Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time

Heiko H. Schütt*

Neural Information Processing Group, Universitäts Tübingen, Tübingen, Germany
Experimental and Biological Psychology, University of Potsdam, Potsdam, Germany

Lars O. M. Rothkegel*

Experimental and Biological Psychology, University of Potsdam, Potsdam, Germany

Hans A. Trukenbrod

Experimental and Biological Psychology, University of Potsdam, Potsdam, Germany

Ralf Engbert

Experimental and Biological Psychology and Research Focus Cognitive Sciences, University of Potsdam, Potsdam, Germany

Felix A. Wichmann

Neural Information Processing Group, Universitäts Tübingen, Tübingen, Germany

Bottom-up and top-down as well as low-level and high-level factors influence where we fixate when viewing natural scenes. However, the importance of each of these factors and how they interact remains a matter of debate. Here, we disentangle these factors by analyzing their influence over time. For this purpose, we develop a saliency model that is based on the internal representation of a recent early spatial vision model to measure the low-level, bottom-up factor. To measure the influence of high-level, bottom-up features, we use a recent deep neural network–based saliency model. To account for top-down influences, we evaluate the models on two large data sets with different tasks: first, a memorization task and, second, a search task. Our results lend support to a separation of visual scene exploration into three phases: the first saccade, an initial guided exploration characterized by a gradual broadening of the fixation density, and a steady state that is reached after roughly 10 fixations. Saccade-target selection during the initial exploration and in the steady state is related to similar areas of interest, which are better predicted when including high-level features. In the search data set, fixation locations are determined predominantly by top-down processes. In contrast, the first fixation follows a different fixation density and contains a strong central fixation bias. Nonetheless, first fixations are guided strongly by image properties, and as early as 200 ms after image onset, fixations are better predicted by high-level information. We conclude that any low-level, bottom-up factors are mainly limited to the generation of the first saccade. All saccades are better explained when high-level features are considered, and later, this high-level, bottom-up control can be overruled by top-down influences.

Introduction

The guidance of eye movements in natural environments is extremely important for our perception of the world surrounding us. Visual perception deteriorates quickly away from the gaze position such that many tasks are hard or impossible to perform without looking at the objects of interest (reviewed by Strasburger, Rentschler, & Jüttner, 2011, section 6; see also Land, Mennie, & Rusted, 1999). Thus, the selection of fixation locations is of great interest for vision researchers, and many theories were developed to explain the selection of fixation locations.

Classically, factors determining eye movements of human observers are divided into bottom-up and top-down influences (Hallett, 1978; Tatler & Vincent, 2008).
Bottom-up influences refer to stimulus parts that attract fixations independent of the internal state of an observer. The existence of bottom-up guidance of eye movements was originally postulated because some stimuli, such as flashing lights, attract subjects’ gaze under well-controlled laboratory conditions even in tasks when subjects were explicitly asked not to look at the stimulus, for example, in the antisaccade task (Hallett, 1978; Klein & Foerster, 2001; Mokler & Fischer, 1999; Munoz & Everling, 2004). How important bottom-up effects are under more natural conditions and especially for static stimuli remains a matter of debate. Top-down influences, on the other hand, refer to cognitive influences on the chosen fixation locations based on the current aims of an observer varying, for example, with task demands and memory (Henderson, Brockmole, Castelhano, & Mack, 2007; Land et al., 1999). The main argument for the involvement of top-down control comes from task effects on fixation locations (Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2007; Underwood, Foulsham, Loon, Humphreys, & Bloyce, 2006; Yarbus, 1967). More recently, systematic tendencies were introduced as a third category (Tatler & Vincent, 2008), which encompasses regularities of the oculomotor system across different instances and manipulations, such as the preference to fixate near the image center (Tatler, 2007), the preference for some saccade directions (Foulsham, Kingstone, & Underwood, 2008), and the dependencies between successive saccades (Rothkegel, Trukenbrod, Schütt, Wichmann, & Engbert, 2016; Tatler & Vincent, 2008; Wilming, Harst, Schmidt, & König, 2013). Although all three aspects seem to contribute to eye-movement control, the debate, how these aspects are combined and how important the different aspects are, continues until today (Borji & Itti, 2013; Einhäuser, Rutishauser, et al., 2008; Foulsham & Underwood, 2008; Hallett, 1978; Harel, Koch, & Perona, 2006; Kienzle, Franz, Schölkopf, & Wichmann, 2009; Schomaker, Walper, Wittmann, & Einhäuser, 2017; Stoll, Thrun, Nuthmann, & Einhäuser, 2015; Tatler, Hayhoe, Land, & Ballard, 2011; Tatler & Vincent, 2009).

Orthogonal to the top-down versus bottom-up distinction, models of eye-movement control can also be categorized by the features they employ. Here, low-level models refer to simple features, which are extracted early in the visual hierarchy, such as local color, brightness, or contrast (Itti & Koch, 2001; Treisman & Gelade, 1980). High-level models, on the other hand, refer to features thought to be extracted in higher cortical areas encoding more complex information, such as the position and identity of objects (Einhäuser, Spain, & Perona, 2008) or the scene category and context (Torrälba, Oliva, Castelhano, & Henderson, 2006).

In the debate on what factors govern eye movements, the two sides typically argued for are, on the one hand, bottom-up control based on low-level features (Itti & Koch, 2001; Kienzle et al., 2009) and, on the other hand, top-down control based on high-level features (Castelhano, Mack, & Henderson, 2009; Yarbus, 1967). This links the question of feature complexity to the question of how much internal goals control our eye movements. These questions are orthogonal, however, and the less usual positions may be sensible as well. Bottom-up control may encompass not only low-level features, such as contrast, color, or edges (Itti & Koch, 2001; Itti, Koch, & Niebur, 1998; Treisman & Gelade, 1980), but also high-level properties of the explored scene, such as object locations (Einhäuser, Spain, et al., 2008), faces (Judd, Ehinger, Durand, & Torralba, 2009; Kümerer, Wallis, & Bethge, 2016), or even locations that are interesting or unexpected in a scene category (Henderson, Weeks, & Hollingworth, 1999; Torralba et al., 2006). This position is implicitly embraced by most modern (computer-vision) models for the prediction of fixation locations in images, which almost all use high-level features computed from the image (Bylinskii et al., 2016; Judd et al., 2009; Kümerer et al., 2016). Similarly, top-down control may not only act on high-level features, but can also act on low-level features, such as contrast, orientations, or color. Such influences are especially important in models of visual attention (Müller & Krummenacher, 2006; Navalpakkam & Itti, 2005; Tsotsos et al., 1995) and visual search (Wolfe, 1994), which often postulate top-down control over low-level features that guide attention and eye movements or top-down influences on the processing of low-level features (Tsotsos et al., 1995).

In addition, inconsistent terminology has added to the confusion. One especially unclear term in this context is “saliency,” which originally stems from attention research. Saliency referred to conspicuous locations that stood out from the remaining display and attracted attention (Koch & Ullman, 1985). As the first computable models for the prediction of fixation locations in images were based on these ideas, saliency was soon associated with these models and became synonymous with bottom-up, low-level control of eye movements (Itti & Koch, 2001; Itti et al., 1998). As it became clear that the prediction of fixation locations benefits from high-level features, they were added to the models, but the models were still referred to as saliency models (Borji & Itti, 2013; Bylinskii et al., 2016; Judd et al., 2009). As a result, saliency in computer vision now refers to any bottom-up, image-based prediction of which locations are likely to be fixated by human observers. To avoid confusion associated with the term “saliency,” we use the term
“saliency model” in the remainder, which refers to any bottom-up model that predicts fixation locations based on an image, independent of feature complexity.

The temporal evolution of eye-movement behavior can be informative to understand the interplay of the different factors as has been noted by researchers early on (Buswell, 1935; Yarbus, 1967). Notably, experimenters changed low-level features and observed that these manipulations were effective only at the beginning of a trial (Anderson, Donk, & Meeter, 2016; Anderson, Ort, Kruijne, Meeter, & Donk, 2015) and could be overruled immediately by task instructions (Einhäuser, Rutishauser, et al., 2008). Furthermore, the predictive power of low-level, bottom-up models was explored over time (Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005), and it was found that the predictive value of low-level features was low but relatively constant over time, and subjects produced more consistent fixation locations early in a trial than later in a trial. This higher consistency is partially explained by the central fixation bias because it is much stronger for the first few fixations (Clarke & Tatler, 2014; Tatler, 2007) pulling fixations closer to the image center and, thus, closer to each other.

Finally, there have been some analyses comparing high- versus low-level features, which include an analysis over time (Onat, Açık, Schumann, & König, 2014; Stoll et al., 2015). These studies found an advantage of high-level features that persisted over time. Similar conclusions were drawn based on models, which include low- and high-level features. By fitting these models to different time points separately, the weights of different features can be evaluated over time (Gautier & Le Meur, 2012; Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014).

In this article, we analyze the temporal evolution of the fixation density and disentangle the contribution of the different factors by evaluating models over time, which include different influence factors.

To allow stable estimates of the fixation density at different ordinal fixation numbers (fixation #s), we employ two large eye-movement data sets, which contain exceptionally many fixations per image. The two data sets were collected using different tasks, which allows us to analyze whether our conclusions hold under different top-down control conditions. The first data set stems from a standard scene-viewing experiment in which participants were asked to explore a scene for a subsequent memory test. Scene viewing has been suggested to minimize top-down control (Itti & Koch, 2001). Nonetheless, subjects could still employ top-down control and the given task probably still influences eye movements (Castelhano et al., 2009), and different subjects could even choose different top-down control strategies (Castelhano & Henderson, 2008; Tatler et al., 2011). For the second data set, subjects searched for artificial targets in natural scenes (Rothkegel, Schüt, Truenbrod, Wichmann, & Engbert, 2019). Because the search target was known in advance, participants had a clear aim and motivation to exploit low-level features in a top-down fashion in the second task.

Using these two data sets, we systematically investigate how fixation locations are controlled by low- and high-level features under different top-down control demands. To quantify the contribution of low- and high-level features, we compare the performance of different computational models using a recently proposed likelihood-based technique (Kümmerer, Wallis, & Bethge, 2015; Schüt et al., 2017). This method avoids ambiguities of typically used ad hoc criteria for saliency model evaluation and, more importantly, provides a unified metric for all models.

To measure the influence of low-level features, we choose three classical low-level models (Harel et al., 2006; Itti et al., 1998; Kienzle et al., 2009). However, the features used by these low-level models were only informally linked to the low-level features used in models of perception. To remove this ambiguity of interpretation, we additionally present a new saliency model, which is based on the representation produced by a model of early spatial vision (Schüt & Wichmann, 2017).

To measure the influence of high-level features, earlier approaches made predictions based on manually annotated object locations (Einhäuser, Spain, et al., 2008; Stoll et al., 2015; Torralba et al., 2006) or experimentally varied low-level features (Açık, Onat, Schumann, Einhäuser, & König, 2009; Anderson et al., 2015; Stoll et al., 2015) or chose specific examples for which low- and high-level features make opposing predictions (Vincent, Baddeley, Correani, Troscianko, & Leondards, 2009). One shortcoming of these classical approaches is that they do not easily make predictions for new images. Fortunately, the idea of object-based saliency models was recently unified with low-level factors due to the advent of deep neural network models (DNNs; see Kriegeskorte, 2015). DNNs contain activation maps, which effectively encode what kind of object can be found where in an image. As with simple low-level features, these object-based features can be used to predict fixation locations (Huang, Shen, Boix, & Zhao, 2015; Kruthiventi, Ayush, & Babu, 2015; Kümmerer et al., 2016; Pan et al., 2017). Saliency models based on this principle currently perform best on the fixation-density prediction benchmarks (Bylinskii et al., 2016). Thus, these DNN-based saliency models provide a better and more convenient quantification than earlier approaches on the question of which information can be predicted by high-level features. As a representative model, we use the...
currently best performing of these models, DeepGaze II (Kümmerer et al., 2016).

**Methods**

**Stimulus presentation**

Sets of 90 (scene-viewing experiment) and 25 (search experiment) images were presented on a 20-in. CRT monitor (Mitsubishi Diamond Pro 2070; frame rate 120 HZ, resolution 1,280 × 1,024 pixels; Mitsubishi Electric Corporation, Tokyo, Japan). All stimuli had a size of 1,200 × 960 pixels. For the presentation during the experiment, images were displayed in the center of the screen with gray borders extending 32 pixels to the top/bottom and 40 pixels to the left/right of the image. The images covered 31.1° of visual angle (dva) in the horizontal and 24.9 dva in the vertical dimension.

**Measurement of eye movements**

Participants were instructed to position their heads on a chin rest in front of a computer screen at a viewing distance of 70 cm. Eye movements were recorded binocularly using an Eyelink 1000 video-based eye tracker (SR-Research, Osgoode, ON, Canada) with a sampling rate of 1,000 Hz.

For saccade detection, we applied a velocity-based algorithm (Engbert & Kliegl, 2003; Engbert & Mangenthaler, 2006). The algorithm marks an event as a saccade if it has a minimum amplitude of 0.5° and exceeds the average velocity during a trial by six median-based standard deviations for at least six data samples (6 ms). The epoch between two subsequent saccades is defined as a fixation. All fixations with a duration of less than 50 ms were removed for further analysis because these are most probably glissades, that is, part of the saccade (Nyström & Holmqvist, 2010). The number of fixations for further analyses was 312,267 in the scene-viewing experiment and 176,828 in the search experiment.

**Scene-viewing data set**

In our scene-viewing experiment, we showed 90 images to 105 participants in three groups with slightly varying viewing conditions, asking them to remember the presented images for a subsequent memory test.

**Participants**

For this study, 105 students of the University of Potsdam with normal or corrected-to-normal vision were recruited. On average, participants were 23.3 years old, and 89 participants were female. Participants received credit points or a monetary compensation of 16€ for their participation. The work was carried out in accordance with the Declaration of Helsinki. Informed consent was obtained for experimentation by all participants.

**Stimuli**

As stimuli, we used 90 photographs taken by the authors, which did not contain text or humans that stand out prominently to avoid the specific eye-motion patterns elicited by these contents. Furthermore, images were selected as six subsets of 15 images each: The first contained photographs of texture-like patterns; the other five contained typical holiday photographs with the prominent structure either at the top, left, bottom, right, or center. The full set of images is available online with the data set (Figure 1, left panel).

For presentation in grayscale, we measured the luminance output of each gun separately and for the sum of all three guns at every value from zero to 255. To convert a stimulus into grayscale, we summed the luminance output for the RGB values and chose the gray value with the most similar luminance.

**Procedure**

Eye movements for our scene-viewing experiment were collected in two sessions. In each session, 60 images were presented, and participants were instructed to memorize them for a subsequent test to report which images they had seen. Subjects were informed that they would be shown images in a second part as well and would then be asked to judge whether they had seen them during the first part or not. Additionally, subjects were informed about the calibration procedure and asked to blink as little as possible during the trials. The memorization task is different from free-viewing and was reported to influence eye movements of subjects (Mills, Hollingworth, Stigchel, Van der Hoffman, & Dodd, 2011). We chose it, nonetheless, to keep subjects motivated to actively view the scenes for the whole observation time of 10 s.

In the first session, all images were new. In the second session, we repeated 30 images from the first session and showed another 30 new images. The 30
repeated images were the same for each observer. For this article, we used all fixations from both sessions, ignoring whether the subject had seen the image before and to which group the subject belonged to maximize the amount of data. Trials began with a black fixation cross presented on a gray background. After successful binocular fixation in a square with a side length of $2.8^\circ$, the stimulus appeared, and subjects had 10 s to explore the image. In the memory test, participants had to indicate for each of 120 images if they had seen it before. Half the images were the ones they saw in the experiment; the other half were chosen randomly from another pool of 90 images we chose according to the same criteria as the images used for the first set of images. Subjects almost perfectly recalled which images they had seen.

The three cohorts of subjects differed in the placement of the fixation cross and whether the images were shown in color or in grayscale:

- For the first 35 subjects, we presented the images in grayscale and placed the start position randomly within a doughnut shape around the center of the screen and stimulus with an inner radius of 100 pixels $= 2.6^\circ$ and an outer radius of 300 pixels $= 7.8^\circ$.
- For the second group of 35 subjects, the images were also presented in grayscale, but the start position was chosen randomly from only five positions: the image center and 20% of the monitor size (256/205 pixels, 5.68/4.55 dva) away from the border of the monitor at the top, left, bottom, and right, centrally in the other dimension.
- For the final group of 35 subjects, the images were shown in color, and the starting position was as for the second group.

**Natural image search**

In our visual search task, participants were asked to look for six targets embedded into natural scenes (Figure 1). The data set has been used for a different purpose in another publication (Rothkegel et al., 2019).

**Participants**

We recorded eye movements from 10 human participants (four female) with normal or corrected-to-normal vision in eight separate sessions on different days. Six participants were students from a nearby high school (age 17–18), and four were students at the University of Potsdam (age 22–26).

**Stimuli**

As natural image backgrounds, we chose 25 images taken by the authors and an additional member of the Potsdam lab in the area surrounding Potsdam. The images contained neither faces nor text.
As targets, we designed six different low-level targets with different orientation and spatial frequency content (Figure 1, right panel). To embed the targets into natural images, we first converted each image to luminance values based on a power function fitted to the measured luminance response of the monitor. We then combined this luminance image $I_L$ with the target $T$ with a luminance amplitude $\alpha L_{max}^T$ fixed relative to the maximum luminance displayable on the monitor $L_{max}$ as follows:

$$I_{fin} = \alpha L_{max} + (1 - 2\alpha)I_L + \alpha L_{max}^T,$$

that is, we rescaled the image to the range $[\alpha(1 - \alpha)]L_{max}$ and then added the target with a luminance amplitude of $\alpha L_{max}$ such that the final image $I_{fin}$ never left the displayable range. After a pilot experiment, we fixed $\alpha$ to 0.15. Thus, contrast was reduced to 70%. We then converted the image $I_{fin}$ back to [0, 255] grayscale values by inverting the fitted power function.

We focus on low-level targets here to avoid expectations about the target location generated from the image background. Furthermore, low-level targets give clear expectations about which low-level features should influence eye-movement control. As low-level features seem to have little influence on eye-movement selection in visual search (see below), these expectations were largely irrelevant for this study. On eye-movement dynamics, however, the low-level target properties had the expected effects as we report in a parallel article (Rothkegel et al., 2019).

**Procedure**

Participants were instructed to search for one of six targets for the upcoming block of 25 images. To do so, the target was presented on a 26th demonstration image, marked by a red square. Each session consisted of six blocks of 25 images for each of the six different targets. The 25 images within a block were always the same presented in a new random order.

Trials began with a black fixation cross presented on a gray background at a random position within the image borders. After successful fixation, the image was presented with the fixation cross still present for 125 ms. This was done to assure a prolonged first fixation to reduce the central fixation tendency of the initial saccadic response (Rothkegel, Trukenbrod, Schütt, Wichmann, & Engbert, 2017; Tatler, 2007). After removal of the fixation cross, participants were allowed to search the image for the previously defined target for 10 s. Participants were instructed to press the space bar to stop the trial once a target was found. In ~80% of trials, the target was present.

At the end of each session, participants could earn a bonus of up to 5€ additional to a fixed 10€ reimbursement, depending on the number of points collected divided by the number of possible points. If participants correctly identified a target, they earned one point. If participants pressed the bar although no target was present, one point was subtracted.

**Kernel density estimation of fixation densities**

To estimate empirical fixation densities, we used kernel density estimation as implemented in the R package spatstat (version 1.51-0). Kernel density estimation requires the choice of a bandwidth for the kernel. The optimal choice for this parameter depends on the shape of the density and on the number of observations available. Thus, it cannot be chosen optimally a priori, but needs to be chosen adaptively for each condition.

To set the bandwidth for our kernel-density estimates, we used leave-one-subject-out cross-validation; that is, for each subject, we evaluated the likelihood of the data under a kernel-density estimate based on the data from all other subjects. For the image-dependent density estimates, we repeated this procedure with bandwidths ranging from 0.5 to 2.0 dva in steps of 0.1 dva. We report the results with the best bandwidth chosen for each fixation # separately. For the image-independent prediction—the central fixation bias—we used the same procedure with bandwidths from 0.2 to 2.2 dva, as these estimates are based on more data, and chose a single bandwidth over all images.

To implement this procedure, we calculated the cross-validated log-likelihood for each fixation using each possible bandwidth. This calculation results in a four-dimensional array with dimensions for the fixation #, the subject, the image, and the chosen bandwidth. We then averaged over subjects and images and report the highest value for each fixation #.

For our analysis over time, we calculated two estimates of the fixation density as upper bounds for the predictability of fixation locations from a static map. For the first, we simply took the cross-validated kernel density estimate based only on the fixations with the same fixation # (labeled “empirical density (each #)”). Fixations typically become fewer later in the trial as fixation durations and, thus, the number of fixations within a 10-s trial vary. Furthermore, we observe that fixations are more dispersed later in the trial such that more fixations are required to estimate the fixation density accurately. Thus, our first estimate declines rapidly over time. To counteract this, we computed a second estimate that uses all fixations on the image from the second to the last fixation to predict the density (labeled “empirical density (all #)”)). This estimate can use more data and performs well because the fixation density converges toward the end of the trial.
The likelihood of the kernel-density estimates always depended smoothly on the bandwidth and showed a single peak such that the bandwidth could be chosen reliably. Furthermore, the chosen bandwidths behaved as expected such that estimates of the same type for later fixation densities were made with larger bandwidths. For the scene-viewing data set, the bandwidth varied from 0.5 to 0.7 dva for the empirical density and from 1.1 to 1.8 dva for the central fixation bias. For the search experiment, the bandwidth for the empirical density varied from 0.7 to 1.1 dva and the bandwidth for the central fixation bias varied from 1.3 to 2.0 dva. Suboptimal choices of the bandwidth could lead to arbitrarily bad performance. Within the evaluated range, we observed differences between bandwidths of up to 0.4 dva for the scene-viewing experiment and up to 0.2 dva for the search experiment. Thus, not choosing the bandwidth optimally, we would noticeably underestimate the absolute performance of the two estimates. The results would not vary qualitatively, however, as we confirmed by repeating all analyses with a constant bandwidth of 1 dva (which roughly represents the size of the foveal visual field).

Comparing fixation densities

To compare two fixation densities, \( p_1, p_2 \), we computed a kernel-density estimate \( \hat{p}_1 \) for one of the fixation densities \( p_1 \) and evaluated the log-likelihood of the fixations \( j_2(\hat{p}_1) \) measured for the other fixation density. As the following equation shows, this is an estimate for the negative of the cross-entropy of the two densities \( H(p_2; p_1) \):

\[
H(p_2; p_1) = -\int p_2(x) \log(p_1(x)) \, dx
\]

\[
= -E_{p_2}(\log(p_1(x)))
\]

\[
\approx -\frac{1}{n} \sum_{i=1}^{n} \log\left( \hat{p}_1(j_2(i)) \right).
\]

This cross-entropy is closely related to the Kullback–Leibler divergence \( KL(p_2||p_1) \), which is simply the cross-entropy minus the entropy of \( p_2 \):

\[
KL(p_2||p_1) = H(p_2; p_1) - H(p_2).
\]

Thus, the log-likelihood we report measures how well \( p_1 \) approximates \( p_2 \) irrespective of the entropy of \( p_2 \), that is, irrespective of the upper limit for predictions of \( p_2 \).

To implement this, we again used leave-one-subject-out cross-validation; that is, for each subject, we computed a separate kernel-density estimate \( \hat{p}_1 \) using only data of the other subjects and evaluated it at the fixation locations of that one subject.

Comparisons over time: Specifically, we compare fixation densities over time taking the distributions of fixations with two given fixation \#s as \( p_1 \) and \( p_2 \); that is, we measure how (dis)similar fixation densities with different fixation numbers are. For the kernel-density estimates necessary in this computation, we tested bandwidths from 1.0 to 5.0 dva in steps of 0.2 dva. We report the results with a bandwidth of 1.6 dva, which results in the highest likelihood for fixation numbers two to 25 averaged over all predicting fixation numbers, images, and subjects.

Comparisons between targets: For the search data, we compared the fixation densities produced by subjects when searching for different targets on the same images. For this comparison, we tested bandwidths from 0.5 to 1.5 dva in steps of 0.05 dva and report values computed with a bandwidth of 0.9 dva, which results in the highest likelihoods averaged over all images, comparisons, and subjects.

Evaluation of saliency models

In our analysis of saliency models, we largely follow Kümerer et al. (2015), who recommend using the log-likelihood of fixations under the model for evaluation after fitting a nonlinearity, blur and center bias for each model to map the saliency map to an optimal prediction for the fixation density. Some transformations cannot be avoided because classical saliency models do not predict a fixation density, but only a saliency map, which is not necessarily a density. Also, saliency maps only aim to be monotonically related to the fixation density when averaged over patches. Thus, fitting a local nonlinearity and a blur allows a fairer comparison between models by fitting the parts of the model, which matter for the evaluation based on likelihoods but not necessarily for the criteria used to design the models. Furthermore, we were interested in how well the fixation density can be predicted with certain predictors, which also argues for an optimal mapping from saliency map to fixation density.

To fit the mapping from the saliency map to the fixation density, we used the DNN framework Keras as included in TensorFlow (Abadi et al., 2015, version 1.3.0) as a back end. In this framework, we fit a shallow network as illustrated in Figure 2 for each saliency model separately after resizing the saliency maps to 128 \( \times \) 128 pixel resolution and rescaling the saliency values to the interval [0,1].

The network contained two conventional 1 \( \times \) 1 convolution layers that first map the original to an intermediate layer with five channels and then to a
single output layer, allowing for a broad range of strictly local nonlinear mappings to the fixation density.

Next, we apply a blurring filter to the activations, which allows saliency to attract fixations to nearby locations. This is not to be confused with blurring the original image, which has entirely different effects on saliency computations, changing the features available to the saliency models, for example. To implement the blur, we used a $25 \times 25$ custom convolution layer in which we set the weights to a Gaussian shape of which we fitted the two standard deviations.

Finally, we apply a sigmoidal nonlinearity to map activations to a strictly positive map and apply a center bias through a custom layer. This layer first multiplies the map with a Gaussian with separately fitted vertical and horizontal standard deviations and then normalizes the sum of the activities over the image to one to obtain a probability density.\(^1\)

As a loss function for training the network, we directly use the log-likelihood as for the kernel-density estimates described above. In Keras, we implemented this by flattening the final density estimate and using the standard loss function categorical_crossentropy to compare to a map with sum one and entries proportional to the number of fixations at each location. For evaluation, we performed fivefold cross-validation over the used images; that is, we trained the network five independent times leaving out one fifth of the data. For training, we ran the Adam optimization algorithm (Kingma & Ba, 2014) with standard parameters until convergence by reducing the learning rate by a factor of two whenever the loss improved less than $10^{-5}$ over 100 epochs and stopping the optimization when the loss improved by less than $10^{-6}$ over 500 epochs. We did not employ a test set here as we did not optimize any hyper parameters and did not use any stopping or optimization rules based on the validation set.

The primary advantage of the likelihood as an evaluation metric is its established role and interpretation in statistics. As we discuss in more detail in a recent article (Schütt et al., 2017), this allows statistically stronger model-fitting and model-evaluation techniques. Furthermore, for the likelihood evaluation, the fixation density is the optimal saliency map (Kümmerer, Wallis, & Bethge, 2017), such that the saliency map can be easily used to generate simulated data from the model. Here, we additionally fit a nonlinearity and central fixation bias for each model, which unifies all commonly used evaluation criteria (Kümmerer et al., 2015) as evaluation criteria primarily differ in whether they expect a correct central fixation bias and/or a correct nonlinearity. For example, the implicit central fixation bias of the graph-based visual saliency (GBVS) model will neither be advantageous nor detrimental under this evaluation scheme.

**Interpretation of the log-likelihood scale**

We measure all model performances in $\text{bit}/\text{fix}$ relative to a model that predicts a uniform distribution of fixation locations. A log-likelihood of $0 \text{ bit}/\text{fix}$ is equal to a uniform prediction over the image—thus, no gain over predicting that any position in the image is equally likely to attract a fixation. If a model is $1 \text{ bit}/\text{fix}$ better than a uniform model, its density is, on average, two times higher than the density of the uniform model. As the density has to integrate to one, this corresponds roughly to reducing the possible area for fixations by the same factor, that is, in the case of $1 \text{ bit}/\text{fix}$ to half the size. Thus, larger $\text{bit}/\text{fix}$ result from an increased accuracy of the prediction. In general, this relation is $2^x$ for a difference of $x \text{ bit}/\text{fix}$. The interpretation of a difference in
log-likelihood per fixation as a factor on the density or predicted area is independent of the absolute performance of the compared models; that is, if one model already reaches a performance of 2 bit/fx, another model reaching 3 bit/fx still predicts densities that are twice as high at fixation locations and, thus, restrict each fixation to roughly half the area.

Other metrics

We expected that different evaluation metrics lead to the same conclusions we draw. To assure that this is the case, we repeated our main analyses, which test saliency model performances over time, using the eight classic metrics from the MIT saliency benchmark (Bylinskii et al., 2016). The authors of this benchmark provide MATLAB code\(^2\) for all used metrics, which we applied with the same cross-validation scheme as for the likelihood-based evaluations, that is, using the fitted parameters from other images and other people for evaluation. For central fixation bias and empirical saliency, we used a fixed kernel bandwidth of 1 dva and evaluated the resulting maps the same way as the saliency model predictions.

Tested saliency models

To get a comprehensive overview over saliency model performance, we chose a few representative models for predicting saliency:

Kienzle: As an example of an extremely simple low-level model of visual saliency, we employ the model by Kienzle et al. (2009), using the original implementation supplied by Felix Wichmann.

Itti and Koch: As the most classic saliency model, we evaluate the original model by Itti et al. (1998). To compute the saliency maps, we used the implementation that accompanies the GBVS saliency model, which performed decisively better than the original implementation from www.saliencytoolbox.net.

GBVS: As a better performing classical hand-crafted saliency model, we use the GBVS model by Harel et al. (2006). Code was downloaded from here.\(^3\)

DeepGaze II: As a representative of the newest DNN-based saliency models, we chose DeepGaze II by Kümerer et al. (2016). This model is currently leading the MIT saliency benchmark (Bylinskii et al., 2016). Saliency maps for this model were obtained from the webservice at deepgaze.bethgelab.org as log-values in a .mat file and converted to linear scale before use.

Early vision: Our early vision saliency model is based on our psychophysical spatial vision model we published recently (Schütt & Wichmann, 2017). This model implements the standard model of early visual processing to make predictions for arbitrary luminance images. As an output, it produces a set of 8 \(\times\) 12 = 96 orientation \(\times\) spatial frequency channel responses, spatially resolved over the image.

To obtain a saliency map from these channel responses, we linearly weighed and added them to form a saliency map. To map this sum to a predicted fixation density, we used the same machinery as the saliency maps for all other models.

The weights for the initial sum were unknown, however, and needed to be fit. Fortunately, we could implement the weighted sum as a 1 \(\times\) 1 convolution layer in TensorFlow as well. This implementation allows us to interpret the weights as additional parameters of the mapping to the fixation density. Thus, we could train an arbitrary weighting for the maps from our early vision model directly while keeping the benefits of a nonlinearity, blur, and center bias as for the other models.

Results: Scene viewing

Overall saliency model performance

Before we analyze the temporal evolution of the fixation density, we compare the overall performance of our saliency model based on a model of spatial vision to a range of classical low-level saliency models, that is, Itti and Koch (Itti et al., 1998), GBVS (Harel et al., 2006), Kienzle (Kienzle et al., 2009), and the currently best DNN-based saliency model DeepGaze II (Kümerer et al., 2016). To make the models comparable, we fitted the same nonlinear map, blur, and center bias for all models (see Methods). As the evaluation criterion, we use the average log-likelihood difference to a uniform model as described by Kümerer et al. (2015) for saliency models and Schütt et al. (2017) for dynamical models (see Methods).

The results of the overall analysis\(^4\) are displayed in Figure 3A. All low-level saliency models predict fixations better than a pure center-bias model. Our early vision–based saliency model performs slightly better than the classical saliency models using only a simple weighted sum of activities as a saliency map. Thus, a simple sum of activities of a realistic early spatial vision model seems to be sufficient for modeling low-level influences.

DeepGaze II clearly outperforms all tested classical saliency models (Harel et al., 2006; Itti et al., 1998; Kienzle et al., 2009) and our early vision–based model by 0.3 bit/fx. But DeepGaze II is not as close to a perfect prediction for our scene-viewing data set as for the MIT saliency benchmark, missing it by roughly 0.4 bit/fx (compare our Figure 3A to Kümerer et al., 2016).
A potential reason for this might be that our data set contains many more fixations per image (≈2,600) than the saliency benchmark (Judd, Durand, & Torralba, 2012; 39 observers × 3 s ≤ 390), which allows a more detailed estimation of the empirical fixation density. An alternative but not exclusive explanation is that the MIT saliency benchmark data set contains (more) humans, faces, and text, which might help DeepGaze II as these are typical high-level properties reported to attract fixations.

These overall performance results suggest that a realistic early vision representation provides similar predictive value for the density of fixations as classical saliency models do. The results do not fully answer the question of whether classical saliency truly represents early visual processing though. To approach this question, we additionally analyzed the similarity of predictions of all saliency models. To compare saliency model predictions, we calculated their performance in predicting each other on the same log-likelihood scale we use to compare how well they predict human fixations.

The resulting cross-entropies between saliency models are shown in Figure 3B. Each cell's color indicates how well the density created by one model (predicting model) predicts draws from another model's density (predicted model). We first look at the diagonal, which represents how well each model predicts itself, that is, the entropies of the different model predictions. The empirical density predicts itself more accurately than any saliency model predicts itself. Also, each of the saliency models is distinct from the others as the diagonal elements have larger values than any corresponding off-diagonal ones.

Next, we can observe some asymmetry in the prediction qualities as the models are sorted according to their prediction quality. Generally, the top-right quadrant is darker than the lower left quadrant; that is, well-performing models predict poorly performing models less accurately than poorly performing models predict well-performing models. For example, the empirical fixation density is predicted reasonably well by all saliency models but is itself not a good predictor of saliency maps. This pattern indicates that even poorly performing models predict nonzero density where people fixate. However, they seem to add density at locations that are never fixated by humans and not predicted by more successful models. Thus, better models reject locations more efficiently.

The tendency that more successful saliency models generally become more specific than less successful ones is partially caused by the link to the fixation density we fit. The local nonlinearity allows the model to adjust how strongly the prediction of the model is weighted, that is, how large the difference between the peaks and valleys of the prediction is. To optimize their performance, weaker models can use this mechanism to increase their predicted density in the valleys as a substantial proportion of human fixations fall into these valleys. This mechanism broadens the prediction of weak models more than that of strong models.

Finally, there is a group of models that make less dissimilar predictions: Low-level saliency models share some common entropy; that is, the off-diagonal values for these are higher than between other models. Especially, classical models predict each other better than our new early vision–based model (entries comparing our model are lower than the square of entries formed by the three classical models). These results imply that the early vision–based saliency is somewhat different from classical saliency models.

Figure 3. (A) Average performance of the models. (B) Similarity of the different saliency maps. Measured in terms of Δ log-likelihood, that is, as the prediction quality when using one map to predict random draws from another.
To evaluate saliency models over time, we computed log-likelihoods for each fixation number of each trial on an image. From these fixations, we computed a kernel-density estimate and evaluated likelihoods for each fixation number using leave-one-subject-out cross-validation. The results of this analysis averaged over images are displayed in Figure 4A. Different rows correspond to using different fixations for constructing the prediction. Different columns correspond to predicting different fixations within a trial. Due to the cross-validation over subjects, the estimates for a fixation number predicting itself are interpretable and comparable to the predictions for other fixations.

Going through the plot in temporal order, we find that (a) the 0th fixation (the starting position) neither predicts the other fixation locations nor is predicted by them well, which was to be expected because the starting position is induced by the experimental design. (b) The first and, to a lesser degree, the following fixations show an asymmetric pattern: They predict other fixations badly but are predicted well by higher fixation numbers, indicating that they land at positions that are fixated later as well but do not cover all of them. (c) This tendency gradually declines from the second fixation until roughly fixation #10, accompanied by a gradual decline in predictability. (d) From fixation #10 onward, the fixation densities of all fixation numbers predict each other equally well, indicating that the fixation density has reached an equilibrium state.

These results suggest a separation into three phases: (a) the first fixation, which seems to be different from all others, (b) the phase with the asymmetric pattern when fixations are well predicted by the later density but have not converged to it yet, and (c) the final equilibrium phase when the fixation density has converged.

Our next aim was to quantify the maximum amount of image-based predictability at different time points after image onset. To quantify upper and lower bounds, we used four limiting cases: first, a central fixation bias, implemented as a kernel-density estimate from fixations with the same fixation number from all trials on all images; second, a central fixation bias based on all fixations from all images; third, the empirical density estimated as a kernel-density estimate from the fixations with the same fixation number on the same image; and fourth, a different estimate of the empirical density estimated from fixation #2 to fixation #25 on the given image to increase the number of fixations available for the kernel-density estimation. All four estimates were again calculated using leave-one-subject-out cross-validation such that only fixations from other subjects were used for estimating the density.

The results of this analysis are displayed in Figure 4B. The central fixation bias declines quickly from the good prediction based on the initial central fixation
bias on the first fixation to a constant level of roughly 0.25 bit/fix, which is retained over the remaining trial. Also, the two estimates of the central fixation bias only differ substantially for the first few fixations affected by the initial central fixation bias.

For the empirical density, both estimates show a gradual decline over time. The estimate based on all fixation numbers flattens out between fixation #10 and fixation #15. The estimate based only on fixations with the same fixation number quickly falls below the estimate based on all fixations and keeps decreasing. This trend is most likely due to the lower and decreasing number of fixations. As trials always lasted 10 s and fixation durations vary, the number of fixations within a trial varies substantially. Thus, we have fewer examples for high fixation numbers (~66% for fixation #25). First fixations are better predicted using only other first fixations despite their smaller count, confirming that the first fixation follows a different density than later ones.

We interpret this observation as further evidence for a separation into a short initial period with a strong initial central fixation bias, a period for which predictability gradually declines, and a late equilibrium period. Additionally, the difference between our two estimates of the maximally predictable information shows that the ~100 fixations we have for each fixation number are not enough for a good estimate of the fixation density. Thus, the fixation density estimate from all later fixations on an image gives a better estimate of the maximally attainable fixation density for all but the first and possibly second fixation. The initial fixations (fixation #1 and #2) seem to deviate from what attracts later fixations.

Influence of low- and high-level features over time

We are interested in the performance of saliency models over time to test whether low-level features play a more important role at the beginning of a trial. The results of this evaluation are displayed in Figure 5A. In general, prediction quality of all saliency models follows the curve for the empirical density with a gradual decline that reaches a plateau between fixation #10 and #15. As expected, all saliency models are better than the central fixation bias, as our fitted mapping includes a central fixation bias, but do not perfectly predict the empirically observed fixation densities.

Differences between models in their overall performance are present throughout the trial. DeepGaze II performs best, and other saliency models run largely in parallel about 0.3 – 0.5 bit/fix below. To investigate the additional contribution of high-level features, we plot the difference between DeepGaze II and the early vision–based model in Figure 5B. This plot emphasizes that DeepGaze II is constantly predicting fixations better than the early vision–based model (all lines are always > 0). This difference is especially large during the initial exploration phase during which the advantage follows the general decline in predictability (fixation #2–#10). In contrast, the advantage of DeepGaze II for the first fixation (#1) is as small as for fixations during the equilibrium phase (~0.3 bit/fix), which results in a substantial jump from fixation #1 to fixation #2, where DeepGaze II has the largest advantage of ~0.5 bit/fix. As the first fixation contains a strong central fixation bias, which varies over the time course of a trial (Rothkegel et al., 2017), and was proposed as the main point in time for low-level, bottom-up effects (Ander-
son et al., 2015), we analyze this first fixation in more detail.

**Density of the first fixation**

To analyze the first fixation in detail, we performed two complementary analyses: First, we display the first fixation location of participants on some example images in Figure 6. Second, we split the data from the first fixation by the latency of the first saccade after image onset. This split allows us to compare the performance of our early vision–based model and DeepGaze II to the performance of the center bias and the empirical density prediction depending on onset of the first saccade. For each predictor, we created two separate fits: one based on only first fixations and one based on all but the first fixation (fixations #2–#25). For the saliency models, we retrained our network, that is, learned a separate blur, nonlinearity, and center bias. For the empirical density and center bias, we generated separate kernel-density estimates. The results of this second analysis are plotted in Figure 7.

Generally, the density of the first fixation shows a pronounced initial center bias (Rothkegel et al., 2017; Tatler, 2007); that is, early saccades almost exclusively move toward the center of an image. This tendency is visible in the raw data (for example, in the upper left image in Figure 6) and in the high prediction quality of the image-independent central fixation bias model for the first fixation (see Figure 7, light gray line). A potential cause of the central fixation could either be that a certain proportion of fixations is placed near the image center independent of image content or that fixation locations depend on image content weighted by the distance to the center. However, exploring first fixations in more detail shows at least two problems with these simple accounts, illustrated by the examples in Figure 6. First, the strength of the central fixation bias differs considerably between images. For some images, fixations are indeed consistent with a Gaussian distribution around the image center (e.g., top left). For other images, fixation locations seem to stem from a mixture of a Gaussian distribution and a distribution depending on image content (e.g., left middle) or are strongly dominated by image content with almost no fixations near the center (e.g., bottom left). Second, when first fixations depend on image content, the distribution of first fixations differs from the distribution of later fixations in some images (e.g., right bottom), in which the distribution of first fixations shows a different peak than later fixations. Thus, the interaction of central fixation bias and image content seems to be more complex than a simple additive or multiplicative relation.

In addition to the central fixation bias, we observe that first fixations are clearly guided by image content. We find that fixations can be predicted much better when knowledge about the image is included (see Figures 4 and 7, difference between empirical density and central fixation bias) and can confirm this by looking at examples in Figure 6 (distributions clearly differ between images and depend on identifiable objects in images). We can also confirm the observation that the first fixation differs from later fixations as all predictions fitted to the first fixation perform much better...
better than predictions fitted to later fixations (compare left and right plots in Figure 7). This benefit is visible in the raw data as the distribution of first fixations generally deviates from the density computed from later fixations (see Figure 6; compare scatter plot to fixation density).

The image guidance is captured by saliency models to some extent. Low-level features as computed in the early vision–based model perform 0.4 bit fix better than the central fixation bias for training based on the first and later fixations, respectively; that is, its predicted density is 1.34 and 1.44 times higher at fixation locations than the density of the central fixation bias. In addition, DeepGaze II performs 0.3 bit fix better than the early vision–based model; that is, its predicted density is, on average, = 1.25 times higher than the early vision–based model. Thus, high-level features predict fixation locations better than low-level features already for the first fixation. These differences are comparable to later fixations, but all estimates are much higher than for later fixations due to the central fixation bias already explaining 1.6 bit fix and 0.85 bit fix, respectively; that is, its density is already up to three times as high at first fixation locations than the uniform distribution.

Analyzing the effect of onset time of the first saccade (saccade latency), all predictions are relatively bad for latencies below 150 ms. These fixation locations appear not to be guided by the image, but represent only 5% of first fixations. After this poor performance follows the bulk of fixations between 200 and 400 ms, which are best predicted by all models. For these fixations, the early vision model performs up to 0.7 bit fix better than the central fixation bias, but the DeepGaze II model is consistently = 0.3 bit fix better than the early vision model. After this, we see a decline in prediction quality of the models trained for the first fixation, emphasizing that late saccades follow a different density than earlier ones. The models trained on the later fixations decline much more slowly. This slower decline of models trained on later fixations could be the earliest part of the general decline in predictability we observe over multiple fixations above. Thus, fixations after a long first saccade latency might already follow the same factors as subsequent fixations.

Interpreting these results, we conclude that high-level information is advantageous for the prediction of eye movements already 200 ms after image onset. However, the central fixation bias and low-level guidance are much better models for the first fixation than for later ones, especially for relatively early saccades.

**Results: Visual search**

**Predictability of fixation densities for different targets**

The first analyses of the visual search data examine whether fixation locations are predictable from the image and if fixation densities differ for different search targets. To investigate this, we calculated kernel-density estimates from the fixation locations for each search target. We evaluate how well these kernel-density estimates predicted the fixations made while subjects searched for the same or other targets using the same bit fix likelihood scale we use for model evaluation, which estimates the (cross-)entropies of the fixation distributions in this case (see Methods).
The results are displayed in Figure 8. In panel A, we plot the performance of the empirical density (black bar) and the central fixation bias (gray bar) of fixation densities for different targets. These estimates were computed the same way as for the scene-viewing data set and are based on a similar number of fixations per image. Comparing these likelihoods to the scene-viewing data reveals that fixation locations during visual search are distributed much broader over the images than in the standard scene-viewing task. Depending on the target, the fixation density contains only \(0.5\) to \(0.6\) bit of predictable information. In contrast, the empirical density in the scene-viewing data explained \(1.4\) bit.

In panel B, we display how well the fixation distributions for the different targets predict each other. The fixation distributions all predict each other to some extent (>0.3 bit for all pairs). Furthermore, the fixation densities of our targets separate into three groups of targets whose fixation distributions predict each other roughly as well as themselves, indicating practically identical fixation distributions. The three high spatial frequency targets lead to similar fixation distributions, and the Gaussian blob and the positive Mexican hat lead to similar distributions, and the negative Mexican hat produces a different distribution from all others. From the perspective of early spatial vision, it is somewhat surprising that the distribution is so different for the two polarities of the Mexican hat as these stimuli have equal spatial frequency content. Thus, this finding might hint at a greater importance of differences between on and off channels in precortical processing (Whittle, 1986).

Nonetheless, log-likelihoods for the fixations of any target were higher under the fixation densities estimated for any other target than for the uniform distribution (all cells \(\gg 0\)). This indicates that some areas attract fixations independent of the search target.

In summary, our results show that fixation locations can be predicted to some extent although fixations are distributed much broader than in the scene-viewing experiment. Although there is some overlap across fixation locations for different targets, fixation locations also depend on specific target features. This corroborates our earlier observation that searchers adjust their eye movements to the target they look for (Rothkegel et al., 2019).

### Influence of low- and high-level features over time

For the analysis of the saliency models to investigate the role of feature complexity, we employed the same techniques as for the scene-viewing dataset. We fit a nonlinearity, blur, and central fixation bias and evaluate the performance of the resulting prediction over time using cross-validation.

As we show in Figure 9, no saliency model predicts the fixation density well during visual search beyond the first few fixations. When we do not adjust the density prediction to the search data, the models are worse than a uniform prediction at most time points. The only time these densities predict fixation locations above chance are the first and second fixations with the initial center bias. When we train the connection
from saliency map to fixation density newly for the search data, the saliency models still explain only a tiny fraction of the fixation density. Even DeepGaze II and the version of the Itti and Koch (2001) model provided with GBVS, which perform best, explain less than 0.2 bits fix; that is, they predict less than a third of the explainable information and increase the average density at fixation locations by a factor of 1.14 at best. Adjusting the link even more strongly, we also trained the connection from saliency to fixation density separately for each target. Such an adjustment had little effect for any of the saliency models, and the early vision–based model did not profit from this adjustment either although its performance changed slightly and at least improved on the training data set (not shown).

Finally, we evaluated the DeepGaze II model—which performed best for free-viewing—without the link we provided (shown as “DeepGaze2 raw”). This evaluation is important to test that our fitting scheme for the mapping to a density works properly for the search data set as well. All other models do not predict a density map themselves. Thus, this evaluation is only possible for DeepGaze II, which already predicts a density as its saliency map. The raw prediction of DeepGaze II is clearly below chance performance, emphasizing that the link we fitted here is not responsible for the failure of this model.

Our results confirm that fixation locations during visual search are not predicted well by any bottom-up model (Najemnik & Geisler, 2008, 2009). Neither high- nor low-level features predict where humans look, whether they are adjusted to the task or not.

Discussion

We explored the temporal dynamics of the fixation density while looking at natural images to investigate how low- and high-level features and top-down and bottom-up control interact over the course of a trial. This analysis is made possible here by the long duration of trials and the large number of viewings for each image.

The temporal evolution of the fixation density

Based on the similarities of fixation densities shown in Figure 4, we suggest a separation of a typical scene-viewing trial into three phases:

1. An onset response, which affects mostly the first saccade.
2. The main exploration, which is characterized by a gradual broadening of the fixation density.
3. A final equilibrium state, in which the fixation density has converged.

We interpret these three phases as an initial orienting response toward the image center, which can be biased by strong bottom-up signals in the image, followed by a brief guided exploration during which observers look at all parts of the image in which they are interested and a final idle phase during which observers look around rather aimlessly.

Exploring the onset response in more detail, we found some guidance beyond a simple movement to the image center. An image-dependent prediction per-
formed substantially better than an image-independent one (see Figure 7). Examples of fixation densities for the target of the first saccade (see Figure 6) confirmed that fixations were guided by the scene sensibly with a bias toward the center.

The main exploration focuses on similar image locations as the subjects fixate when the fixation density is converged (see Figure 4). The fixations during this phase are even better predicted by later fixation densities than later fixations themselves. During this phase, the fixation density gradually broadens, becoming less and less predictable. Correspondingly, the performance of all saliency models is maximal at the beginning of this phase and decreases over time. Importantly, DeepGaze II, a model that includes high-level features, has the largest advantage at the beginning of this phase; that is, the advantage of including high-level features starts immediately and reaches its peak already at fixation #2. As all predictions decline in parallel, a reason for the decline might be an increase of fixations that are not guided by the scene at all.

Finally, in the last phase, the fixation density reaches an equilibrium, and all fixation numbers predict each other equally well. Although subjects preferentially look at the same locations they look at during the main exploration, they are overall less predictable.

In the search data, we find a qualitatively similar temporal evolution of the fixation density as for memorization. We again see an onset response with initial central fixation bias, a period of marginally better predictability, and a final equilibrium state. However, the fixation density is much less predictable in general, there is virtually no central fixation bias after the onset response, and all saliency models perform much worse in predicting fixation locations, especially when we reuse the mapping from saliency map to fixation density from the scene-viewing data set. The initial central fixation bias is weaker in this data set as we delayed the onset of the first saccade (Rothkegel et al., 2017).

Low- versus high-level

At first glance, the observation that low-level models predict fixations well at the beginning and worse later in the trial fits well with the classical saliency model idea that the initial exploration is driven by low-level, bottom-up factors. However, the performance decline of low-level models resembles the decline of the empirical density, early fixations are well predicted by later fixation densities, and adding high-level features as in DeepGaze II improves predictions throughout the trial. These findings rather suggest that, during the initial main exploration, fixations are driven by the same high-level features as later fixations.

Indeed, even within the first fixation, adding high-level information improves predictions. Starting 200 ms after image onset, DeepGaze II performs better than the early vision–based model. Nonetheless, low-level models perform best for the first fixation and have the largest advantage over the central fixation bias for the first fixation. This increase in early vision–based model performance for the first fixation corroborates earlier findings that low-level guidance influences mainly the first fixation (Anderson et al., 2016; Anderson et al., 2015).

After the initial onset response, our data are even compatible with the extreme stance that low-level features have no influence on eye-movement behavior. This account agrees well with a range of literature that shows influences of objects (Einhäuser, Spain et al., 2008; Stoll et al., 2015) and other high-level features (Henderson et al., 1999; Torralba et al., 2006) on eye movements. The predictive value of low-level features, such as contrast at a location, could then be explained by their correlation with being interesting in a high-level sense. Such correlations are expected because very low contrast areas are devoid of content. As such, this explanation would also work to explain high-level influences based on low-level features. However, high-level features are better at predicting such that they necessarily have some predictive value beyond low-level features. Also, manipulations of contrast seem to have little influence on the fixation distribution beyond the first fixation (Açık et al., 2009; Anderson et al., 2015) such that the part of the fixation distribution that could be explained both by low- and high-level features is more likely to be explained by high-level features.

Adjusting the early vision–based model to the target subjects search for barely improves model performance. This finding implies that merely reweighing low-level features is insufficient for modeling eye movements in visual search. This failure argues against models in which simple top-down control operates on low-level features to guide eye movements (Itti & Koch, 2000; Treisman & Gelade, 1980; Wolfe, 1994). More complex processing of low-level features resulting, for example, in optimal control during visual search (Najemnik & Geisler, 2008) is compatible with our data. However, even for low-level targets, the feature of how well the target could be detected by the observer at each location is not low-level anymore. Thus, optimal control in search is not a low-level feature theory in our nomenclature even when the targets are defined by low-level features only.

Based on these considerations, low-level features seem relatively unimportant for eye-movement control in natural scenes and are largely restricted to an early bottom-up response during a trial. One reason to explain this lack of effect might be that we used stationary scenes. Instead, onsets or movements might
be necessary to attract fixations against top-down control (Jonides & Yantis, 1988; Yantis & Jonides, 1990). Moving scenes can produce much higher coherence among eye movements of participants (Dorr, Martinetz, Gegenfurtner, & Barth, 2010), and the classic experiments showing bottom-up control all used sudden onsets (Hallett, 1978, for example).

A normative reason why eye-movement control should focus on high-level features in stationary stimuli might be that fast responses are only required if the stimulus changes. When the stimulus changes, a fast response based on simple features might be advantageous, but if the stimulus is stationary, the eye-movement control system has sufficient time for more complex computations.

**Bottom-up versus top-down**

Based on the search results, we can confirm earlier reports that fixation locations during visual search are hardly predicted by saliency models (Chen & Zelinsky, 2006; Einhäuser, Rutishauser, et al., 2008; Henderson et al., 2007), which shows that top-down control can overwrite bottom-up control when subjects view static natural scenes. We even see some influence of the target in our visual search data, which argues for a fairly detailed adjustment of eye movements to the specific task at hand. Another observation supporting the conclusion that top-down control can override bottom-up control during visual search is that previewing the image and additional information about the target improve visual search performance (Castelhano & Heaven, 2010; Castelhano & Henderson, 2007). This result also fits well with earlier observations we made on this data set (Rothkegel et al., 2019), which showed that subjects adjusted their saccade lengths and fixation durations to the target for which they searched. Thus, our overall observations argue for a strong, detailed, top-down influence on eye-movement control during visual search. These observations do not imply that bottom-up influences are always overruled or never have predictive value in visual search. In other contexts, bottom-up features might be more predictive, especially when they are informative about target location. As these effects change with the task and the search target, they clearly represent top-down control, however.

This explanation implies that bottom-up factors can be overruled by top-down control signals as present in visual search. The only exception to this argument might be the first fixation chosen by the observer as the first chosen fixation follows a different density than later fixations, is best predicted by saliency models, and can even be predicted reasonably well in the visual search condition. Complicating the analysis of the first chosen fixation, we observe a temporal evolution within the first fixation from bottom-up to top-down control. This transition has been reported before as earlier saccades are more strongly biased toward the image center (Rothkegel et al., 2017) and might be driven more by bottom-up features (Anderson et al., 2016; Anderson et al., 2015). Furthermore, faces seem to attract fixations mostly early in a trial and later during a trial only when the fixation is preceded by a short fixation (Mackay, Cerf, & Koch, 2012), which fits the general trend we observe for high-level, bottom-up factors and shows another case in which the preferred saccade target changes within a single fixation duration. The transition from bottom-up effects to more value-driven saccades within a single fixation duration was also observed in single-saccade tasks with artificial stimuli (Schütz, Trommershäuser, & Gegenfurtner, 2012). Thus, the transition from bottom-up to top-down control might occur early, most likely already within the first fixation.

**Saliency metric**

To make sure that our conclusions do not depend on our choice of metric, we repeated our analyses over time with the eight metrics used in the MIT saliency benchmark (Bylinskii et al., 2016, see Appendix, Figure A1). All our conclusions hold independent of the metric used confirming earlier reports that the adjustment of blur, center bias, and nonlinearity unifies the commonly used saliency metrics (Kümmerer et al., 2015). We still present most of our results only in terms of the likelihood because it has the strongest connection to formal statistical model comparisons and evaluations (Schütz et al., 2017). Also some of the classically used metrics show a noticeable dependence on the number of fixations.

**Future prospects**

As we observe that the fixation density changes over the course of a trial, a single fixation density seems to be an insufficient description of eye-movement control. Instead, we found that exploring the temporal dynamics of eye-movement behavior throughout a trial provides interesting insights into the control of eye movements. Eye-movement dynamics have been studied earlier (e.g., Over, Hooge, Vlaskamp, & Erkelens, 2007; Tatler & Vincent, 2008) already providing interesting insights.

Additionally, investigating systematic tendencies in eye-movement behavior (Tatler & Vincent, 2008, 2009) could be informative. Such tendencies might include behaviors such as “scanning” or other systematic ways to search for a target. How these tendencies arise and
interact with the image content are upcoming challenges for eye-movement research.

To combine these observations into a coherent theory, models of eye movement behavior will have to evolve to incorporate predictions over time and with dependencies between fixations. So far, there are few models that produce dependencies between fixations (see Clarke, Stainer, Tatler, & Hunt, 2017; Engbert, Trukenbrod, Barthelmé, & Wichmann, 2015; Le Meur & Liu, 2015; Tatler, Brockmole, & Carpenter, 2017, for notable exceptions), and even those that do are rarely evaluated regarding their abilities to produce natural dynamics and generally do not handle a connection to the explored images. Here, we only scratch the surface of the possibilities to check models more thoroughly using the dynamics of eye movements. We believe that we now have the statistical methods (Barthelmé, Trukenbrod, Engbert, & Wichmann, 2013; Schütt et al., 2017) and data sets to pursue this research direction further.

To facilitate the exploration of the dynamic aspects of eye-movement behavior, we share the data from our scene-viewing experiment at http://dx.doi.org/10.17605/OSF.IO/N3BYQ and the data from our search experiment at http://dx.doi.org/10.17605/OSF.IO/CAQT2. We hope that such shared large data sets may provide a strong basis for the exploration of the dynamics of eye movements.

Keywords: saliency, fixations, natural scenes, visual search, eye movements

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft through grants to F.A.W. (grant WI 2103/4-1) and R.E. (grant EN 471/13-1). We acknowledge support of the Open Access Publishing Fund of the University of Tübingen.

*Heiko H. Schütt and Lars O. M. Rothkegel contributed equally to this work.

Commercial relationships: none.

Corresponding author: Heiko H. Schütt.

Email: Heiko.schuet@uni-tuebingen.de.

Address: Neural Information Processing Group, Universität Tübingen, Tübingen, Germany.

Footnotes

1 We also implemented an additive center bias, which performed worse than the multiplicative version for all models.

2 Link for paper version: https://github.com/cvzoya/saliency/tree/master/code_forMetrics

3 Link for the paper version: http://www.vision.caltech.edu/~harel/share/gbvs.php

4 We restrict ourselves to fixations #1–#25 here for consistency with later plots over time. Including later fixations does not qualitatively change the results displayed here. The only change is that later fixations are predicted worse by all models, decreasing the absolute performance of all models slightly.

5 As we did not use the same images for the two experiments, we cannot rule out an effect of image content entirely. However, 87 of the 90 images in the scene-viewing experiment had higher average likelihoods than the best predictable image from the search experiment; that is, the distributions were almost nonoverlapping. Thus, this explanation is unlikely to explain the whole effect.

References


Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F.,


Rothkegel, L. O. M., Truyenbrod, H. A., Schütt, H.


Appendix

Figure A1. Results over time for all metrics used in the MIT saliency benchmark (Bylinskii et al., 2016). The top two rows show the results for the corpus data set. The bottom two rows show the results for the search data set. The differences between models appear to have slightly different sizes and SIM and KL-divergence show an additional trend over time, which is caused by the change in the number of available fixations.