

Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories

Alper Açıık^{a,*}, Selim Onat^a, Frank Schumann^a, Wolfgang Einhäuser^b, Peter König^a

^a University of Osnabrück, Institute of Cognitive Science, Albrechtstr. 28, 49069 Osnabrück, Germany

^b Philipps-University Marburg, Department of Neurophysics, Renthof 7, 35032 Marburg, Germany

ARTICLE INFO

Article history:

Received 11 August 2008

Received in revised form 12 March 2009

Keywords:

Attention

Eye movements

Luminance contrast

Phase features

Image categories

ABSTRACT

During viewing of natural scenes, do low-level features guide attention, and if so, does this depend on higher-level features? To answer these questions, we studied the image category dependence of low-level feature modification effects. Subjects fixated contrast-modified regions often in natural scene images, while smaller but significant effects were observed for urban scenes and faces. Surprisingly, modifications in fractal images did not influence fixations. Further analysis revealed an inverse relationship between modification effects and higher-level, phase-dependent image features. We suggest that high- and mid-level features – such as edges, symmetries, and recursive patterns – guide attention if present. However, if the scene lacks such diagnostic properties, low-level features prevail. We posit a hierarchical framework, which combines aspects of bottom-up and top-down theories and is compatible with our data.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

During daily life, two to four eye movements per second bring portions of a scene to the human fovea. The part of the visual field falling onto the fovea is represented with the highest spatial acuity and compared to the periphery receives disproportionately more cortical processing (Tootell, Silverman, Switkes, & De Valois, 1982). Both top-down mechanisms that are task-specific or goal-driven (Land, Mennie, & Rusted, 1999; Yarbus, 1967) and bottom-up mechanisms considering only sensory-driven aspects such as image features (Dickinson, Christensen, Tsotsos & Olofsson, 1997; Koch & Ullman, 1985) are known to contribute to the fixation selection process. However, the relative roles and the interaction of these mechanisms are still not understood. As such, the principles of fixation selection are still not clear.

Research emphasizing bottom-up aspects of overt attention generally employs the concept of a saliency map (Itti & Koch, 2000; Koch & Ullman, 1985; Parkhurst, Law, & Niebur, 2002). The essential assumption behind this approach is to see fixation probability as a monotonically increasing function of several feature contrasts (e.g. luminance contrast, texture contrast, and color contrast). Feature contrasts are independently computed at different spatial scales in parallel channels. The resulting contrast maps for individual features are then weighted and summed to generate the final feature-independent saliency map. A straightforward

winner-take-all mechanism then selects the peaks of the saliency map as foci of overt visual attention. Even though saliency maps are popular, since they predict fixation locations to a reasonable degree (Parkhurst & Niebur, 2003, 2004; Peters, Iyer, Itti, & Koch, 2005), there are still fundamental questions that remain unanswered (Einhäuser & König, 2003; Henderson, Brockmole, Castelhano, & Mack, 2007; Pomplun, 2006; Tatler, 2007; Vincent, Troscianko, & Gilchrist, 2007).

One major issue concerns the relation between image features and fixation selection. The major evidence supporting saliency map models comes from eye-tracking studies (Krieger, Rentschler, Hauske, Schill & Zetsche, 2000; Parkhurst & Niebur, 2003; Parkhurst et al., 2002; Peters et al., 2005; Reinagel & Zador, 1999). These report that various feature contrasts, especially luminance contrast (LC), are found to be higher at fixated points than at other parts of an image viewed by the participants. Even though this is conclusive evidence for a correlation between certain image features and overt attention, the relation is not necessarily causal. LC might just be correlated to some other property – low-level (Baddeley & Tatler, 2006) or high-level (Einhäuser & König, 2003) – that guides fixation selection. One way to address this issue is to consider the correlations between features as was done by Baddeley and Tatler (2006). Employing a Bayesian model, the authors showed that among the features tested, high-frequency edges had the highest correlation with fixation probability. After this contribution to fixation selection was accounted for, the further improvement in the prediction of fixation location provided by the addition of the contrast information was negligible. In order

* Corresponding author. Fax: +49 541 9692596.
E-mail address: aacik@uos.de (A. Açıık).

to address the causality directly, our group recently introduced a paradigm of local LC modification (Einhäuser & König, 2003; Einhäuser, Kruse, Hoffmann & König, 2006; Einhäuser, Rutishauser, et al., 2006). We found that both local increases and decreases of LC attract attention. The latter finding speaks against a simple form of the causality hypothesis, which predicts that negative modifications should repel attention, instead of attracting it. Parkhurst and Niebur (2004) criticized several aspects of the first Einhäuser and König (2003) study. The main critique concerned changes in texture contrast (a second-order contrast, and hence a mid-level feature), which were an unintended and inevitable result of LC modifications. They showed that both positive and negative modifications led to increases in texture contrast. A computational model selected fixations from high texture contrast regions, and those points were shown to fall on modified regions, matching human data remarkably well. In an analysis of image features at human-selected fixations however, Einhäuser, Kruse, et al. (2006) failed to find the same principle of fixation selection based on high texture contrast. Further examination of the local changes resulting from these local modifications is therefore needed in order to clarify whether the relation between overt attention and image features is only correlative, or causal as well.

Not all saliency models view fixation probability as a monotonically increasing function of feature values. For instance the Statistical Saliency Model (Rosenholtz, Li, & Nakano, 2007) selects points characterized by feature values that deviate largely from the mean, independent of the direction of the deviation. Tatler, Baddeley, and Gilchrist (2005) also weigh both ends of the feature distributions similarly while computing local saliency. The LC modifications discussed above, by definition, alter LC values in an image at a subset of pixels. But whether these new LC values are different from the LC of the unaltered parts of the same image depends on the statistics of the modified image. For instance, in the case of an image with average LC values distributed rather uniformly over space, both LC decreases and increases would create values that differ from the mean of the LC map. But if the LC of the base image has high variance, then the modified LC values can still be in the central portions of the LC distribution, due to the occurrence of similar values elsewhere in the image. That is, regarding the effect of feature alterations on fixation behavior, different types of saliency models have different predictions.

A further disagreement in the literature concerns the effects of time on the relation between image features and overt attention. Initial reports (Parkhurst et al., 2002; Reinagel & Zador, 1999) suggested that during the first few seconds of presentation of a static image, the effects of low-level features are most prominent. Einhäuser and König (2003) reported that LC modifications also exerted their effects on attention in a temporally dependent manner. Tatler et al. (2005) however argued that observations by Parkhurst et al. (2002) are due to a methodological artifact. The first fixations may differ from later fixations in terms of location (e.g. central) or the distance between them (e.g. close), giving rise to a temporal bias that is independent of the image under consideration. Controlling for such temporal biases, the authors found that the effects of low-level features were constant over time. In their study, images of office interiors and urban outdoor scenes were used as stimuli. It remains to be tested if the constancy of feature effects holds for different image categories, especially for natural environments containing no man-made objects.

In the visual attention literature, 'natural scene' is a term that encompasses a wide variety of scenes, including urban scenes, faces, or natural environments characterized by a lack of man-made objects. Torralba and Oliva (2003) analyzed the spectral signatures of scene types as specific and diverse as mountains, fields, beaches, streets, and high buildings, and concluded that these global second-order statistics can be used for image classification.

When the images were pooled into the superordinate classes of man-made and natural environments, the power spectra – in particular, the values found along the cardinal directions – led to a clear separation of the two classes. If such global statistics can be computed early in the brain, this could be used to guide the local analysis that is mediated by sequential eye movements. Category-specific effects of low-level features on overt attention have already been reported (Einhäuser, Kruse, et al., 2006; Einhäuser, Rutishauser, et al., 2006; Frey, Honey, & König, 2008; Parkhurst et al., 2002). Given physiological (Bar et al., 2006) and psychophysical (Hochstein & Ahissar, 2002) evidence that a scene is analyzed first globally and then locally, studying scene-content dependencies and category differences offers a nice means of answering attention-related questions.

Notwithstanding the importance of attempts to describe and model global scene configuration of natural scenes, an interesting possibility that is only rarely addressed concerns the relationship between the 'total amount' of image structure in an image and the magnitude of local feature effects on attention. In other words, to what degree does the presence of global image features determine whether local feature values will guide attention? The lack of research in this direction is partly due to the difficulty of quantifying the total amount of global scene configuration. Traditionally, the image phase spectrum, computed from the whole image, is used to capture the informative image content. This approach is validated by the classical finding that an image with intact phase content but randomized amplitude is still correctly recognized by human observers (Piotrowski & Campbell, 1982). Using the amount of phase noise added to the image in the spatial frequency domain as a measure of image content removal, Einhäuser, Rutishauser, et al. (2006) demonstrated that the correlative effect of luminance contrast drops with phase noise addition. Using similar modifications of phase, Wichmann, Braun, and Gegenfurtner (2006) demonstrated that natural image classification performance drops with phase noise, and that this result cannot be explained by changes in local contrast alone. These and other phase modification studies (e.g. Kayser, Nielsen & Logothetis, 2006) suggest that the phase spectrum is a good starting point for quantifying the amount of global scene configuration in an image.

Phase spectrum and scene configuration are also related to each other at different spatial scales. Kovessi (1997, 1999, 2000a) has demonstrated that edges, corners, symmetries, and asymmetries in images can be extracted via local phase analysis. Phase congruency measures the phase alignment of different frequency components of a signal. Intuitively, at step-edges for instance, all frequency components of a signal have the same phase. Kovessi's (1997, 1999) phase-congruency measure is blind to first- and second-order statistics, and this differentiates his work from the saliency map approach where luminance contrast is of special importance. If computed at different spatial scales for the whole image, the average phase congruency in an image can thus be used as a measure of total image structure, since it returns the average feature-related response that is independent of contrast.

In the present study we gathered fixation data while human observers freely viewed unmodified or locally LC-modified images from four categories: natural environments, urban scenes, fractal images, and human faces. We first focused on how image features and modification levels were characterized at fixated points. LC, texture contrast (TC), and the modifications at fixations were compared to carefully selected control points. A pre- and post-modification analysis of TC was conducted in the case of modified images. Where appropriate, the effects of features and modifications were tracked over time. We then computed the average of phase-related features for each image, and correlated that with the effects of modifications on attention. A discussion on how

these analysis methods address the interplay between top-down and bottom-up processes is provided.

2. Methods

2.1. Participants

The participants were 14 university students (five female), ages 21–27. They were either granted extra course credits or received monetary compensation for their participation. All had normal or corrected-to-normal vision and were naïve to the purpose of the study. Written consent was obtained from all participants. The experimental procedure conformed to the Declaration of Helsinki and national guidelines.

2.2. Stimuli

The basis images (see below) were chosen from four categories (for representative examples see Fig. 1a–d). The images in the first category (*'naturals'*) depicted natural environments such as forests and open landscapes, with an absence of any man-made object. These came from the Zurich Natural Image Database ([http://](http://www.klab.caltech.edu/~wet/ZurichNatDB.tar.gz)

www.klab.caltech.edu/~wet/ZurichNatDB.tar.gz) used in our earlier studies (Einhäuser & König, 2003; Einhäuser, Kruse, et al., 2006). Images from the second category (*'manmades'*) depicted either urban scenes like streets and house exteriors, or man-made objects such as barrels and hats. These were part of the McGill Calibrated Colour Image Database (Olmos & Kingdom, 2004). The third category (*'fractals'*) consisted of software-generated fractal pictures taken from a World Wide Web fractal database Chaotic *N*-space Network (<http://www.cnspace.net/html/fractals.html>). The last category (*'faces'*) included frontal, close-up shots of faces, taken especially for the purpose of this study. They were taken with a high-resolution camera (Sony DSC-V1 Cyber-shot, Tokyo, Japan) in enclosed public places (e.g. university library) using the available light in the room. All images from all categories were downsampled or cropped to a resolution of 1024 × 768 pixels and converted to grayscale. In the case of face pictures, care was taken that the face filled the image as much as possible in the vertical direction. 16 images were selected from each of the four categories, making a total of 64 basis images.

To introduce local contrast changes, we followed the procedure used in previous studies (for details see Einhäuser & König, 2003) with one improvement. Briefly, a random point was chosen from

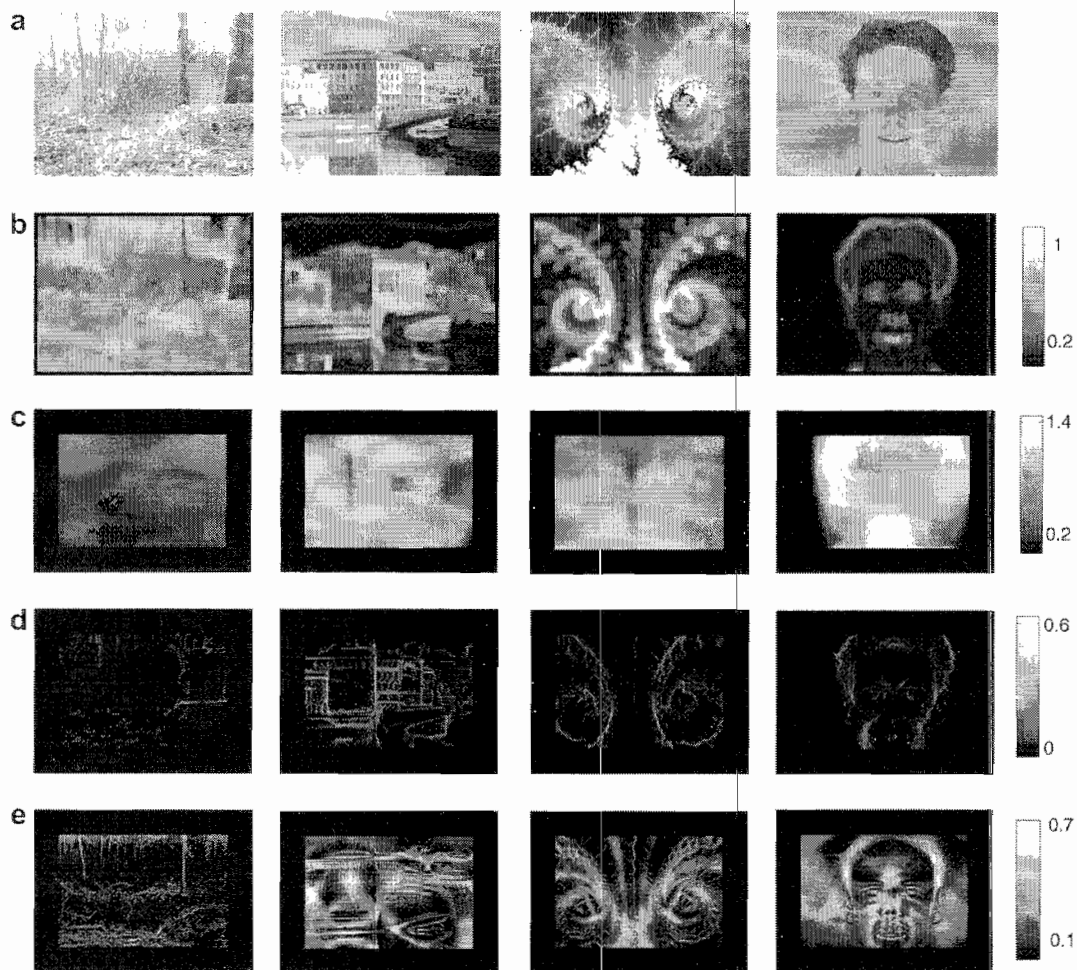


Fig. 1. Stimuli and feature map examples. (a) Representative examples from *'naturals'*, *'manmades'*, *'fractals'*, and *'faces'* categories, shown from left to right, respectively. (b) Luminance contrast maps (computed in 1.1° patches) for the example images shown in (a). (c) Texture contrast maps (computed in 5.5° patches) for the example images shown in (a). (d) Phase congruency maps for the example images shown in (a). (e) Phase symmetry maps for the example images shown in (a). Color code specified at the left end of each panel holds for all elements in that row.

an image-sized, all-zero matrix, and a two-dimensional Gaussian ($\sigma = 80$ pixels, $\approx 2.2^\circ$) was centered on this point. In order to generate multiple local modifications on a single image, further image-sized matrices were created using this method, and the maximum value of each entry over these matrices was then taken to create the mask G . A new mask was created for each image, and was used to modify the pixel intensity of an image according to the following equation (Fig. 2):

$$I(x, y) = [1 + \alpha G(x, y)] [I_0(x, y) - \hat{I}(x, y)] + \hat{I}(x, y) \quad (1)$$

where I_0 denotes the basis image, I the modified image, α the peak modification level, and \hat{I} the mean intensity map of the basis image. This intensity map is the result of convolving the image with a two-dimensional Gaussian ($\sigma = 80$ pixels). After applying Eq. (1) to the image, the pixel values smaller than 0 were set to 0 and the values larger than 255 were set to 255, such that the images could be displayed on the monitor. The percentage of pixels that had to be subjected to this clipping was 2% of all modified pixels.

Each modified image contained three modifications. The randomly chosen points used as centers of modification were at least 160 (2σ) pixels away from each other, 160 pixels (2σ) from the vertical image boundary, and 336 ($1/4$ of the horizontal image size plus σ) pixels from the horizontal image boundary. This allowed the modification of local luminance contrast without the introduction of any boundary artifacts, and the higher horizontal limit ensured that modifications did not fall onto the uniform background regions in the case of faces. The peak modification levels used in our study ranged from $\alpha = -0.6$ to -0.2 and $+0.2$ to $+1.0$ with intervals of 0.2 (see Fig. 2 for examples of maximum positive and negative modifications).

2.3. Eye tracking

The Eyelink II system (SR Research, Ontario, Canada) was used to record participants' eye movements. It uses infrared pupil tracking and corneal reflection based tracking at a sampling rate of 250 Hz and can compensate for head movements. In this study,

monocular eye position data were sampled with infrared-based tracking only.

To calibrate eye position and to validate this calibration, participants made saccades to a grid of fixation spots on the screen, which appeared one by one in a random order. As soon as the mean calibration error for any eye was below 0.24° for two successive validations, tracking of that eye started. After each stimulus presentation a fixation spot appeared in the middle of the screen and the drift error of the measured eye movement to this spot was computed. If this error was larger than 1° , calibration and validation were repeated.

Fixation locations and times were calculated online by the eye tracker. The default system parameters that are used to define saccades were: an acceleration threshold of $8000^\circ/\text{sec}^2$, a velocity threshold of $30^\circ/\text{sec}$, and a deflection threshold of 0.1° . These values had to be maintained for at least 4 ms. Fixations were defined as the complement of a saccade, i.e. periods without saccades. The first fixation of each trial was excluded from analysis, since its location was an artifact of the preceding central fixation spot.

2.4. Experimental procedure

During the course of the experiment, each subject saw 576 pictures (16 images per category \times 4 categories \times 9 modification levels), each of which was displayed on the screen for 5 s. The presentation order was randomized, with the single constraint that two modifications of the same basis image were never shown consecutively.

In order to maintain participants' alertness, we limited the duration of eye-tracking sessions to a maximum of 45 min. For that purpose, we divided our experiment in four sessions, in each of which only 144 of the pictures were shown. The sessions were conducted on different days within one week. Experiments were conducted in a darkened room. Participants' only instruction was to "study the images carefully" (Einhäuser & König, 2003). In each session, a 5-min break was given after 72 image presentations and the eye tracker was removed from the subject's head. After the break, tracking was begun again with calibration and valida-

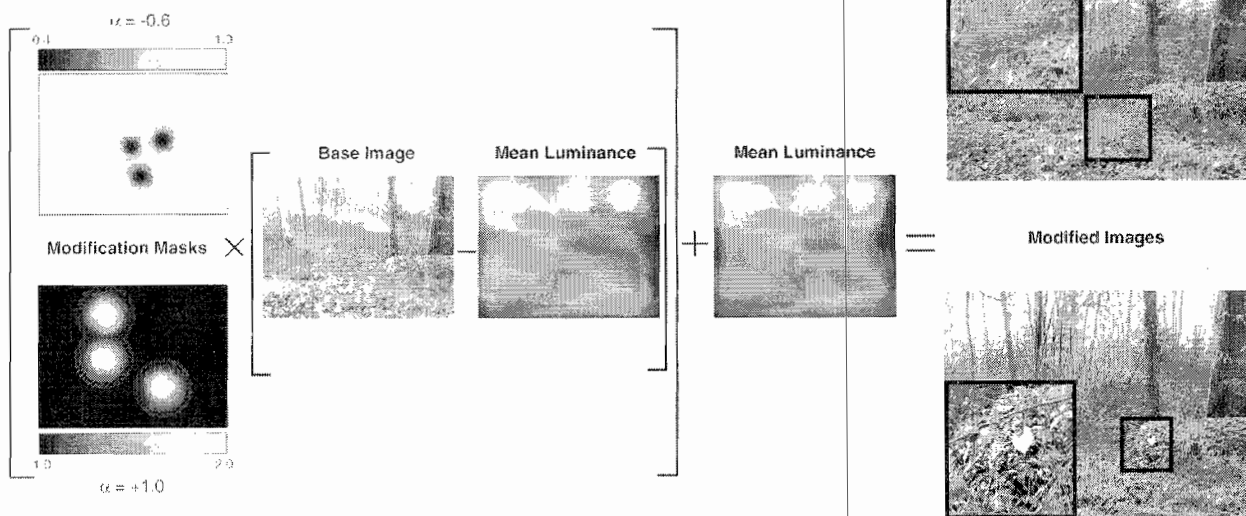


Fig. 2. Illustration of the local luminance contrast modification process for the maximum negative ($\alpha = -0.6$; upper panels) and maximum positive ($\alpha = +1.0$; lower panels) modification levels. The image is local mean luminance normalized (inside the smaller brackets) and then multiplied pixel wise with the mask in which the Gaussians are embedded (inside the larger brackets). The addition of the mean luminance map back to the resulting map returns the modified images. Please note that after this process pixel values smaller than zero are set to zero, and the pixel values larger than 255 are set to 255.

tion. At the end of the fourth session, the participants were informed about the purpose and details of the study.

2.5. Luminance contrast and texture contrast

The image features of interest were luminance contrast (LC) and texture contrast (TC). For LC, we employed the definition of Reinagel and Zador (1999), according to which LC at a given point is the standard deviation of luminance in a square region surrounding this point, divided by the mean luminance of the image. In the present study, the length of the square region was 40 pixels or 1.1° . Parkhurst and Niebur (2003) found the most striking effects of luminance contrast for 1° patches, which is very close to our choice. Furthermore, the images were low-pass filtered (Einhäuser & König, 2003) at different cut-off frequencies (1.42, 0.71, 0.35, and 0.18 cycles per degree, CPD), and the LC was computed on these versions as well. The cut-off frequency is defined as one standard deviation of the Gaussian kernel with which the image was convolved. Following Einhäuser, Rutishauser, et al. (2006) we generalized the definition of LC to TC, i.e. TC was calculated as the standard deviation of LC in a square region, divided by the mean LC of the image. We tested different patch sizes but our primary analysis concerned a patch length of 200 pixels (5.5°). This scale faithfully covers one modified region and hence is sensitive to the influence of LC alterations on TC. Fig. 1b and 1c show examples of LC and TC maps for each category.

2.6. Actual and control fixations

The main aim of this study is to characterize the actually fixated points by their feature values and the changes in those feature values that are a result of the LC modifications. This requires the generation of a set of control points in order to make comparisons. Previous research conclusively have shown that due to biases observed during scene viewing, such control points cannot be sampled randomly from the whole image (Parkhurst, Law, & Niebur, 2002; Tatler, 2007; Tatler et al., 2005). For instance, the central bias, the tendency to fixate central regions of a displayed scene more often than the peripheral regions, has been shown to be independent of image content and the location at which the viewing starts (Tatler, 2007). Here in order to make comparisons to the actually fixated points, we choose the so-called 'control fixations' according to the type of analysis we employ.

In the case of the analyses on unmodified images, fixations measured during the presentation of an image (actual fixations) were compared to the fixations gathered from the rest of the images in the same category (Fig. 3a). This controls for any viewing bias and is especially important given the inclusion of a *faces* category. As the images in this category contain uniform background regions, randomly chosen control points and fixations from other categories would mostly fall on these areas of the image, which would then lead to an overestimation of effect. Thus, for a given image feature map, the values encountered at actually fixated points were compared to values encountered at control fixations that came from the other images in the same category.

The control fixations for modified images were defined differently. While investigating the effects of modifications on fixation behavior, one has to quantify whether there are fixations in these modified regions more than expected by chance. In this case we took as control points all the fixations that come from the same image at different α -levels (Fig. 3b). This is justified since every modified image has a unique modification mask (Fig. 2). This controls for image content related biases, isolating the effects of modifications. To sum up, for a given modified image, values encountered at the actual fixations were compared to the values encountered

at the fixations that came from the presentation of the same basis image at the remaining α -levels.

Furthermore, whenever an analysis was performed on fixations occurring only within a certain time window of presentation, the same temporal criterion was applied to the generation of control fixations.

2.7. Data analysis

Most of our analyses concerned whether the feature values at actual and control fixations differed. Following Tatler and colleagues (2005), we employed the area under the Receiver Operating Characteristic (ROC) curve as a measure to quantify how well a given feature (or the change in a feature) discriminates between actual and control fixations. The advantages of this measure include being non-parametric, having upper and lower bounds (0.5 for random discrimination and 1.0 or 0 for perfect discrimination, depending on whether actual or control feature values are higher, respectively), and the ability to compare different area under curve (AUC) values gathered from different types of analysis. In order to allow statistical inference, we calculated 95% bootstrap confidence intervals (CIs) for the AUC measures as described by Tatler et al. (2005).

In order to address the correlation between fixation probability and a given feature, we used the data gathered from the presentations of unmodified images. For each image, the corresponding actual and control fixations were laid upon the feature map and the values at those locations were gathered. After the repetition of this process for each image, the actual and control feature values were concatenated separately. Thus, computing the AUC with actual and control feature values reveals the correlation between features and fixation likelihood.

Another matter of concern was whether the modifications had an effect on overt attention. For this analysis, the values in the modification mask at the actual and control fixations of modified images were used. Again, pooling the data over images led to the actual and control modification distributions. Together with the computation of the AUC, the similarity of these distributions was quantified using a two-sided Kolmogorov–Smirnov (KS) test too. In eye-tracking data, two subsequent fixation points are not independent, because the location of a fixation is affected by the previous location of the gaze. To address this violation of KS-test assumption, we computed an effective sample size by examining the autocorrelation function of the vertical and horizontal locations of successive fixations (see Einhäuser & König, 2003 for a detailed account). We decreased our sample size accordingly by a factor of two and used these reduced samples in the KS-tests. To sum up, if the modification mask values at actual fixations differ from those at control fixations, AUC would deviate from 0.5 and the KS-test would result in significant p -values.

Furthermore, in order to demonstrate the role of modifications on fixations for each image separately, we have repeated the AUC analysis before the pooling step. The computation of this modification effect for each image allows us to investigate factors that can predict whether the fixations on a given unmodified image will change with the arrival of LC changes.

Modifications of the type used in the present study have effects not only on LC values, but on TC values as well. As such, unintended and inevitable increases in TC as a result of both LC increases and decreases have the potential to explain the observed effects. This issue was addressed as follows: For points fixated on modified images, TC was calculated once on the underlying modified image and once on its unmodified counterpart. The difference of these two maps, which we will call ΔTC , has non-zero values only at those regions where TC was changed as a result of modification. By using the values in ΔTC at the actual and control

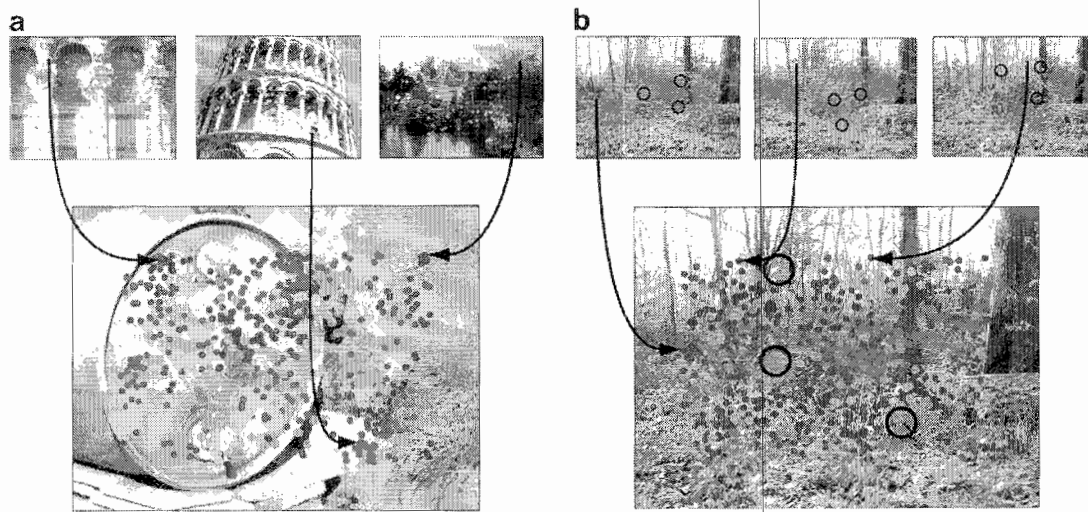


Fig. 3. Selection of actual and control fixations. (a) Unmodified image fixations. Actual fixations, the fixations that are recorded during the viewing of a given image, are shown as green circles for different images. Control fixations (shown as red circles) for a given image (in this example the larger image) are the fixations that are recorded during the viewing of other images in the same category. Accordingly, the actual fixations for a given image serve as a subset of control fixations for the other images in the same category. This is illustrated with the arrows connecting the same fixation between its actual and control conditions. For illustration purposes, here we only show three sets of control fixations that come from three different images. Please note that since there are 16 images per category, eventually 15 images provide control fixations for any given image. (b) Modified image fixations. Actual fixations, the fixations that are recorded during the viewing of a given modified image, are shown as green circles for different images. Control fixations (shown as red circles) for a given modified image (in this example the larger image) are the fixations that are recorded during the viewing of the same basis image at different modification levels (α). Accordingly, the actual fixations for a given modified image serve as a subset of control fixations for the modified generated from the same basis image. This is justified since every modified image has a unique modification mask, as shown by different regions that are modified (black circles). The arrows connect the same fixation between its actual and control conditions. Again we only show three images, but since we have nine modification levels (including $\alpha = 0$), eight images provide control fixations for a given modified image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

fixations for modified images, we again computed the AUC. As an attempt to explain the changes in fixation behavior with alterations in TC, the relation between ΔTC -AUC and the modification-AUC was quantified by linear regression.

Do modifications at the same α -level influence image statistics similarly for all images and categories? To address this question we first standardized all LC maps by converting all values to z-scores. In order to consider both LC increases and decreases as deviations from the mean, the absolute score of the standardized map was taken. These operations are captured by Eq. (2):

$$LC_{norm} = \left| \frac{LC - \mu}{\sigma} \right| \quad (2)$$

LC stands for the given LC map; μ and σ correspond to the mean and standard deviation of this LC map, respectively. The difference of the LC_{norm} maps of the modified and unmodified images was taken. From this ΔLC_{norm} map, distributions of the values that were found within the half maximum of the modification kernels were computed. The percentile-based properties of these distributions were compared across categories and α -levels, in order to quantify to what degree the altered LC values deviated from the image mean.

Finally, we addressed whether category and α -level differences in the influence of modifications on fixation behavior could be explained by the above explained image statistics analysis. AUC was computed comparing the values in the ΔLC_{norm} maps at actual and control fixations for modified images. This was again done separately for each category and α -level. A match between the two types of ΔLC_{norm} analyses would suggest that the influence of modifications on fixation behavior is explained by low-level statistics alone, without any role for high-level factors.

All analyses were restricted to a 784×528 pixels region in the center, so that the features of interest could be computed faithfully, without the introduction of boundary effects.

2.8. Image features from local phase

For the quantification of the total amount of scene configuration, the local phase of unmodified images was analyzed. This type of analysis requires a local frequency transformation and a subsequent analysis of the phase relations between different frequency components. Phase congruency is a measure of how similar the phases of different frequency components of a signal are, and identifies lines, edges, and corners (Kovesi, 2003). For instance, in the case of a step function, all components will be in phase at the step, revealing an edge. Phase symmetry reveals symmetrical and asymmetrical luminance patterns by looking at arrangements of phase. An example would be the mid-point of a rectangular wave where all components will be either at their maximum or minimum, revealing symmetry. In this work, Peter Kovesi's (1997, 1999, 2000a) phase congruency and symmetry definition and algorithms (Kovesi, 2000b; available at <http://www.csse.uwa.edu.au/~pk/research/matlabfns/>) are used. These measures are appealing because they are dimensionless (values range from zero to one), contrast-independent, and robust against noise. Local frequency content is analyzed by applying a set of oriented log Gabor wavelet filters of different scales to the image. Here we used 2-octave bandwidth filters at 6 scales, with the smallest scale filter having a wavelength of 8 pixels (0.22°), and for two successive filters this was increased by a factor of two (Fig. 1d–e). The rest of the parameters in the phase-congruency computation were kept the same as in Kovesi's original work (Kovesi, 1997, 2000b, 2003), but it is worth noting that these parameters are known not to influence the results (Kovesi, 1999). Phase congruency itself is a mid-level, local feature due to its computation at each pixel in the image. Nevertheless taking the average phase congruency of the whole image corresponds to a global analysis, where the image is assigned a signature defined by the total amount of phase-related features it includes. These average phase feature values were correlated with modification effects by applying linear regression.

The procedure was repeated for average LC and TC. As such, the global phase or contrast-related features were tested for their effectiveness in predicting the modification effect.

3. Results

3.1. Luminance contrast (LC)

We first compared the mean luminance contrast (LC) of basis images across categories. To calculate the mean LC of a basis image, an image-specific control fixation set was generated. This involved taking the fixations of all participants to all basis images in the same category, except the image of interest. The mean LC for an image was defined as the average of the LC values of the image at those control points. For our categories, the grand mean values were 0.29 (± 0.07 , standard deviation), 0.22 (± 0.08), 0.34 (± 0.18), and 0.15 (± 0.04) for *naturals*, *manmades*, *fractals*, and *faces*, respectively. A one-way ANOVA showed that these differences were significant ($F(3, 60) = 9.57$, $p < 0.0001$). Post-hoc tests (Tukey–Kramer tests) revealed that the mean LC of *fractals* is significantly different from the LC means of *manmades* and *faces*. The category differences in these grand means of LC and the high standard deviation in the case of *fractals* reveal that second-order statistics differ from image to image, which must be considered while comparing LC values from different images.

Is the relation between fixation behavior and LC category-specific? To answer this question, for each unmodified image and subject, we compared LC values at actually fixated points to control points obeying same viewing biases as actual fixations. For *naturals*, *manmades*, *fractals*, and *faces* the AUC values were 0.52 (95% lower CI/upper CI, 0.51/0.53), 0.59 (0.58/0.60), 0.57 (0.56/0.58), and 0.59 (0.58/0.60), respectively (an AUC value of 0.5 indicates chance level; values above 0.5 indicate feature is attracting attention, and values below 0.5 indicate feature is repelling attention). This very small correlative effect of LC in the case of *naturals* was not unexpected. Einhäuser and König (2003) failed to find any significant effect for similar images when LC was computed at presentation resolution, but managed to do so when LC was computed on low-pass filtered versions of the same images (please note that low-pass filtered images were never shown to participants). We performed the same analysis on low-pass filtered images with cut-off frequencies of 1.42, 0.71, 0.35, and 0.18 CPD and observed that the AUC for *naturals* improved (the AUC measures for these respective frequencies were: 0.55 (0.54/0.56), 0.55 (0.54/0.56), 0.56 (0.54/0.57), 0.55 (0.53/0.56) for *naturals*; 0.58 (0.57/0.59), 0.57 (0.57/0.59), 0.56 (0.55/0.57), 0.54 (0.53/0.55) for *manmades*; 0.57 (0.56/0.59), 0.57 (0.56/0.58), 0.56 (0.55/0.57), 0.54 (0.53/0.55) for *fractals*; 0.57 (0.56/0.58), 0.56 (0.55/0.57), 0.53 (0.52/0.54), 0.50 (0.49/0.51) for *faces*). For further analysis and visualization, we chose the cut-off frequency of 0.71 CPD, as its use resulted in the most similar amounts of mean actual LC values higher than their control counterparts (Fig. 4). Although the overall size of the effect is dependent on image category, when only lower spatial frequencies are taken into account, the effect of LC is actually comparable across image categories.

3.2. Texture contrast (TC)

Next, we investigated the effects of the second-order contrast, namely texture contrast (TC), identically to the analysis of LC, again using the unmodified images only. Generalizing our LC definition, we defined TC as the deviation of LC in local patches, normalized by the mean LC of the image under consideration (Einhäuser, Kruse, et al., 2006). We computed TC in different patch sizes (patch lengths of 2.2°, 3.3°, 4.4°, and 5.5°). The largest patches were cho-

sen such that they match the size of modified regions. The fixation-based analyses were identical to the LC measures introduced above. The AUC values for *naturals* in increasing patch-size order were 0.56 (0.55/0.57), 0.55 (0.54/0.56), 0.54 (0.53/0.55), and 0.54 (0.53/0.55). Total scale-independence was observed for the rest of the categories, and for *manmades*, *fractals*, and *faces* the AUC values were 0.54 (0.53/0.55), 0.51 (0.50/0.52), and 0.51 (0.50/0.52), respectively. For display purposes (Fig. 4b) we have chosen 5.5° patches, since we will be using this scale in the analysis of modification effects. Thus, even though TC is correlated with fixated points less than LC, for all image categories used in this study, its effect is remarkably scale-invariant.

Since the definition of TC is based on LC, one might expect a correlation between these features. To address this issue we pooled the control fixations for each category separately and computed the correlation coefficient of LC and TC values encountered at those points. For LC computed on original images, the means (standard deviation) of correlation coefficients for different images were found to be 0.27 (± 0.24), 0.41 (± 0.16), 0.09 (± 0.39), and 0.40 (± 0.13) for *naturals*, *manmades*, *fractals*, and *faces*, respectively, when the TC was computed in patches of 2.2°. The mean correlation coefficients dropped to 0.15 (± 0.34), 0.25 (± 0.16), -0.04 (± 0.37), and 0.24 (± 0.14) when the TC was computed in patches of 5.5°. Similar reductions in correlation coefficients as a result of TC scale increase were also observed for LC computed on low-pass images (0.71 CPD). Thus, even though LC is correlated with second-order contrast computed at small scales, these correlations drop when the scale is increased to the level that will be used in further analysis.

3.3. Effects of modifications on image statistics

Before analyzing fixated points, we first characterized the effects of modifications on overall image statistics. For each modified image, we calculated the mean LC and TC at control points, as explained above. These means were divided by the mean feature value of the corresponding unmodified image and converted to percentages. From the maximum negative modification ($\alpha = -0.6$) to the maximum positive modification ($\alpha = +1.0$), the mean LC was observed to lie between 90% ($\pm 1\%$, standard deviation) and 114% ($\pm 2\%$) of the mean LC of the basis images for *naturals*, 91% ($\pm 2\%$) to 113% ($\pm 4\%$) for *manmades*, 88% ($\pm 4\%$) to 112% ($\pm 5\%$) for *fractals*, and 87% ($\pm 5\%$) to 112% ($\pm 6\%$) for *faces*, as a monotonic function of α . Thus, any given modification level had very similar effects on the mean LC for all categories. The change in mean TC, on the other hand, was category specific and characterized by variability. From negative to positive modification, the percentages changed from 109% ($\pm 9\%$) to 118% ($\pm 11\%$), 97% ($\pm 4\%$) to 107% ($\pm 5\%$), 103% ($\pm 12\%$) to 107% ($\pm 13\%$), and 96% ($\pm 5\%$) to 109% ($\pm 7\%$), respectively. Even though the same modification level had very similar and expected effects on LC for all categories, TC was altered in a category-specific manner, and differences were found both in the magnitude and direction of TC changes.

3.4. Effects of modifications on fixation selection

The next step was to analyze the effects of modifications on fixation selection. For each peak modification level and category, the distribution of modifications at actual fixations was compared to that at control points, again using the AUC measure. Fig. 5a shows AUC as a function of modification level for each category. The figure is arranged such that points above 0.5 indicate the attractive effects of modifications, and points below zero attention-repelling effects (i.e. modifications at control fixations are higher). Fig. 5b shows the category means of image-specific AUC values. The two curves are nearly identical, suggesting that pooling fixations over

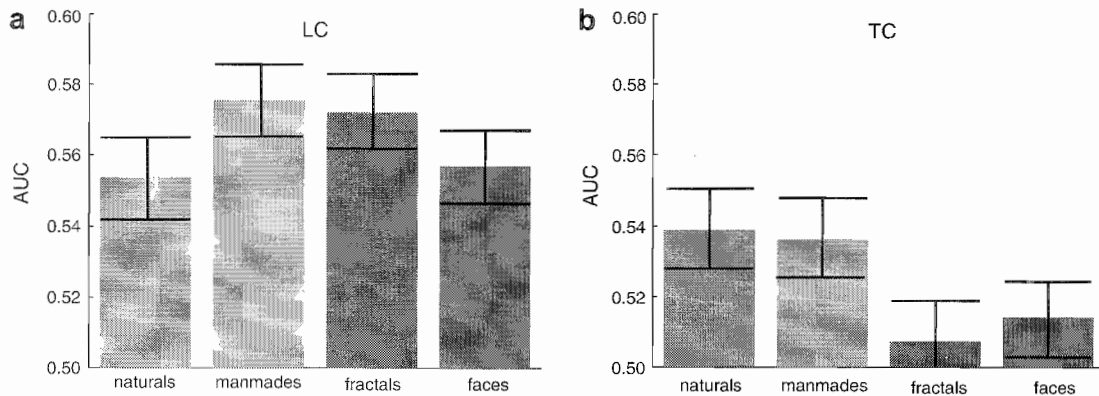


Fig. 4. Luminance contrast (LC) and texture contrast (TC) and fixation relationship. (a) Area under ROC curve (AUC) measures of LC for the four categories. These LC maps are computed from the low-pass filtered versions of our image with a cut-off frequency of 0.71 CPD (defined as the standard deviation of the Gaussian filter used for the filtering) using 1.1° patches. Results on LC computed from original images and other low-pass filtered images at different cut-off frequencies, see text. In original AUC values greater than 0.5 indicate higher LC values at fixated points than at other points in the image. The error bars indicate 95% bootstrap confidence intervals (CIs). (b) AUC measures of TC computed from 5.5° patches. In the case of *fractals* the confidence interval includes 0.50. For results on TC computed at other scales, see text.

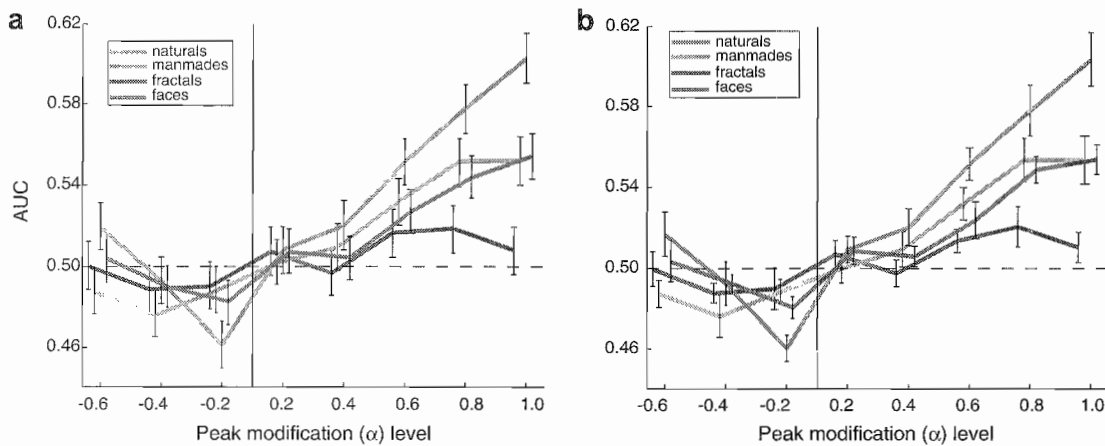


Fig. 5. Effects of LC modifications of varying magnitude and direction on fixation behavior. (a) AUC measure as a function of peak modification level (α) is plotted separately for each category. Error bars denote 95% bootstrapping CIs. The plot is scaled such that when modifications attract attention, AUC is above 0.5 and when they repel attention AUC is below 0.5. (b) The category means of the AUC measures computed separately for each image. Error bars denote the standard error of the mean. Please note the good agreement between the two measures, even though the former measure pools the fixations over images, and the latter keeps them separate and then takes later the mean. The data is moved slightly along the x-axis to avoid the overlap of error bars.

images do not have a crucial influence on the metrics of interest. There are several noteworthy properties of these curves. The *fractals* curve is flat, negative modifications have an influence in the case of *naturals* only, and this effect is rather weak, and the effects of positive modifications are category dependent. In order to check the significance of these observations directly, we follow Einhäuser and König (2003) and perform KS-tests. In the case of *fractals*, KS-statistics do not reach significance ($p < 0.05$) for any of the modification levels. For other categories, on the positive modification side (LC increase), the KS-test gives significant results for the three highest modification levels ($\alpha = +0.6, +0.8, +1.0$), but the p -values change according to category and magnitude of modification (for *naturals*: $p < 10^{-4}, 10^{-8}, 10^{-9}$; *manmades*: $p < 0.05, 10^{-4}, 10^{-4}$; *faces* $p < 0.05, 10^{-3}, 10^{-4}$, for increasing α). The results were different for LC decreases, with an absence of effect ($p > 0.05$) for all categories except *naturals*. In this case, minor LC decreases ($\alpha = -0.2$) repelled attention ($p = 0.001$) and major decreases ($\alpha = -0.6$) attracted attention ($p = 0.016$), even though the latter effect is weak, as shown by both the low AUC of 0.52 (0.51/0.53) and the borderline significance of the KS-test. To summarize, except for

fractals (where none of the modifications changed fixation behavior), positive modifications attracted attention; and high negative modifications were attractive only in the case of *naturals*.

3.5. Effects of time

Do the observed effects of features on fixation selection remain constant over time? To monitor the effect of time we performed sliding window analyses for LC, TC, and the maximum positive modifications ($\alpha = 1.0$). For this analysis actual and control fixations were chosen from the same time intervals of the stimulus presentation. Other criteria for control fixation selection remained the same. The windows were of one-second length with window overlap of half a second. In the case of *naturals*, the LC-AUC is highest in the first second of presentation (Fig. 6a). For other categories, however, the effect is either constant or oscillates without a clearly discernible behavior. Results were qualitatively similar for TC (Fig. 6b) and $\alpha = 1.0$ (Fig. 6c), that is, only in the case of the *naturals* the effect was highest in the first second. Therefore, we conclude that except for natural environment images, where the effects

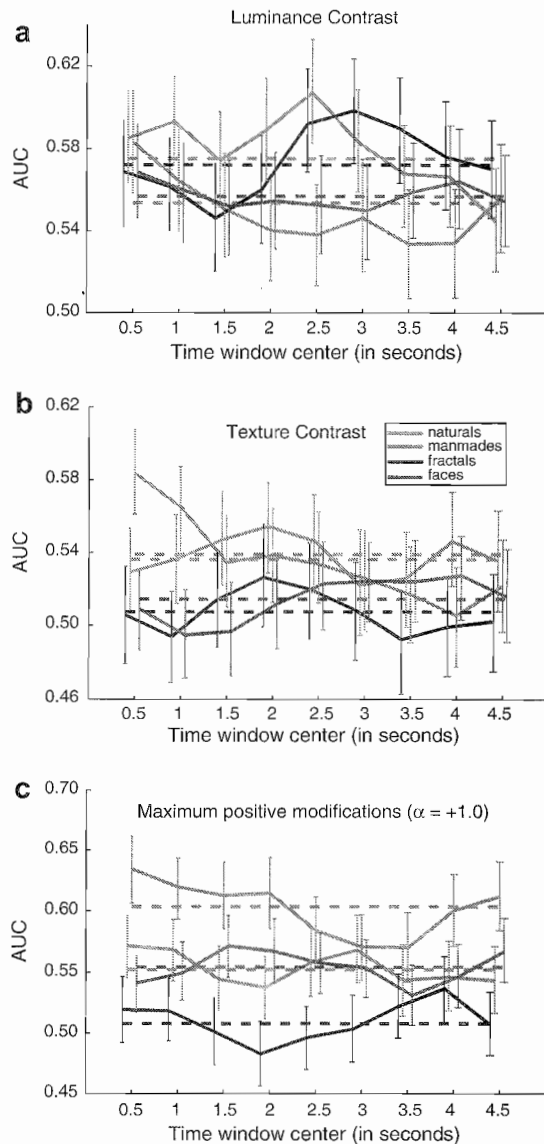


Fig. 6. Effects over time. (a) Sliding window analysis of the AUC measure for LC computed on low-pass filtered images (cut-off 0.71 CPD) and in 1.1° patches. A one-second long window was shifted with a window overlap of 0.5 s and AUC is computed using the actual and control fixations falling in each of those windows. The results are shown as solid lines. The error bars indicate 95% bootstrap confidence intervals. The AUCs computed for the whole presentation duration are plotted as dashed lines, using the same color code. (b) Sliding window analysis of the AUC measure for TC computed in 5.5° patches. All conventions are identical to panel (a). (c) Sliding window analysis of the AUC measure for maximum positive ($\alpha = +1.0$) modifications. All conventions are identical to panel (a) and (b). Please note that all AUC measures start to drop after the first second in the case of *naturals* only.

are largest early during the presentation, there is no difference between the early and late parts of the presentation in terms of the low-level feature–fixation relationship.

3.6. Modification effect and changes in TC

Is it the changes in TC that influence fixation behavior, rather than changes in LC? To answer this question we introduced a ΔTC -AUC measure (see Methods), quantifying how much TC was

increased at the actual fixations compared to the control fixations. In cases where TC is guiding attention, one would expect the ΔTC -AUC to lie above 0.50. In Fig. 7 we plot this TC change effect as a function of the peak modification level separately for each category. For the three image categories where modifications had an effect on fixations, a selective TC increase was found for positive modifications only.

The above analysis could be criticized for the choice of the scale at which TC is computed (5.5°). One could argue that on such a large scale the differences between actual and control points might be underestimated. Even though the large difference between actual and control TC change on the positive modification side speaks against this possibility, we also performed the same analysis with TC calculated at smaller scales (2.2° , 3.3° , and 4.4°). For *naturals* at $\alpha = -0.6$, where modifications attracted attention, there was indeed a difference between the TC values at actual and control fixations, but this difference was in the opposite direction, i.e. ΔTC_C was greater than ΔTC_A . This is probably due to a higher correlation between LC and TC values when calculated in small patches. Thus, changing the scale at which TC is computed did not lead to results indicating a causal relation between TC and overt attention.

3.7. Absolute standardized LC analysis

How can we explain the absence of modification influence on fixations in the case of *fractals*? One possibility is that modifications created regions with LC statistics that deviated from the surroundings in all categories but *fractals*. To address this we first quantified how much the absolute standardized LC was increased at actual and control points as a result of modifications. For this analysis we took the difference between the absolute values of standardized LC maps of modified and original images. We refer to this difference map as ΔLC_{norm} . If during viewing of modified images, the people have selectively attended to those regions, where the LC values had changed such that they landed in the LC distribution away from the central portion, the ΔLC_{norm} values at actual fixations must be higher than the values at control fixations. As seen in Fig. 8a, the AUC computed from these actual and control

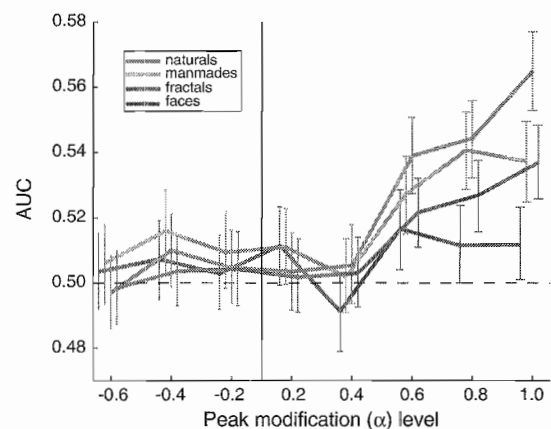


Fig. 7. Changes in TC at fixations as a result of LC modifications. For this analysis the TC maps (5.5° patches) of basis images were subtracted from their modified counterparts, which returns what we call ΔTC maps. ΔTC -AUC is then computed using actual and control fixations for control images identical to the previous analysis with modified images. That is, the only difference is modification masks are replaced by ΔTC -maps. In the figure, ΔTC -AUC is plotted as a function of α separately for each category. If the effects of modifications on fixations could be explained by changes in TC, this figure should be similar to the panels of Fig. 7. But the similarities are observed only for $\alpha > 0$, that is, the slight but significant attraction observed in the case of *naturals* for $\alpha = -0.6$ cannot be explained by ΔTC .

distributions bears close similarities to the effects of modifications. For instance in the case of *fractals*, the fixation locations do not differ from control locations in terms of their normalized LC changes. But this analysis alone does not prove that modifications on *fractals* were ineffective in terms of creating LC values. There can be still modified regions with these properties, but they are not being selected by the attention system. To show that this is indeed the case, from the ΔLC_{norm} map, distributions of the values that were found within the half maximum of the modification kernels were computed. Fig. 8b shows the category mean of the 75th-percentiles computed for each image separately. Obviously, the *fractals* curve is similar to *manmades* and *faces* curves, and at $\alpha = 1.0$ its value is higher than the value of these two categories. This shows that for each category, the LC values changed with modifications fell into similar regions of the LC distributions of the images. Thus, just like ΔTC analysis, this relative LC based attempt to explain category differences with low-level differences, fails too.

3.8. Phase features and modification effect

In order to approximate the total amount of structure in images, we computed local phase-related feature maps for each image for the features of phase congruency and phase symmetry (Kovesi, 1999, 2003), and computed the mean over the map (for example maps see Fig. 1d and 1e). This analysis revealed that *fractals* were characterized by especially high amounts of structure, and *naturals* displayed the lowest values. The category averages of mean phase congruency were $0.019 (\pm 0.011, SD)$, $0.035 (\pm 0.008)$, $0.035 (\pm 0.009)$, and $0.026 (\pm 0.004)$ for *naturals*, *manmades*, *fractals*, and *faces*, respectively. For phase symmetry, the values were $0.039 (\pm 0.025)$, $0.116 (\pm 0.028)$, $0.122 (\pm 0.041)$, and $0.113 (\pm 0.021)$, with the same category order. As can be seen in Fig. 9, image-specific AUC for $\alpha = 1.0$ is inversely related to these two features ($r^2 = 0.34$, $p < 10^{-4}$ for phase symmetry, $r^2 = 0.40$, $p < 10^{-6}$ for phase congruency). Additionally the category averages mostly exhibit the same inverse relationship. Thus, images with high average phase-related features are less susceptible to the effects of modifications.

But do contrast-related features (LC and TC) have a similar relationship to modification effects? This is plausible given that LC is correlated with edges, just like phase congruency. Furthermore, since TC is defined as the standard deviation of LC, it can have a scaling effect on the influence of LC modifications. In images with low TC and a relatively narrow range of LCs, the modifications could result in LC values not encountered in the basis image more often than in the case of images with high TCs and wide ranges of LC values. In the case of global LC, neither LC at original resolution nor LC at 0.71 CPD cut-off low-pass filtered images had a significant relation to modification effect. TC had a small but significant relationship with the modification effect of $\alpha = 1.0$ only ($r^2 = 0.15$, $p = 0.002$). But *fractals* coming second after *naturals* in the category average of global TC show that this measure does not capture the striking absence of modification effect in the case of *fractals*. Compared to the results obtained with phase-related features, global LC and TC are not predictive of the effects of modifications.

4. Discussion

In order to study the correlational and causal roles of luminance contrast (LC), we measured the eye movements of human observers as they viewed unmodified or LC-modified images from four different categories, depicting natural environments, urban scenes, fractal images, or human faces. Our results for unmodified images confirmed earlier findings showing higher LC values at fixated points when compared to control points (Krieger et al., 2000; Parkhurst et al., 2002; Rajashekar, van der Linde, Bovik, & Cormack, 2007; Reinagel & Zador, 1999). Interestingly, the correlation between fixation and LC value varied across categories when LC was computed on the image at its original resolution – LC was correlated with fixation locations in all image categories except natural scene images. However, we found that the influence of LC became independent of image category after low-pass filtering the images. As a result of low-pass filtering, LC explained fixation locations better in the case of natural scene images, and a slight decrease in the correlation of LC to fixations was observed for urban

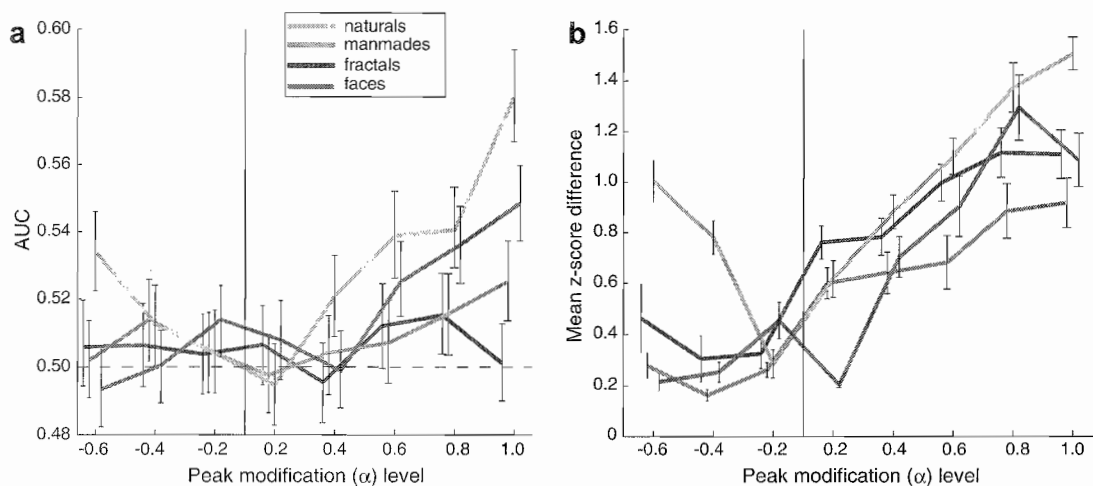


Fig. 8. Absolute z-score analysis as an attempt to explain observed effects of LC modifications. (a) AUC analysis for the change in absolute z-scores of LC. For this analysis absolute of the z-score transformed LC maps of basis images were subtracted from their modified counterparts. Then the same analysis in the case of modifications and ΔTC was performed. This analysis demonstrates that the normalized change in LC in either direction correlates well with the effect of LC modifications on fixation behavior. Error bars denote 95% CIs for the AUC. (b) The changes in normalized LC inside the modification kernels. Again, absolute of the z-score transformed LC maps of basis images were subtracted from their modified counterparts. Next, the regions in these maps that fell into the half maximum of the modification kernels were taken. Then the probability density functions for the values inside these regions were computed. Plotted are the 75th percentiles of those distributions. As can be seen, high positive modifications lead to the appearance of new LC values that deviate from the mean LC values in the case of *fractals* as well. Error bars denote standard error of the mean. As such, this figure illustrates the hypothetical scenario of what would happen if the fixations were drawn to locations that now have LC values that deviate highly from the mean LC of the image, which is a result of LC modifications. But panel (a) shows that the data do not fit this hypothetical scenario.

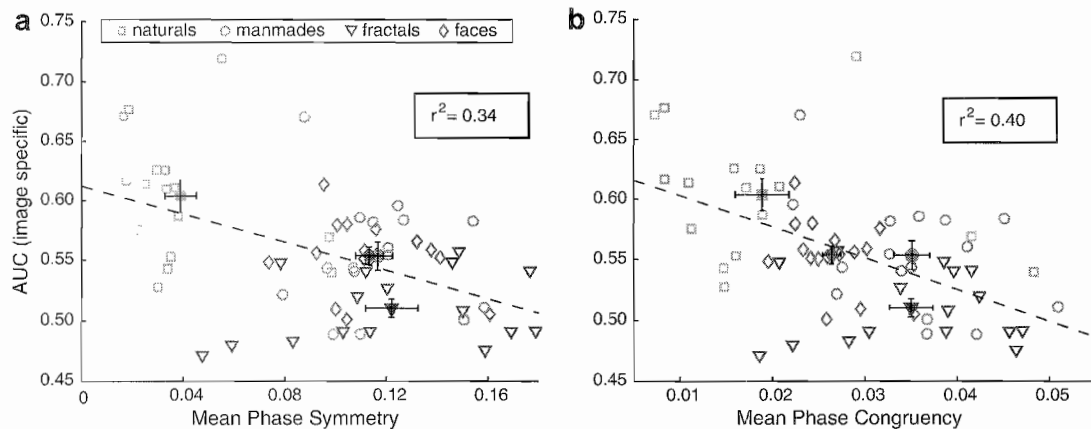


Fig. 9. Phase features and modification effect relationship. (a) The modification effect for each image is plotted against the mean phase symmetry of the same image. Empty symbols denote individual images; filled symbols denote the category averages. The linear regression fit to the data is plotted as a dashed line ($r^2 = 0.34$, $p < 10^{-5}$, 95% CI 0.15/0.53). The error bars display the standard error of the mean for both data dimensions. (b) Same as (a) but for phase congruency ($r^2 = 0.40$, $p < 10^{-4}$, 95% CI 0.21/0.58). The category averages mostly follow the pattern observed for the individual images.

scene images. Rajashekar et al. (2007), using natural scene images very similar to ours also observed that LC computed at original resolution was less correlated with fixations than LC computed at space-variant low-pass filtered versions. Reminiscent of the decrease observed in the case of our urban scenes, Tatler and colleagues (Tatler et al., 2005), using similar scenes with man-made objects in them, have shown that high-frequency content was more important for fixation selection. As such, for this image category, it remains to be investigated whether the larger effect size obtained in the presence of high-frequency information is causal, or correlational only. However, as fixation targets at higher eccentricities are not represented in the visual system in full resolution, but rather spatially low-pass filtered (Séré, Marendaz, & Héroult, 2000), it would be not surprising that fixation selection preferentially makes use of low frequencies. Although the cut-off frequency for eccentricities corresponding to the average saccade length of our study ($\approx 3.5^\circ$) is found to be higher in recognition and visual acuity paradigms (Séré et al., 2000) than in our case, the attention system might still consider reasonably low frequencies only, in order to cover the full range of saccade lengths. To conclude, even though it is still possible that the spatial frequencies of interest change while viewing images from different categories, at least in our data, correlation between fixation and LC is very similar for all the categories tested if the spatial scale of LC reflects the resolution of peripheral vision.

Tatler et al. (2005) argue that the seemingly increased saliency effects found in the initial phase of stimulus presentation (Parkhurst et al., 2002; Reinagel & Zador, 1999) can be attributed to methodological artifacts. They showed that early fixations differ from late fixations regardless of the image being viewed, and concluded that an appropriately chosen baseline is needed to deal with this temporal bias, as was done in our analysis. The correlation of fixation to LC, TC, and positive modifications decreased over time for natural scenes, but remained stable or fluctuated non-monotonically in other categories. This does not contradict the results of Tatler et al. (2005), who used mainly urban scenes and home interiors in their study. Interestingly, Kayser et al. (2006) found no effect of time on macaque fixations, using a stimulus set of animal pictures, and natural scenes like ours. It has yet to be tested whether this is due to species differences (Einhäuser, Kruse, et al., 2006) or to the particularities of the stimuli used. Overall, the LC effect remains stable over time for most image

types, but drops in the case of natural scene images, whose possible special properties are discussed below.

In order to test the assumption of the classical saliency map models (Itti & Koch, 2000; Parkhurst et al., 2002) that LC is causally related to overt attention, we locally increased or decreased LC (Einhäuser & König, 2003; Einhäuser, Kruse, et al., 2006). If LC is causally related to fixation selection in a linear manner, increases should draw attention and decreases should repel it. The former prediction was validated for all categories but *fractals*, given an adequate level of modification. LC decreases, on the other hand, had no significant effect on fixation behavior, except during viewing of natural scenes. In that case, we observed repulsion when the LC decrease was minor and attraction when the decrease was of high magnitude. We first attempted to explain the observed effects with respect to changes in local features, namely LC and texture contrast (TC), the latter being a mid-level feature that has previously been suggested to sufficiently explain the attractive effects of LC decreases (Parkhurst & Niebur, 2004). In line with the results of Einhäuser, Kruse, et al. (2006), we failed to find a selective TC increase at points fixated during the viewing of LC-decreased images. Einhäuser, Kruse, et al. (2006) suggested that this might necessitate the consideration of even higher-level properties. Whether these higher-order properties are image-related (bottom-up) features or top-down influences – or indeed, if such a distinction is meaningful – will be discussed below. At this point, we conclude that the changes in fixation behavior during viewing of modified images are at least partly due to high-level properties such as top-down influences or changes in the global scene configuration.

But how noticeable are the luminance contrast modifications of the type discussed here? In Einhäuser and König (2003), with a second experiment it was shown that modifications in the case natural scene photos could be detected, even if the magnitude of the modification was relatively low. Even though in the current study we did not perform such an experiment, most subjects reported at the end of the experiment that they did not notice the changes in the case of fractal images. But as shown by the z-score analysis, modified regions in that case were still different from the unmodified regions in terms of their luminance contrast statistics, and this difference was comparable to the differences obtained with the other image categories tested. The relation of the noticeability of the modifications to their effectiveness in guiding attention is interesting and should be addressed in future work.

The category and image dependence of modification effects in the current study poses the question of whether one can predict that modifications will attract attention based on image statistics. Here, this issue was approached by correlating the modification effects with several average feature values in the image. Whereas the average LC and TC of images were uninformative about the extent of the modification effect, a negative correlation was found for phase congruency and for phase symmetry, features that are computed from local phase spectrum analysis (Kovesi, 1997, 1999, 2000a). These contrast-invariant metrics reveal image features such as lines, edges, symmetries, and blobs, which are the building blocks of mid-to-high-level and meaningful vision. Furthermore, phase scrambling, a method that is used to remove global scene configuration (Einhäuser, Rutishauser, et al., 2006; Kayser et al., 2006; Wichmann et al., 2006) obviously removes most of the correlations between the phases of different frequency components. We argue that the average phase-related feature response is related to the amount of global structure and regularities in the image, and if such structure is present in the scene, as in the case of fractals, the changes in low-level features such as LC will be ineffective in directing attention.

Although purely bottom-up models have been influential in research aiming at predicting fixation locations, in recent years (Itti & Koch, 2001; Parkhurst et al., 2002), there is a growing interest in top-down aspects (Einhäuser & König, 2003; Hayhoe & Ballard, 2005; Henderson et al., 2007; Hochstein & Ahissar, 2002; Land & Hayhoe, 2001; Land & Tatler, 2001; Navalpakkam & Itti, 2005; Pomplun, 2006; Rao, Zelinsky, Hayhoe, & Ballard, 2002; Torralba, 2003; Tsotsos, Culhane, Wai, Davis, & Nuflo, 1995; Underwood, Crundall, & Hodson, 2005). In the present study, we employed different techniques to explain the effects of low-level image alterations on human fixation behavior. The failure of contrast-related features and the success of phase spectrum analysis lead us to the contemplation of a framework for overt visual attention that can describe the observed results by combining aspects of top-down and bottom-up models. We argue that overt attention is guided by a hierarchy of features, in which higher-order properties have precedence over lower ones (see Tsotsos et al., 1995 for the implementation of a similar idea). We see the highest-order features as being identical to top-down influences such as memory or expectations. In the presence of higher-order properties, the lower-order features have only a correlative effect. In the absence of high-level features, on the other hand, lower-order features guide attention to some degree, i.e. they are in fact causally related to fixation selection. This is certainly the case when it comes to natural scenes that are characterized by bushes, undergrowth or branches, where the global scene configuration lacks diagnostic properties. The low average phase congruency and symmetry values in these images support this argument. Fractal images, on the other hand, contain symmetries (for the relation between symmetry and human fixation selection, see Privitera & Stark, 2000) and are characterized by repetitive and recursive patterns, which is partly demonstrated here by the high phase alignment of local spatial frequency components. These properties are exactly what we mean by high and mid-level features. Accordingly, earlier assumptions that fixation behavior in the case of fractals is bottom-up dominated (Parkhurst & Niebur, 2003; Parkhurst et al., 2002) must be reconsidered. Such features also exist in urban scenes and images of faces, albeit to a lesser degree than in the case of fractals. In the former category, these features are constituted by a clear distinction between sky and ground, together with the presence of vertically and horizontally directed edges – as revealed in the high power found in the cardinal directions of the images' power spectra (Torralba & Oliva, 2003). When it comes to faces, the vertical symmetry, the standard orientation of the face, and the alignment of parts (eyes, mouth etc.) correspond to the highest

features, as revealed by categorization tasks (Donnelly, Humphreys, & Sawyer, 1994) or by powerful illusions (Thompson, 1980). This hierarchical framework, where higher-order features have priority, combines aspects of the attention models stressing bottom-up (Koch and Ullman (1985); Itti & Koch, 2001) and top-down (Hochstein & Ahissar, 2002) influences, and matches the results of the present study.

Obviously, the greatest challenge for future research is to uncover the principles governing fixation selection during viewing of images that differ in their higher and lower-level feature content. Recently, Einhäuser, Rutishauser, et al. (2006) showed that the correlative effect of LC is inversely proportional to the amount of phase noise that was artificially added to images. The findings indicated that the correlative effects of low-level properties are bound to the higher-order statistics of images. However, to test the hierarchical theory proposed here, one needs to investigate the causal role of low-level features in such stimuli lacking high-order features. The theory predicts that in stimuli like pink noise ($1/f$ -noise), local LC decreases and increases will attract attention, probably even more so than in the case of natural scene images. Thus, employing stimuli differing only in terms of higher-order properties, one can introduce systematic changes in lower-order features and quantify their influence on overt attention.

What could be the neural locus of saliency-related effects as those described here? The brain regions that have been proposed to encode saliency include LIP (Gottlieb, Kusunoki, & Goldberg, 1998), pulvinar (Posner & Petersen, 1990), and the frontal eye field (Thompson & Schall, 2000), none of which, to our knowledge, displays LC-related firing. Contrast invariance increases while moving up in the cortical hierarchy, with V1 changing its activity according to the LC of stimulus, mid-level areas such as V4 showing moderate levels of contrast invariance, and lateral occipital and posterior fusiform being the most contrast invariant (Avida, Harel, Hendler, Ben-Bashat, & Zohary, 2002). Two recent studies investigating brain regions involved in saliency encoding are of particular interest for our study. Using parts of natural scenes as stimuli, Mazer and Gallant (2003) showed that half of the monkey V4 neurons they recorded displayed saliency-encoding behavior. Neurons with receptive fields at the targets of the next saccades increased their firing rate before the saccade onset. Additionally, the activity of some of these neurons changed according to the target identity, suggesting top-down modulation. Gardner and colleagues (Gardner et al., 2005) recorded fMRI signals from humans as they viewed simple grating stimuli, which underwent contrast modifications. After adaptation to a certain contrast level, V1, V2, and V3 activity increased with increasing contrast, and decreased with decreasing contrast. The activity in V4, on the other hand, increased with the magnitude of the change, regardless of the direction of the change. This suggests that V4 encodes for salient changes in stimuli. If spatial changes were coded in a similar way to these temporal changes, V4 would be a good candidate for guiding eye movements to our modified regions, regardless of the direction of modification. Furthermore, projections from FEF modulate firing rates (Moore & Armstrong, 2003) and receptive field properties (Armstrong, Fitzgerald, & Moore, 2006) in V4, which might be the substrate for the top-down modulation. In the absence of signals carrying information about higher-order properties, processing in V4 – an area in which saliency is defined by lower-order features – would lead to the selection of salient regions. It remains to be established how such saliency-encoding regions will act in the presence of powerful highest-order features.

Acknowledgments

We thank Christian Kaul for the improvements in contrast modification code. We are grateful to Clíodhna Quigley for thorough

discussions on an earlier draft of this article and to two anonymous reviewers for comments and suggestions on an earlier version of the paper. This work is related to the Perception on Purpose STREP project and was supported by EU Grant 027268.

References

- Armstrong, K. M., Fitzgerald, J. K., & Moore, T. (2006). Changes in visual receptive fields with microstimulation of frontal cortex. *Neuron*, *50*, 791–798.
- Avida, G., Harel, M., Hendler, T., Ben-Bashat, D., & Zohary, E. (2002). Contrast sensitivity in human visual areas and its relationship to object recognition. *Journal of Neurophysiology*, *87*, 3102–3116.
- Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Visual Research*, *46*, 2824–2833.
- Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proceedings of National Academy of Science USA*, *103*, 449–454.
- Dickinson, S. J., Christensen, H. I., Tsotsos, J. K., & Olofsson, G. (1997). Active object recognition integrating attention and viewpoint control. *Computer Vision and Image Understanding*, *67*, 239–260.
- Donnelly, N., Humphreys, G. W., & Sawyer, J. (1994). Stimulus factors affecting categorization of faces and scrambled faces. *Acta Psychologica*, *85*, 219–234.
- Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, *17*, 1089–1097.
- Einhäuser, W., Kruse, W., Hoffmann, K., & König, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research*, *46*, 1194–1209.
- 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.
- Frey, H., Honey, C., & König, P. (2008). What's color got to do with it? The influence of color on visual attention in different categories. *Journal of Vision*, *8*(14), 6(1–17).
- Gardner, J. L., Sun, P., Waggoner, R. A., Ueno, K., Tanaka, K., & Cheng, K. (2005). Contrast adaptation and representation in the human early visual cortex. *Neuron*, *47*, 607–620.
- Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual saliency in monkey parietal cortex. *Nature*, *391*, 481–484.
- Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*, 188–194.
- Henderson, J. M., Brockmoie, J. R., Castelano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain*. Oxford, UK: Elsevier.
- Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*, 791–804.
- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.
- Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *National Reviews Neuroscience*, *2*, 194–203.
- Kayser, C., Nielsen, K. J., & Logothetis, N. K. (2006). Fixations in natural scenes: Interaction of image structure and image content. *Vision Research*, *46*, 2535–2545.
- Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.
- Kovesi, P. (1997). Symmetry and asymmetry from local phase. In: *AI'97, Tenth Australian joint conference on artificial vision, proceedings – poster papers* (pp. 185–190). Perth, Australia.
- Kovesi, P. (1999). Image features from phase congruency. *VIDERE: Journal of Computer Vision Research*, *1*, 1–26.
- Kovesi, P. (2000a). Phase congruency: A low-level image invariant. *Psychological Research*, *64*, 136–148.
- Kovesi, P. (2000b). MATLAB and Octave functions for computer vision and image processing. School of Computer Science and Software Engineering, The University of West Australia. <<http://www.csse.uwa.edu.au/~pk/research/matlabfns/>>.
- Kovesi, P. (2003). Phase congruency detects corners and edges. In C. Sun, H. Talbot, S. Ourselin, & T. Adriaansen (Eds.), *Proceedings of seventh digital image computing: Techniques and applications* (pp. 309–318). Sydney, Australia.
- 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.
- Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559–3565.
- Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*, 1311–1328.
- Land, M. F., & Tatler, B. W. (2001). Steering with the head: The visual strategy of a racing driver. *Current Biology*, *11*, 1215–1220.
- Mazer, A. M., & 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.
- Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of the frontal cortex. *Nature*, *421*, 370–373.
- Navalpakkam, V., & Itti, L. (2005). Modeling the role of task on attention. *Vision Research*, *45*, 205–231.
- Olmos, A., & Kingdom, F. A. A. (2004). McGill Calibrated Colour Image Database. <<http://tabby.vision.mcgill.ca/>>.
- Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*, 107–123.
- Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*, 125–154.
- Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt attention in natural scenes. *European Journal of Neuroscience*, *19*, 783–789.
- Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*, 2397–2416.
- Piotrowski, L. N., & Campbell, F. W. (1982). A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, *11*, 337–346.
- Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research*, *46*, 1886–1900.
- Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42.
- Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *22*, 970–982.
- Rajasekar, U., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2007). Foveated analysis of image features at fixations. *Vision Research*, *47*, 3160–3172.
- 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.
- Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network-Computation in Neural Systems*, *10*, 1–10.
- Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*, 1–22.
- Séré, B., Marendaz, C., & Héroult, J. (2000). Nonhomogeneous resolution of images of natural scenes. *Perception*, *29*, 1403–1412.
- 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), 4(1–17).
- Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of time and scale. *Vision Research*, *45*, 643–659.
- Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, *9*, 482–484.
- Thompson, K. G., & Schall, J. D. (2000). Antecedents and correlates of visual detection and awareness in macaque prefrontal cortex. *Vision Research*, *40*, 1523–1538.
- Tootell, R. B., Silverman, M. S., Switkes, E., & De Valois, R. L. (1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science*, *218*, 902–904.
- Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A-Optics Image Science and Vision*, *20*, 1407–1418.
- Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network-Computation in Neural Systems*, *14*, 391–412.
- Tsotsos, J. K., Culhane, Y. L., Wai, N., Davis, N., & Nuflo, F. (1995). Modelling visual attention via selective tuning. *Artificial Intelligence*, *78*, 507–547.
- Underwood, G., Crundall, D., & Hodson, K. (2005). Confirming statements about pictures of natural scenes: Evidence of the processing of gist from eye movements. *Perception*, *34*, 1069–1082.
- Vincent, B. T., Troscianko, T., & Gilchrist, I. D. (2007). Investigating a space-variant weighted saliency account of visual selection. *Vision Research*, *47*, 1809–1820.
- Wichmann, F. A., Braun, D. L., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, *46*, 1520–1529.
- Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum Press (Translated by Haigh, B.).