Attention, Perception, & Psychophysics
2009, 71 (6), 1337-1352
doi:10.3758/APP.71.6.1337

Saliency on a natural scene background: Effects of color and luminance contrast add linearly

Sonja Engmann
University of Osnabrück, Osnabrück, Germany
and University of Montreal, Montreal, Quebec, Canada

Bernard M. 't Hart
Philipps University Marburg, Marburg, Germany

Thomas Sieren, Selim Onat, and Peter König
University of Osnabrück, Osnabrück, Germany

and

Wolfgang Einhäuser
Philipps University Marburg, Marburg, Germany

In natural vision, shifts in spatial attention are associated with shifts of gaze. Computational models of such *overt* attention typically use the concept of a *saliency map*: Normalized maps of center–surround differences are computed for individual stimulus features and added linearly to obtain the saliency map. Although the predictions of such models correlate with fixated locations better than chance, their mechanistic assumptions are less well investigated. Here, we tested one key assumption: Do the effects of different features add linearly or according to a max-type of interaction? We measured the eye position of observers viewing natural stimuli whose luminance contrast and/or color contrast (saturation) increased gradually toward one side. We found that these *feature gradients* biased fixations toward regions of high contrasts. When two contrast gradients (color and luminance) were superimposed, linear summation of their individual effects predicted their combined effect. This demonstrated that the interaction of color and luminance contrast with respect to human overt attention is—irrespective of the precise model—consistent with the assumption of linearity, but not with a max-type interaction of these features.

While inspecting complex natural scenes, human observers sequentially allocate attention to subsets of the stimulus (James, 1890). Under natural conditions, shifts in attention are typically associated with shifts of gaze (Rizzolatti, Raggio, Dascola, & Umiltà, 1987). Several factors guide this *overt* attention (Buswell, 1935; Yarbus, 1967), such as the task, the observer's experience, and the features of the stimulus. Models of the latter, *bottom-up* factors are often based on the concept of a so-called *saliency map* (Koch & Ullman, 1985): Various feature channels (luminance, color, orientation, etc.) are analyzed independently, local center–surround filters yield maps of differences (*contrasts*) in these features, and these maps are added up. Following the saliency map literature, such maps in a single feature are referred to as *conspicuity* maps. These conspicuity maps are then added linearly across features to obtain the saliency map, which represents the likelihood that a location will be attended. Various studies have demonstrated that implementations of this model predict human fixations in natural scenes at levels above chance (Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005; Tatler, Baddeley, & Gilchrist, 2005). In addition, luminance contrast (LC) is significantly elevated at fixation points (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Mannan, Ruddock, & Wooding, 1997; Reinagel & Zador, 1999). This correlative effect of contrast depends, however, on spatial frequency (Mannan et al., 1997; Tatler et al., 2005) and acts mostly indirectly through correlations with higher order scene structure (Einhäuser & König, 2003), which may include texture contrast (Parkhurst & Niebur, 2004), edge density (Baddeley & Tatler, 2006), or objects (Einhäuser, Spain, & Perona, 2008; Elazary & Itti, 2008) and faces (Cerf, Harel, Einhäuser, & Koch, 2008). In sum, the predictions of saliency map models can correlate with the actual fixations of human observers freely viewing natural scenes under laboratory conditions (Parkhurst et al., 2002; Peters et al., 2005). Such correlation is, however, absent under some conditions (e.g., during search; Einhäuser, Rutishauser, & Koch, 2008; Henderson, Brockmole,

W. Einhäuser, wet@physik.uni-marburg.de

Castelhano, & Mack, 2007). More and more evidence has accumulated that the saliency map's fixation prediction is mostly indirect, which undermines the causal and mechanistic implications of the model. In spite of this absence of a direct causal effect of low-level features on fixation locations, understanding the indirect correlative link (through objects or another higher order structure) will nonetheless benefit from knowledge as to how low-level features interact. Independently of existing models, such data will constrain future approaches toward a better mechanistic understanding of the neural basis of attention.

Despite a large body of data on the neural representation of saliency (Gottlieb, Kusunoki, & Goldberg, 1998; Horwitz & Newsome, 1999; Kustov & Robinson, 1996; Mazer & Gallant, 2003; McPeek & Keller, 2002; Posner & Petersen, 1990; Robinson & Petersen, 1992; Thompson, Bichot, & Schall, 1997), the mechanistic principles underlying its computation are less well understood. Koch and Ullman's (1985) model was founded on neural principles but did not make any explicit reference to the nature of interactions between feature channels. In contrast, most later saliency map implementations (Itti, 2005; Itti & Koch, 2000; Peters et al., 2005) made the critical assumption that feature effects added linearly. First, the conspicuity maps for each feature are linearly summed, and second, possible dependencies between features are neglected when obtaining the final saliency map. In addition, most models of visual attention that are *not* based on the saliency map still implicitly share the assumption of linearity (Wolfe, Butcher, Lee, & Hyle, 2003). Several studies have tested this assumption, using well-controlled, albeit artificial, stimuli. Using grids of bars in a matching task, Nothdurft (2000) found that different features are additive, although their interaction may be sublinear. Along these lines, for the features of color and orientation, Li (2002) contradicted the assumption of linearity and, instead, proposed that the overall saliency of an item is defined by the most salient feature alone. This implies a maximum operation, rather than a linear summation across features, to compute saliency. Recently, research from the same lab has suggested that this maximum operation might also apply to human overt attention in natural scenes (Lewis & Zhaoping, 2005) and has suggested a computation of saliency as early in the visual hierarchy as V1. In contrast, Navalpakkam and Itti (2005) argued that linear summation is more compatible with performance in conjunction search experiments. Complementary to the question of under which conditions low-level features influence fixations at all, it has remained open how the effects of different features interact. Irrespective of whether the features' effects are causal or correlative, the answer will constrain models of attention.

In addition to linearity, the second major assumption of most saliency models is the independence of different feature channels. In a discrimination task on grating stimuli, Morrone, Denti, and Spinelli (2002) found that the features of color and luminance recruited independent attention channels. However, the extent to which such results can be transferred to natural stimuli—where higher order dependencies between features not only exist, but also are exploited by the visual system (Golz & MacLeod, 2002)—remains to be investigated. When it comes to natural scenes, stimulus features are not independent but highly correlated. In the context of overt attention, Baddeley and Tatler (2006) showed that conditional on edge density, other feature maps have little predictive power; that is, one feature can "explain away" the effect of others. Consequently, when attention in natural scenes is measured directly, such stimulus-inherent correlations need to be considered.

Here, we combined the usage of natural scenes with modifications that were independent along the two stimulus dimensions under investigation (color and luminance). We adopted a previously proposed paradigm (Einhäuser, Rutishauser, et al., 2006) to bias attention by increasing contrast toward one side of the stimulus (*feature gradients*). We compared effects on fixated locations of gradients in color contrast (CC), which was modulated by varying saturation, and in LC to the effect of the feature gradients applied simultaneously. This allowed us to test directly how well a linear interaction of CC and LC would predict their combined effect against a natural scene background.

METHOD

Participants

Eight students at the Philipps University Marburg (3 of them female and 5 male; age, 20–27 years, $M = 22.3$) participated in the study. All the participants had normal or corrected-to-normal vision and normal color vision, as assessed by the Ishihara 16-plate color blindness test. They were naive as to the purpose of the study and had not previously viewed the stimuli used. All the procedures conformed to national and institutional guidelines for experiments on human observers and to the Declaration of Helsinki. All the participants gave informed written consent for participation in this study and received paid compensation.

Experimental Setup

The experiments were conducted in a dark room with negligible ambient light levels. The stimuli were presented using a 19.7-in. EIZO FlexScan F77S CRT monitor located at an 85-cm distance from the participant, and the stimulus subtended an angle of 26º × 18º. The display resolution was set to 1,280 × 1,024 pixels, and its refresh rate to 100 Hz. The monitor was characterized (*calibrated*) using a PR-650 spectrometer (Photo Research, Chatsworth, CA) and, for low luminance values, an S370 photometer (UDT Instruments, San Diego, CA). Gun CIE coordinates of the monitor were at $x = 0.610$, $y = 0.339$ (red), $x = 0.282$, $y = 0.601$ (green), and $x = 0.151$, $y = 0.065$ (blue); the maximum luminance was at 36.9 cd/m^2; and the luminance of the dark screen (black) was at 0.001 cd/m^2.

During the experiment, the observers' eye position was recorded at 2000 Hz, using an infrared, noninvasive Eyelink-2000 eyetracking system (SR Research Ltd., Mississauga, ON, Canada). Standard procedures, as recommended by the manufacturer, were used to calibrate the eyetracker and to validate the eye position. In brief, 13 fixation points were presented before each experimental block in order to compute the mapping from eyetracker signal to screen coordinates. The calibration was then verified with a similar display and was 0.4º root mean square on average and never larger than 1º. Before each trial, the observers were asked to fixate a fixation point in the center of the screen for at least 300 msec. If they failed to do so within 5 sec, the eyetracker was recalibrated.

All stimulus presentation and eye position recording was programmed in MATLAB (MathWorks, Natick, MA), using its psychophysics and eyelink toolbox extensions (Brainard, 1997; Cornelis-

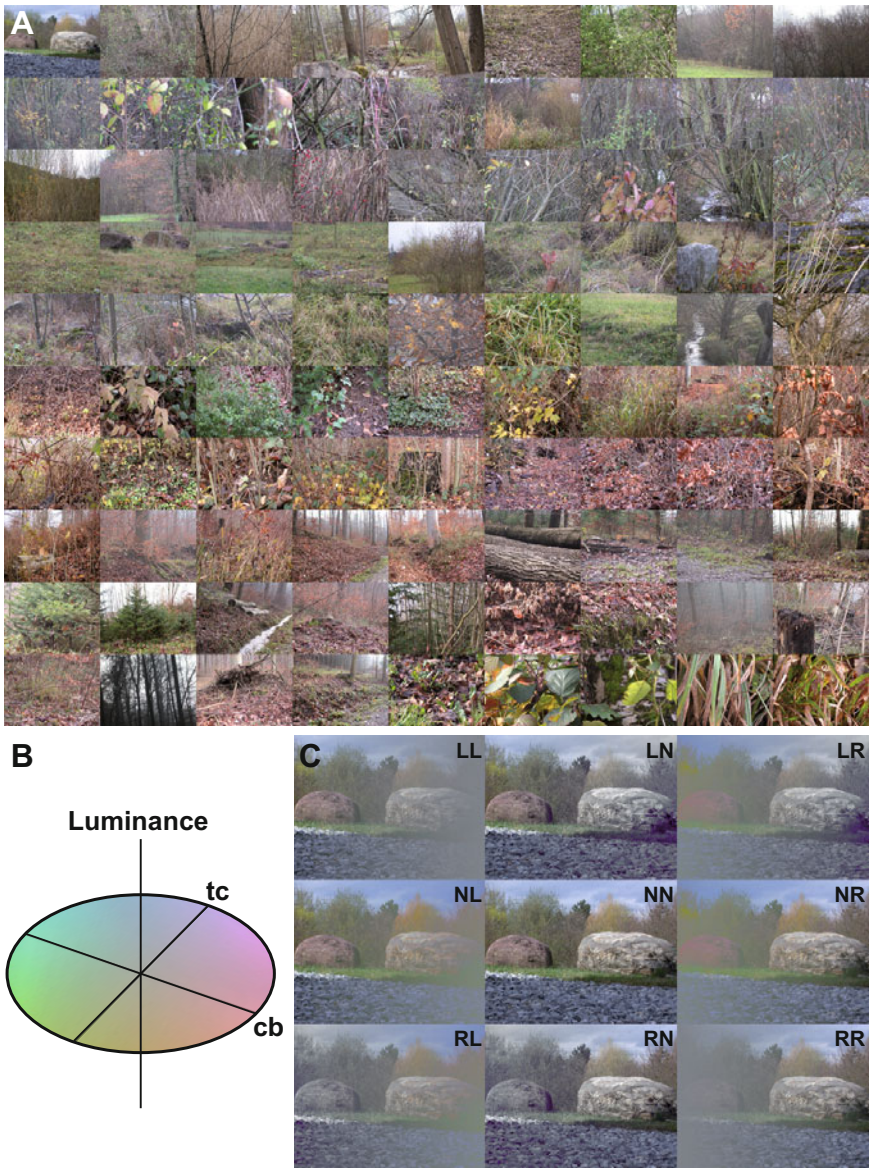

Figure 1. Stimuli. (A) The 90 stimuli from the Zürich Natural Image database used for the experiment. (B) Schematic representation of the DKL color space. (C) *Columns*: Modification of luminance contrast; increase to left, no increase, increase to right. *Rows*: Modification of color contrast; increase to left, no increase, increase to right. Letters denote condition abbreviation (gradient increasing to left [L] or right [R] or neutral [N]; first letter, color; second, luminance contrast). tc, tritanopic confusion; cb, contrast blue.

sen, Peters, & Palmer, 2002, psychtoolbox.org; Pelli, 1997). The data were preprocessed in Python 2.5 (www.python.org), and statistical analysis was performed in R 2.5.1 (www.r-project.org).

Stimulus Database

All the stimuli were based on a set of 90 photographs of natural scenes selected from the Zurich Natural Image database (Einhäuser, Kruse, Hoffmann, & König, 2006), which are available from the authors at www.klab.caltech.edu/~wet/ZurichNatDB.tar.gz. The images depict natural outdoor scenes, which only rarely contain isolated nameable objects or man-made artifacts (Figure 1A). The images were captured using a digital camera (3.3 Mega pixel color mosaic CCD, Nikon Coolpix 995, Tokyo, Japan) with high-quality settings. The stimuli were stored at a resolution of 2,048 × 1,536 pixels and a color depth of 24 bits in RGB format. To fit the screen resolution, images were down-sampled to 1,280 × 960 pixels, using bicubic interpolation in MATLAB, and were presented at the center of the 1,280 × 1,024 pixel screen. Without clearly nameable objects in the images, the influence of higher order structures (objects) was reduced, while at the same time, a realistic, naturalistic "background" was preserved, on which saliency manipulations could be superimposed.

Color Space

Stimuli were characterized and modified in the DKL color space (Derrington, Krauskopf, & Lennie, 1984; see Figure 1B). This space is defined physiologically, using the relative excitations of the three types of retinal cones. It is spanned by the orthogonal axes of *luminance*, *constant blue* (cb; the difference between L and M cone exci-

tations), and *tritanopic confusion* (tc; L + M − S cone excitations). Hue in DKL space is given by the azimuth, luminance by the respective axis, and saturation by the projection on an isoluminant plane.

In DKL space, we defined LC as variation along the space's luminance axis. CC—as used in saliency map models—is inspired by the excitation of color-opponent cells in the retina and thalamus. Hence, it scales linearly with saturation, and we modified CC by varying saturation. The mapping from DKL space used the known parameters of the screen's guns—in particular, correcting for their nonlinearities (*gamma*). Since the camera parameters were unknown, they were assumed to be the inverse of the screen. This guaranteed that an unmodified stimulus looked natural and all the stimuli fitted within the gamut of the screen.

Stimulus Modification: Feature Gradients

To modify the stimulus features of interest (LC and CC) without introducing novel local image structure, we adapted the feature gradient technique introduced in Einhäuser, Rutishauser, et al. (2006). Here, images were first converted into DKL color space. To modify luminance contrast, we first subtracted the mean image luminance $<I^0>$ from the luminance values $I^0(x,y)$ of the original image. We then multiplied the luminance with a value depending on the horizontal position (gradient). For contrast increase to the right ("R"), this factor ranged linearly from 0 on the left to 1 on the right, and the converse held for contrast increase to the left ("L"). Finally, the original mean value was added:

$$\text{Modification "R"}: I(x,y) = x/w[I^0(x,y) - <I^0>] + <I^0>$$

$$\text{Modification "L"}: I(x,y) = (1 - x/w)\,[I^0(x,y) - <I^0>] + <I^0>,$$

where w denotes the image width (w = 1,280 pixels). Intuitively, the low end of the gradient reduced the contrast to 0, since it clamped all luminance values to the mean image luminance [for an "R" gradient, $I(x{=}0,y) = <I^0>$; for an "L" gradient, $I(x{=}w,y) = <I^0>$]. At the other side of the image (the gradient's high end), the contrast remained unaffected [$I(x{=}w,y) = I^0(x{=}w,y)$ or $I(x{=}0,y) = I^0(x{=}0,y)$, for "R" and "L" gradients, respectively]. Both extremes are most easily exemplified by an image consisting only of an equal number of black and white pixels. In Appendix A, we provide a detailed analysis as to how the gradient definition relates to common definitions of luminance contrast. As a consequence of the orthogonality of the DKL space, this modification did not affect physical color at any point (neither hue nor saturation).

To modify color contrast, we similarly subtracted the means along the tc and cb axes, multiplied the result by the gradient from 0 to 1 ("R") or 1 to 0 ("L"), and shifted back to the original mean:

Modification "R":

$$T(x,y) = x/w\,[T^0(x,y) - <T^0>] + <T^0>$$

$$C(x,y) = x/w\,[C^0(x,y) - <C^0>] + <C^0>$$

Modification "L":

$$T(x,y) = (1 - x/w)\,[T^0(x,y) - <T^0>] + <T^0>$$

$$C(x,y) = (1 - x/w)\,[C^0(x,y) - <C^0>] + <C^0>,$$

where C and T denote the values along the cb and tc axes, respectively, superscript 0 the original image, and $<.>$ the image mean as above. This varied the saturation of each pixel from 0 to its original value across the image. Intuitively, the usage of saturation as a proxy for CC can be understood by considering an isoluminant red–green grating, which, at 0 saturation, would be a mere gray patch (0 CC) and would take maximum color contrast whenever saturation is at 100%. To formalize this, we demonstrate in Appendix A that this modification affected color conspicuity in the expected way.

Taking advantage of the orthogonality of DKL space, both gradients could be combined without interaction on the physical stimulus. The modified stimuli were converted back to RGB space, using the screen gun's specifications.

Notation for modifications. As a shorthand notation, we will denote conditions by two-letter abbreviations, where the first characterizes the color modification, the second the luminance modification, with "L" implying a contrast increase to the left, "R" an increase to the right, and "N" no modification. For example, LN denotes a stimulus modified solely in color, with the gradient increasing to the left and no changes in luminance, and NN denotes an unmodified stimulus (Figure 1C). Where there is no risk of ambiguity, the same abbreviations will also be used to denote the corresponding effect sizes. We will refer to the conditions in which a single gradient is applied (LN, NL, RN, NR) as *single-feature conditions* and to the conditions in which two gradients are superimposed (LL, RR, LR, RL) as *dual-feature conditions*. For part of the analysis, we considered the effects of each modification relative to the modulation of eye position in the unmodified condition (NN). As a short-hand notation, we used brackets [.] to denote subtraction of NN (e.g., [LN] := LN − NN).

Paradigm

For all observers, each of the 90 images was presented in each of the nine conditions exactly once. The experiment was split into nine blocks of 90 trials. Trials were balanced such that, per block, each image appeared once and each condition 10 times. Since pilot experiments had demonstrated little change in effect after 2 sec, each stimulus was presented for 2 sec. A trial started with a fixation cue at the center of the screen. As soon as the participant's gaze was steady on this cue for at least 300 msec, stimulus presentation was triggered. The observers were instructed to "study the images carefully," be "free to move [their] eyes naturally," and "reduce head movements as much as possible." None of our previous studies, which used the same instruction of "studying images carefully," showed any evidence that this induced a top-down bias. To the contrary, eye movement patterns were indistinguishable from an explicit instruction of "free viewing" (Steinwender & König, 2007).

Data Analysis

Fixations. The main body of the analysis was based on periods of fixation, which accounted for 77.3% of the total data (see Appendix B for an analysis based on raw eye-position data). Fixations were defined by the default algorithm implemented in the Eyelink system as periods between saccades. Saccades were defined as movements that exceeded an acceleration threshold (9,500 deg/sec^2) and a velocity threshold (35 deg/sec). Although no explicit lower limit for the duration of a fixation was used, 96.1% of the fixations lasted longer than 100 msec. The initial central fixation for each stimulus originated from the fixation cue and was not used for any analysis.

Statistical analysis. Since, in 64.5% of the 2-sec trials, there were at least five fixations, but six or more fixations were reached only in 35.2%, we restricted fixation analysis to the first five. For each observer, the average horizontal coordinate of each of the first five fixations was calculated. A linear model ANOVA was performed with this dependent variable, using fixation number (1–5), CC condition (L, N, R), and LC condition (L, N, R) as factors. Linear model ANOVAs were also performed over the two sets of single-feature data in which the data with manipulations in the other feature were left out (using only the data from LN, NN, and RN or from NL, NN, and NR). To see how the effect of the single-feature manipulations developed over time, post hoc t tests were done on the average horizontal coordinate for each fixation, comparing LN with RN and NL with NR.

RESULTS

Number of Fixations

We recorded the eye movements of 8 observers while they were viewing natural scenes upon which a gradient in LC, CC, or both had been superimposed. For each of

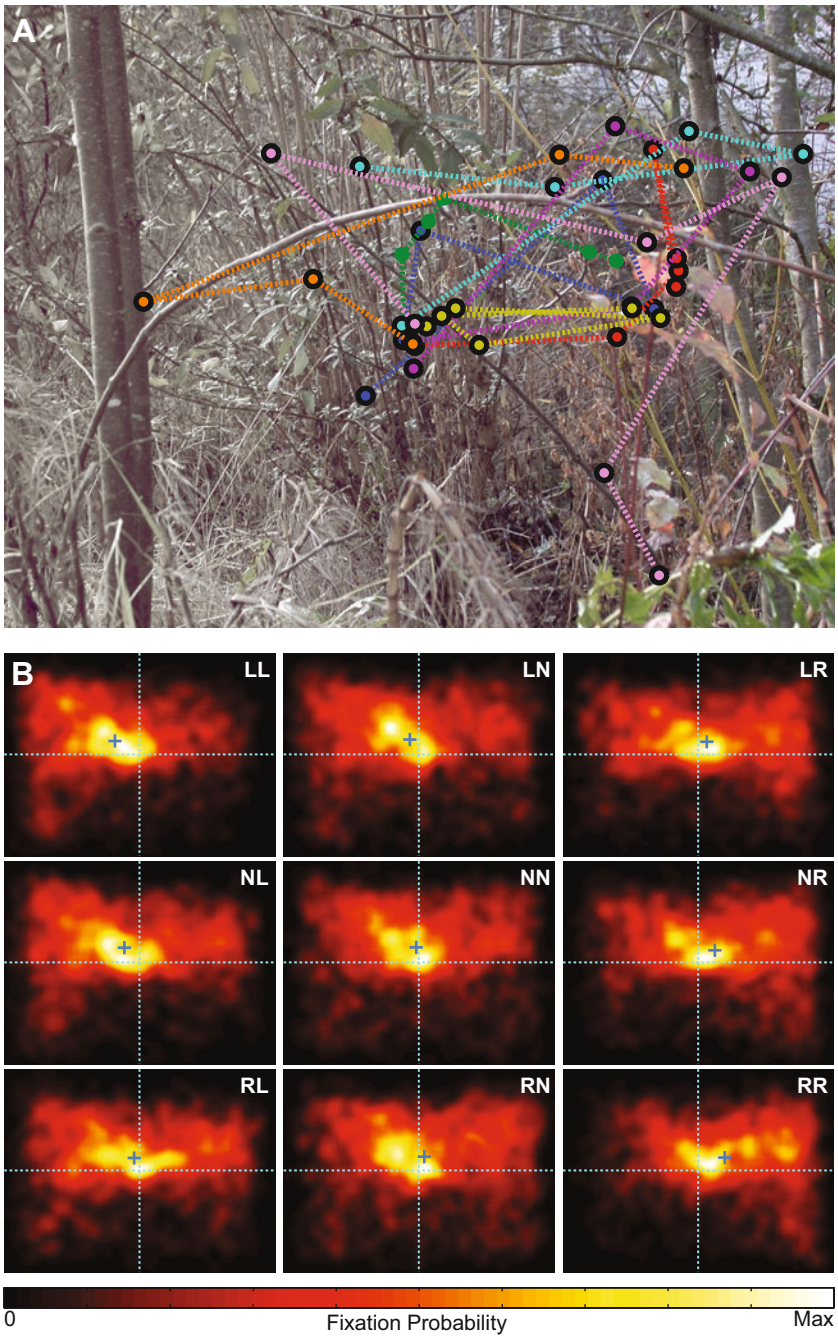

Figure 2. Effect of modifications. (A) Example stimulus with a color contrast increase to the right (RN). The fixations of all the observers are superimposed; color identifies the observer. (B) Average fixation maps (spatial distribution of fixated location) for each condition, sorted as in Figure 1C. For display, maps are smoothed with a 27-pixel-wide (0.5º at the center) Gaussian kernel; the extension of each map corresponds to the full image size. Dashed lines indicate midlines, cyan crosses the center-of-mass locations.

the nine conditions defined by the directions of those gradients, we recorded 90 trials for each observer. On average, the observers made 5.2 fixations on each stimulus. We did not find evidence that this number was dependent on the luminance condition, the color condition, or their interaction (all *p*s $>$ .07). Consequently, we could directly compare different conditions on the basis of fixation locations.

Fixation Maps

In the image of Figure 2A, CC increased to the right (the RN condition). In this example, the observers' fix-

ations exhibited a bias toward the right, the side of increased CC. To visualize this effect across all observers and images, we computed an average fixation map for each condition. That is, we computed the histograms of fixated locations and aggregated the histograms over all fixations (excluding the initial, central one), observers, and images (Figure 2B). In all the conditions, the center of mass of these maps was slightly (1.2º–1.5º) above the midline. That is, the horizontal gradient had little effect on vertical eye position. In contrast, the horizontal location depended on the condition. For unmodified images, there was a slight (0.3º) bias to the left. If both gradients pointed to the left (LL), however, the center of mass was shifted 2.5º to the left; if both gradients pointed to the right (RR), the shift was 2.7º to the right. For single-feature gradients (LN, RN, NL, NR), the center-of-mass shifts were smaller but always to the higher (color or luminance) contrast side (0.5º, 0.9º, 1.5º, and 1.7º, respectively). The incongruent gradients showed a slight bias toward the higher LC, consistent with the somewhat larger effect of this feature, as compared with color. This first qualitative and aggregate analysis of horizontal eye position was suggestive of a superposition between the effects of CC and LC gradients on horizontal eye position, on which the further quantitative analysis was based.

Overall Effect of Gradients

We performed a three-way ANOVA to characterize the dependence of average horizontal fixation location on fixation number (1–5), LC condition (L, N, R), and CC condition (L, N, R). Each factor had a significant effect [$F(4,315) = 33.3$, $F(2,315) = 78.0$, and $F(2,315) = 20.5$, respectively; all ps $< .0001$]. There were no two-way interactions between LC and CC [$F(4,315) = 0.18$, $p = .95$], between CC and fixation number [$F(8,315) = 0.38$, $p = .93$], or between LC and fixation number [$F(8,315) = 1.31$, $p = .24$]. There was no three-way interaction between all three factors [$F(16,315) = 0.063$, $p = 1$]. Hence, we could analyze the effects separately.

Single-Feature Conditions

Color contrast. First, we analyzed the effect of single-feature modifications: whether CC and LC gradients alone induced biases in fixated locations. To quantify this bias, we compared the average horizontal eye position at each fixation in the RN condition with that in the LN condition (Figure 3A). Taking the NN, RN, and LN conditions into account, there were main effects of color [$F(2,105) = 5.38$, $p = .006$] and of fixation number [$F(4,105) = 17.30$, $p < .0001$], but there was no interaction [$F(8,105) = 0.21$, $p = .99$]. Post hoc paired t tests comparing the effects of gradients to the left (LN) and gradients to the right (RN) by individual showed a significant effect for all the fixations tested [first fixation, $t(7) = 3.24$, $p = .01$; second, $t(7) = 6.61$, $p = .0003$; third, $t(7) = 3.75$, $p = .007$; fourth, $t(7) = 4.68$, $p = .002$; fifth, $t(7) = 3.00$, $p = .02$]. This demonstrated a robust and prolonged effect of the CC gradient on fixation location.

Luminance contrast. For the gradients in LC, we observed a pattern similar to that for CC modifications (Figure 3B). Considering the NL, NN, and NR conditions, there was a main effect of LC [$F(2,105) = 24.29$, $p < .0001$] and of fixation number [$F(4,105) = 9.04$, $p < .0001$] and no interaction [$F(8,105) = 0.50$, $p = .86$]. Post hoc paired t tests showed a significant effect starting at the first fixation [first, $t(7) = 2.49$, $p = .042$; second, $t(7) = 2.70$, $p = .03$; third, $t(7) = 2.99$, $p = .02$; fourth, $t(7) = 4.38$, $p = .003$; fifth, $t(7) = 5.76$, $p = .0007$]. This showed—consistent with our earlier results (Einhäuser, Rutishauser, et al., 2006)—that gradients in LC induced robust biases in fixated locations.

Normalized analysis. The NN condition showed a modulation with fixation number (Figure 3B). To measure the effects that gradients had on top of this general bias, we normalized horizontal fixation locations by subtracting the respective values of the NN condition. The normalized data showed the reported effects as even more pronounced, for both CC (Figure 3C) and LC (Figure 3D). In all cases and for all fixations, single-feature gradients biased the condition in the direction of higher (color or luminance) contrasts, relative to the general bias, which is revealed by the NN condition.

Average position. So far, we had analyzed the data separated by fixation number. The mean positions exhibited the biases in the direction consistent with the gradient (rightmost data points in each panel of Figure 3), whose significance had already been quantified by the aforementioned two-way-ANOVA main effects of CC and LC, respectively. Other averaging schemes (e.g., weighting fixations with their duration or using all data including periods of saccades) yielded the same result.

In sum, the single-feature gradients induced robust biases, especially relative to a neutral (NN) condition, which held for the average eye position but also for individual fixations.

Dual-Feature Conditions

In the dual-feature conditions, the interaction between the effects of CC and LC was examined. If the effects of LC and CC add linearly, there will be the following predictions as to how the effects of superimposed gradients can be computed from the single-feature gradients with a correction for the unmodified (NN) condition:

$$(1)\ [LL] \sim [LN] + [NL],$$

$$(2)\ [RR] \sim [RN] + [NR],$$

$$(3)\ [RL] \sim [RN] + [NL],$$

and

$$(4)\ [LR] \sim [LN] + [NR],$$

where the shorthand notation [.] for NN subtraction was used. In terms of raw data, these relations can be equivalently expressed by adding NN on each side as

$$(1')\ LL \sim LN + NL - NN,$$

$$(2')\ RR \sim RN + NR - NN,$$

$$(3')\ RL \sim RN + NL - NN,$$

and

$$(4')\ LR \sim LN + NR - NN.$$

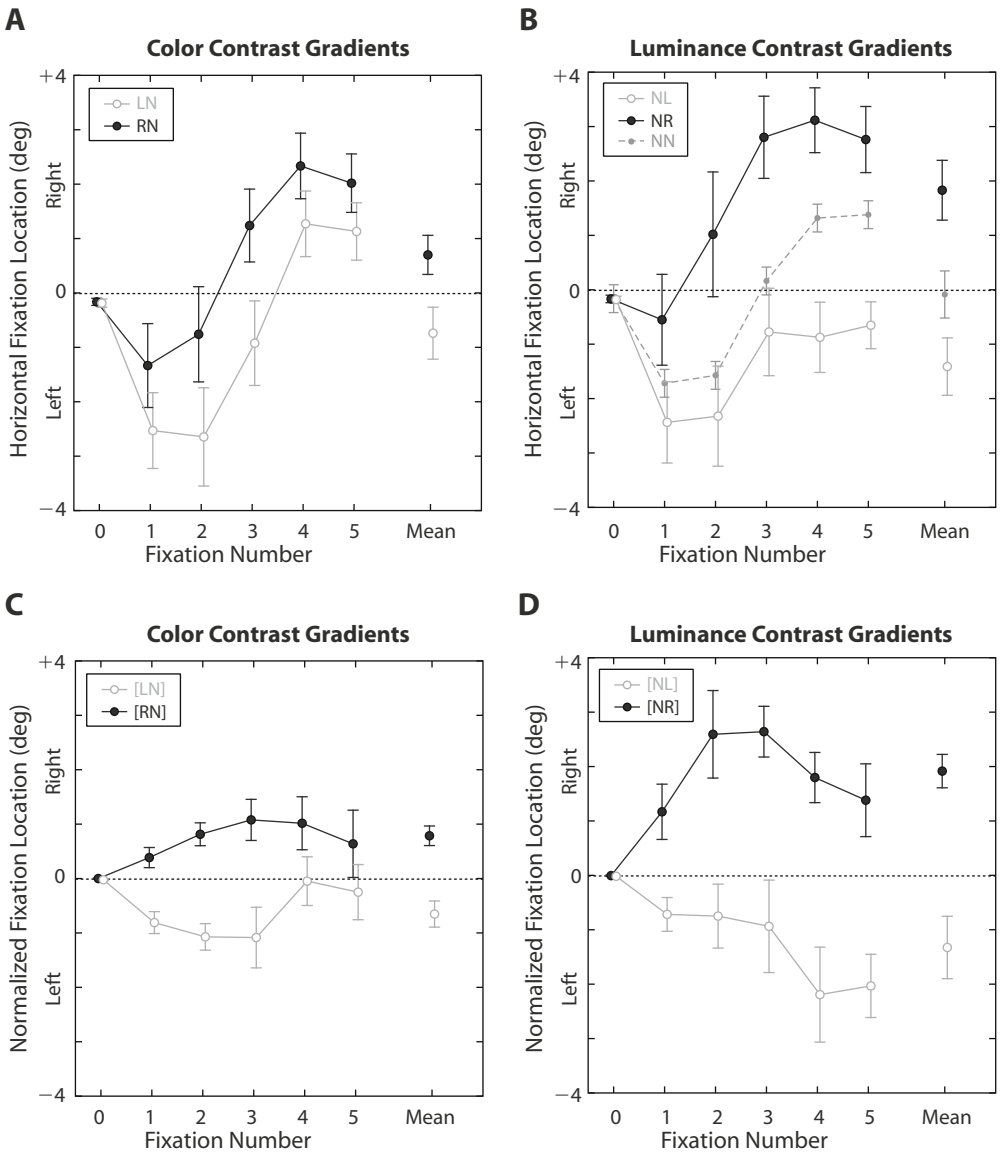

Figure 3. Single-feature gradients. (A) Effect of color contrast gradients. Mean ± *SEM* over participants of horizontal fixation location, positive values to the right, negative values to the left of screen center. *Black*, the RN condition; *gray*, the LN condition. The 0th (initial) fixation, which starts before stimulus onset is central by instruction, was not used for analysis. Note that all statistics are based on paired tests, and the standard errors of unnormalized locations include differences in general observer biases. Overlap in error bars thus does not contradict a significant effect in paired tests. (B) Effects of luminance contrast gradients. *Black*, NR; *gray solid*, NL; *gray dashed*, general bias without modification (NN) for comparison (omitted in panel A). (C) Normalized effect of color contrast gradients. *Black*, [RN] = RN − NN; *gray*, [LN] = LN − NN. (D) Normalized effect of luminance contrast gradients. *Black*, [NR] = NR − NN; *gray*, [NL] = NL − NN.

Linearity now predicts that the left-hand sides (*data*) are statistically indistinguishable from the right-hand sides (*model*). The difference between model and data was tested by means of a two-sided paired t test over observers, first considering the average effect over all images, but separated by fixation number. The right-hand sides of all the relations were indistinguishable from the respective left-hand sides for any of Fixations 2–5 ($p_{\min} = .25$; see Table 1, gray shaded rows). Furthermore, for Relations 1 and 4, this also held for the first fixation ($p = .81$ and $p = .55$, respectively). Hence, the dual-feature data were consistent with linear summation of single-feature effects, for both congruent (Figures 4A and 4C) and incongruent (Figures 4B and 4D) gradients.

When the average fixation location rather than individual fixations (rightmost data point in each panel of Figure 4) were considered, even the remaining deviations from a linear model vanished: For all the conditions, the linear model's prediction was indistinguishable from the corresponding data ($p_{\min} = .20$; see Table 1, rightmost column).

Table 1
Analysis of How Well the Linear "Model" From the Single-Feature Conditions Deviates From the "Data" Obtained in the Respective Dual-Feature Conditions

Data	Model	First Fixation	Second Fixation	Third Fixation	Fourth Fixation	Fifth Fixation	Mean
[LL]	[LN] + [NL]	**.81**	**.27**	**.53**	**.85**	**.51**	**.20**
	[LN]	**.07**	**.08**	.04	.003	.006	.004
	[NL]	.005	.002	.007	**.66**	**.25**	.0003
[RR]	[RN] + [NR]	.006	**.94**	**.57**	**.56**	**.27**	**.46**
	[RN]	.01	.009	.02	.01	.004	.006
	[NR]	.0007	**.16**	**.15**	.02	.046	.02
[RL]	[RN] + [NL]	.005	**.94**	**.55**	**.70**	**.71**	**.48**
	[RN]	**.75**	**.41**	**.16**	.001	.007	.03
	[NL]	.005	**.055**	**.056**	.001	.043	.001
[LR]	[LN] + [NR]	**.55**	**.25**	**.9997**	**.51**	**.71**	**.74**
	[LN]	.02	.02	.01	.02	.002	.005
	[NR]	.04	.0004	.02	**.23**	**.994**	.0003

Note—Each entry denotes the p value of a paired t test. Linearity predicts that there is no evidence for differences of data from model in the gray-shaded rows. Significant effects in the other rows (*control models*) show that we would have sufficient power to recognize a deviation if it occurred. Note that there were two equivalent ways of testing, using either the normalized or the raw positions. For example, the distance between [LL] and [LN] + [NL] is equivalent to the distance between LL and LN + NL − NN, as is directly seen by adding NN on each side of the latter relation.

To evaluate our statistical power, we tested the individual right-hand side summands as alternative models (white rows in Table 1). For these controls, the average fixation location was always statistically different from the left-hand sides. This indicates that we had sufficient statistical power to find a deviation of model from data, if there were any. For the analysis of individual fixations, the results were less clear, especially in the case of incongruent gradients. Nonetheless, with few exceptions, the linear model was in general more consistent with the dual-feature data than with any individual single-feature effect, even for individual fixations (Table 1). Hence, the compatibility between the linear summation model and the dual-feature data cannot be attributed to a lack of statistical power.

Alternative Model: Max Norm

So far, we have argued merely that a linear model was *consistent* with our data and that we would have had sufficient power to discriminate linearity from each feature alone. A maximum operation presents a frequently proposed alternative model. It predicts that the combined effect of two features corresponds to the larger individual effect. That is, the combined effect is predicted to have the magnitude of the larger individual effect and also to point in the direction of the effect with larger magnitude. Let a and b be the individual normalized effects (e.g., a = [LN], b = [NL]); then, the predicted combined normalized effect f (e.g., f prediction for [LL]) is given by

$$(5a)\ f(a,b) = \max(|a|,|b|)\ \mathrm{sign}(a) \text{ if } |a| > |b|$$

and

$$(5b)\ f(a,b) = \max(|a|,|b|)\ \mathrm{sign}(b) \text{ if } |b| > |a|,$$

where $\mathrm{sign}(x) = 1$ *f.* $x > 0$ and $\mathrm{sign}(x) = -1$ *f.* $x < 0$. For ease of notation, we will denote $f(a,b)$, as defined in Equation 5, as $\mathrm{smax}(a,b)$ (for *signed maximum*).

As for the linear model, we tested whether the mean fixation location in the dual-feature gradients was distinguishable from this max-norm prediction, using paired t tests across observers:

$$(6)\ [\mathrm{LL}] \sim \mathrm{smax}([\mathrm{LN}],[\mathrm{NL}]),$$

$$(7)\ [\mathrm{RR}] \sim \mathrm{smax}([\mathrm{RN}],[\mathrm{NR}]),$$

$$(8)\ [\mathrm{RL}] \sim \mathrm{smax}([\mathrm{RN}],[\mathrm{NL}]),$$

and

$$(9)\ [\mathrm{LR}] \sim \mathrm{smax}([\mathrm{LN}],[\mathrm{NR}]).$$

In all but one case, we found significant differences between the model and the data: [LL], $t(7) = 6.53$, $p = .0003$; [RR], $t(7) = 2.71$, $p = .03$; [RL], $t(7) = 1.36$, $p = .22$; [LR], $t(7) = 6.56$, $p = .0003$. In conclusion, whereas the linear combination of single-feature effects was indistinguishable from the dual-feature data in all four cases, the maximum norm was significantly different in three out of four. This not only confirmed that our statistical power sufficed to exclude alternative models, but also clearly demonstrated that a linear addition of single-feature effects explained the data better than did a max-norm model.

Image-by-Image Results

Up to here, we have considered aggregate data across images. This was motivated by the fact that the images primarily served as "background" and image structure by itself probably has had a substantial effect on fixation allocation. To quantify this, we analyzed data averaged over participants and fixations for each image individually. For the congruent gradients 85/90 (LL) and 89/90 (RR), images showed a bias to the left and right (relative to NN), respectively. For the single-feature conditions, this bias to the higher contrast side was slightly less pronounced (LN, 60/90; RN, 60/90; NL, 75/90; NR, 82/90), but the fraction was still significantly above chance (all ps $< .003$, sign

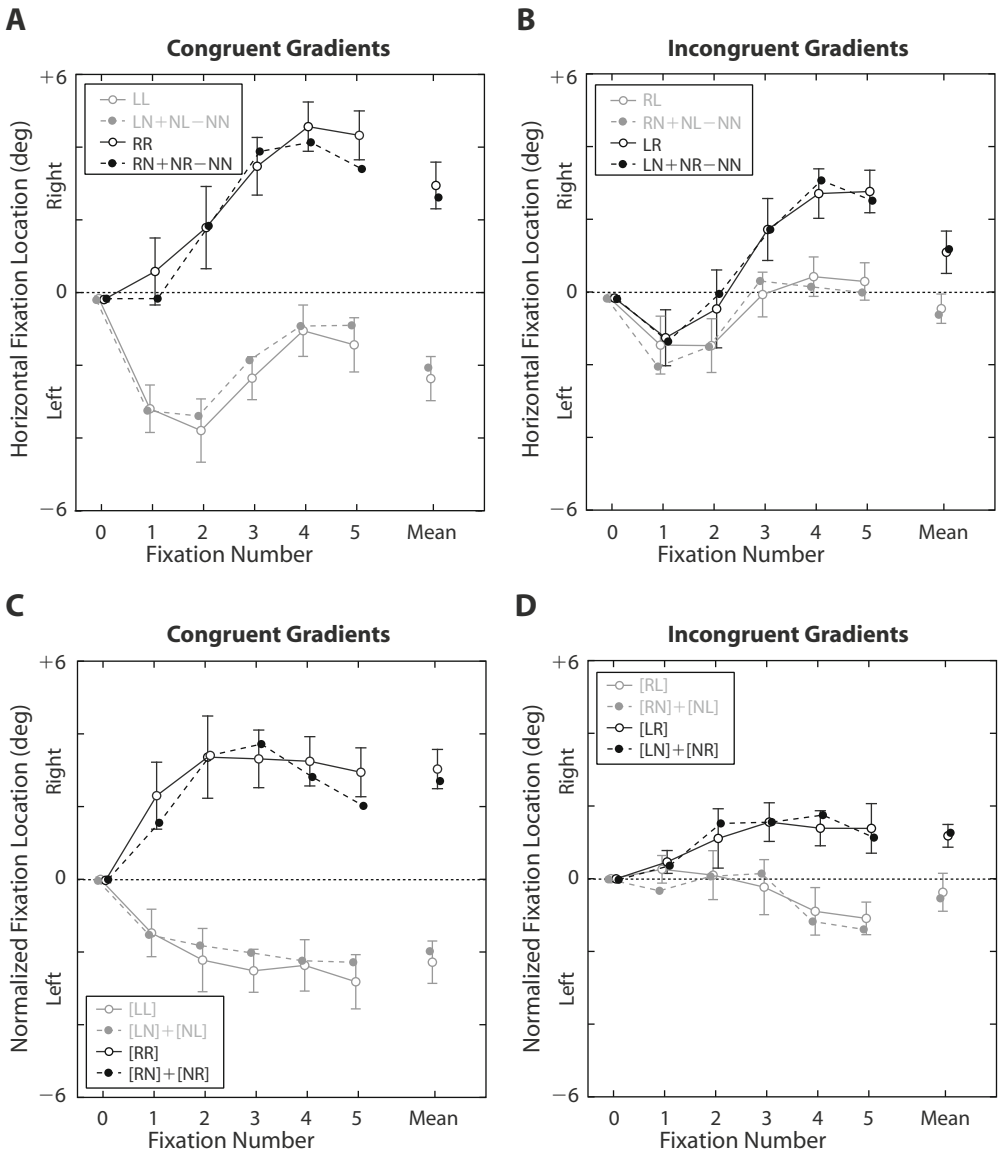

Figure 4. Dual-feature gradients. (A) Mean ± *SEM* for dual-feature conditions with same direction of gradients (congruent gradients). *Solid black*, RR; *solid gray*, LL. Dashed lines denote predictions from single-feature trials; *dashed black*, RN + NR − NN; *dashed gray*, LN + NL − NN. (B) Mean ± *SEM* for dual-feature conditions with opposing directions of gradients (incongruent gradients). *Solid black*, LR; *solid gray*, RL. Dashed lines denote predictions from single-feature trials; *dashed black*, LN + NR − NN; *dashed gray*, RN + NL − NN. (C and D) Analogous to A and B, using normalized data instead. These normalized representations more evidently visualize the time course over fixations, relative to the general bias (NN).

tests). Consequently, the biases were robust across images. Finally, we tested the predictions of the two models (linear addition and max-norm) on these imagewise data. In all cases, the linear model was indistinguishable from the data—[LL], $t(89) = 1.65$, $p = .10$; [RR], $t(89) = 1.02$, $p = .31$; [RL], $t(89) = 0.77$, $p = .45$; [LR], $t(89) = 0.89$, $p = .38$—whereas the max-norm model showed significant differences for the congruent dual-feature gradients—[LL], $t(89) = 6.08$, $p = 3\times10^{-8}$; [RR], $t(89) = 5.70$, $p = 1\times10^{-7}$; [RL], $t(89) = 0.05$, $p = .96$; [LR], $t(89) = 0.67$, $p = .51$. Hence, the image-by-image analysis confirmed the main finding: Dual-feature effects were consistent with a linear addition of single-feature effects in all the conditions. In contrast, a max-norm was consistent with the data only when individual effects were too small to clearly distinguish between the models. Our results therefore provided clear evidence that the interaction of CC and LC on a natural scene background is more consistent with linear summation than with a maximum operation.

DISCUSSION

The present study has investigated human overt attention on a natural scene background. We have demonstrated

that LC and CC gradients that are superimposed over a scene affect the selection of fixation points: Fixations were biased toward regions of high contrasts. Most notably, the combined effects of LC and CC gradients were consistent with a linear summation of feature effects, but not with a maximum operation.

The effects of gradients operated on top of a general bias in viewing direction when unmodified stimuli (the NN condition) were inspected, which started to the left and then rebounded to the right of the midline. Although this had not been the aim of the present study, it might be interesting to speculate whether this bias reflects a general strategy, possibly related to reading direction, as has been observed for other attentional phenomena, such as inhibition of return (Spalek & Hammad, 2005).

In order to encourage participants to pay attention to the stimuli, we asked them only to "study the images carefully." We had used this instruction in earlier studies and expected it to bias fixation allocation in a bottom-up driven mode and to operate in sharp contrast to explicit top-down tasks, such as search (Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2007). A recent experiment (Steinwender & König, 2007) indeed showed that "study carefully" yielded the same result with respect to low-level features as the explicit instruction of "free-viewing," whereas, for example, "subjective assessment" yielded distinct fixation behavior. Although we cannot exclude the possibility that the size of the effects for single-feature conditions depended on the particular choice of instruction, we clearly saw a bottom-up (i.e., feature-driven) component. In the present context, we built on this observation of a systematic shift of fixation locations induced by single-feature gradients. The prediction of linear interaction of different features was tested by comparing these measured single-feature shifts with those measured in dual-feature conditions. This test was therefore independent of the size of single-feature effects, as long as single-feature effects were different from 0 and sufficiently small to avoid having the image boundaries come into play: If single-feature effects were too large, combined effects could run out of room before hitting the left or right edge of the image. In particular, the test did not depend on whether or not the effect of LC and/or CC modification itself was linear in gradient strength, although we had observed linearity, at least for LC, earlier (Einhäuser, Rutishauser, et al., 2006). Consequently, as long as the instruction allows for shifts in the single-feature conditions that are sufficiently robust for the comparison with dual-feature effects, their precise size is not critical, nor is the exact choice of instruction.

Since Koch and Ullman's (1985) original proposal, the saliency map model has repeatedly been used to predict fixation behavior in natural scenes (Itti & Koch, 2000; Parkhurst et al., 2002; Peters et al., 2005; Tatler et al., 2005). In all these studies, however, prediction remained well below the theoretical optimum for any bottom-up model (i.e., a model taking into account only the current stimulus's features); this upper limit is given by the inter-observer prediction—that is, by the prediction derived from the fixation locations of a (large) set of other observers (Peters et al., 2005). Furthermore, the reasonable success of predictions on the system level neither implies causality nor provides support for the model's mechanistic assumptions. This raises the question of the extent to which individual features are indeed correlated with overt attention. With respect to LC, various studies (Krieger et al., 2000; Reinagel & Zador, 1999) showed this feature to be elevated at fixation points. Depending on presentation conditions, however, the correlative effect of LC was observed only after correcting for general biases in fixation pattern and depended on spatial frequency (Einhäuser & König, 2003; Mannan, Ruddock, & Wooding, 1996, 1997; Tatler et al., 2005); its size depended on the image material used (Parkhurst et al., 2002; Privitera & Stark, 2000). In addition, the effect of LC was often small, as compared with other luminance-related features, such as edge density (Mannan et al., 1996), texture contrast (Einhäuser & König, 2003; Parkhurst & Niebur, 2004), higher order geometric kernels (Privitera, Fujita, Chernyak, & Stark, 2005), and image-category-specific features (Privitera & Stark, 2000). With respect to the relative effects of the features under investigation here, Tatler et al. (2005) found LC and "edge-content" to contribute consistently more strongly to human fixation location than "chromaticity" and luminance itself. Since measuring the additivity of features had been the main aim of the present study, our single features had to fulfill two conditions: They had to be sufficiently large and robust to allow statistical analysis of their interaction (in the limit of no effect, all summation schemes are equally valid), but to be sufficiently small that image boundaries did not artificially cut the dual-feature effect. Therefore, we chose gradients that induced a robust effect for single-feature conditions. The fact that at least the effect of luminance gradients is linear in gradient slope (Einhäuser, Rutishauser, et al., 2006) renders it likely that our results on linearity can be generalized to weaker contrast changes, as found in natural contrast variations.

Most of the aforementioned studies measured the influence of each feature in its natural context. This, however, did not allow the isolation of the effects of each feature. If a feature were correlated with a higher order structure in natural scenes, increased fixation probability might result from the higher order structure or from correlation with other features, rather than from the feature itself (Baddeley & Tatler, 2006). To overcome this confound, Einhäuser and König (2003) locally increased or decreased LC in natural scenes. They found that reduced local contrast attracts human attention and concluded that this was inconsistent with saliency map model predictions. Although Parkhurst and Niebur (2004) reconciled this particular finding with saliency map models by incorporating higher order contrasts, local modifications were suboptimal in the present experimental context.

Strong local modifications introduce local deviations from global context, which are likely to attract attention. This is most evidently seen in the phenomenon of pop-out (Treisman & Gelade, 1980) and has recently entered the saliency map literature as the notion of *surprise*, an information-theoretic measure of deviations from the temporal context (Itti & Baldi, 2005). This issue of local

deviations from context becomes especially prominent when the applied modifications extend beyond the naturally occurring range of the feature. To avoid this potential confound in analyzing the interaction between features, we used large-scale gradients rather than local modifications. This procedure neither introduced local deviations nor modified higher order contrasts *locally*.

Obviously, the contrast gradient did not leave higher order structure unaffected; for example, reducing contrast also reduced edge density (if there is zero contrast, there are also no edges) and affected texture contrast. In any case, since we compared the effects of LC and CC in isolation from their combined effects, correlations to higher order structure *within a feature channel* would not confound our findings. One needed to ensure, however, that modifying LC did not affect CC and vice versa. By using definitions of LC and CC that were orthogonal in DKL space, this requirement was fulfilled, although it is conceivable that *perceived* LC varied with CC and vice versa. Although our gradients may affect higher order structure to some extent, their large scale, as well as the physical independence of the modified features, means that the linearity of LC and CC effects is also likely to hold in the natural context.

The rationale for using natural scenes as a *quasi-background* for the observed effects is twofold. First, the effect of the gradients is independent as to whether the scene is perceived as natural, at least as long as the amplitude spectrum is conserved (Einhäuser, Rutishauser, et al., 2006). Second, if we used a noise background instead, it could be argued that the interaction would be different if objects distract from the superimposed low-level effects. Hence, observing a linear interaction of CC and LC on—or maybe despite—the natural scene background strengthens our argument. Our data do, however, not address the issue of whether or not feature biases that are inherent in a scene affect fixated locations. Tatler (2007) found that those biases do not influence fixation. Similarly, our data are agnostic with respect to whether features such as color and luminance naturally occurring in natural scenes drive attention causally and, thus, do not contradict the large body of recent work that has failed to show a causal effect under realistic conditions.

Any model of attention that incorporates different features needs a mechanism to appropriately combine those features. Contemporary implementations of saliency maps usually solve this issue by using a sophisticated normalization scheme to achieve comparable saliency measures for each individual feature (see Itti & Koch, 2000, for a thorough discussion of normalization schemes). Subsequently, these models linearly combine the resulting conspicuity maps into the final saliency map. Here, we directly measured the individual effects (conspicuities) of each feature by using single-feature conditions and then tested whether linearity between these effects would hold. We found that CC and LC interacted linearly. Using a model based on psychophysical and physiological data, Li (2002) proposed that the saliency of an item was given by "the salience of its most salient component" (p. 12). This implied a maximum operation. Lewis and Zhaoping (2005) suggested that this model might also be applicable to the interaction of color and orientation in human overt attention for natural scenes. Given the different features and different methodology, our data do not contradict these findings directly. Instead, it will be an interesting issue, for future research, whether our results can be extended to other features, such as color and orientation, on a natural scene background. For the case of CC and LC, however, our data clearly falsify the max-norm hypothesis.

Since the saliency map model was originally designed as a purely bottom-up model of attention, by construction, it does not capture top-down influences such as the observer's experience or the task. The task plays a decisive role for human overt attention in inspecting pictures (Buswell, 1935; Yarbus, 1967) or search displays (Bacon & Egeth, 1997) or in everyday activities (Land & Hayhoe, 2001). When memorizing objects, for example, observers tend to replicate their own scan-paths, a feature not adequately captured by bottom-up saliency alone (Foulsham & Underwood, 2008).

Visual search constitutes a task frequently used to quantify the performance of attention models. Predictive performance of the original bottom-up saliency map model reduces or vanishes in search tasks (Einhäuser, Rutishauser, & Koch, 2008; Henderson et al., 2007), but inclusion of contextual or task-dependent information can improve the predictions of saliency map algorithms (Navalpakkam & Itti, 2005; Oliva, Torralba, Castelhano, & Henderson, 2003; Torralba, 2003). For evaluating the performance of saliency-map-type models in predicting search in natural scenes, the intuitive strategy of fixating the point of highest saliency is usually suboptimal; instead, the discriminability between target and distractor on the basis of the full map should be utilized (Gao, Mahadevan, & Vasconcelos, 2008; Vincent, Troscianko, & Gilchrist, 2007). For specific search tasks, such as searching for a pedestrian in a street scene, the task-modulated prior alone may predict search patterns better than do bottom-up signals (Torralba, Oliva, Castelhano, & Henderson, 2006). This approach, however, requires the prior distribution of potential target locations to be nonuniform and known. Such knowledge may be learned from scene statistics, and joint learning of bottom-up and top-down saliency in a Bayesian framework seems a promising approach (Zhang, Tong, Marks, Shan, & Cottrell, 2008).

Visual search models often use the selective upregulation of target features (Pomplun, 2006; Wolfe, Cave, & Franzel, 1989), of the corresponding visual filters (Rao, Zelinsky, Hayhoe, & Ballard, 2002), or statistical knowledge of target location (Najemnik & Geisler, 2005) to predict human performance. Rao et al.'s model bears some similarity to Itti and Koch's (2000) saliency map, but instead of adding different feature maps linearly, it computes a single map, which is modulated on the basis of the distance to the target template, rather than treating features individually. As Navalpakkam and Itti (2005) pointed out, this approach predicts that search for targets differing in one feature (pop-out) should be as efficient as conjunction search, contrary to experimental evidence (Treisman & Gelade, 1980). Although not contesting the approach of

Rao et al. per se, this argues in favor of different feature channels that need to be appropriately integrated.

We are well aware that the seeming mechanistic implications of the saliency map model have to be interpreted with care. In fact, we consider it likely that its predictive power for fixations stems entirely from correlations of its constituents with higher order structure inherent in natural scenes, such as *interesting objects* (Einhäuser, Spain, & Perona, 2008; Elazary & Itti, 2008). Furthermore, we are just beginning to understand how context and top-down information can be integrated in computational models of attention. Nevertheless, the original, purely bottom-up model is widely used and, up to now, other models that reach similar correlations with fixation probability (under the constraints of free viewing, laboratory setup, etc.) are rare. Independently of the precise model and its prediction on a systems' level, a sound understanding of human attention on a mechanistic level will always require a rigorous test of its assumptions. Irrespective of the exact nature of a future model that finally supersedes the saliency map for fixation prediction, it will be constrained by the present finding: Effects of CC and LC—under laboratory conditions and on a natural scene background—add linearly. The extent to which this finding transfers to other low-level features and to spatial distributions of higher order scene structures thus remains an exciting issue for future research, no matter one's take on the original saliency map.

AUTHOR NOTE

The work was supported financially by Deutsche Forschungsgemeinschaft Research Training Group 885–NeuroAct (to B.M.tH. and W.E.), the German Academic Exchange Service (to S.E.), and by IST-027268-POP (Perception On Purpose; to P.K.). The authors are grateful to K. Libertus and H.-P. Frey for technical assistance and to C. Quigley for editorial assistance. S.E. and B.M.tH. contributed equally to this article. Correspondence concerning this article should be addressed to W. Einhäuser, AG Neurophysik/FB Physik, Philipps University Marburg, Renthof 7, D-35032 Marburg, Germany (e-mail: wet@physik.uni-marburg.de).

REFERENCES

Bacon, W. F., & Egeth, H. E. (1997). Goal-directed guidance of attention: Evidence from conjunctive visual search. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 948-961.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, **46**, 2824-2833. doi:10.1016/j.visres.2006.02.024

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289-300. Available at www.jstor.org/stable/2346101.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436. doi:10.1163/156856897X00357

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology of perception in art*. Chicago: University of Chicago Press.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing*, **20**, 241-248.

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

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology*, **357**, 241-265.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, **17**, 1089-1097. doi:10.1046/j.1460-9568.2003.02508.x

Einhäuser, W., Kruse, W., Hoffmann, K.-P., & König, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research*, **46**, 1194-1209. doi:10.1016/j.visres.2005.08.032

Einhäuser, W., Rutishauser, U., Frady, E. P., Nadler, S., König, P., & Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision*, **6**, 1148-1158. doi:10.1167/6.11.1

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, **8**(2, Art. 2), 1-19. doi:10.1167/8.2.2

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, **8**(14, Art. 18), 1-26. doi:10.1167/8.14.18

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, **8**, 1-15. doi:10.1167/8.3.3

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, **8**(2, Art. 6), 1-17. doi:10.1167/8.2.6

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center–surround hypothesis for visual saliency. *Journal of Vision*, **8**(7, Art. 13), 1-18. doi:10.1167/8.7.13

Golz, J., & MacLeod, D. I. A. (2002). Influence of scene statistics on colour constancy. *Nature*, **415**, 637-640. doi:10.1038/415637a

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, **391**, 481-484. doi:10.1038/35135

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 537-562). Amsterdam: Elsevier.

Horwitz, G. D., & Newsome, W. T. (1999). Separate signals for target selection and movement specification in the superior colliculus. *Science*, **284**, 1158-1161. doi:10.1126/science.284.5417.1158

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

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 631-637). Los Alamitos, CA: IEEE Computer Society Press.

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

James, W. (1890). *Principles of psychology*. New York: Holt.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, **4**, 219-227.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, **13**, 201-214. doi:10.1163/156856800741216

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature*, **384**, 74-77. doi:10.1038/384074a0

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, **41**, 3559-3565. doi:10.1016/S0042-6989(01)00102-X

Lewis, A., & Zhaoping, L. (2005). Saliency from natural scene statistics. *Abstract Viewer/Itinerary Planner* (Program No. 821.11). Washington, DC: Society for Neuroscience.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, **6**, 9-16. doi:10.1016/S1364-6613(00)01817-9

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The re-

lationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, **10**, 165-188. doi:10.1163/156856896X00123

MANNAN, S. K., RUDDOCK, K. H., & WOODING, D. S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception*, **26**, 1059-1072.

MAZER, J. A., & GALLANT, J. L. (2003). Goal-related activity in V4 during free viewing visual search: Evidence for a ventral stream visual salience map. *Neuron*, **40**, 1241-1250. doi:10.1016/S0896-6273(03)00764-5

MCPEEK, R. M., & KELLER, E. L. (2002). Superior colliculus activity related to concurrent processing of saccade goals in a visual search task. *Journal of Neurophysiology*, **87**, 1805-1815.

MICHELSON, A. A. (1927). *Studies in optics*. Chicago: University of Chicago Press.

MORRONE, M. C., DENTI, V., & SPINELLI, D. (2002). Color and luminance contrasts attract independent attention. *Current Biology*, **12**, 1134-1137. doi:10.1016/S0960-9822(02)00921-1

NAJEMNIK, J., & GEISLER, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, **434**, 387-391. doi:10.1038/nature03390

NAVALPAKKAM, V., & ITTI, L. (2005). Modeling the influence of task on attention. *Vision Research*, **45**, 205-231. doi:10.1016/j.visres.2004.07.042

NOTHDURFT, H. (2000). Salience from feature contrast: Additivity across dimensions. *Vision Research*, **40**, 1183-1201. doi:10.1016/S0042-6989(00)00031-6

OLIVA, A., TORRALBA, A., CASTELHANO, M. S., & HENDERSON, J. M. (2003). Top-down control of visual attention in object detection. *IEEE Proceedings of the International Conference on Image Processing*, **1**, 253-256.

PARKHURST, D., LAW, K., & NIEBUR, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, **42**, 107-123. doi:10.1016/S0042-6989(01)00250-4

PARKHURST, D., & NIEBUR, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, **19**, 783-789. doi:10.1111/j.0953-816X.2003.03183.x

PELI, E. (1997). In search of a contrast metric: Matching the perceived contrast of Gabor patches at different phases and bandwidths. *Vision Research*, **37**, 3217-3224. doi:10.1016/S0042-6989(96)00262-3

PELLI, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442. doi:10.1163/156856897X00366

PETERS, R. J., IYER, A., ITTI, L., & KOCH, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, **45**, 2397-2416. doi:10.1016/j.visres.2005.03.019

POMPLUN, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research*, **46**, 1886-1900. doi:10.1016/j.visres.2005.12.003

POSNER, M. I., & PETERSEN, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, **13**, 25-42. doi:10.1146/annurev.ne.13.030190.000325

PRIVITERA, C. M., FUJITA, T., CHERNYAK, D., & STARK, L. W. (2005). On the discriminability of hROIs, human visually selected regions-of-interest. *Biological Cybernetics*, **93**, 141-152. doi:10.1007/s00422-005-0586-7

PRIVITERA, C. M., & STARK, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **22**, 970-982.

RAO, R. P. N., ZELINSKY, G. J., HAYHOE, M. M., & BALLARD, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, **42**, 1447-1463. doi:10.1016/S0042-6989(02)00040-8

REINAGEL, P., & ZADOR, A. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, **10**, 341-350.

RIZZOLATTI, G., RAGGIO, L., DASCOLA, I., & UMILTÀ, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, **25**, 31-40.

ROBINSON, D. L., & PETERSEN, S. E. (1992). The pulvinar and visual salience. *Trends in Neurosciences*, **15**, 127-132. doi:10.1016/0166-2236(92)90354-B

SPALEK, T. M., & HAMMAD, S. (2005). The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science*, **16**, 15-18. doi:10.1111/j.0956-7976.2005.00774.x

STEINWENDER, J., & KÖNIG, P. (2007, August). *Context dependency of overt attention in natural scenes*. Poster presented at the 14th European Conference on Eye Movements, Potsdam.

TATLER, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, **7**(14, Art. 4), 1-17. doi:10.1167/7.14.4

TATLER, B. W., BADDELEY, R. J., & GILCHRIST, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, **45**, 643-659. doi:10.1016/j.visres.2004.09.017

THOMPSON, K. G., BICHOT, N. P., & SCHALL, J. D. (1997). Dissociation of visual discrimination from saccade programming in macaque frontal eye field. *Journal of Neurophysiology*, **77**, 1046-1050.

TORRALBA, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A*, **20**, 1407-1418. doi:10.1364/JOSAA.20.001407

TORRALBA, A., OLIVA, A., CASTELHANO, M. S., & HENDERSON, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, **113**, 766-786.

TREISMAN, A. M., & GELADE, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, **12**, 97-136. doi:10.1016/0010-0285(80)90005-5

VINCENT, B. T., TROSCIANKO, T., & GILCHRIST, I. D. (2007). Investigating a space-variant weighted salience account of visual selection. *Vision Research*, **47**, 1809-1820. doi:10.1016/j.visres.2007.02.014

WOLFE, J. M., BUTCHER, S. J., LEE, C., & HYLE, M. (2003). Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons. *Journal of Experimental Psychology: Human Perception & Performance*, **29**, 483-502.

WOLFE, J. M., CAVE, K. R., & FRANZEL, S. L. (1989). Guided search: An alternative to the feature integration model of visual search. *Journal of Experimental Psychology: Human Perception & Performance*, **15**, 419-433.

YARBUS, A. L. (1967). *Eye movements and vision* (B. Haigh, Trans.). New York: Plenum.

ZHANG, L., TONG, M. H., MARKS, T. K., SHAN, H., & COTTRELL, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, **8**(7, Art. 32), 1-20. doi:10.1167/8.7.32

(*Continued on next page*)

APPENDIX A
Effects of Modifications on Luminance Contrast, Saliency, and Color Conspicuity

Here, we will address the relation of our proposed modifications to common definitions of contrast, feature conspicuity, and saliency. There are plenty of possible ways to define LC (Peli, 1997). Most definitions are originally based on the comparison of a single foreground intensity with a single background intensity, such as the Weber contrast (the difference of foreground and background, divided by the background) or the Michelson contrast (the difference of foreground and background, divided by their sum; Michelson, 1927) and have been extended to account for arbitrary stimuli. Among the possible variants, we here will focus on those that are of common use in the context of eyetracking and attention studies.

1. The standard deviation of luminance in a local patch, divided by the image mean (e.g., Reinagel & Zador, 1999).

2. The standard deviation of intensity in a local patch, divided by the patch mean (also suggested, and dismissed as suboptimal in the present context, by Reinagel & Zador, 1999).

3. The difference of maximum and minimum in a local patch, divided by their sum in the same local patch. This is a definition most closely related to the Michelson contrast. Note that Mannan, Ruddock, and Wooding's (1996) usage of the mean of intensity in a local patch as foreground and the mean of the image as background in their calculation of Michelson contrast is a measure of luminance, rather than of LC, in the present context. The denominator is also commonly scaled or replaced by the mean (akin to the Weber contrast) or by the maximum alone.

We computed all these contrasts for the luminance channel of our images in DKL space, which we shifted and scaled to a range from 0 to 1 (rather than from −1 to 1), and for squared patches with a width of 24 pixels (corresponding to 0.5º at the screen center). Comparing the conditions in which luminance contrast increased to the left or right or remained unmodified for a single image (the one in Figure 2A) and averaging over rows, we saw the intended effect of modification clearly, and the differences between the various contrast definitions were minute (Figures A1A–A1C). Note that, by definition, the LC profile was not affected by modifications to the CC; for example, the LL, NL, and NR conditions had the same LC profile. The example profiles show that highly noticeable structures, such as the tree on the left-hand side of the example image, are still visible, although the modification dominates this contrast profile.

To quantify how "unnatural" the contrast modifications were, we assessed the additional variation of contrast introduced by the gradients, using the Contrast Definition 1 above. In the unmodified condition, the mean contrast within an image amounted to 0.61 ± 0.13 (mean ± *SD* across images). As one would expect by construction, this value was about halved for modifications, no matter whether the increase was to the left (LL, NL, RL) or to the right (LR, NR, RR), with values of 0.30 ± 0.07 in both cases. It should be noted, however, that a *large-scale* single-feature modification of contrast always biased toward the higher contrast side, no matter whether the gradient decreased or increased the contrast, relative to unmodified (Einhäuser, Rutishauser, et al., 2006), which was different from local modifications (Einhäuser & König, 2003). More important, the gradients lowered the variation of contrast within each image, quantified as standard deviation of contrast values, only by about 25% (0.33 ± 0.10 for unmodified, 0.25 ± 0.06 for gradients to the right, and 0.25 ± 0.07 for gradients to the left). This indicated that a sufficient amount of image-inherent variability remained in the low-level features, which could, in principle, drive attention. The fact that the gradient nonetheless dominated the fixation allocation was consistent with a minute (or absent) effect of image-inherent low-level features.

Next, we considered the effect of our LC modifications on the Itti and Koch (2000) model for visual saliency. For the model, we used the implementation provided at ilab.usc.edu with no normalization, but otherwise, default parameter settings. To be closer to the typical scenario for the application of these algorithms, we here used the image in the RGB version sent to the screen, rather than the original DKL definition; that is, luminance was nonlinearly scaled. By performing the same analysis as that for the contrast definitions, we found that our LC modifications strongly modulated model saliency in the expected direction: Saliency increased to the right in the NR condition and increased to the left in the NL condition (Figure A1D).

Finally, we addressed the effect of our color modification on the color channel of the saliency map model with the same settings as above. As was expected, we found color conspicuity to increase to the right in the RN condition and to the left in the LN condition (Figure A1E). The original image structure, however, was conserved more than in the luminance case, and the luminance conspicuity dominated the overall saliency map with default weighting (not shown). That is, the effect of color modification on image structure was weaker, consistent with the slightly weaker bias induced by color. Although this bias difference is worth investigating—in particular, with respect to feature-weighting schemes for saliency maps—it was not of relevance for the present article. When both gradients induced a robust fixation bias (Figure 3), we could compare their effects (Figure 4). Here, we verified that the modifications leading to these effects were indeed consistent with common definitions of LC and color conspicuity.

APPENDIX A (Continued)

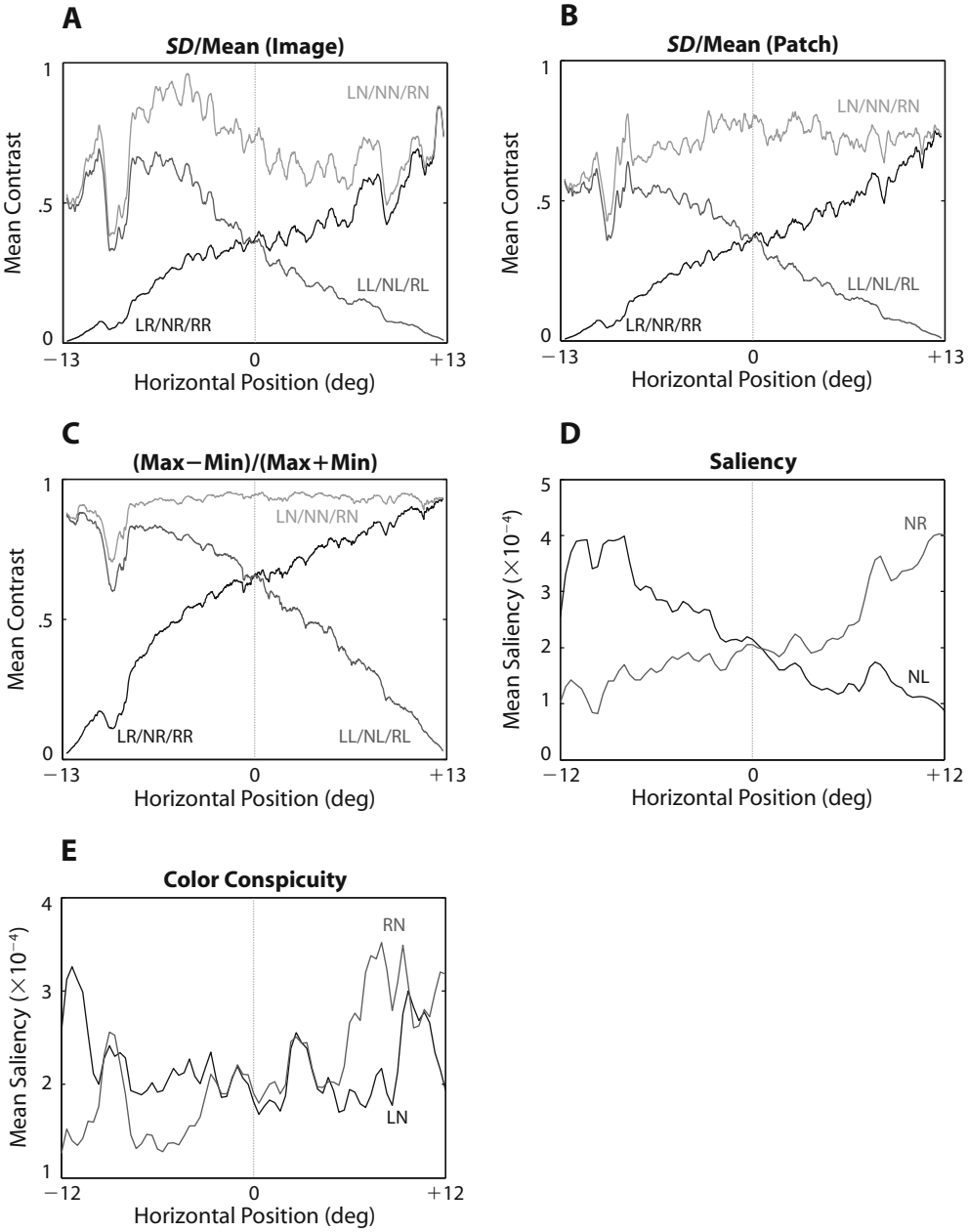

Figure A1. Effect of gradients on common definitions of contrast, conspicuity, and saliency. (A–C) Effect of luminance contrast gradient on different definitions of luminance contrast along horizontal scanline for the example image in Figure 2A, averaged over image rows. (A) Standard deviation of luminance in a 1º × 1º patch, divided by image mean. (B) Standard deviation of luminance in a patch, divided by patch mean. (C) Difference between maximum and minimum luminance in a patch, divided by their sum. (D) Saliency according to Itti and Koch (2000), maps linearly normalized to unit integral. (E) Color conspicuity according to Itti and Koch, maps linearly normalized to unit integral. Note that the maps in panels D and E have a lower resolution and additional cutoff at the image boundary.

(*Continued on next page*)

APPENDIX B
Raw Eye Position

In order to be independent of the fixation definition, we repeated our main analysis, using 1-msec bins instead of individual fixations. Since subsequent time points fell on the same fixation and were thus not independent, we could not perform the equivalent statistical analysis. Instead, we used paired *t* tests to test for the significance of difference but adjusted the alpha level to match an expected false discovery rate (FDR) of 5%, using the procedure proposed by Benjamini and Hochberg (1995). A result was called significant if it fell below this adjusted level (denoted as $FDR_{.05}$). For color-only gradients, we found a significant difference between LN and RN ($p < .019 = FDR_{.05}$) on 773 sample points between 364 and 1,198 msec. Similarly, for LC-only gradients, there was a significant difference between NL and NR ($p < .036 = FDR_{.05}$) on 1,435 sample points between 117 and 2,000 msec. This confirmed that during the majority of the presentation time, gradients affect eye position. At an expected FDR of .05, LL was at no time point different from LN + NL − NN (Figure B1A, gray). Subtracting the NN condition on both sides by construction did not alter the results; that is, [LL] was not different from [LN] + [NL] (Figure B1C). Neither was [RR] different from [RN] + [NR] anywhere (Figures B1A and B1C, black). Similarly, the incongruent gradient data did not exhibit significant differences from their respective models at any time point; [LR] was indistinguishable from [NR] + [LN] (Figures B1B and B1D, black) and [RL] from [NL] + [RN] (Figures B1B and B1D, gray). In sum, the analysis of the raw eye position data confirmed the fixation analysis, ruling out the possibility that the observed effects depended on the definition or timing of fixations.

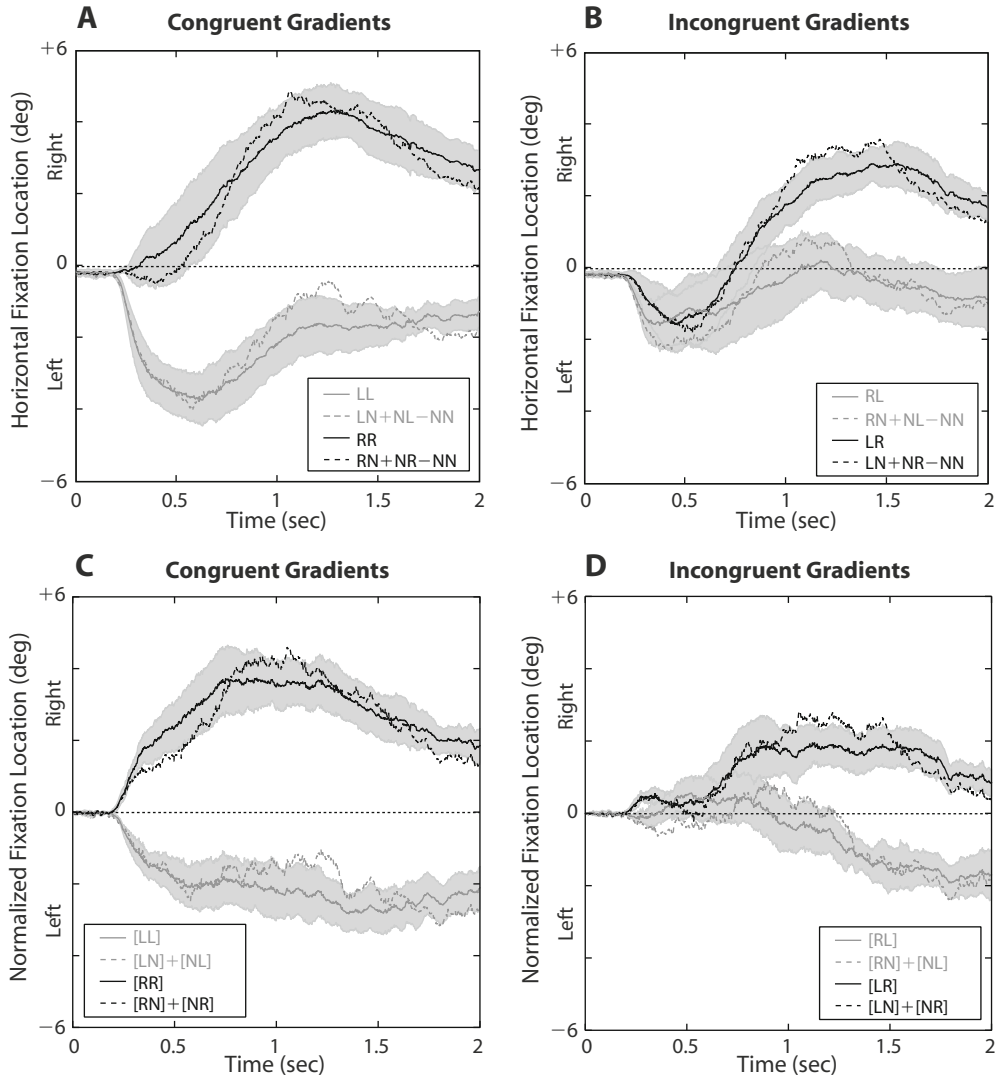

Figure B1. Analysis over time. Analogous to Figure 4, time into trial, rather than fixation number, is used as parameter. Shaded areas denote *SEM* of data.

(Manuscript received September 24, 2008;
revision accepted for publication March 25, 2009.)