detection events but hypothesized a difference for inspections and returns. Indeed, observers reliably detected the objects in the scene with high PfC in both conditions, resulting in similar statistics for detection events (Fig. 4b). The average time spent on return events across all objects independent of PfC score was, on the other hand, 33 ms longer in the frozen-video condition compared with the static-image condition, whereas it was only 8 ms longer for inspections. The objects that were inspected, however, systematically differed between the conditions, with high-PfC objects receiving significantly more overt attention if a change could be expected (see Fig. 4c). Hence, observers in the frozen-video blocks were more likely to inspect the details within an object with high PfC and to return more often to these objects in anticipation of the movement onset.

Effect of animacy and saliency

We confirmed in an additional analysis that the observed effect is actually best explained by PfC and not by established concepts like animacy or saliency, which are correlated with PfC (see Fig. S3 in the Supplemental Material) and which have systematic effects on viewing behavior (Linardos et al., 2021; New et al., 2007). To analyze the effect of animacy, we split the data into the 40 scenes that contained only inanimate objects and

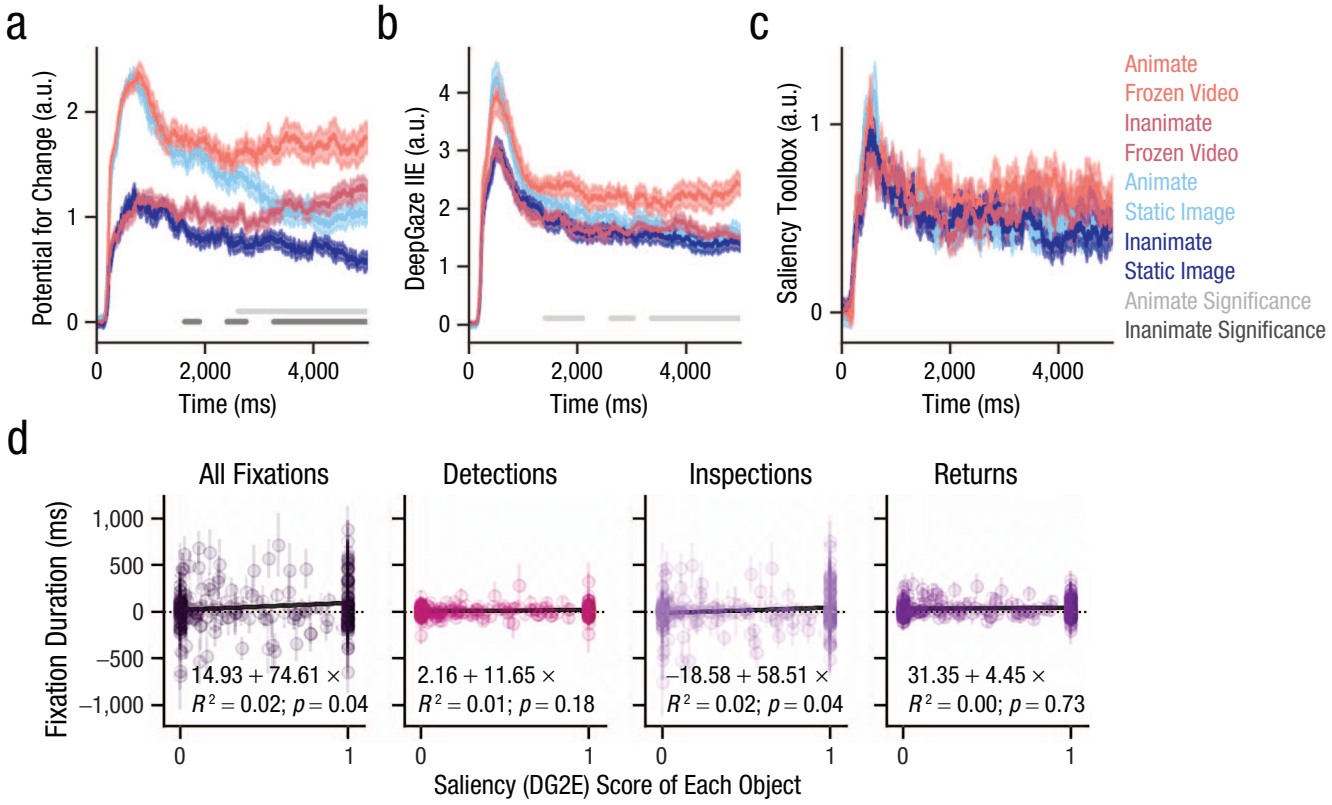

Fig. 5. Viewing differences quantified in terms of animacy and saliency of the scenes. We show comparison of (a) average potential for change (PfC) score, (b) visual saliency (quantified using DeepGaze IIE (DG2E); Linardos et al., 2021), and (c) SaliencyToolbox results (Walther & Koch, 2006) over time (analogous to Fig. 3b), plotted separately for scenes containing only inanimate objects (darker colors) and scenes containing animate objects. Significant differences between conditions for the inanimate and animate scenes are indicated by the dark gray or light gray lines at the bottom, respectively. The difference in total fixation duration between conditions, depending on the saliency score of each object, is illustrated in (d), with the results of the linear regression in each panel (analogous to Fig. 4). a.u. = arbitrary units.

the 40 scenes that included at least one human or animal (Fig. 5a). If the systematic viewing differences could be explained by the animate-monitoring hypothesis (Calvillo & Jackson, 2014; New et al., 2007), the discrepancies between static images and frozen videos would occur only in the animate scenes. Instead, we found that the inanimate scenes showed an equally sized effect. Overall, PfC was more predictive of where people look in animate scenes because such scenes contain features that tend to have a high PfC rating and consistently attract attention in particular faces (Broda et al., 2023; Hershler & Hochstein, 2005). The time course and actual effect size measured by the discrepancy in PfC between the static-image and frozen-video stimulus, however, was similar for animate and inanimate scenes (Fig. 5a). This was supported by a 2×2 analysis of variance of the accumulated PfC after 5 s in the frozen-video versus static-image experimental condition, $F(1, 76) = 23.2$, $p < 10^{-15}$, and for animate versus inanimate scenes, $F(1, 76) = 106.1$, $p < 10^{-15}$; there was no significant interaction, $F(1, 76) = 0.5$, $p = .47$.

Finally, we explored the alternative hypothesis that in anticipation of the scene-change onset, observers simply orient their gaze behavior more to the most salient parts of the scene. When we, as with the PfC score, plot the average saliency score quantified with state-of-the-art methods at the gaze positions over time (see Method: Visual Saliency in the Supplemental Material), the effect completely disappears for the inanimate scenes (Fig. 5b). The effect observed in the animate scenes is a consequence of the higher correlation of PfC and high-level saliency for animate objects (see Fig. S3 in the Supplemental Material). This was also confirmed by the object-based salience analysis (Fig. 5d), in which we plotted the same difference in total fixation duration for each object in the dataset, but as a function of the objects' saliency score. In comparison to the PfC score (Fig. 4), the variance in total difference in fixation durations explained by the linear regressions is small (2%, compared with 15% for PfC). Low-level saliency, computed on the basis of color, luminance, and orientations within an image, does not explain any viewing differences in our scenes (Fig. 5c).

Discussion

We showed that expectations about potential scene changes lead to systematically predictable viewing behavior. We disentangled the effect of the top-down expectation about potential scene changes from any bottom-up motion cues by presenting real-world scenes either as static images (10-s presentation time) or as videos (the initial frames were the same as the static images, shown for 5 s, but the videos then were unfrozen and turned into 5-s videos). Crucially, we evaluated the eye-tracking data for only the first 5 s of presentation time, when the visual stimulus is identical in both experimental conditions.

We found that there was almost no difference in the initial detection of objects in the scene (Fig. 4b) and that observers were highly coherent within the first second of exploration, irrespective of the experimental condition (Fig. 3a). This is consistent with previous findings of early saccades being strongly dependent on saliency, even if it is not relevant for a given task (Anderson et al., 2015), and with the "onset effect" (Dorr et al., 2010), in which the center bias leads to high interobserver coherence for the first few saccades. After this initially high interobserver coherence, our results showed a systematic difference in gaze behavior between the static-image and the frozen-video conditions. When expecting scene changes, observers consistently allocated their gaze more toward the parts of the scene that had a higher PfC, as assessed by independent raters (Fig. 3b). Similarly, objects with high PfC ratings were more thoroughly inspected and were returned to more often in the frozen-video condition (Fig. 4).

These findings imply that the PfC measure we introduced here plays a crucial role in affecting eye movements in real-world situations, where visual alterations in the environment might happen, even when it appears static at present. We calculated the PfC scores on the basis of the assessment of independent labelers who mapped where motion or scene changes could occur. This scoring is potentially correlated with other measures—for example, the highest information content, assuming the scene is dynamic. We expect all such semantic variations of metrics that intuitively capture the top-down expectations in this experiment to be highly correlated with PfC. Hence, we did not propose further alternative metrics. We did, however, ensure that concepts like implied motion, animacy, and saliency did not explain the observed effect.

Previous studies have shown that implied motion can attract gaze even in static scenes (Açık et al., 2014). Implied motion, which refers to motion deduced from static cues in the absence of real motion, did not occur in the investigated stimuli. The central frame, used for the first 5 s in the static-image and frozen-video conditions, was extracted before the change in the original video (see Video S1 in the Supplemental Material). Hence, the presented scenes show, by construction, only expected motion and not implied motion. Moreover, any implied motion would have been identical in static images and frozen videos, as they were visually identical. Thus, although the concept of implied motion is related to PfC, it cannot explain the observed viewing differences.

The animate-monitoring hypothesis (Calvillo & Hawkins, 2016; Calvillo & Jackson, 2014; New et al., 2007) suggests that animate objects capture attention because of their evolutionary importance for humans, and animate motion, in particular, has been shown to capture attention (Pratt et al., 2010). Under this rationale, one would expect increased monitoring of animate objects in the frozen-video condition compared with the static-image condition. Interestingly, however, the effect size in inanimate scenes (containing no animate object) was as large as in the animate scenes. Because animate objects can generate movement, PfC and animacy often correlate in our everyday experience, but we purposely decoupled the two in the recording of our dataset. Hence, we conjecture that effects typically ascribed to the animate-monitoring hypothesis may be better explained by a more general monitoring of PfC. For our dataset, we concluded that the PfC causes different top-down expectations between the conditions independent of an object's animacy.

Finally, we found that biologically inspired, low-level saliency maps (Itti et al., 1998; Walther & Koch, 2006) capture little variance in observed gaze behavior and cannot explain the difference between the experimental conditions (Fig. 5c). We also computed high-level saliency using the DeepGaze IIE network (Linardos et al., 2021), which predicts where people look on the basis of its training on human-gaze data on static scenes. The model predicted well where observers in this experiment look (see the high scores in Fig. 5b) but did not explain the viewing difference between the static-image and frozen-video conditions. The difference between conditions for animate scenes is a result of the correlation between high DeepGaze IIE predictions and PfC scores for animate objects—for instance, for faces (see Fig. S1 and Fig. S3 in the Supplemental Material). In the absence of task instructions, computational models of visual attention typically do not consider top-down influences (Kümmerer & Bethge, 2023). To reproduce the effect of experience-based expectations on viewing behavior, statistical models (e.g., Kümmerer et al., 2022; Linardos et al., 2021) could be retrained or fine-tuned on the gaze data presented here. Mechanistic models of gaze behavior (e.g., Roth et al.,

2023) can explicitly include the PfC maps to reproduce the anticipatory saccades in the frozen-video condition, but not take PfC into account for reproducing the behavior for the static-image condition.

A related concept to the PfC presented here is the construction of *meaning maps*. Henderson and Hayes (2017) introduced meaning maps as the spatial representation of semantic richness in a scene, obtained from context-free assessments by many raters of local scene patches in terms of "low and high meaning." Meaning maps are, therefore, defined only for static images and do not take the global scene context for individual objects into account. Similar to manipulations in which objects were presented in atypical contexts (Pedziwiatr et al., 2021), meaning maps would—by construction—not capture the expectation about movement or sudden onsets as quantified by PfC. Consequently, we would expect meaning maps to be both similarly predictive of where people look and incapable of explaining the viewing differences between the static-image and frozen-video conditions as high-level salience maps (cf. Fig. 5b and d).

Although expectations about the spatial structure in natural environments have been investigated extensively in recent years (Summerfield & De Lange, 2014; Vo et al., 2019; Wolfe, 2021), expectations about the temporal structure of dynamic real-world scenes have received less attention. It has been shown that humans make use of temporal regularities to guide behavior (Nobre & Van Ede, 2018). The interaction between anticipatory eye movements and object motion has been studied in smooth-pursuit experiments, revealing a significant impact of motion-direction expectancy on the initiation of anticipatory movements (Carneiro Morita et al., 2023; Damasse et al., 2018; Stewart & Fleming, 2023) and an increased prediction of visual motion through pursuit eye movements (Spering et al., 2011). Eye-tracking experiments during the interaction with the environment, like preparing food (Land & Hayhoe, 2001) or intercepting a ball in flight (Binaee & Diaz, 2019; G. Diaz et al., 2013; Fooken et al., 2021), show that in everyday life, eye movements are "proactive, anticipating actions rather than just responding to stimuli" (Land & Furneaux, 1997, p. 1231). This also holds true for gaze behavior during the free viewing of dynamic natural scenes, when observers' reactions to salient events often preceded or coincided with events in the videos (Vig et al., 2011). The contextual cues used to guide predictive eye movements have been studied systematically when following the puck in an ice hockey game (Goettker et al., 2021). In this example, kinematic cues, experience with the stimuli, and the amount of context information were critical factors for successful predictive eye movements (Goettker et al., 2023). The differences in top-down effects for real-world images and videos have not yet been explored. We deliberately designed our experiments to eliminate the influence of the task, dynamic (or implied) motion cues, and expertise by recording free-viewing gaze on initially static everyday scenes. Hence, we show that the experience-based expectations evoked by potentially dynamic scenes alone can lead to anticipatory eye movements.

The primary focus of this study was to demonstrate the existence of different expectations between static and (potentially) dynamic scenes and to show their effect on gaze behavior. More specific characteristics of this effect remain to be investigated. In our experimental design, the duration of the static first frame in the frozen-video blocks was always 5 s. Varying this freeze duration in future experiments to higher, lower, or unpredictable values would provide additional information about the time course of the expectation. Furthermore, follow-up experiments could investigate to what extent the experience-based expectations about dynamic scenes can be experimentally manipulated. In this study, we measured the gaze behavior of the observers (i.e., their allocation of overt attention). Although visual attention does not necessarily coincide with overt gaze behavior (Carrasco, 2011; Posner, 1980; Spering & Carrasco, 2015), the tight link between attention allocation and eye movements during free viewing is well established (Henderson, 2003). Hence, we expect that if we actively probed covert attention in this paradigm (cf. Dorr & Bex, 2013), the attended locations would correlate to the gaze behavior measured here.

In summary, our results demonstrate that the top-down expectations based on PfC explain substantial variance in gaze behavior in natural scenes, beyond the impact of implied motion, animacy, or visual saliency. These results reveal that images evoke different experience-based expectations than the (potentially dynamic) real world, emphasizing the significance of using dynamic scenes in our pursuit of understanding attention in ecologically valid environments. Indeed, given the robustness of the effect, it provides a potentially insightful marker in the context of developmental and educational psychology (Kaakinen, 2021; Kirkorian & Anderson, 2017) as well as for investigations of anticipation in sports (G. J. Diaz et al., 2012; Loffing & Cañal-Bruland, 2017).

Transparency

Action Editor: Zhicheng Lin
Editor: Simine Vazire

Author Contributions

Nicolas Roth: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing – original draft; Writing – review & editing.

Jasper McLaughlin: Data curation; Investigation; Methodology; Writing – review & editing.

Klaus Obermayer: Funding acquisition; Methodology; Project administration; Writing – review & editing.

Martin Rolfs: Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, or DFG) under Germany's Excellence Strategy (EXC 2002/1 "Science of Intelligence" Project No. 390523135). M. Rolfs was supported by the Heisenberg program of the DFG (Grant No. RO3579/8-1 and Grant No. RO3579/12-1).

Artificial Intelligence

No artificial intelligence assisted technologies were used in this research or the creation of this article.

Ethics

The ethical review board of the Department of Psychology at Humboldt-Universität zu Berlin approved the experimental procedure.

Open Practices

Preregistration: The study was preregistered (https://doi.org/10.17605/OSF.IO/K2JTE) on 2022-09-30. The research hypotheses (https://doi.org/10.17605/OSF.IO/24GSB/), stimuli recording (https://doi.org/10.17605/OSF.IO/FWS93/), data acquisition (https://doi.org/10.17605/OSF.IO/95VXK/), and measured variables (https://doi.org/10.17605/OSF.IO/2M6D7) were not *formally* preregistered, however, the timestamps indicate that the contents have not been changed since 2022-09-30, prior to data collection which began on 2022-11-24. There were minor deviations from the preregistration (for details, see Supplementary File Table S1). The data preprocessing and statistical analysis methods were not preregistered. Materials: All study materials are publicly available (https://osf.io/vj5dr/files/osfstorage). Data: All primary data are publicly available (https://osf.io/x2gaz). Analysis scripts: All analysis scripts are publicly available (https://github.com/rederoth/LPA_experimental_code, registered at https://osf.io/vj5dr/files/osfstorage). Computational reproducibility: The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

ORCID iDs

Nicolas Roth https://orcid.org/0000-0001-5382-5354
Martin Rolfs https://orcid.org/0000-0002-8214-8556

Acknowledgments

We want to thank Clara Kuper for the fruitful discussions about this project, especially regarding the cluster-based permutation test; Richard Schweitzer and Olga Shurygina, for their contributions to the experimental code; and Conrad Blau and Mindia Wichert for their help with collecting the eye-tracking data.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976241279198

References

Açık, A., Bartel, A., & König, P. (2014). Real and implied motion at the center of gaze. *Journal of Vision*, *14*(1), Article 2. https://doi.org/10.1167/14.1.2

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), Article 9. https://doi.org/10.1167/15.5.9

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314.

Binaee, K., & Diaz, G. (2019). Movements of the eyes and hands are coordinated by a common predictive strategy. *Journal of Vision*, *19*(12), Article 3. https://doi.org/10.1167/19.12.3

Broda, M. D., Haddad, T., & de Haas, B. (2023). Quick, eyes! Isolated upper face regions but not artificial features elicit rapid saccades. *Journal of Vision*, *23*(2), Article 5. https://doi.org/10.1167/jov.23.2.5

Calvillo, D. P., & Hawkins, W. C. (2016). Animate objects are detected more frequently than inanimate objects in inattentional blindness tasks independently of threat. *The Journal of General Psychology*, *143*(2), 101–115.

Calvillo, D. P., & Jackson, R. E. (2014). Animacy, perceptual load, and inattentional blindness. *Psychonomic Bulletin & Review*, *21*, 670–675.

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, *46*(26), 4333–4345.

Carneiro Morita, V., Souto, D., Masson, G. S., & Montagnini, A. (2023). Anticipatory smooth eye movements scale with the probability of visual motion: Role of target speed and acceleration [Preprint]. *bioRxiv*.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 613–617.

Damasse, J.-B., Perrinet, L. U., Madelain, L., & Montagnini, A. (2018). Reinforcement effects in anticipatory smooth eye movements. *Journal of Vision*, *18*(11), Article 14. https://doi.org/10.1167/18.11.14

Dar, A. H., Wagner, A. S., & Hanke, M. (2021). Remodnav: Robust eye-movement classification for dynamic stimulation. *Behavior Research Methods*, *53*(1), 399–414.

de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary

along semantic dimensions. *Proceedings of the National Academy of Sciences, USA*, *116*(24), 11687–11692.

Diaz, G., Cooper, J., Rothkopf, C., & Hayhoe, M. (2013). Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. *Journal of Vision*, *13*(1), Article 20. https://doi.org/10.1167/13.1.20

Diaz, G. J., Fajen, B. R., & Phillips, F. (2012). Anticipation from biological motion: The goalkeeper problem. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 848–864.

Donk, M., & Van Zoest, W. (2008). Effects of salience are short-lived. *Psychological Science*, *19*(7), 733–739.

Dorr, M., & Bex, P. J. (2013). Peri-saccadic natural vision. *Journal of Neuroscience*, *33*(3), 1211–1217.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), Article 28. https://doi.org/10.1167/10.10.28

Ehinger, B. (2016). *Statistics: Cluster permutation test*. https://benediktehinger.de/blog/science/statistics-cluster-permutation-test/

Fooken, J., Kreyenmeier, P., & Spering, M. (2021). The role of eye movements in manual interception: A mini-review. *Vision Research*, *183*, 81–90.

Goettker, A., Borgerding, N., Leeske, L., & Gegenfurtner, K. R. (2023). Cues for predictive eye movements in naturalistic scenes. *Journal of Vision*, *23*(10), Article 12. https://doi.org/10.1167/jov.23.10.12

Goettker, A., Pidaparthy, H., Braun, D. I., Elder, J. H., & Gegenfurtner, K. R. (2021). Ice hockey spectators use contextual cues to guide predictive eye movements. *Current Biology*, *31*(16), R991–R992.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October 22-29). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2961–2969). IEEE. doi: 10.1109/ICCV.2017.322

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743–747.

Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research*, *45*(13), 1707–1724.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*(6), 1093–1123.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Kaakinen, J. K. (2021). What can eye movements tell us about visual perception processes in classroom contexts? Commentary on a special issue. *Educational Psychology Review*, *33*(1), 169–179.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776.

Kirkorian, H. L., & Anderson, D. R. (2017). Anticipatory eye movements while watching continuous action across shots in video sequences: A developmental study. *Child Development*, *88*(4), 1284–1301.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*, *36*(14), 1–16.

Kümmerer, M., & Bethge, M. (2023). Predicting visual fixations. *Annual Review of Vision Science*, *9*, 269–291.

Kümmerer, M., Bethge, M., & Wallis, T. S. (2022). DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, *22*(5), Article 7. https://doi.org/10.1167/jov.22.5.7

Land, M. F., & Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *352*(1358), 1231–1239.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25–26), 3559–3565.

Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021, October 10-17). DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12919–12928). IEEE. doi: 10.1109/ICCV48922.2021.01268

Linka, M., & de Haas, B. (2023). *Detection, inspection, return: A functional classification of fixations in complex scenes* [Preprint]. PsyArXiv.

Loffing, F., & Cañal-Bruland, R. (2017). Anticipation in sport. *Current Opinion in Psychology*, *16*, 6–11.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, *3*, 5–24.

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences, USA*, *104*(42), 16598–16603.

Nobre, A. C., & Van Ede, F. (2018). Anticipated moments: Temporal structure in attention. *Nature Reviews Neuroscience*, *19*(1), 34–48.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204.

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition*, *206*, Article 104465.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Pratt, J., Radulescu, P. V., Guo, R. M., & Abrams, R. A. (2010). It's alive! Animate motion captures visual attention. *Psychological Science*, *21*(11), 1724–1730.

Roth, N., Rolfs, M., Hellwich, O., & Obermayer, K. (2023). Objects guide human gaze behavior in dynamic real-world scenes. *PLOS Computational Biology*, *19*(10), 1–39. https://doi.org/10.1371/journal.pcbi.1011512

Rothkegel, L. O., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2017). Temporal evolution of the

central fixation bias in scene viewing. *Journal of Vision*, *17*(13), Article 3. https://doi.org/10.1167/17.13.3

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, *13*(8), Article 16. https://doi.org/10.1167/13.8.16

Spering, M., & Carrasco, M. (2015). Acting without seeing: Eye movements reveal visual processing without awareness. *Trends in Neurosciences*, *38*(4), 247–258.

Spering, M., Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Keep your eyes on the ball: Smooth pursuit eye movements enhance prediction of visual motion. *Journal of Neurophysiology*, *105*(4), 1756–1767.

Stewart, E. E., & Fleming, R. W. (2023). The eyes anticipate where objects will move based on their shape. *Current Biology*, *33*(17), R894–R895.

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756.

Vig, E., Dorr, M., Martinetz, T., & Barth, E. (2011). Eye movements show optimal average anticipation with natural dynamic scenes. *Cognitive Computation*, *3*, 79–88.

Vo, M. L.-H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2