Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *Competitive Visual Processing Across Space and Time*

A new approach to modeling the influence of image features on fixation selection in scenes

Antje Nuthmann[1] and Wolfgang Einhäuser[2]

[1]Psychology Department, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, United Kingdom. [2]Neurophysics Department, Philipps-University Marburg, Germany

Address for correspondence: Antje Nuthmann, Psychology Department, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, UK. Antje.Nuthmann@ed.ac.uk

Which image characteristics predict where people fixate when memorizing natural images? To answer this question, we introduce a new analysis approach that combines a novel scene-patch analysis with generalized linear mixed models (GLMMs). Our method allows for (1) directly describing the relationship between continuous feature value and fixation probability, and (2) assessing each feature's unique contribution to fixation selection. To demonstrate this method, we estimated the relative contribution of various image features to fixation selection: luminance and luminance contrast (low-level features); edge density (a mid-level feature); visual clutter and image segmentation to approximate local object density in the scene (higher-level features). An additional predictor captured the central bias of fixation. The GLMM results revealed that edge density, clutter, and the number of homogenous segments in a patch can independently predict whether image patches are fixated or not. Importantly, neither luminance nor contrast had an independent effect above and beyond what could be accounted for by the other predictors. Since the parcellation of the scene and the selection of features can be tailored to the specific research question, our approach allows for assessing the interplay of various factors relevant for fixation selection in scenes in a powerful and flexible manner.

Keywords: naturalistic scenes; image features; eye movements; fixation probability; GLMM

Introduction

Research using simple displays has shown that attention and memory are coupled, as evidenced by interference between attention and features of items in visual working memory.[1,2] In the context of natural scenes, however, it has remained controversial which features drive attention. Since eye movements are highly correlated with the path of visual attention,[3] this question can be operationalized by asking which properties make a region of a complex scene likely to be fixated.

The dominant theoretical and computational framework to emerge has been image salience, in which low-level properties of the stimulus play a crucial role in guiding attention and the eyes.[4,5] Empirical studies on salience maps have addressed the questions of what features should be part of the map and how these features should be combined.[6] The typical approach has been to test if there are any differences between visual characteristics at locations that were fixated by observers and control locations.[7,8] The basic picture that emerged is that image features, including luminance, contrast, and edge density all differ between fixated locations and control locations.[9] However, such explorations of visual features at fixation suffer from several limitations. First, such analyses compare average feature values for two post hoc created groups of fixated and control locations in a scene. A more informative approach would be one that distinguishes between fixated and nonfixated scene regions and directly describes the relationship between fixation probability and local feature values on a continuous scale. For example, a mixture model approach suggested that observers do not actively fixate luminance extremes,[10] which is suggestive of a nonmonotonic

doi: 10.1111/nyas.12705

© 2015 The Authors. *Annals of the New York Academy of Sciences* published by Wiley Periodicals Inc. on behalf of The New York Academy of Sciences.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

relationship between luminance and fixation probability. Such relationships cannot be uncovered by existing analysis approaches. Second, for a particular location, different features tend to be correlated, possibly shadowing the true effect (or true null-effect) of one feature by image-inherent correlations to other features.[11] Similarly, features and viewing behavior are both associated with generic biases. Importantly, it is a well-established finding that observers fixate more often toward the center of the image than the edges.[7,9,12] This central bias is not fully explained by centrally located features or initial fixation location.[13] To be of interest, any effects of image features need to be above and beyond such generic biases. In turn, considering their relevance for viewing behavior, such biases should be treated as "features" in their own right,[14] rather than just being accounted for by baseline choice.

Here, we introduce a novel analysis approach, which overcomes these issues. First, we present a scene-patch analysis that allows for fully describing the relationship between continuous feature values and fixation probability. Second, we utilize a statistical control approach to assess each feature's unique contribution to fixation selection.

Five candidate image features were chosen. First, three common measures of local image statistics that characterize different properties of image luminance were examined: luminance, luminance contrast, and edge density. Luminance contrast, arguably the best investigated feature, has been found to be elevated at fixated scene patches in grayscale images.[7,8,12,15,16] In addition, edge density has been found to be greater at fixated than nonfixated locations.[7,17]

In addition, we examined visual clutter[18] as a surrogate measure for objects and synergistic image segmentation[19] as an approximation of local object density in the scene. Clutter is an image-based feature of visual complexity, which has been studied mostly in the context of a search task. A frequently adopted model of clutter is the feature congestion model,[18] which estimates clutter in terms of the density of luminance contrast, color, and orientation. In a study investigating the influence of clutter on real-world scene search, it was found that the first fixation, but not subsequent fixations, tended to be centered on a region of significantly higher clutter than would be predicted by chance.[20] One goal of image segmentation is to break up the image into meaningful "chunks," approximating the beginnings of an object-based representation. Here, we use synergistic image segmentation,[19] which combines image segmentation based on the mean shift procedure[21] with a confidence-based edge detector.[22]

Luminance and luminance contrast are low-level features. Edge density is operationalized as a mid-level feature, as it can be defined independent of object content, but is not contained in the second-order scene structure.[23] We operationalize clutter and synergistic segmentation as higher-level features (but not high-level features, as their computation does not include any contextual component or task demand).

This study presents a statistical modeling framework to simultaneously test the influence of image features on fixation selection in scenes during a memorization task. Our approach requires three steps of image and data processing. First, feature maps for each image and feature are constructed via image processing. Second, to obtain local image statistics, each photograph is parcellated into local image regions. Unless otherwise stated, we use a 8 × 6 grid, yielding 48 quadratic scene patches (see Fig. 1A) with each grid cell spanning 3.2° × 3.2° (100 × 100 pixels). For each patch, local image statistics are extracted from the image feature maps. Third, the empirical eye-fixation data are mapped onto the scene analysis grid: For each observer and image it is coded whether a given image patch was fixated (1) or not (0) throughout the trial. Generalized linear mixed models (GLMMs) are then used to assess the impact of various image features on selecting image patches for fixation.

Methods

Participants, apparatus, and materials

Analyses were based on a large corpus of eye movements during scene viewing.[24,25] Seventy-two participants (mean age = 22.6 years, 34 males) each viewed 135 color photographs of real-world scenes from a variety of categories (indoor and outdoor). The 92 indoor scenes came from different subcategories, ranging from common rooms in one's house (e.g., living room, kitchen) to images from shops, garages, etc. Scenes were presented on a 21-inch CRT monitor with a screen resolution of 800 × 600 pixels and subtended 25.78° horizontally × 19.34°

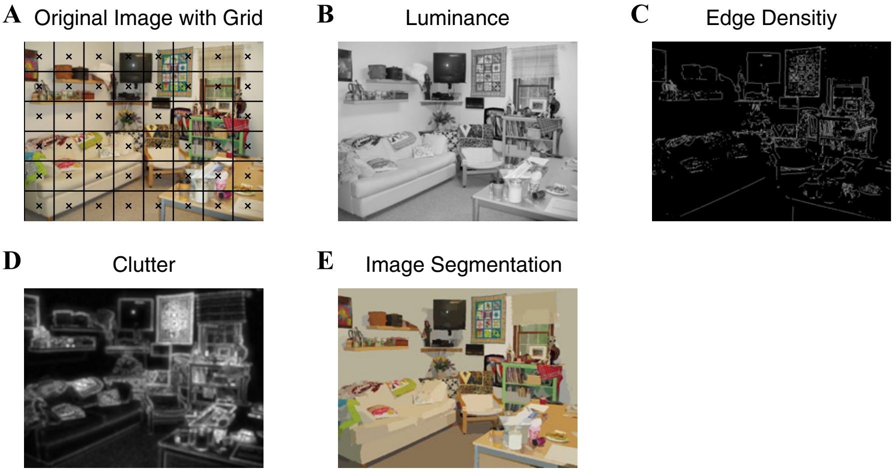


Figure 1. Example image and feature maps. (A) The original image with the analysis grid overlaid. (B) Luminance map. (C) Edge density map after filtering the image with a Sobel operator. (D) Feature-congestion visual clutter map. (E) Synergistic segmentation of the scene, resulting into 2,277 homogenous tiles.

vertically at a viewing distance of 90 cm. Eye movements were recorded using an SR Research EyeLink 1000/2K system. Data from the right eye were analyzed.

Design and procedure

The 135 scenes were divided into three blocks of 45 scenes. In each block, participants performed one of three viewing tasks: scene memorization, preference judgment, or scene search.[24] For the purpose of this paper, only data from the memorization task were analyzed. Participants were instructed to encode the scene in preparation of an old/new recognition test administered at the end of the experiment. Each trial started with a centrally located pretrial fixation marker, which acted as a fixation check. Afterwards, the scene was presented for 8 seconds. Scenes were rotated through task and task order across groups of participants.

Data analysis

Gaze raw data were converted into a fixation sequence matrix using SR Research Data Viewer. Data were further processed and analyzed using MATLAB 2009b (The MathWorks, Natick, MA, USA) and the R system for statistical computing (version 3.1; R Development Core Team, 2014) under the GNU General Public License (Version 2, June 1991). Image processing was performed in MATLAB.

Computation of image features. For each image, five different features were defined at each of the 8 × 6 grid locations.

Luminance. Luminance of each pixel was defined by converting the sRGB values of the image (assuming IEC 61966–2–1 specification) to CIE $L^*a^*b^*$ space and retaining only luminance (L^*) information. For each image, luminance was then mapped linearly to the interval [0, 1]. This scaled version will be referred to as luminance throughout (Fig. 1B). The feature value of each grid cell was defined as mean luminance over all 100 × 100 pixels in the cell. Greater luminance is associated with a higher degree of subjectively perceived brightness.

Luminance contrast. Based on the luminance map (Fig. 1B), each local image patch was labeled with its local contrast value. The contrast for each grid cell was defined as a version of root-mean-square contrast:[26] the standard deviation of luminance values of all pixels in the grid cell divided by the mean luminance of the image.[8,15] In general, more uniform patches have less contrast.

Edges. Edges were defined as boundaries between regions of distinctly different mean luminance. The locations of edges in an image were determined by applying a Sobel operator to the luminance map, which extracts an approximation to the luminance gradient at each point in the image.[7,17] Thresholds were applied using the adaptive procedure implemented in the *edge* function in the Image Processing

Toolbox for MATLAB, resulting in a binary image with 1's where the function finds edges in the image and 0's elsewhere. Thus, the procedure produced a black and white image, with white representing the edges (see Fig. 1C). Edge density was then defined as the mean over all pixels in a grid cell for this binary image; that is, the proportion of edges in the cell. These proportions ranged from 0 to 0.339 (mean: 0.043, standard deviation: 0.034). To "stretch out" proportions that are close to 0, edge densities were submitted to a logit transformation ($\text{logit}(p) = 0.5 \times \ln(p/(1-p))$),[27] after regularizing 0 to the smallest possible nonzero value in the data (10^{-4}) for numerical reasons.

Clutter. A feature congestion map of visual clutter was computed for each scene, using the algorithms described by Rosenholtz *et al.*[18] and MATLAB code provided at http://dspace.mit.edu/handle/1721.1/37593. For each such feature map, the range of feature values was normalized linearly to [0, 1]. Figure 1D depicts the feature congestion map of visual clutter for the example scene shown in Figure 1A. Local feature values for clutter were defined as the mean over this feature map's values within each grid cell.

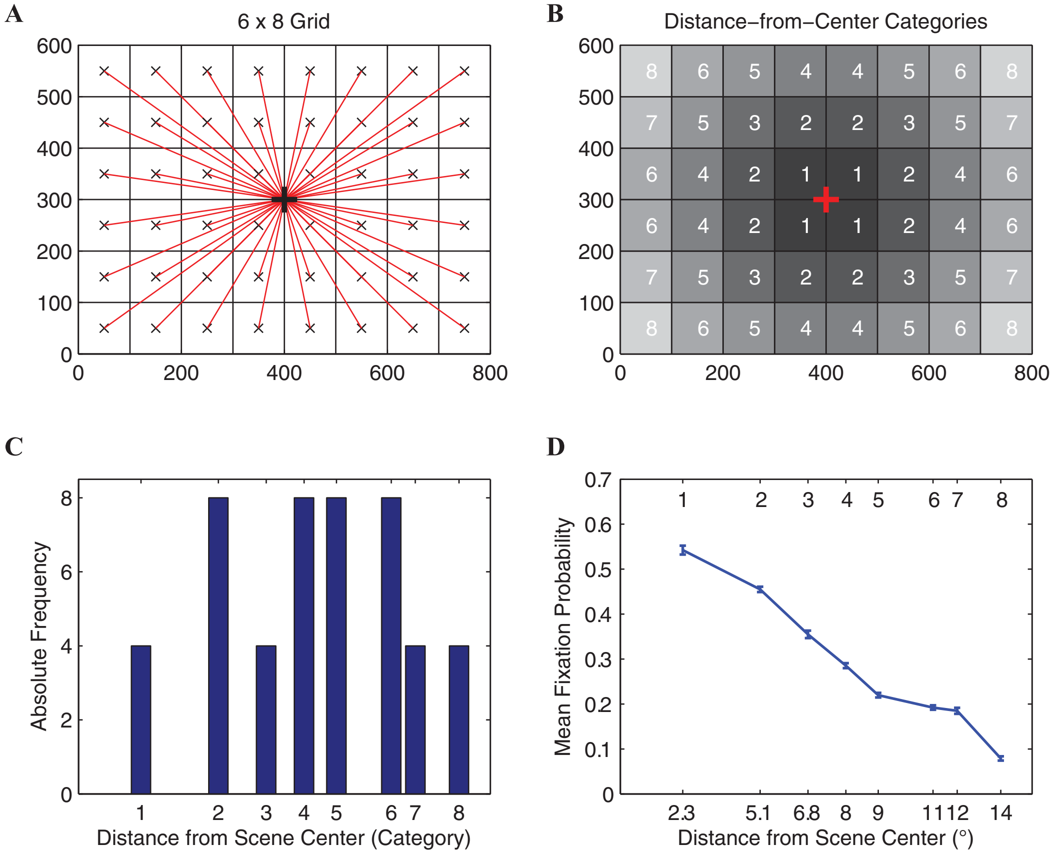
Synergistic image segmentation. The goal of image segmentation is to break up the image into meaningful or perceptually similar regions. We used the synergistic segmentation,[19] which combines mean shift based color image segmentation[21] with edge confidence and gradient maps.[22] The algorithms, implemented in C++, are available via the Edge Detection and Image SegmentatiON (EDISON) System,[19] as is a MEX wrapper for MATLAB (http://www.wisdom.weizmann.ac.il/~bagon/matlab.html). Each image was subjected to the synergistic image segmentation by using the default parameters (mean shift: spatial resolution parameter $h_s = 7$, range bandwidth parameter $h_r = 6.5$, minimum region size $M = 20$). On average, 2,947 segments per scene were obtained (see Fig. 1E for an example). For each grid cell, the number of homogenous segments was determined.

We did not analyze low-level color features since neither the stimuli nor display used in this study were designed to capture low-level chromatic properties. By design, however, clutter and synergistic image segmentation make use of chromatic information; these composite features are rather insensitive to the precise color space or color representation.

Central bias. To explicitly model the central bias of fixation in the GLMM framework, a central-bias predictor was created as follows. For each cell of the image grid, the distance between the center of the grid cell and the center of the image was determined (red vectors in Fig. 2A). This resulted in eight distinct distance categories; each of them comprised either four or eight cells (Fig. 2C). By definition of the grid, these categories are not equidistant. In Figure 2B image grid cells are numbered according to the distance category they belong to (from 1 = proximal to 8 = distal), while absolute distance is color-coded such that the color of more distant cells becomes progressively brighter. Statistical models included the central-bias predictor as distance from scene center in degrees of visual angle.

Generalized linear mixed models. Our response variable is binary—for a given observer and image a given grid cell was either fixated (1) or not (0). The observation matrix comprised 155,520 entries of zeros and ones (45 images × 72 subjects × 48 grid cells). GLMM[28–30] were used to determine the impact of various image features on fixation probability in scenes. An advantage of GLMM is that they do not require any form of data reduction; hence we can model the data at the level of individual observations, that is, the zeros and ones. The probabilities are modeled through a link function. For binary data, this link function is the logit transformation of the probability.[28–30] For our analyses, we used the *glmer* program of the *lme4* package[31] supplied in *R*, with the bobyqa optimizer. For the GLMMs, we report regression coefficients (bs), standard errors (SEs), and z-values ($z = b/\text{SE}$). Predictors were centered to have mean 0 and scaled to have standard deviation 1.

Mixed models are statistical models that incorporate both fixed-effects parameters and random effects. Our models included subjects (subject ID) and scenes (scenes ID) as random effects to capture variance attributed to the randomness of subject and item sampling. All models included random intercepts for subjects and items. To determine whether random slopes should be included, we pursued a data-driven approach. For each predictor, four models that differed in their random effects



8	6	5	4	4	5	6	8
7	5	3	2	2	3	5	7
6	4	2	1	1	2	4	6
6	4	2	1	1	2	4	6
7	5	3	2	2	3	5	7
8	6	5	4	4	5	6	8

Figure 2. Central bias analysis. (A) Image grid with vectors (in red) connecting the center of the grid cell with the center of the image. (B) Assignment of the resulting eight distinct distance categories to image grid cells. Absolute distance is color-coded such that the color of more distant cells becomes progressively brighter. (C) Frequency of occurrence of categorical distances. (D) Mean fixation probability as a function of distance from scene center. Error bars are 95% binomial proportion confidence intervals, obtained using the score confidence interval.[51] In panels (C) and (D) the spacing on the *x*-axis preserves relative distances between distance categories.

structure were compared. The first model included random intercepts for subjects and items only. The second model added random slopes for subjects; the third model added random slopes for items. The fourth model included random intercepts and slopes for subjects and items, that is, the maximal random effects structure. The models were compared using likelihood ratio tests to identify the best random effects structure by taking both goodness of fit and model parsimony into account.

Results

The first valid fixation in each trial was defined as the first fixation that began after the onset of the scene image. In a given trial, the fixation on the pretrial fixation marker fell on one of the centrally located image patches, and this patch was excluded from analysis for this image and observer, irrespective of whether it was revisited or not. This was done because the fixation on the fixation marker typically extended into the period of scene presentation.

Across scenes and participants, 28.3% of image patches were selected for fixation; 16.2% received exactly one fixation, and 12.1% were fixated more than once during the course of the 8-s viewing. Here, we modeled the probability of fixation, not distinguishing between single and multiple fixations.

To explore the empirical data, for each image feature we calculated fixation probability as a function

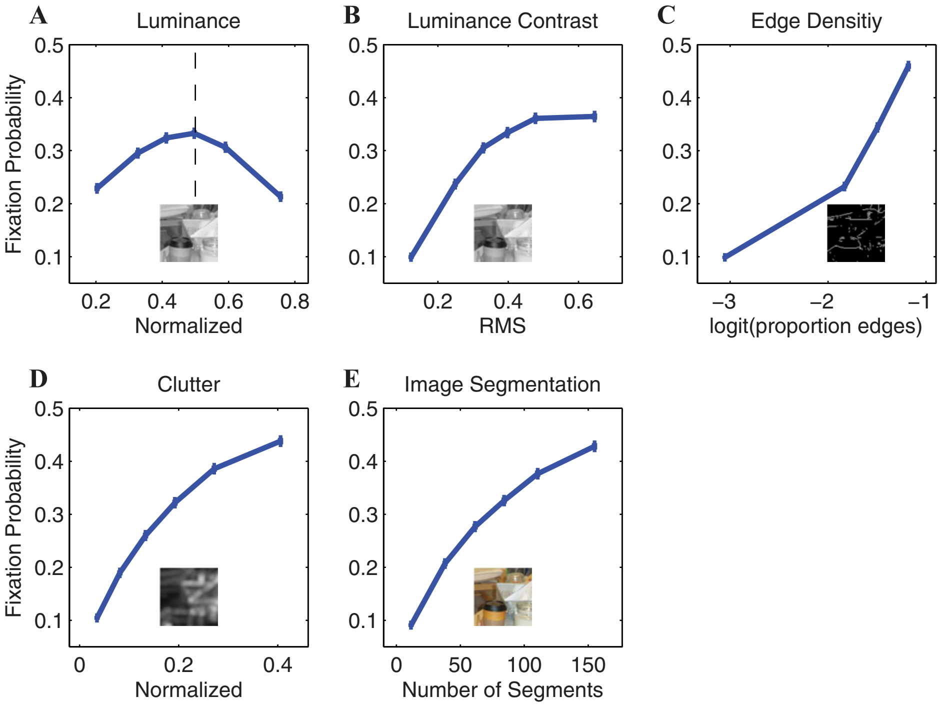


Figure 3. Five main effects of local image statistics on fixation probability in a scene memorization task. Predictors are (A) luminance, (B) luminance contrast, (C) edge density, (D) clutter, and (E) synergistic segmentation. Error bars are 95% binomial proportion confidence intervals. Data are from right eye.

of the respective feature. The panels in Figure 3, one for each feature, display observed mean fixation probabilities over suitably binned category means. For each feature, categories were created using quantiles of the continuous variable, resulting into approximately equal-sized data subsets. The data are suggestive of a negative quadratic relationship between luminance and fixation probability (Fig. 3A) and a monotonically increasing relationship between luminance contrast and fixation probability (Fig. 3B). Furthermore, as the number of edges in a patch increases, fixation probability increases (Fig. 3C). Likewise, as the visual clutter in a patch increases, fixation probability increases as well (Fig. 3D). Finally, the more meaningful "chunks" there are in a patch, the higher fixation probability (Fig. 3E). With regard to the central bias of fixation, the averaged empirical data suggest that fixation probability linearly decreases with increasing distance from scene center (Fig. 2D).

In sum, the averaged empirical data depicted in Figure 3 suggest that all tested visual features predict whether image patches are fixated or not. However, to be of interest any effects need to be (1) above and beyond what can be accounted for by other features and (2) above and beyond a general preference for fixating the center of the image.

Before embarking on the statistical model building, we consider the distribution of features within images (Fig. S1) and the correlations between image features (Fig. S2). Our composed scenes tended to show the common bias toward having more visual features in their center.[13] The only exception was luminance, for which there was a slight increase/decrease in mean luminance for image patches in upper/lower scene regions, respectively (Fig. S1). The feature bias toward the center of the image also shows in significant negative correlations between the cells' local feature values and their distance from scene center. The strength of the correlation between local image statistics and central bias ranged between −0.31 (for edge density, $P < 0.001$) and −0.03 (for luminance, $P < 0.05$). As noted earlier, in natural images different visual features tend to be correlated for a particular location.[11] For the images and features considered here, the largest correlations involve edge density, which correlates both with luminance contrast ($r = 0.60$), clutter ($r = 0.62$), and the number of homogenous segments ($r = 0.61$). Further, the correlation between clutter and number of homogenous segments is 0.58; the matrix of pairwise scatter plots in Figure S2 provides a full account. The purpose of linear mixed models is to factor in the correlations between predictors.

We pursued an incremental model building strategy. Luminance and luminance contrast are fundamental stimulus dimensions encoded by the visual system. Therefore, we first modeled the effects of luminance (luminance-only model) and contrast (contrast-only model) separately, before assessing their unique effects in a model including them both (LumCon model). A final model in this series adds the central-bias predictor. We conclude by reporting the results for the full model, which included all image features along with the central-bias predictor.

Luminance-only model

The averaged empirical data suggest that fixation probability is highest for medium-luminance patches (Fig. 3A). As local luminance moves toward extreme values, fixation probability decreases, while this drop is not linear but shows negative acceleration. Accordingly, the luminance-only GLMM included an intercept and a quadratic term for luminance (lum^2)[a] as fixed effects. In addition, the model included random intercepts for subjects and items, and random slopes for items. The fixed effect of lum^2 was significant ($b = -0.31$, SE $= 0.03$, $z = -9.87$, $P < .001$). The corresponding partial GLMM effect is displayed in Figure 4B. Parameter estimates are obtained on the log-odds or logit scale, which is symmetric around zero, corresponding to a probability of 0.5, and ranges from negative to positive infinity. Thus, negative log-odds correspond to probabilities $P < 0.5$. For computation of the partial GLMM effect, random factors variance was removed using the *remef* function provided by Hohenstein and Kliegl.[32]

Contrast-only model

According to Figure 3B, fixation probability increases as local luminance contrast increases. The contrast-only GLMM included contrast as the only fixed effect (in addition to the intercept), and by-item random intercepts and slopes along with by-subject intercepts. The effect of contrast was significant ($b = 0.71$, SE $= 0.04$, $z = 17.83$, $P < 0.001$), and the corresponding partial GLMM effect is displayed in Figure 4C.

Luminance-and-contrast model

For our complex color scenes, local luminance and contrast are not independent, but show a nonlinear relation. Patches with medium luminance are associated with a large variability in local contrast, but they tend to have higher contrast (Fig. 4A). Conversely, darker and brighter patches tend to have lower contrast, with minimal/maximal luminance leaving little room for variability in local contrast. For comparison, we also fit a linear regression to the data (Fig. 4A). The linear correlation between local luminance and contrast is -0.18 ($P < 0.001$); the correlation is larger for outdoor scenes ($r = -0.3$, $P < 0.001$) than for indoor scenes ($r = -0.12$, $P < 0.001$), probably owing to the fact that sky regions tend to be both bright and low in contrast.[33] The LumCon GLMM included both a quadratic term for luminance and a linear term for contrast as fixed effects. In addition to the random intercepts, the model included by-item random slopes for lum^2 and contrast. Both fixed effects remained significant (lum^2: $b = -0.13$, SE $= 0.03$, $z = -4.17$, $P < 0.001$; contrast: $b = 0.68$, SE $= 0.05$, $z = 14.94$, $P < 0.001$), and their partial GLMM effects are displayed in Figure 4B (luminance) and Figure 4C (contrast). Comparing the results from the three models, it becomes clear that the addition of contrast to the luminance model diminishes the quadratic effect of luminance.

In the next step, the central-bias predictor was added to the LumCon model. The significant negative estimate for the central-bias predictor ($b = -0.60$, SE $= 0.03$, $z = -21.84$, $P < 0.001$) confirmed that fixation probability decreases with increasing distance from image center. Importantly, the effects of both luminance (lum^2: $b = -0.07$, SE $= 0.03$, $z = -2.44$, $P < 0.05$) and contrast ($b = 0.59$, SE $= 0.04$, $z = 13.95$, $P < 0.001$) continued to be reliable.

[a] For the mean-centered luminance predictor, the *x*-coordinate of the parabola's vertex coincides with 0 such that the linear term vanishes.

Table 1. **Final generalized linear mixed model fitting fixation probability for a scene memorization task, fit by Laplace approximation: means, standard errors, and *z*-values of fixed effects on fixation probability; variances of the random effects**

	Fixed effects			Random effects, variance	
Predictor	*b*	SE	*z*	By-items	By-subjects
Intercept	−1.217	0.059	−20.51	0.349	0.046
Luminance quadratic	**−0.012**	**0.027**	**−0.46**	0.084	–
Luminance contrast	**0.033**	**0.052**	**0.62**	0.334	–
Edge density	0.609	0.065	9.38	0.492	–
Clutter	0.184	0.043	4.31	0.217	–
Number of segments	0.389	0.046	8.40	0.260	–
Central bias	−0.490	0.030	−16.28	0.057	0.031

NOTE: Nonsignificant coefficients are set in bold ($|z| < 1.96$, $P > 0.05$).

Full model

The question arises whether these relationships hold once additional image-feature predictors are included in the model. The full model included the central-bias predictor and all five image features as fixed effects. The maximal random effect structure[34] would require estimating 56 random effects parameters (28 by subject and 28 by item), and this model failed to converge. Data-driven exploration of random effects suggested that by-subject random slopes for image features are not needed. Consequently, the full model included 31 random effects parameters (by subject: 2 random effects, 1 correlation term; by item: 7 random effects, 21 correlation terms). The results for the fixed effects and the variances of the random effects are summarized in Table 1. The central bias is the strongest predictor of where observers fixate in a scene. Notably, after taking central bias into account, edge density, visual clutter, and the number of homogenous segments can still independently predict whether image patches are fixated or not. The *z*-statistics are suggestive of particularly strong effects of edge density and the number of segments in a patch. Importantly, neither luminance nor contrast have an independent effect above and beyond what can be accounted for by edge density and the two higher-level features approximating local object density in the scene. To illustrate this point, Figure 4 includes the partial GLMM effects for luminance (panel 4B) and luminance contrast (panel 4C) in the full model (blue long-dashed line).

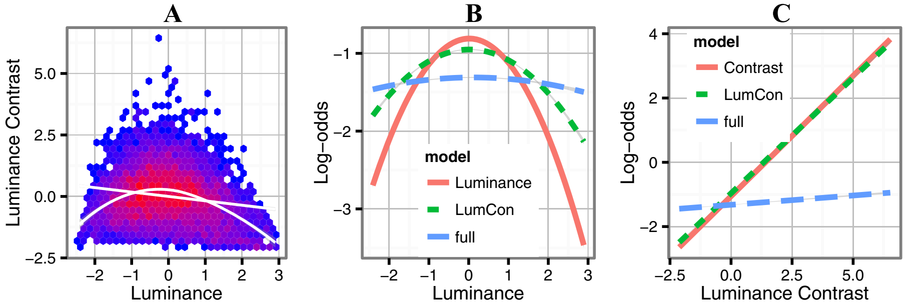


Figure 4. Statistical analysis of local luminance and luminance contrast. (A) Joint distribution of local luminance and contrast values. Hexagonal binning was used to avoid overplotting of 6,480 data points (135 images × 48 grid cells). Frequency information is displayed as variations in color, with colors ranging from blue (few data points) to red (many data points). The curved white line is an approximation of the data by a polynomial spline, and the straight white line represents a linear regression fit. (B) Partial quadratic effect of local luminance on fixation probability in log-odds scale for the luminance-only model (red solid line), the LumCon model (green dashed line), and the full model (blue long-dashed line). (C) Same for local luminance contrast. Feature values are *z*-scores. See text for more details.

Additional analyses

To explore in more detail whether the results from the full model converge with previously reported findings,[11] we tested which image features made the effects of luminance and luminance contrast disappear by incrementally extending the GLMM that included lum^2 and luminance contrast along with the central-bias predictor. If only edge density was added to this model, the effect of contrast was much reduced but remained just significant ($b = 0.09$, $\text{SE} = 0.05$, $z = 1.98$, $P = 0.048$). However, if both edge density and visual clutter were added, the effect of luminance contrast was no longer significant ($b = 0.01$, $\text{SE} = 0.05$, $z = 0.12$, $P = 0.901$). If only edge density was added, the quadratic effect of luminance was no longer significant ($b = 0.01$, $\text{SE} = 0.03$, $z = 0.33$, $P = 0.741$); the same was true when only the number of homogenous segments was added ($b = -0.02$, $\text{SE} = 0.03$, $z = -0.89$, $P = 0.375$).

Our central-bias predictor was calculated as the Euclidean distance from image center (Fig. 2), which is an isotropic measure. Clarke and Tatler[35] recently proposed that the central bias is best modeled by an anisotropic two-dimensional Gaussian distribution whose vertical variance is less than half the horizontal variance. Therefore, for exploratory purposes, we applied an alternative measure of central bias for each grid cell based on a two-dimensional Gaussian distribution centered over the image center. Following Clarke and Tatler,[35] the horizontal variance of the Gaussian was set to 0.23 (in units of half the image width), and the vertical variance to 0.10 (0.23×0.45). We then reran the full GLMM with the anisotropic Gaussian central-bias predictor rather than the isotropic Euclidean distance-to-center predictor. To ease comparison, the anisotropic predictor was entered with a negative sign, such that increasing values correspond to more peripheral locations. The fixed-effect estimate for the anisotropic predictor ($b = -0.50$, $\text{SE} = 0.03$, $z = -17.19$, $P < 0.001$) was very similar to the estimate for the isotropic predictor ($b = -0.49$, $\text{SE} = 0.03$, $z = -16.28$, $P < 0.001$, Table 1). We conclude that the Euclidean distance-to-center predictor does not substantially underestimate the contribution of the central bias in our model.

Furthermore, it is important to confirm that the results do not depend on the choice of grid-cell size. Therefore, we repeated the analyses for a fine grid that had four times as many grid cells as the original grid. Compared to the original grid, the sides of the squared patches were cut in half (50×50 pixels $= 1.6° \times 1.6°$), leading to a 16×12 grid with 192 cells. For all models reported in this paper, the qualitative pattern of results was the same as for the 8×6 grid. The results for the full GLMM are summarized in Table S1. Significant effects for image features and the central-bias predictor were somewhat stronger than for the original 8×6 grid, most likely owing to the finer resolution of the 16×12 grid. We also tested a very coarse grid by doubling the sides of the squared patches (200×200 pixels $= 6.4° \times 6.4°$), which led to a 4×3 grid with only 12 cells. Not surprisingly, effects were weaker for the coarse 4×3 grid, but the overall pattern of results did not change (Table S1).

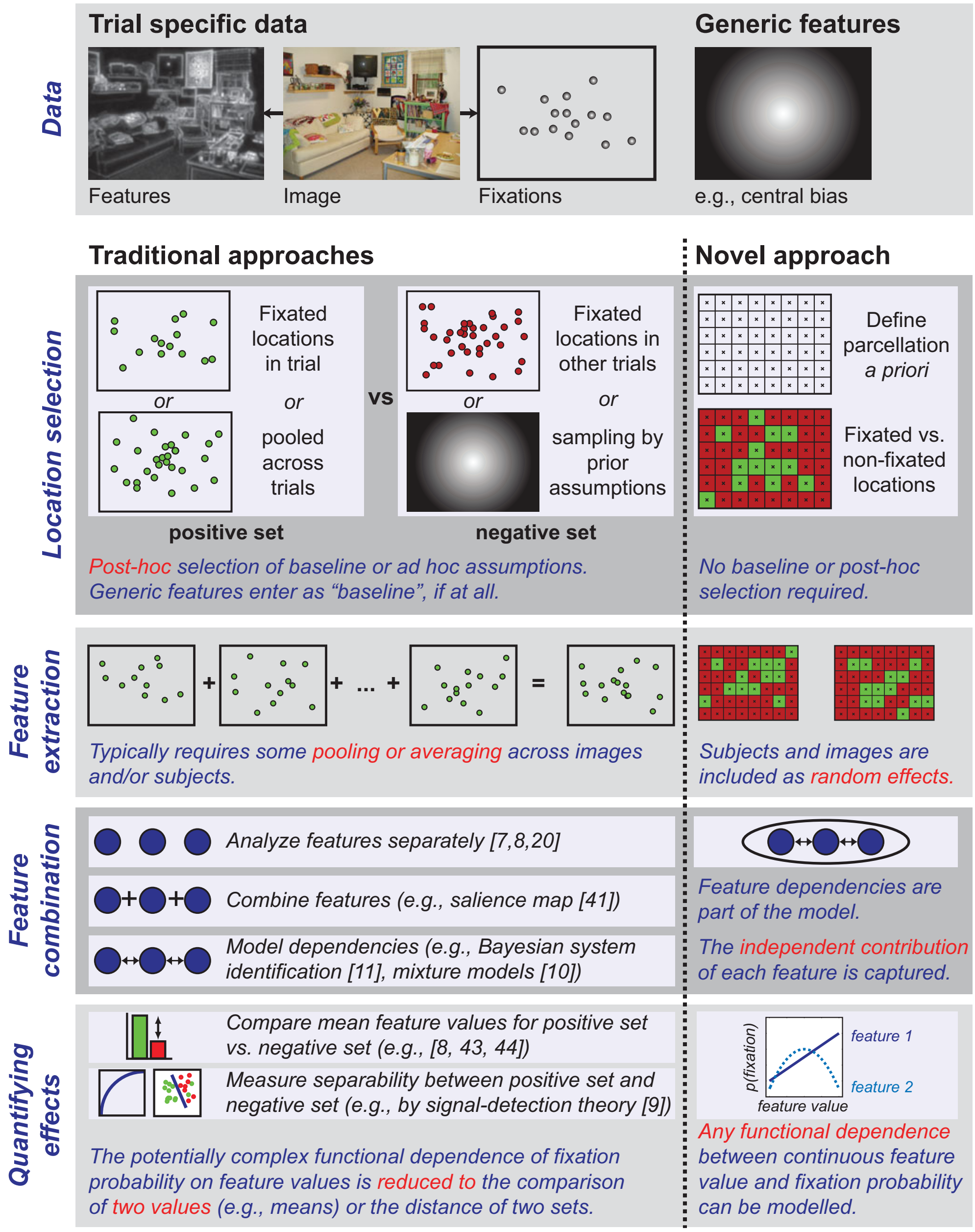
Discussion

This paper introduces a new analysis approach to assess quantitatively the extent to which image features predict the probability with which scene regions are selected for fixation. Specifically, we combine a scene-patch analysis with a statistical modeling approach that allows for directly describing the relationship between continuous feature values and fixation probability. The approach unites four desirable properties by *explicitly* accounting for (1) generic biases (e.g., central bias), (2) inter-item and inter-subject variability, (3) nonmonotonic (e.g., quadratic) effects of feature values on fixation probability, and (4) dependencies between features (see Fig. 5 for a visual summary).

Our approach bears some similarities with analysis techniques for eye guidance in reading. In this analogy, the image corresponds to a paragraph of text, the grid cell to a word, the features within the grid cell to word properties, and the probability of fixating a cell to the probability of fixating a word. In reading, such analysis techniques revealed, for example, that the probability of fixating (i.e., not skipping) a word decreases linearly with word frequency.[36] Here, we asked whether such relationships exist between local image statistics and fixation selection in scenes.

Unit of analysis

Our approach allows and requires an *a priori* parcellation of the scene (Fig. 5). We used a grid, which is a natural choice whenever homogeneous, nonoverlapping, and exhaustive image coverage is required



Post-hoc selection of baseline or ad hoc assumptions. Generic features enter as "baseline", if at all.

No baseline or post-hoc selection required.

Typically requires some pooling or averaging across images and/or subjects.

Subjects and images are included as random effects.

Analyze features separately [7,8,20]

Combine features (e.g., salience map [41])

Model dependencies (e.g., Bayesian system identification [11], mixture models [10])

Feature dependencies are part of the model.

The independent contribution of each feature is captured.

Compare mean feature values for positive set vs. negative set (e.g., [8, 43, 44])

Measure separability between positive set and negative set (e.g., by signal-detection theory [9])

The potentially complex functional dependence of fixation probability on feature values is reduced to the comparison of two values (e.g., means) or the distance of two sets.

Any functional dependence between continuous feature value and fixation probability can be modelled.

Figure 5. A new approach to modeling the influence of image features on fixation selection in scenes—summary of its main properties including a comparison to existing approaches. References are indicated in brackets.

or desirable. Provided the scale-invariance of natural scenes,[37,38] the precise choice of grid dimensions is not critical as long as the scale is above the eye-tracker's precision and accuracy and as long as a sufficient number of cells are available per image. Here, we verified this by showing that a factor of 2 on the linear grid dimension (i.e., a factor of 4 with regard to area) did not alter the qualitative pattern of results. Rather, effects tended to be stronger with increasing resolution of the grid.

We ourselves have argued that attentional selection in scenes has a strong object-based component.[24,25,39] Accordingly, an alternative parcellation may be based on object outlines. Unlike the grid, such a parcellation is typically neither exhaustive nor necessarily a tiling (i.e., gaps and overlap between objects are acceptable), but this presents no challenge to the proposed method. Objects typically distinguish themselves from scene background in the visual features relevant for fixation selection.[39] For the present purposes, our higher-level features were chosen as surrogate measures of objects, rendering an object-based parcellation circular. An object-based parcellation would be appropriate if one, for example, wishes to investigate the relative prioritization of objects by their features.[40] This example illustrates how the parcellation of the scene can be adapted depending on the experimental question asked.

Choice of baseline and generic biases

In particular in the context of salience map models, which aggregate multiple low-level features,[41] there has been considerable debate regarding the evaluation of the prediction performance of a given feature or model.[4,42] Typically, a positive (fixated) and negative (control) set of locations is defined and then feature values in these sets are compared, either directly,[8,43] after image-wise *z*-score normalization[44] or in terms of discriminability as quantified by the area under the receiver-operating characteristic (ROC) curve (AUC or A′)[9] (Fig. 5). The requirement to choose a negative set, however, is critical to all these approaches and the choice of this baseline is prone to generic biases. For example, when sampling values uniformly from all locations, generic biases that are shared between fixation selection and feature distributions across all images yield an overestimation of prediction performance.[9,13] In simple designs, this issue can be alleviated by a "shuffle" baseline, sampling map values of one image at locations that were fixated in different images (baseline set). In experimental designs with multiple conditions (image modifications, tasks, etc.), however, it is not clear *a priori* which images should constitute the baseline set, and recent suggestions involve using a generic baseline across all stimuli, tasks, and conditions.[35] In the present approach, the issue of choosing an appropriate baseline is overcome by reversing the traditional logic of fixation prediction: rather than evaluating differences in scene statistics at fixated and control locations, we distinguish between fixated and nonfixated scene regions and directly describe the relationship between image features and fixation probability. As a consequence, generic biases, such as the central bias, can be treated naturally as a predictor akin to image features.

Advantages of GLMM

Binary response variables are oftentimes analyzed using subject and item ANOVAs (F1 and F2) over proportions or percentages. In the context of scene perception, the proportion of fixations directed to regions of interests has been analyzed in that way (for example, see Ref. 45). However, applying ANOVA to fixation probabilities is associated with a number of serious problems, which can be overcome by using GLMM.[30] We would like to highlight one specific advantage of using GLMM in the present context. When using pictures of real-world scenes as stimuli, it is conceivable that the extent to which image features influence where people look may depend on the selected images. We took this into account by including by-item random slopes for each image feature. Indeed, we consistently found that including by-item random slopes significantly improved the model fit. Images also varied considerably in their intercept,[b] representing the overall fixation probability. A larger by-item intercept means that more scene patches were fixated, and this is of course associated with a smaller central fixation bias (by-item random slope for central-bias predictor). As a general observation, item variances were much larger than subject variances (Table 1).

Nonmonotonic relationships

Unlike previous approaches, the present approach explicitly considers fixation probability as the dependent variable and can deal with nonmonotonic relationships between fixation probability and feature values. Nonmonotonic functions can, for example, occur when medium levels of a feature are preferred relative to either extreme (or vice versa). For the case of luminance, our analyses revealed a negative quadratic relationship between

[b] By-item random effects describe items' deviations from the fixed-effect parameters. Thus, the intercept for a given image is obtained as the sum of the fixed-effect estimate for the intercept and the conditional mode of the random effect for the intercept.

luminance and fixation probability. The negative quadratic coefficient in the luminance-only GLMM suggested that fixation probability is maximal for image patches with medium-luminance, and is considerably reduced for both very dark and very bright scene regions. This result qualifies the previously reported finding that observers do not actively fixate luminance extremes.[10]

Feature dependencies

When considered in isolation, the features selected for our analyses have been found to predict where people look in natural scenes (with the exception of synergistic image segmentation, which has not been previously assessed).[7,8,20] The chosen set of features exemplifies a major issue in identifying associations between features and fixation: features are not independent.

A key strength of the present approach is its explicit handling of feature dependencies (Fig. 5). The generalized linear mixed modeling approach used here allows for assessing each feature's unique contribution to fixation selection, and its relative importance. The LumCon model took the quadratic dependency between luminance contrast and luminance in our stimulus set into account (Fig. 4A) and demonstrated that, once contrast was added to the luminance model, the nonmonotonic effect of luminance was still significant but reduced in size (Fig. 4B). However, neither luminance nor contrast had an independent effect on fixation probability once the remaining image features edge density, visual clutter, and the number of local segments were included in the model (Table 1). Additional analyses substantiated that this pattern of results is in principle agreement with findings from a Bayesian model analysis, suggesting that high-frequency edges are more predictive of fixation behavior than luminance or contrast.[11] The results from the full model suggest that, when correlations between image features were accounted for, edges provided the best ability to predict fixation selection, closely followed by the synergistic image-segmentation predictor.[21] According to the z-statistics, the independent effect of the feature-congestion measure of visual clutter[18] was smaller in size (Table 1). Importantly, the effects of image features were above and beyond a general preference for fixating the center of the image. With regard to the relative contributions of the different sources of variability in the data, the central bias was the strongest predictor.

Feature dependencies are unavoidable for at least two reasons. First, dependencies can be a consequence of the hierarchical definition of features: luminance is a pixel property, contrast is its difference measure, edge density exploits gradient information and is thus related to contrast, and synergistic image segmentation incorporates edge information. Second, feature dependencies oftentimes arise from structural properties of natural scenes. For example, there is a negative quadratic relationship between luminance and contrast (Fig. 4A) because brightness and darkness extremes tend to be associated with uniform surfaces (sky, walls, etc.) rather than being distributed as salt and pepper sprinkles. The central-bias predictor is correlated with local image features because of the photographer's bias in scene composition. Object boundaries are likely to co-occur with luminance edges.

The present results demonstrate that it would be inadequate to consider the higher-level features only. Notably, while the mid-level feature of edge density almost completely explains away any influence of the low-level features,[11] the higher-level features do *not* render edge density redundant. These results suggest that edges can attract attention even if they are not object (or in our case segment) boundaries. We conclude that any method that seeks to relate image features to fixation selection needs to take feature dependencies into account, or it will eventually fall short.

Features and objects

In research on both attention and memory, the number of objects or items in a display (i.e., set size) is an index of cognitive load. In a natural scene, however, an "object" is a perceptual and hierarchical construct that can change depending on the task, context, and mindset of the observer. This renders quantifying the number of items in a natural scene difficult, and some even argue that the problem of object segmentation is ill-conceived.[46] There are two main strategies for experimentally relating fixation patterns to objects in a scene. First, one can ask observers to label the scene and relate this ground truth to fixation patterns.[24,25,39] Second, one can delineate preattentive objects based on features that can be computed from the stimulus.[46–48] The resulting entities, which can potentially gain objecthood, are frequently identified as proto-objects.[49] The conditions under which such proto-objects behave like real objects for fixation selection is an interesting issue in itself.[24,46,47] In this study, we restricted ourselves to features that can be computed from the current stimulus and refrain from relating them to any semantic scene content or context. However, we included features that have been proposed as proxies or surrogates of objects in the literature. Our data show that all mid-level and higher-level features, though tightly related by design, have independent effects on fixation selection. We may speculate that they indeed capture partially complementary aspects of object presence: the feature congestion measure has been explicitly introduced as a proxy for set size,[18] the synergistic segmentation measure[19] counts the number of individuated parts, and edge density counts the number of edges in a local region. Therefore, the synergistic segmentation may approximate the number of items weighted by the number of their parts, and edge density may approximate the number of items weighted by the complexity of their boundaries. Hence, one possible interpretation of the present results is that the three features capture complementary aspects of objecthood and that their independent effects on fixation probability result from this complementarity. It should be noted, however, that the present data do not speak directly to what the relevant entities for guiding attention in scenes are (objects or features).

Correlation versus causality

The profound effect of feature dependencies also points to a deeper issue for any natural-scene analysis. Even if interfeature dependencies are modeled, observational results on natural scenes will remain correlational. There could be an unconsidered feature with mutual dependencies, and correlation of fixation probability to a feature does not imply the feature's causal effect. Even if features are experimentally manipulated to test for causality,[15] care must be taken to not introduce new feature dependencies.[50] Unlike earlier modeling attempts, however, the proposed statistical approach will allow quantitative predictions about the effects that targeted manipulation of stimulus features should have on fixation probability. Therefore, it will be able to inform experimental designs that manipulate the correlation structure among low-level features and between low-level features and high-level content. By fostering the interplay between observation on large sets of natural scene fixation data, and experimental designs that manipulate scene statistics, the proposed approach has the potential to pave the way from correlations between fixation probability and feature values to the causal effect of features.

Future research

In this paper, we have considered the main effects of a set of five image features on fixation selection in a scene memorization task. Future research may examine additional image features as well as their interactions, and interactions with viewing task. Previous research suggests that color is not a strong correlate of fixation location.[9] However, since the features used here differ with respect to how they capture color information, additional research is needed to fully explore the role of color. The approach taken here was to assess the relative contribution of various image features to fixation selection. Our method can thus be used to single out features that may be combined into a computational model of salience. However, the proposed method may also be used to evaluate the performance of salience models, simply by including a measure of local salience—based on a salience map—as a predictor in the GLMM. In conclusion, the present approach might be a good way forward in understanding scene exploration, since it allows for assessing the interplay of image features relevant for fixation selection in scenes in a powerful and flexible manner.

Acknowledgment

This paper reports partial analyses of data collected under a grant from the Economic and Social Research Council (ESRC) of the UK (RES-062-23-1092).

Conflicts of interest

The authors declare no conflicts of interest.

Supporting Information

Additional supporting information may be found in the online version of this article.

Figure S1. Average feature maps for grid cells across images.

Figure S2. Matrix of pairwise scatter plots given the image features considered in the study. Panels above the diagonal plot the actual data points, which are mean feature values for a given scene and analysis grid cell (135 scenes × 48 grid cells = 6,480 data points in each panel). Hexagonal binning was used to avoid overplotting; the *x*–*y* plane is tiled using hexagons which are then colored to indicate the number of points that fall inside (see Ref. 1 in supplementary online file); darker colors indicate a larger number of data points. Each feature combination additionally shows a regression line (in green) and a polynomial spline (red) that were fit to the data. Panels below the diagonal display the linear correlation coefficients (scaled according to their absolute size) and corresponding *P* values. The diagonal panels show the distributions of feature values.

Table S1. Generalized linear mixed models fitting fixation probability for a scene memorization task for a fine 16 × 12 grid and a coarse 4 × 3 grid: means, standard errors, and *z*-values of fixed effects on fixation probability; variances of the random effects.

References

1. Kasper, R.W. *et al.* 2015. Multimodal neuroimaging evidence linking memory and attention systems during visual search cued by context. *Ann. N. Y. Acad. Sci.* **1339:** 176–189.
2. Souza, A. S., L. Rerko & K. Oberauer. 2015. Refreshing memory traces: thinking of an item improves retrieval from visual working memory. *Ann. N. Y. Acad. Sci.* **1339:** 20–31.
3. Deubel, H. & W.X. Schneider. 1996. Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Res.* **36:** 1827–1837.
4. Borji, A. & L. Itti. 2013. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35:** 185–207.
5. Tatler, B. W. *et al.* 2011. Eye guidance in natural vision: reinterpreting salience. *J. Vis.* **11**(5):5**:** 1–23.
6. Schütz, A. C., D. I. Braun & K. R. Gegenfurtner. 2011. Eye movements and perception: a selective review. *J. Vis.* **11**(5):9**:** 1–30.
7. Mannan, S. K., K. H. Ruddock & D. S. Wooding. 1996. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spat. Vis.* **10:** 165–188.
8. Reinagel, P. & A. M. Zador. 1999. Natural scene statistics at the centre of gaze. *Network: Comput. Neural Syst.* **10:** 341–350.
9. Tatler, B. W., R. J. Baddeley & I. D. Gilchrist. 2005. Visual correlates of fixation selection: effects of scale and time. *Vis. Res.* **45:** 643–659.
10. Vincent, B. T. *et al.* 2009. Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Vis. Cogn.* **17:** 856–879.
11. Baddeley, R. J. & B. W. Tatler. 2006. High frequency edges (but not contrast) predict where we fixate: a Bayesian system identification analysis. *Vis. Res.* **46:** 2824–2833.
12. Parkhurst, D. J. & E. Niebur. 2003. Scene content selected by active vision. *Spat. Vis.* **16:** 125–154.
13. Tatler, B. W. 2007. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* **7**(14):4**:** 1–17.
14. Judd, T. *et al.* 2009. Learning to predict where humans look. In *Proceedings of the IEEE 12th International Conference on Computer Vision*: 2106–2113. Kyoto, Japan.
15. Einhäuser, W. & P. König. 2003. Does luminance-contrast contribute to a saliency map for overt visual attention? *Eur. J. Neurosci.* **17:** 1089–1097.
16. Krieger, G. *et al.* 2000. Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spat. Vis.* **13:** 201–214.
17. Mannan, S. K., K. H. Ruddock & D. S. Wooding. 1997. Fixation patterns made during brief examination of two-dimensional images. *Perception* **26:** 1059–1072.
18. Rosenholtz, R., Y. Z. Li & L. Nakano. 2007. Measuring visual clutter. *J. Vis.* **7**(2):17**:** 1–22.
19. Christoudias, C. M., B. Georgescu & P. Meer. 2002. Synergism in low level vision. In *Proceedings of the 16th International Conference on Pattern Recognition*, Vol. 17: 150–155. Quebec, Canada.
20. Henderson, J. M., M. Chanceaux & T. J. Smith. 2009. The influence of clutter on real-world scene search: evidence from search efficiency and eye movements. *J. Vis.* **9**(1):32**:** 1–8.
21. Comaniciu, D. & P. Meer. 2002. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24:** 603–619.
22. Meer, P. & B. Georgescu. 2001. Edge detection with embedded confidence. *IEEE Trans. Pattern Anal. Mach. Intell.* **23:** 1351–1365.
23. Franz, M. O. & B. Schölkopf. 2005. Implicit Wiener series for higher-order image analysis. In *Advances in Neural Information Processing Systems.* Vol. **17**. L. K. Saul, Y. Weiss & L. Bottou, Eds.: 465–472. Cambridge, MA: MIT Press.
24. Nuthmann, A. & J. M. Henderson. 2010. Object-based attentional selection in scene viewing. *J. Vis.* **10**(8):20**:** 1–19.
25. Pajak, M. & A. Nuthmann. 2013. Object-based saccadic selection during scene perception: evidence from viewing position effects. *J. Vis.* **13**(5):2**:** 1–21.
26. Moulden, B., F. Kingdom & L. F. Gatley. 1990. The standard deviation of luminance as a metric for contrast in random-dot images. *Perception* **19:** 79–101.
27. Cohen, J. & P. Cohen. 1975. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Hillsdale, NJ: Erlbaum.
28. Barr, D. J. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J. Mem. Lang.* **59:** 457–474.
29. Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: Cambridge University Press.

30. Jaeger, T. F. 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* **59:** 434–446.

31. Bates, D. M. *et al.* 2014. lme4: Linear mixed-effects models using Eigen and S4 (version 1.1–7). http://CRAN.Rproject.org/package=lme4.

32. Hohenstein, S. & R. Kliegl. 2014. Semantic preview benefit during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* **40:** 166–190.

33. Mante, V. *et al.* 2005. Independence of luminance and contrast in natural scenes and in the early visual system. *Nat. Neurosci.* **8:** 1690–1697.

34. Barr, D. J. *et al.* 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* **68:** 255–278.

35. Clarke, A. D. F. & B. W. Tatler. 2014. Deriving an appropriate baseline for describing fixation behaviour. *Vis. Res.* **102:** 41–51.

36. Kliegl, R., *et al.* 2004. Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* **16:** 262–284.

37. Field, D. J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A Vis.* **4:** 2379–2394.

38. Hancock, P. J. B., R. J. Baddeley & L. S. Smith. 1992. The principal components of natural images. *Network: Comput. Neural Syst.* **3:** 61–70.

39. Einhäuser, W., M. Spain & P. Perona. 2008. Objects predict fixations better than early saliency. *J. Vis.* **8**(14):**18:** 1–26.

40. Stoll, J. *et al.* 2015. Overt attention in natural scenes: objects dominate features. *Vis. Res.* **107:** 36–48.

41. Itti, L. & C. Koch. 2001. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2:** 194–203.

42. Wilming, N. *et al.* 2011. Measures and limits of models of fixation selection. *PLoS ONE* **6:** e24038.

43. Parkhurst, D., K. Law & E. Niebur. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vis. Res.* **42:** 107–123.

44. Peters, R. J. *et al.* 2005. Components of bottom-up gaze allocation in natural images. *Vis. Res.* **45:** 2397–2416.

45. Smith, T. J. & P. K. Mital. 2013. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *J. Vis.* **13**(8):**16:** 1–24.

46. Yu, C.-P., D. Samaras & G. J. Zelinsky. 2014. Modeling visual clutter perception using proto-object segmentation. *J. Vis.* **14**(7):**4:** 1–16.

47. Russell, A. F. *et al.* 2014. A model of proto-object based saliency. *Vision Res.* **94:** 1–15.

48. Walther, D. & C. Koch. 2006. Modeling attention to salient proto-objects. *Neural Net.* **19:** 1395–1407.

49. Rensink, R. A. 2000. The dynamic representation of scenes. *Vis. Cogn.* **7:** 17–42.

50. Parkhurst, D. J. & E. Niebur. 2004. Texture contrast attracts overt visual attention in natural scenes. *Eur. J. Neurosci.* **19:** 783–789.

51. Agresti, A. & B. A. Coull. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. *Am. Stat.* **52:** 119–126.