Available online at www.sciencedirect.com
SCIENCE DIRECT
Vision Research 46 (2006) 1194–1209
Vision Research
www.elsevier.com/locate/visres

Differences of monkey and human overt attention under natural conditions

Wolfgang Einhäuser [a,b,*], Wolfgang Kruse [c], Klaus-Peter Hoffmann [c], Peter König [a,d]

[a] Institute of Neuroinformatics, University of Zurich and Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
[b] Division of Biology, California Institute of Technology, Pasadena, CA, USA
[c] Department of Zoology and Neurobiology, Ruhr-University Bochum, Germany
[d] Department of Neurobiopsychology, Institute of Cognitive Science, University of Osnabrück, Germany

Received 3 February 2005; received in revised form 11 May 2005

Abstract

Rhesus monkeys are widely used as animal models of human attention. Such research rests upon the assumption that similar mechanisms underlie attention in both species. Here, we directly compare the influence of low-level stimulus features on overt attention in monkeys and humans under natural conditions. We recorded eye-movements in humans and rhesus monkeys during free-viewing of natural images. We find that intrinsic low-level features, such luminance-contrast, texture-contrast and saliency—as predicted by a standard model, are elevated at fixation points in the majority of images. These correlative effects are not significantly different between species. However, local image modifications affect both species differently: moderate modifications, which are in the range of natural fluctuations, attract overt attention in monkeys significantly stronger than they do in humans. In addition, humans show a higher inter-individual consistency regarding which locations they fixate than monkeys, in spite of the similarity for intrinsic low-level features. Taken together, these data demonstrate that—under natural conditions—low-level stimulus features affect attention in monkeys and humans differently.

Keywords: Saliency-map; Natural visual stimuli; Attention; Eye-movements; Luminance-contrast

1. Introduction

Rhesus monkeys are the experimental animal of choice to study the physiological basis of human visual attention. In these studies it is assumed that humans and monkeys share common underlying mechanisms of visual processing. Many aspects of the anatomy and functional organization of the visual system are indeed similar in the two species (Astafiev et al., 2003; Van Essen, Drury, Joshi, & Miller, 1998). In a speed task categorizing novel complex stimuli, monkeys perform with only slightly decreased accuracy, but greater speed, than human observers (Fabre-Thorpe, Richard, & Thorpe, 1998; Thorpe, Fize, & Marlot, 1996). These results reveal a striking similarity between rhesus monkeys and humans regarding rapid processing of natural scenes. It remains to be investigated, however, whether or not this similarity generalizes to prolonged viewing under natural conditions, when attentional processes and eye-movements take effect.

Covert and overt attention are widely studied using well-controlled artificial stimuli. From such studies, several brain regions have been implicated in the encoding of stimulus saliency, i.e., the likelihood that the stimulus will attract attention and possibly subsequent eye-movements. Representations of saliency have been reported in several brain regions projecting directly or indirectly to the oculomotor system, such as the pulvinar (Posner & Petersen, 1990; Robinson & Petersen, 1992) the frontal eye-fields (Thompson, Bichot, & Schall, 1997), the superior colliculus (Horwitz & Newsome, 1999; Kustov & Robinson, 1996; McPeek & Keller, 2002; Posner & Petersen, 1990) and the lateral interparietal cortex (Gottlieb, Kusunoki, & Goldberg, 1998). Evidence for the encoding of saliency

* Corresponding author. Tel.: +1 626 395 8967; fax: +1 626 796 8876.
E-mail address: wet@klab.caltech.edu (W. Einhäuser).

0042-6989/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2005.08.032

has also been found in visual areas as early in the visual pathways as primary visual cortex (Li, 2002).

Adopting a free viewing visual search task that used fragments of natural scenes, Mazer and Gallant (2003) claim that saliency is encoded in V4 of rhesus monkeys. Area V4 encodes a variety of stimulus features and activity in this area is strongly modulated by attention. According to Gallant and Mazer, saliency is computed across various brain regions and area V4 serves to integrate information from both higher (top–down processing) and lower visual (bottom–up processing) areas. A key question concerns how top–down scene interpretation and bottom–up stimulus properties are weighted against each other under the more natural viewing conditions in which no explicit (search) task is involved.

Theoretical models of bottom–up attention are frequently based on the concept of a so-called saliency map (Koch & Ullman, 1985). Different feature channels (orientation, color, luminance, etc.) are analyzed independently; maps of local differences (contrasts) in these features are summed up and attention is allocated to the location of highest activity. Ever improving variants of this concept describe human bottom–up attention with improved reliability on the system level (Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005). In addition to local contrasts, deviations from the global image structure are a strong predictor of visual attention. This is most evidently seen in the phenomenon of "pop-out" (Treisman & Gelade, 1980), in which an odd item immediately attracts attention. This effect is not caused by the item being more salient in terms of local features, but because it differs from the global context. This idea has entered saliency-map modeling recently as the notion of "surprise" (Itti & Baldi, 2005), which is an information-theoretic measure to define deviations from the global (temporal) context as salient. The authors find that "surprise" better predicts human eye-movements in dynamic scenes than classical saliency map models, which only use local (in space and/or time) features. For static natural scenes, however, saliency maps according to Koch and Ullman's (1985) notion of local feature-contrasts still remain the predominant model of human overt attention.

Evidence for the validity of saliency map scheme arises mainly from studies of human performance at the system level. Few studies have investigated, however, the extent to which different features are correlated with overt attention and still fewer have examined whether these effects are directly causal in nature. In humans, several studies (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999) have found a correlation between fixation probability and luminance-contrast in natural scenes. Confirming this correlative result, Einhäuser and König (2003) show that luminance-contrast does not *causally* attract human overt attention. Instead, higher order properties appear to guide fixation in natural images. Taking one particular higher order effect into account, Parkhurst and Niebur (2004) provide an extension of the saliency map model to explain the data of Einhäuser and König (2003). In their model, a strong effect of a second order (or "texture-") contrast dominates a residual effect of first order ("luminance-") contrast for human overt attention. In monkeys, eye-movement studies with naturalistic stimuli are receiving increasing interest (Sheinberg & Logothetis, 2001). Guo, Robertson, Mahmoodi, Tadmor, and Young (2003) present faces and scrambled versions thereof to rhesus monkeys and conclude that the saliency of a specific facial feature does not only depend on its low-level appearance, but also on "higher levels of perceptual processing". In summary, there is evidence for the importance of higher order stimulus features to visual attention in humans and—for special subsets of naturalistic stimuli—also in monkeys. It is unclear, however, whether the observed correlations of low-level stimulus features to overt attention in monkeys are qualitatively and quantitatively similar to those in humans. Furthermore, it is an unresolved question whether monkeys and humans employ the same processing strategies for overt attention under natural viewing conditions.

In the present study, we record eye-movements of human subjects and rhesus monkeys while they freely view natural scenes and modified versions thereof. We compare the value of various stimulus features at fixation locations to the values expected by chance. In addition, we quantify to what extent the "classical" Itti and Koch (2000) saliency map model predicts the data. To assess the influence of local image features independent from their natural context, we evaluate the effect of contrast modifications in natural scenes on overt attention. Finally, we analyze how well the fixations of one individual predict the fixation of another individual within and across species.

2. Methods

2.1. Stimuli

Stimuli for this study were based on 108 images of natural scenes, which were taken with a Coolpix 995 (Nikon, Tokyo, Japan). Original resolution of the images was 2048×1536 in three colour channels. Images were converted to greyscale by MatLab's (Mathworks, Natick, MA, USA) *rgb2gray.m* function using default settings and down-sampled to a resolution of 1024 and 768 using bi-cubic interpolation. Twelve representative example images are shown in Fig. 1A. The complete dataset is available from the authors on request.

2.1.1. Modification

In addition to unmodified images, contrast-modified versions were also presented. Modification was performed as first described in Einhäuser and König (2003). To locally increase or decrease luminance-contrast, five points (x_i,y_i) were randomly chosen, such that each point had a distance of at least 160 pixels (23.5°—all values in ° refer to center) from the other points and from the image boundary. A two-dimensional Gaussian $G_i(x,y) = \exp\left(-\frac{\left((x-x_i)^2+(y-y_i)^2\right)}{\lambda^2}\right)$ with $\lambda = 80$ pixel (12°) was centered over each point. Taking the maximum over G_i resulted in the mask $G(x,y) := \max_{i\in\{1,..,5\}}[G_i(x,y)]$. At each image point the original pixel intensity $I_0(x,y)$ was then modified to $I(x,y) = I_0(x,y) + \alpha G(x,y) \times$

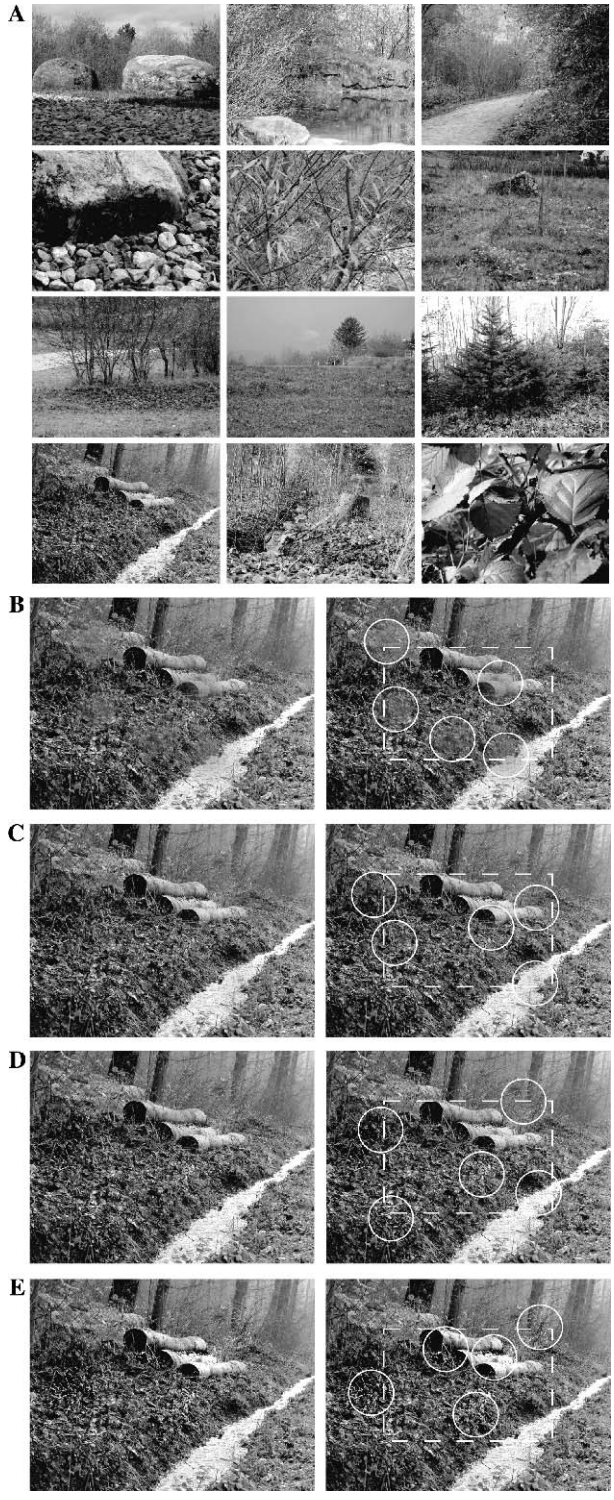

Fig. 1. Stimuli. (A) Twelve examples out of the 108 basis images used in the present study. (B) Left: image modified with peak modification level $\alpha = -0.6$; right: same stimulus with modifications marked by white circles at a distance of λ (80 pixels) from the modification center. For details of modification see Section 2.1.1. (C) Same image as in (B), but using peak modification level $\alpha = -0.2$. (D) Same image as in (B), but using peak modification level $\alpha = +0.2$. (E) Same image as in (B), but using peak modification level $\alpha = +1.0$. In (B–E) the white dashed rectangle indicates the 600×400 pixel ($76° \times 55°$) wide central region, which was used for analysis. Note that this figure is intended only to illustrate the modification procedure; printouts and screen images depend on the system and might differ from the actual stimulus.

$(I_0(x,y) - \langle I_0 \rangle)$, where $\langle . \rangle$ denotes the mean over the image and α the peak contrast modification level. If $I(x,y)$ exceeded the 8-bit range of possible pixel values, the result was clamped to the maximum or minimum possible value, respectively. Within a given stimulus, the modifications had the same peak modification level, α (Figs. 1B–E). Eight different peak modification levels were used ranging from −60 to −20% (locally decreased contrast) and 20 to 100% (locally increased contrast).

2.1.2. Presentation

For presentation of the stimuli we used the MatLab's psychophysics toolbox (Brainard, 1997; Pelli, 1997) on a Pentium 3 Laptop running Windows ME. Stimuli were projected from the back on a transparent screen with a black circular aperture (diameter: 110 cm) using a Nec LT 157 (NEC Solutions, Itasca, IL) projector (resolution: 1024×768 pixel at 60 Hz). Each stimulus was presented for 6 s interleaved by blanks of medium intensity and viewed at a distance of 50 cm. The image spanned 133×100 cm, which corresponds to $106° \times 90°$ of visual angle. Since part of the field of view was obstructed by the setup, we only used the central 600×400 pixels ($76° \times 55°$) for analysis. As a further advantage of this restriction, potential effects of the image boundaries cannot confound the analysis. Note, that we take the fixations outside the analyzed region into account for analyses that count fixations. This means e.g., if the first and the third fixation are inside the region, while the 2nd is not, the 3rd fixation would still be regarded as the 3rd fixation on that stimulus and this particular stimulus would contribute no data on the 2nd fixation.

While the mapping from real-world to image brightness might contain unknown non-linearities imposed by the CCD-camera, the γ-factor of the projector was corrected for, ensuring a linear relation between pixel-values and presented luminance. Peak luminance ("white") on the screen was 240 cd/m^2.

Stimulus presentation was separated into 18 sessions. In each session 54 stimuli were presented. Presentation was balanced such that

(1) Each of the nine modification levels (including unmodified) occurred exactly six times per session.
(2) Each image was used in each modification level once in the course of the whole experiment (18 sessions).
(3) 54 different images were used in each session. Half of the 108 images were used in odd sessions only, the other half in even sessions.

Since each basis image reoccurs only after at least 54 different stimuli and the same subjects never views the same image at the same modification level more than once, this balanced design minimizes potential effects of stimulus repetitions.

2.2. Monkey subjects

Two rhesus monkeys (*Macaca mulatta*) were used for this experiment. Each performed all 18 sessions (972 trials). To the best of our knowledge, the monkeys had no prior exposure to naturalistic images in the laboratory setup. Monkey N was head-fixed, while monkey C was not head-restrained for technical reasons. Neither before, nor during the period in which the present experiments were performed, the monkeys were involved in tasks dealing with natural visual stimuli. While monkey C had been involved in eye–hand coordination experiments prior to the present study, monkey N had received no task-specific training.

Monkeys were rewarded with a drop of water after each trial irrespective of the eye-movements they performed. In the course of the experiment apple juice was added to the water to keep motivation at an approximately constant level. Fixation performance in calibration blocks between experimental sessions did not worsen noticeably, which indicates that monkeys remained motivated throughout. Trial onset was triggered manually, as soon as the monkey had stopped drinking the reward from the previous trial. Monkey N performed sessions 1–3 on the first day, sessions 4–12 on the second day and sessions 13–18 five months later. Monkey C performed sessions 1–13 on his first day and the remaining sessions the day after.

Eye-position was monitored using the scleral search coil technique (Judge, Richmond, & Chu, 1980) and the output of the eye monitor system (Primelec, Regensdorf, Switzerland) was sampled at 75 Hz. The coordinate transform from the coil's output to screen coordinates was computed analogously to Einhäuser and König (2003). Before each session monkeys had to fixate points presented on the screen and were rewarded manually for correct fixation. From these fixation points we computed the bi-linear transform that yielded the best match between screen coordinates and coil output. Comparison between the calibration trial before and after a presentation session was used to verify the stability of the recording.

All animal handling was performed in full compliance with the guidelines of the National Institutes of Health for the care and use of laboratory animals and of the European Community (EUVD86/609/EEC).

2.3. Human subjects

Seven subjects (undergraduate students, three male, and four female) with normal or corrected-to normal vision participated in the experiment. Six subjects performed 6 sessions each—two subjects performed sessions 1–6 (CK, KF), two subjects sessions 7–12 (AN, IZ) and two subjects sessions 13–18 (JB, TH) yielding 324 trials per subject. This yields equivalent amounts of data for humans and monkeys and will allow us a stimulus-by-stimulus comparison of species. One additional subject (SH) performed all 18 sessions (972 trials); from her data sessions 3 to 6 are missing due to setup failure.

Human eye-position was recorded using a head-mountable "Eyelink I" eye-tracking system (SR Research, Mississauga, Ontario, Canada). Before each session a calibration block was performed. The calibration protocol was identical to the monkey experiment and was verified with the internal calibration of the Eyelink system. Subjects used a chin-rest to maintain a constant distance from the screen and were allowed to have a break between sessions if they desired. Subjects were instructed to "study the images carefully." In separate control experiments we demonstrate that instructions of "free viewing" and "study carefully" yield equivalent results with respect to the correlation of luminance-contrast and the selection of fixation points (Steinwender et al., unpublished results).

The experiment conformed to national and institutional guidelines for experiments with human subjects and with the Declaration of Helsinki. All subjects gave written informed consent to participate in the study.

2.4. Data analysis

2.4.1. Luminance-contrast

In line with earlier studies and as the straightforward generalization of two-point contrast, we defined luminance-contrast at a point as the standard-deviation of luminance in an 80×80 pixel ($12.5° \times 12.5°$) region around that point normalized by the mean luminance of the image. Note that changes to the exact size of the region (we also tested 40×40 pixels, $6° \times 6°$) and the use of a different normalization method (we also tested normalization by the patch mean) had no qualitative effect on the results.

2.4.2. Texture-contrast

A recent model (Parkhurst & Niebur, 2004) suggests that "texture-contrast," not luminance-contrast, dominates the guidance of overt attention in grey-scale images. To define texture-contrast without any further model assumptions, we canonically generalized our definition of luminance-contrast. We defined the texture-contrast at a given location as the standard-deviation of luminance-contrast in an 80×80 patch around the point normalized by the mean luminance-contrast of the image.

2.4.3. IK-saliency

To compare our results to established computational models of overt attention, we computed a saliency map for each stimulus. We used the model of Itti and Koch (2000), which we—as there is no color in our stimuli—restricted to the orientation and luminance channel. Both channels were used at equal weight and the parameters were set as originally published. This results in a saliency map, which we rescaled linearly to take values between 0 and 1 for each stimulus. To avoid confusion with the saliency measures that we define for each feature below (Section 2.4.7), we will hereafter refer to this measure as "IK-saliency" throughout. For analysis we treat IK-saliency analogously to the other features.

2.4.4. Modification

All the measures defined so far (luminance-contrast, texture-contrast, and IK-saliency) are defined agnostic about the modifications that we imposed to the modified stimuli. As additional feature, we measured the "modification" $\alpha G(x,y)$ and treated it in analogy to the intrinsic features (luminance-contrast, texture-contrast, and IK-saliency). This provides us with a measure in how far a deliberate deviation from the natural stimulus influences overt attention behavior.

2.4.5. Baseline ("control") contrasts

To achieve an unbiased estimate of the distribution a certain measure (luminance-contrast, texture-contrast, IK-saliency or modification) would take at fixation points, if fixation and this measure were unrelated, we defined the following baseline: we defined as "control fixations" all fixations that result from all presentations of all stimuli in the same subject or species. For analyses in which only unmodified images were concerned, control fixations were taken from all unmodified stimuli only. For each given stimulus we compared the measure at the fixations in that particular stimulus ("actual" fixations) to the same measure at the control fixations. For each stimulus, the medians at actual/control fixations are then referred to as actual/control luminance-contrast, actual/control texture-contrast, actual/control IK-saliency, and actual/control modification, respectively. If the actual measure significantly differs from the control measure, the measure is related to the likelihood of fixation. Note that we take the control fixations from all stimuli including the actual stimulus. Although excluding the actual fixations from the control set would seem a cleaner definition, this difference is only a matter of concern for small datasets. Since in our case actual fixations account only for 1/108 (analysis of unmodified stimuli alone) or 1/972 (other analyses) of the control fixations, we may safely include them in the control set for computational efficiency.

2.4.6. Saliency measures

To quantify the effect of luminance-contrast further, we defined the saliency of luminance-contrast (S_{LC}) as the difference between actual luminance-contrast and control luminance-contrast. If luminance-contrast was not related to fixations, this measure, on average, would not be different from 0. This is so because actual luminance-contrast would not differ from control luminance-contrast. If S_{LC} is positive, luminance-contrast along fixation points will be larger than predicted if luminance-contrast had no effect. If S_{LC} is negative, luminance-contrast along fixation points will be smaller than one would predict if luminance-contrast had no effect. In complete analogy to S_{LC} we defined the saliency of texture-contrast (S_{TC}) and the saliency of IK-saliency (S_{IK}). One should note however, that S_{LC}, S_{TC}, and S_{IK} are correlative measures. Positive $S_{LC}/S_{TC}/S_{IK}$ does not imply that luminance-/texture-contrast, as such, causally attracts overt attention. It only implies that luminance-/texture-contrast is higher at fixated locations than would be expected from a random distribution of fixation points, which may also be caused by other features that are correlated to both attention and luminance-/texture-contrast.

In addition, we directly assessed the saliency of the modification (S_{mod}). In analogy to S_{LC}, S_{TC}, and S_{IK} we defined S'_{mod} as difference between actual and control modification (i.e., between the median of modifications $\alpha G(x,y)$) at the actual fixations compared to the median of modifications at control fixations. Since the total modification in a stimulus scales linearly with the peak modification level α, this measure would depend trivially on α. Hence, we normalized by the peak modification level $S_{mod} = S'_{mod}/\alpha$. An additional effect of the normalization is that for all values of α (positive and negative), a *positive* S_{mod} implies that a modification *attracts* overt attention. This is consistent with the definitions of S_{LC}, S_{TC}, and S_{IK} and justifies the definition of S_{mod} as the *saliency* of a modification.

2.4.7. Statistical analysis

For the saliency measures S_{LC}, S_{TC}, S_{IK}, and S_{mod} we assume a normal distribution across images. Hence, we performed a *t*-test to test whether the mean of each of these measures across unmodified images was different from 0. A *t*-test was also used to test whether these means across images differ between species. In all cases two-tailed tests were used.

When comparing the saliency measures for modified images, we must consider two factors, the species and the peak modification level, α. Since different modifications of the same image are not independent from each other, we treated α as a repeated measurement factor. Hence, we performed a general linear model repeated measures analysis of variance (with each basis image corresponding to one level) to test whether there are significant differences in our saliency measures between subjects, and/or between peak modification levels. If we found such a difference we computed post hoc comparisons of interest. Furthermore, if we did not find a dependence on a factor we pooled over this factor for further testing.

In cases where normal distributions cannot be assumed, we compared the medians of two distributions by using a Wilcoxon-test (ranked sign-test). This measure is justified for comparing actual contrasts versus control contrasts. While the Wilcoxon-test takes the size of the values in the distributions into account, we may also compare the distribution medians using a sign-test. Here, we ask whether an actual contrast (luminance- or texture-) is larger (or smaller) than the respective control contrast in a significant number of images. This means we test the following null hypothesis: That there are as many images in which actual contrast is larger than control contrast as there are images in which actual contrast is smaller than control contrast. Since this measure is independent of the size of the effect, it does not weight different images independently. Hence, it is the most robust measure with respect to any potential particularities of a given stimuli.

For the GLM repeated measures ANOVA and its post hoc tests, the SAS 8.02 (SAS-Institute Inc., Cary, NC, USA) Software package was used. All other tests were performed using MatLab's statistics toolbox.

2.4.8. ROC analysis

Tatler et al. (2005) recently suggested a measure to assess the saliency of a feature based on signal detection theory, the area under the receiver-operating-characteristics (ROC) curve. We computed the ROCs for each feature and each peak modification level separately: for each stimulus we normalized the feature values from 0 to 1; we varied a threshold and determined the fraction of fixations that fall on pixels (or bins in case of IK-saliency) above threshold ("hits"). This was compared to the fraction of pixels (bins) above threshold without fixations ("false alarms"). Hits and false alarms are accumulated across all stimuli. Plotting hits vs. false alarms results in the ROC. If the ROC is the diagonal (ROC-area: 0.5), the respective feature does not allow any prediction on fixations. If curve is above the diagonal (ROC-area large 0.5), fixated points can be partly discriminated from non-fixated points on basis of this feature. Perfect discrimination on the basis of the feature under investigation would yield an ROC area of 1.

To derive confidence limits we apply a bootstrap technique: we generate 1000 surrogate data sets by drawing with replacement from the stimuli used to compute the respective ROC. We compute ROC areas for these surrogate data and determine the confidence limits such that 99% of this surrogate areas are within these limits.

2.4.9. Model- and feature independent distance measure

To obtain a model-independent measure of how well humans and monkeys tend to direct their attention to similar locations in each image, we measured the Euclidian distance between fixation locations within and across species. We compared the fixation locations for each pair of subjects that share common stimuli (N to C, all monkeys to all humans, CK to KF, AN to IZ, JB to TH, and all to SH). To account for the effects of prolonged viewing, we performed this comparison separately for each fixation on each stimulus, i.e., we compared the 1st fixations, 2nd fixations, etc. separately. For each pair of subjects the distances obtained are first averaged over all stimuli, which both subjects contribute data for. This measure could still be confounded by biases that are not image-specific (like e.g., one subject tends to look to the right, the other more to the left). Hence, we subtracted the obtained average distance from and divided it by the mean distance of all respective fixations (separated according to 1st, 2nd,... fixation) of the same pair of subjects. For each pair of subjects this "normalized distance" yields the fraction of how much the distance between corresponding fixations is smaller than expected by their general biases. Hence, the normalized distance provides a reasonable, model-independent measure of how much any two subjects tend to look at the same items, irrespective of their low-level features.

3. Results

3.1. Image statistics

We measured eye-movements of 2 monkeys and 7 human subjects, while they were freely viewing images of natural scenes and similar images that were locally modified in contrast. To relate the imposed contrast modifications to contrast fluctuations naturally occurring in natural scenes, we first determined the mean luminance-contrast of each unmodified image. The average luminance-contrast across all images was 0.28 (*SD* over images: 0.09). Across all images, at 63% (*SD*: 19%) of pixels the luminance-contrast was within 20% of the mean luminance-contrast in the image, at 90% (*SD*: 12%) within 40% of the mean and at 97% (*SD*: 5%) within 60% of the mean luminance-contrast. Based on these numbers we can define, that a peak modification of $\pm 20\%$ is well within the range of natural contrast fluctuations, while a modification of $\pm 60\%$ or stronger is outside the range of natural contrast fluctuations. Throughout the paper, we consequently will refer to peak modifications of $\pm 20\%$ as "modifications in the range of natural contrast fluctuations."

3.2. Eye-tracking data

The analysis was based on successfully recorded fixation data of 1944 trials from monkeys (2 monkeys $\times$ 18 sessions $\times$ 54 trials/session) and 2700 trials from human subjects ($6 \times 6 \times 54 + 1 \times 14 \times 54$). These trials resulted in a total of 32,225 fixations in humans and 33,695 fixations in monkeys, yielding on average 11.9 fixations per trial in humans and 17.3 fixations per trial in monkeys. On average a fixation lasted 370 ± 293 ms (mean $\pm$ SD over all fixations) in humans, and 294 ± 215 ms in monkeys. This implies that fixations account for 74% (human) and 85% (monkey) of total stimulus presentation time (6 s per stimulus). These data justify restriction of the analysis to the points of eye fixation. The average distance of subsequent fixation points on the screen was 95 pixels (14°) in monkeys and 127 pixels (19°) in humans. Both distances were larger than the radius of the contrast modifications, which was 80 pixels (12.5°). This justifies treating subsequent fixations independently with respect to the analysis of modifications.

To avoid a potential confound due to obstruction of the outermost regions of the stimulus by the setup, higher

order non-linearities of the eye-tracking systems or head-movement artifacts (humans and monkey C), we restricted analysis to fixations within a 600×400 pixel ($76° \times 55°$) central stimulus region. This analyzed region contained 92% (29,725) of all fixations in humans and 75% (25,135) in monkeys. The comparably low number for monkeys mainly results from them looking at parts of the setup at the fringes of their field of view, as they were rewarded irrespective of whether or not they actually looked at the stimulus. However, 99.7% (humans) and 97.7% (monkeys) of trials contained at least one fixation in the analyzed region; in 98.2% (humans) and 79.5% (monkeys) of trials more than half of the fixations were in the analyzed region. These data justify restriction of the analysis to the central region of the image.

To test whether the overall properties of eye-movement remained constant throughout the experiment, we compared the first and last session of each recording day. We did not find a difference in fixation duration between the first and last session in either species (humans, $p = 0.51$, t-test; monkeys, $p = 0.53$, t-test). Furthermore, the fraction of fixations in the analysed region also remained constant between the first and last session of each day ($p = 0.93$ for humans and $p = 0.33$ for monkeys). Hence the spatial and temporal properties of eye-movements can be assumed to be constant. This furthermore suggests that the motivational state of a subject remained about constant throughout the experiment.

We tested whether there is a general bias in fixation to one side of the screen. While—averaged over all fixations—three humans had a center of mass of fixation slightly right from the center, for the four others the center was slightly left and for no subject the deviation was larger than 4.9°. The same held for monkeys: while one monkey had a slight bias to the left (4.6°) the other monkey had about the same bias to the right (5.7°). We observed a slight vertical bias above the midline in monkeys ($5.3° \pm 2.3°$) and in humans ($6.0° \pm 2.7°$). The between-species difference was non-significant ($p = 0.74$, t-test). Thus, we found no significance in the bias between the two species. The systematic deviation from the center of the display was small compared to size of the analysed region ($76° \times 55°$).

These data also show, however, that monkey fixations are different from human fixations in several respects, including duration and spatial distribution. This justifies the use of the "saliency" measures (S_{LC}, S_{TC}, S_{IK}, and S_{mod}), which use an intra-species control and are thus insensitive to such inter-species differences, to perform an unbiased comparison between monkeys and humans.

3.3. Effects of luminance-contrast in experiments using unmodified images

In a first analysis we tested whether there is a relation between luminance-contrast and the likelihood of fixation at a given location.

The left panel of Fig. 2A displays all fixations recorded from all monkeys on the unmodified stimulus (green points). Since these fixations were actually recorded on that particular stimulus, we refer to them as "actual" fixations. The red points in the same panel show the fixations of all monkeys pooled over all unmodified stimuli, i.e., fixations unrelated to this particular stimulus. The image properties at these "control fixations" serve as the baseline for analysis. The right panel of Fig. 2A shows the corresponding data for all human subjects. It is evident that the horizontal spread of control fixations is larger in humans than in monkeys (horizontal standard deviation of position: 106 pixels (16°) in monkey; 143 pixels (21°) in human). This further justifies the use of control fixations within species (rather than across) as baseline.

Fig. 2B shows the measure of the distribution of luminance-contrast at the position of actual and control fixations. The non-Gaussian shape of the control distribution suggests the median (and not the mean) should be used to summarize the distribution in a single value. This is why, we use these medians as "actual luminance-contrast" and "control luminance-contrast," respectively. Please note that the control luminance-contrast was highly correlated to the mean luminance-contrast in the analyzed region of the image for monkeys ($r = 0.979$) as well as for humans ($r = 0.989$), which shows that our analysis is not critically dependent on the choice of this baseline. For the example image the actual luminance-contrast in monkeys (0.253, green line in Fig. 2B) was larger than the control luminance-contrast (0.246, red line). The same relation held for humans (Fig. 2B, right). This trend was conserved for the majority of images in both species. In monkeys, actual luminance-contrast was larger than control in 69 images and smaller in 39 images (Fig. 2C, left). In humans this relation was 74–34 images (Fig. 2C, right). Using a sign-test we tested whether the number of images in which actual contrast was larger than control, was significantly larger than the number of images in which the opposite relation held. This means we tested against the null hypothesis that in an equal amount of images the actual contrast is greater than the control contrast and vice versa. We found this number to be highly significant in both species (human: $p = 0.0002$; monkey: $p = 0.005$). In addition, we tested by a ranked sign-test (Wilcoxon-test) whether the medians of the distributions differed significantly. In humans the median of actual contrast was 0.285 and significantly larger than the median of control contrast of 0.282 ($p = 0.0009$, Wilcoxon-test) and the same was true for monkeys (0.290 vs. 0.286, $p = 0.002$). In conclusion, pooled over subjects we found a consistent relationship between luminance-contrast and selection of fixation points in both species.

Next, we evaluated whether the observed effect of luminance-contrast was consistent among individual subjects. We therefore performed the same analysis for each individual, i.e., we computed actual fixations and control fixations separately for each subject. Both monkeys showed a very similar quantitative effect: actual contrast was larger than

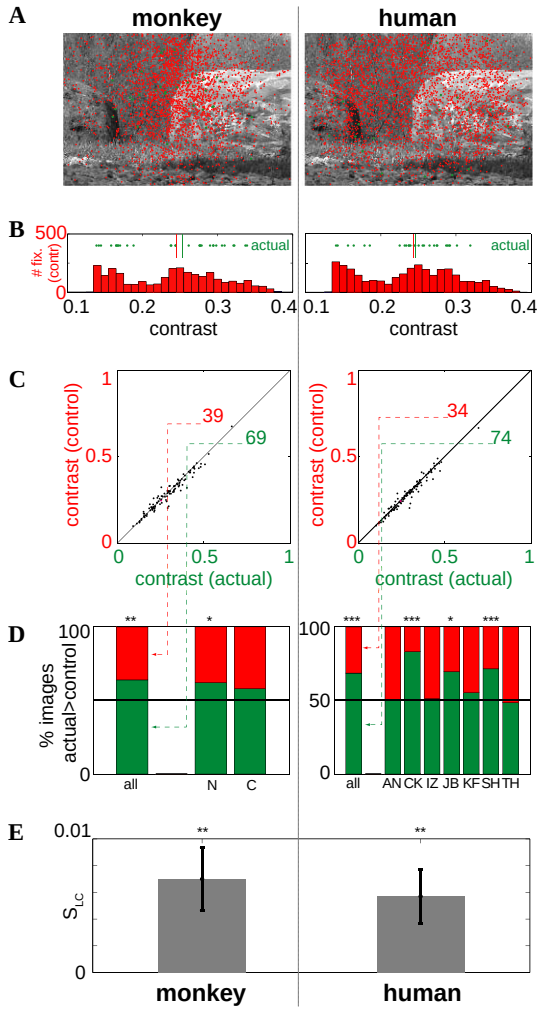

Fig. 2. Luminance-contrast. In all panels the left column refers to monkeys, the right column to human subjects. Except for part of (D), data are pooled over all subjects of within a species. (A) Green: "Actual" fixations for the example image shown. Red: "Control" fixations from all other unmodified images pooling data from the same species (see Section 2.4.5). Only the central 600×400 ($76° \times 55°$) region used for analysis is shown. (B) Red histogram: distribution of luminance-contrast at control fixations on the stimulus shown in (A). Green points: luminance-contrast at actual fixations in the stimulus shown in (A). Green vertical line: median of luminance-contrast at actual fixations ("actual luminance-contrast") for example stimulus shown in (A); red vertical line: median of luminance-contrast at control fixations ("control luminance-contrast") for example stimulus shown in (A). (C) Control luminance-contrast plotted versus actual luminance-contrast. Each data-point corresponds to one stimulus. The example from (A and B) is marked in magenta. For images with data-points below the diagonal, actual luminance-contrast is larger than control luminance-contrast. The number of points below the diagonal is given in green, the number of points above in red. (D) Percentage of images in which actual luminance-contrast is larger than control luminance-contrast (green), or smaller (red). Left bar for each species represents data from all subjects and thus represents the numbers given in (C). For the bars representing individuals, actual *and control* distributions are computed exclusively on a per individual basis. Significance levels refer to sign-test (see text for details). (E) Mean and standard-error across images for saliency of luminance-contrast (S_{LC}).

control contrast in 67 (monkey N) and 61 (monkey C) out of the 108 images (Fig. 2D, left). In monkey N both aforementioned tests yielded significant results (sign-test: $p = 0.02$, Wilcoxon-test: $p = 0.002$). In monkey C the hypothesis that an equal number of images shows higher actual than control luminance-contrast could not be fully rejected ($p = 0.12$, sign-test). However, the medians were also significantly different in this monkey ($p = 0.03$, Wilcoxon-test).

Humans showed a larger inter-subject variation than monkeys. However, no subject showed a tendency for control contrast to be larger than actual contrast in a significant fraction of images. Subject TH took the minimum value with actual luminance-contrast being smaller than control luminance-contrast in 17 out of 36 images; in all other subjects this relation held for at least half of the images. Three subjects (AN, IZ, and TH) were at about chance level, i.e., actual contrast was larger than control in approximately half the images (50% ± 1image, $p > 0.86$, sign-test, Fig. 2D, right). In three of the remaining subjects (CK, JB, and SH) actual contrast was larger than control in a significant fraction of images ($p = 0.0001$, $p = 0.03$, and $p = 0.0001$, respectively, sign-test). The medians were also significantly different for subjects CK and SH ($p < 0.001$, Wilcoxon-test).

For the interpretation of inter-subject variability in humans, one has to note that—with the exception of SH—humans were tested only on a small number of unmodified images (6 sessions × 6 unmodified images/session = 36 unmodified images) as compared to the monkey subjects and human subject SH (108 unmodified images). However, this does not imply that subject SH biases the population analysis: with the exception of 24/108 unmodified images, in which no data from SH are available, three human subjects (SH and two others) contributed data to each image of the population analysis. This allows us to conclude that both humans and monkeys tend to fixate regions of high contrast in a majority of unmodified images.

To further quantify relation of luminance-contrast to fixation and to compare this effect between the species, we use the saliency measure S_{LC}, which we defined as difference between actual and control luminance-contrast. In unmodified images the mean S_{LC} across images was 0.0070 in monkeys and 0.0057 in humans (Fig. 2E). Consistent with the Wilcoxon-test on the medians above, the means of both S_{LC} are also significantly different from 0 ($p = 0.004$, $p = 0.006$, respectively). In addition S_{LC} allows us to directly probe the difference between human and monkey. We found no difference between mean S_{LC} in monkeys and mean S_{LC} in humans ($p = 0.67$, t-test). In summary, we confirmed earlier results that luminance-contrast was related to fixation in human subjects. In addition, we showed that the same is true for monkeys. Finally, we demonstrated that for this correlative measure no difference can be observed between the two species.

3.4. Effect of texture-contrast in experiments using unmodified images

Since a recent modelling study (Parkhurst & Niebur, 2004) had suggested that texture-contrast was more

relevant to the guidance of human overt attention than luminance-contrast, we measured the relation between texture-contrast, which we defined canonically as 2nd order luminance-contrast, and fixation. We found in monkeys and in humans that actual texture-contrast was larger than control texture-contrast in the majority of images (Fig. 3A). This majority was significant in humans (72:36, $p = 0.0008$, sign-test). Although we found the same trend for monkeys the effect was not significant (60:48, $p = 0.29$, sign-test). A test on the difference between the medians was also in line with this result: in humans the median actual texture-contrast of 0.076 was significantly larger than median control texture-contrast of 0.069 ($p = 0.0003$, Wilcoxon-test), while these medians were not significantly different in monkeys (0.074 vs. 0.069, $p = 0.15$). By directly comparing actual to control texture-contrast we therefore only found a significant effect in humans, but not in monkeys.

With respect to individual subjects, we found that actual texture-contrast was larger than control in approximately the same fraction of images in each monkey (Fig. 3B, left). This effect was not significant in either monkey ($p = 0.21$ and $p = 0.33$, sign-test, monkey N and C, respectively). In humans we found a significant difference for two individuals ($p = 0.004$ and 0.001, sign-test, in KF and SH, respectively). No subject had a larger number of images with actual texture-contrast smaller than in the control images (Fig. 3B, right). This result confirms an effect of texture-contrast in humans, while a significant effect cannot be revealed for monkeys using a sign-test.

To directly compare both species, we defined the saliency of texture-contrast (S_{TC}) in an analogous fashion to S_{LC} as the difference of actual texture-contrast minus control texture-contrast (Fig. 3C). Mean S_{TC} was significantly larger than 0 in humans ($p = 0.001$, t-test) and also achieved significance in monkeys ($p = 0.04$, t-test). The latter may indicate that a small effect of texture-contrast is also present in monkeys. Indeed, a direct comparison between species could not reject the hypothesis that S_{TC} is identical for humans and for monkeys ($p = 0.18$, t-test). In conclusion, although there seems to be some indication of a quantitative difference between the two species, the analysis of unmodified images did not reveal any significant difference between monkeys and humans.

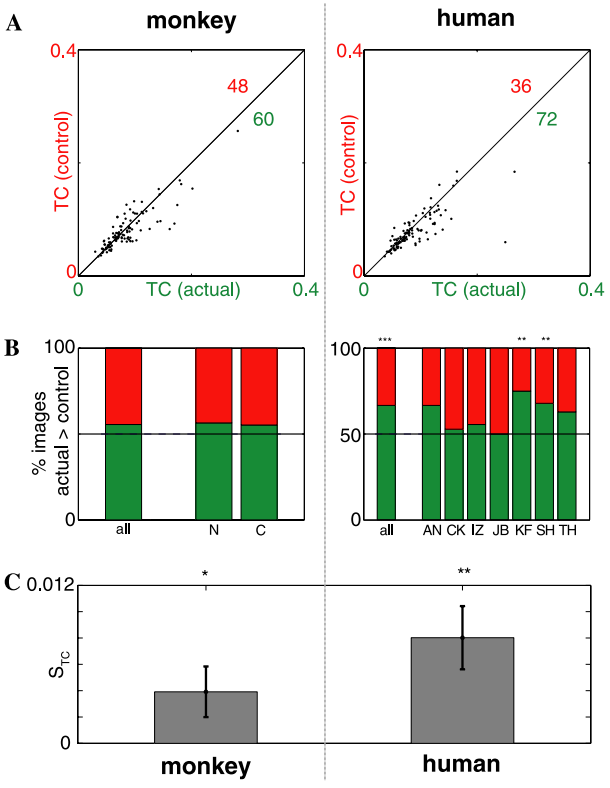

Fig. 3. Texture-contrast. (A) Control texture-contrast plotted versus actual texture-contrast (analogous comparison to Fig. 2C for texture-contrast). For images with data-points below the diagonal actual texture-contrast is larger than control texture-contrast. (B) Percentage of images in which actual texture-contrast is larger than control texture-contrast (green), or smaller (red). Analogous to Fig. 2D for texture-contrast. Significance levels refer to sign-test. (C) Analogous to Fig. 2E for texture-contrast, mean and standard-error across images for saliency of texture-contrast (S_{TC}).

3.5. Effects of IK-saliency in experiments using unmodified images

Most contemporary models of bottom–up saliency are based on the architecture proposed in Itti and Koch (2000). Hence, we repeated the analysis done for luminance-contrast and texture-contrast on an implementation of this model that is restricted to the luminance and orientation channel. In monkeys and humans significantly more images have higher actual IK-saliency than control IK-saliency (humans: 73:35; monkeys: 71:35 (remaining 2 images show no difference); $p < 10^{-3}$ sign-test; Fig. 4A). The medians across images are also significantly between actual and control IK-saliency ($p < 10^{-4}$, Wilcoxon-test). Regarding individuals, for both monkeys and for all but one human subject (TH) more images have higher actual IK-saliency than control IK-saliency (Fig. 4B). This difference is significant in monkey N ($p = 0.009$, sign-test) and in humans KF ($p = 0.03$) and SH ($p < 10^{-4}$). Wilcoxon-tests are consistent with this result as they reveal median actual IK-saliency to be larger than control in both monkeys (N, $p < 10^{-4}$; C, $p = 0.02$) and in two humans (KF, $p = 0.001$; SH, $p < 10^{-4}$). Computing S_{IK} in analogy to S_{LC} and S_{TC} shows—consistent with the aforementioned Wilcoxon-test on medians—the means of S_{IK} to be significantly different from 0 ($p < 10^{-4}$, t-test). As with S_{LC} and S_{TC} there is no significant difference between humans and monkeys for S_{IK} ($p = 0.19$, t-test). In summary—although IK-saliency considers orientation and different spatial scales—the results on IK-Saliency are well in line with the results obtained for luminance-contrast alone. In particular, there is a correlation to overt attention and there are no significant differences between humans and monkeys.

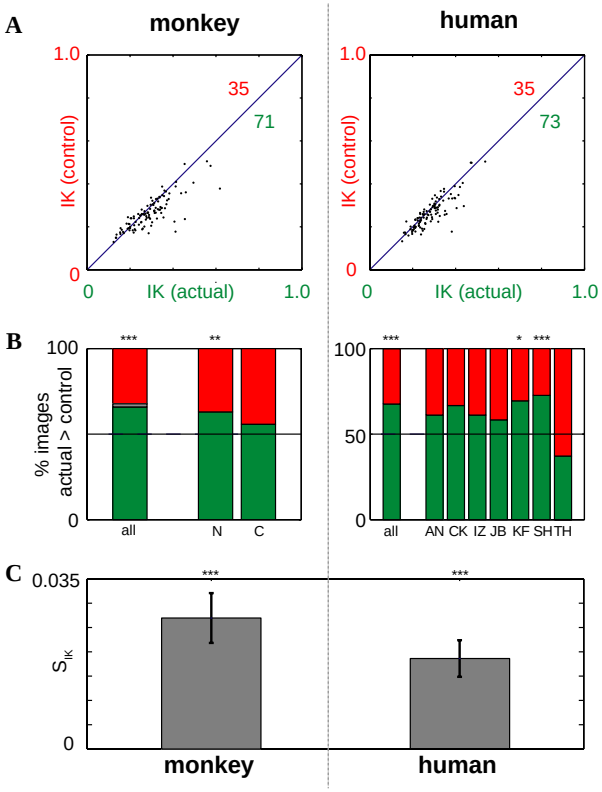

Fig. 4. IK-saliency. (A) Control IK-saliency plotted versus actual IK-saliency (analogous comparison to Figs. 2C and 3A). For images with data-points below the diagonal actual IK-saliency is larger than control IK-saliency. (B) Percentage of images in which actual IK-saliency is larger than control IK-saliency (green), or smaller (red). Analogous to Figs. 2D and 3B, significance levels refer to sign-test. (C) Analogous to Figs. 2E and 3C for IK-saliency, mean and standard-error across images for saliency of IK-saliency (S_{IK}).

3.6. Modified images

Does the apparent similarity between species observed for unmodified images imply that the underlying mechanisms that guide overt attention are similar in both species? To address the question in more detail, we analyzed the data from the stimuli that were locally modified in contrast. Fig. 5A shows the actual fixations of monkeys and human subjects on such a modified stimulus. The control fixations were in this case taken from all trials (modified and unmodified) within the same species. First, we performed the same analysis as for the unmodified images, i.e., we computed the correlative measures S_{LC} and S_{TC} in dependence on peak modification levels. In addition, we computed the saliency of the modifications (S_{mod}). This measure is independent of those effects of luminance-contrast and texture-contrast on overt attention that arise from correlations of both to a third local property. Therefore, it better captures the effect of the modification than the intrinsic measures S_{LC}, S_{TC}, and S_{IK}.

For statistical analysis of our saliency measures (S_{LC}, S_{TC}, and S_{mod}) we performed a general linear model repeated measures analysis of variance (GLM-repeated measures

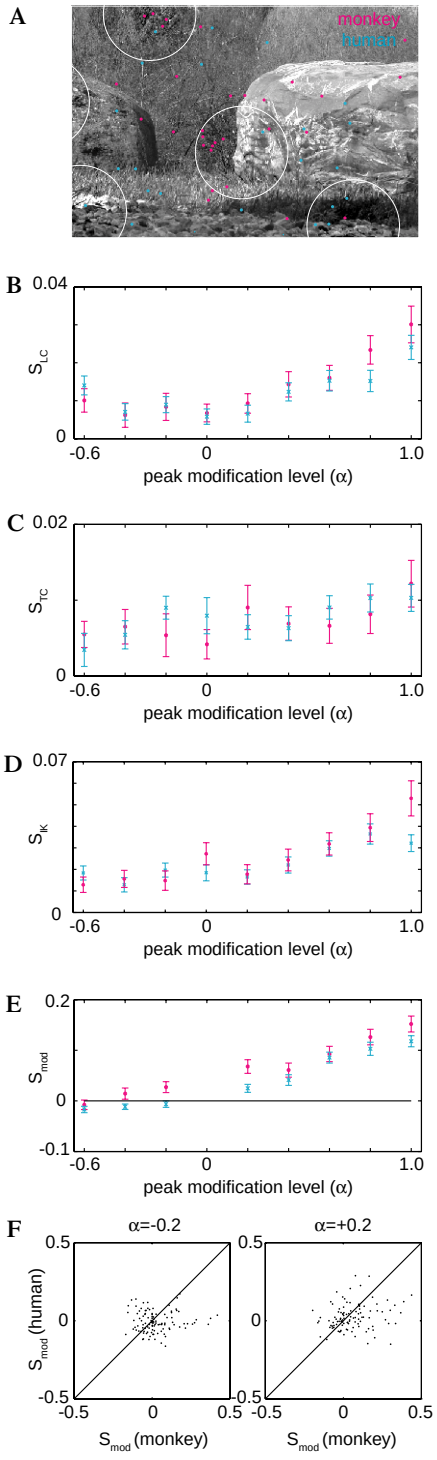

Fig. 5. Contrast modification. (A) Actual fixations of monkeys (magenta) and humans (cyan) on a modified image ($\alpha = 1.0$). (B) Saliency of luminance-contrast (S_{LC}) in monkeys (magenta dots) and humans (cyan crosses) for different peak modification levels (α). Error-bars indicate standard-errors across images. (C) Saliency of texture-contrast (S_{TC}) in monkeys and humans for different peak modification levels (α). Markers as in (B). (D) Saliency of IK-saliency (S_{IK}) in monkeys and humans for different peak modification levels (α). Markers as in (B). (E) Saliency of contrast modifications (S_{mod}) in monkeys and humans for different peak modification levels (α). Markers as in (B). (F) Stimulus by stimulus species comparison of saliency of modifications (S_{mod}) at moderate peak modifications levels ($\alpha = \pm 0.2$). Each data-point corresponds to one stimulus.

ANOVA) across images on the factor species (monkey vs. human) and the repeated measures factor peak modification level ($\alpha = -0.6, \ldots, 1.0$). This allowed us to test whether there is an effect of these two factors. In addition we tested whether the saliency measures are different from zero. The details of this analysis are described in the Section 2.4.7.

We found that S_{LC} showed a strong dependence on peak modification level (Fig. 5B; $p < 10^{-4}$, $F(8,207) = 7.83$ for the factor peak modification level). This result is not surprising given that the modifications affect the luminance-contrast distribution within one image. When the modification remained in the range of natural contrast-fluctuations (i.e., $\alpha = \pm 0.2$), S_{LC} was indeed not different from unmodified images ($p = 0.29$ post hoc contrast for $\alpha = -0.2$ vs. $\alpha = 0$, $p = 0.42$ for $\alpha = 0.2$ vs. $\alpha = 0$). These data based on the intrinsic measure S_{LC} provided no indication that modifications within the range of natural contrast fluctuations differentially affect the saliency of luminance-contrast.

In line with the results on unmodified images, no differences were found between monkeys and humans in the values for S_{LC} across modification levels ($p = 0.40$ for the factor species). Since there was also no interaction between the factors species and peak modification level ($p = 0.74$), S_{LC} can be regarded as species independent across modification levels.

Mean S_{LC} was positive for all peak modification levels (Fig. 5B). Using the independence between peak modification levels, we tested whether S_{LC} is different from 0 for each peak modification levels separately. Using a t-test, we found that for $\alpha \geqslant +0.4$ all p values in both species were smaller than 10^{-4}. For the remaining α, S_{LC} was significantly larger than 0 in both species for all but one peak modification level (monkeys: $p = 0.002$ at $\alpha = -0.6$, $p = 0.057$ at $\alpha = -0.4$, $p = 0.021$ at $\alpha = -0.2$, $p = 0.005$ at $\alpha = 0$ and $p = 0.0005$ at $\alpha = +0.2$; humans: $p < 10^{-4}$ at $\alpha = -0.6$, $p = 0.002$ at $\alpha = -0.4$, $p < 10^{-4}$ at $\alpha = -0.2$, $p = 0.005$ at $\alpha = 0$, and $p = 0.004$ at $\alpha = +0.2$). Using F statistics instead of t statistics yielded the same results. This result confirms the positive relationship between fixation probability and luminance-contrast, which we have described for unmodified images above, for the modified images.

When performing the same analysis as for S_{LC} for S_{TC} (Fig. 5C) we found no dependence on peak modification level ($p = 0.11$, GLM-MANOVA for factor peak modification level). Hence, the modifications had no impact on the strength of this attraction over the range tested. We observed no difference between species across modification levels ($p = 0.79$ for the factor species) and no interaction between species and peak modification level ($p = 0.50$). This allows us to conclude that on the phenomenological level there is no difference between species regarding the correlative effect of texture-contrast on overt attention. S_{TC} was positive for all peak modification levels tested. Since S_{TC} did not depend on peak modification level, we pooled over all peak modification levels and found the mean S_{TC} to be significantly larger than 0 for both species ($p < 10^{-4}$, t-test for both species). Hence, texture-contrast is positively correlated to overt attention.

We performed the same analysis for S_{IK} (Fig. 5D) and found a highly significant dependence on peak modification level ($p < 10^{-4}$; $F(8,207) = 7.44$). For all peak modification levels, S_{IK} is significantly larger than 0 (monkey: $p < 0.002$ for all α; human: $p < 2 \times 10^{-4}$ for all α, t-tests). There is no dependence on species ($p = 0.29$) and no interaction between species and peak modification level ($p = 0.17$). Hence, S_{IK} can be regarded as species independent across modification levels. In summary, although S_{IK} also incorporates the orientation domain, results are similar to the results obtained on S_{LC}. Most importantly, there is no dependence on species in neither S_{LC}, S_{TC} nor S_{IK}.

Next, we directly measured the saliency of the modification (S_{mod}). Performing the GLM analysis for S_{mod} showed a strong dependence on modification level ($p < 10^{-4}$, $F(7,208) = 40.41$, Fig. 5E). More importantly, S_{mod} was also strongly dependent on species ($p < 10^{-4}$, $F(1,214) = 16.10$, Fig. 4D). Saliency in monkeys was larger at all modification levels than in humans. Post hoc t-tests revealed that this species dependence is significant at moderate modification levels ($p = 0.04$ at $\alpha = -0.4$; $p = 0.008$ at $\alpha = -0.2$; $p = 0.006$ at $\alpha = +0.2$). At the other more extreme peak modification levels, no significant difference was observed ($p = 0.41$ at $\alpha = -0.6$; $p = 0.26$ at $\alpha = +0.4$; $p = 0.70$ at $\alpha = +0.6$; $p = 0.24$ at $\alpha = +0.8$; $p = 0.08$ at $\alpha = +1.0$). Since the distribution of S_{mod} cannot necessarily be assumed to be normal across images (Fig. 5F), we in addition tested the difference of medians at the modification levels −20% and +20% using a Wilcoxon-test. Indeed the medians across images were also significantly different ($p = 0.01$ at $\alpha = -0.2$; $p = 0.02$ at $\alpha = +0.2$). This confirms that at moderate modification levels lying in the range of natural contrast fluctuations, there is a different effect of modifications on monkeys compared to humans.

3.7. Signal-detection theory based analysis

On a large data-set, such as ours, observing a significant relation of a feature to fixation does not necessarily imply that this feature is indeed a good predictor of where monkeys and humans fixate on a trial-by-trial basis. To address this issue Tatler et al. (2005) recently suggested the ROC area as a more meaningful measure. This measure takes a value of 1 if a feature perfectly predicts fixation. It takes 0.5 if there is no relation at all, which we verified by using the control fixations. We measured the ROC area for each feature (luminance-contrast, texture-contrast, IK-saliency, and modification) at each peak modification level. Luminance-contrast, texture-contrast, and IK-saliency are larger than chance (0.5) for all peak modification levels in both species (Figs. 6A–C). In addition, for luminance-contrast and IK-saliency there is no peak modification level, for which the lower 99% confidence limit drops below chance. This also holds for the majority of peak-modification levels

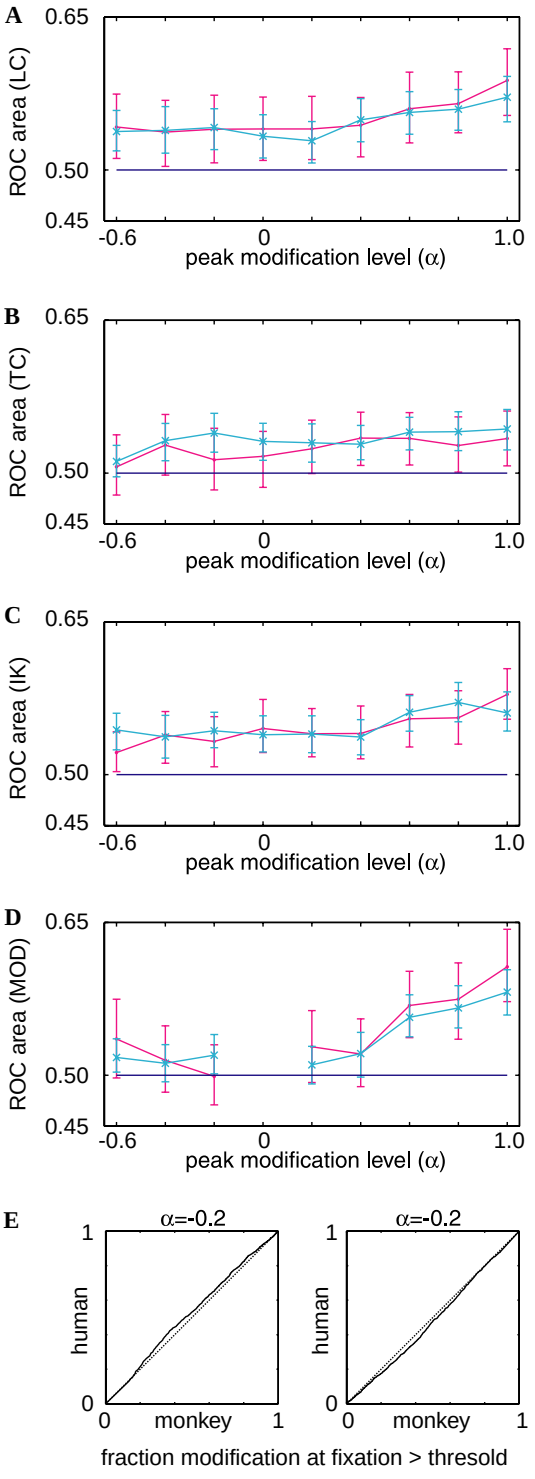

Fig. 6. ROC analysis. (A) ROC areas for luminance-contrast in monkeys (magenta dots) and humans (cyan crosses) for different peak modification levels (α). Error-bars indicate 99% confidence intervals as estimated by bootstrap method; chance at 0.5. (B) ROC areas for texture-contrast in monkeys and humans. Markers as in (A). (C) ROC areas for IK-saliency in monkeys and humans. Markers as in (A). (D) ROC areas for contrast-modification in monkeys and humans. Markers as in (A). (E) "Hits" (fixations above threshold) in humans plotted against hits in monkeys for varying threshold. Left: peak modification level $\alpha = -0.2$; right: peak modification level $\alpha = +0.2$. Dotted line indicates diagonal (chance).

in the case of texture-contrast. While this analysis confirms the finding above that these features have a positive correlation to overt attention, it also shows that the effects are—though significant—small: the maximum ROC area stays below 0.59 for luminance-contrast (0.59), texture-contrast (0.54) and IK-saliency (0.57). In addition the ROC area is directly comparable across features. Comparing luminance-contrast to texture-contrast we find for all but one peak modification level ($\alpha = 0.2$ in humans), ROC area to be larger for luminance-contrast. This suggests that—across all peak modification levels—texture-contrast is slightly less related to overt attention than luminance-contrast.

ROC areas for contrast-modification are above chance for all but one modification level ($\alpha = -0.2$ in monkeys, 0.499), and reach a maximum of 0.61 for monkeys and 0.58 for humans at $\alpha = 1$ (Fig. 6D). Although these data show that an effect of modifications is weak to absent for small modification levels on a trial-by-trial basis, we may nevertheless analyze whether humans or monkeys are more susceptible to the modifications using a similar method of signal-detection analysis. Instead of plotting "hits" for each species compared to "false alarms" in the same species, we plot hits of humans versus the hits of monkeys for varying thresholds of contrast-modification. For $\alpha = -0.2$ (Fig. 6E, left) we find the resulting curve being above the diagonal. This implies that for the same modification threshold more fixations of humans are above threshold than for monkeys. Consequently—as the modifications are of negative value—monkeys are more susceptible to the modifications than humans at $\alpha = -0.2$. For $\alpha = +0.2$ (Fig. 6E, right) the curve is slightly below the diagonal. Hence, more fixations of monkeys are above threshold than for humans. As the modification is positive, again monkeys are more susceptible to modifications at $\alpha = +0.2$. In both cases the effects are small with the areas under the curve being 0.52 and 0.48, respectively, but consistent with the data obtained with the statistical analysis above. In summary, the ROC measure confirms the result of the statistical analysis regarding the difference between species. However, it also points out that the prediction of any simple model for static scenes on fixation is rather poor. This will be the rationale to employ a model-independent measure of how well human and monkey fixations are interrelated below (Section 3.9).

3.8. Time-course of saliency

Regarding human data, there has been considerable debate on whether or not there is a difference between the early fixations on a given stimulus and the later fixations (Parkhurst et al., 2002; Tatler et al., 2005). Hence, we here analyze the development of the saliency measures S_{IK} and S_{mod} over the course of the 10 first fixations on each stimulus.

In the case of S_{IK} a 3-factor ANOVA over the factors fixation number, peak modification level and species

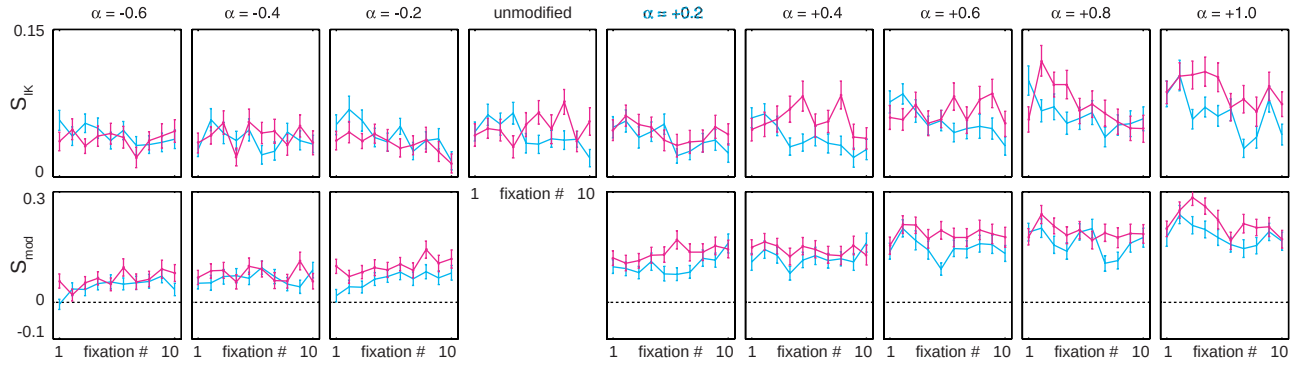

Fig. 7. Fixation-by-fixation analysis. Saliency measures plotted over increasing fixation numbers for different peak modification levels (increasing from left to right). Top row, saliency of IK-saliency (S_{IK}) bottom row, saliency of modifications (S_{mod}). In each panel cyan marks human data and magenta monkey data. Errorbars denote standard errors over images.

reveals an interaction between those 3 factors ($p = 0.03$). Consequently, we analyze the effect of fixation number and species separately for each peak modification level (Fig. 7, top). For unmodified scenes, we do not observe any effect, neither of fixation number ($p = 0.12$), species ($p = 0.17$) nor of interaction between those two ($p = 0.42$). In case of modified images, we find a significant species effect for $\alpha = -0.2$ ($p = 0.04$), $\alpha = +0.4$ ($p = 0.0002$), $\alpha = +0.8$ ($p = 0.02$) and $\alpha = +1.0$ ($p < 10^{-4}$; $F(1,8) = 13.53$). Fixation number shows a significant effect at $\alpha = +0.8$ ($p = 0.03$) and $\alpha = +1.0$ ($p = 0.008$). A significant interaction between species and fixation is found at $\alpha = +0.4$ ($p = 0.02$) and $\alpha = +0.8$ ($p = 0.04$). In the cases, where there is a significant effect of species, this effect mostly arises from monkeys having smaller S_{IK} values for the first fixations, but a slower decay of these values over time. In cases where there is significant dependence on fixation number, S_{IK} is high for the first 1 or 2 fixations and then gradually decays and the decay is faster for humans than it is for monkeys. This is in line with the trend that humans fixate the single most salient spot—as defined by IK-saliency—on average slightly earlier (after 6.4 ± 1.3 fixations) than monkeys (7.3 ± 0.6), a trend which is, however, not significant ($p = 0.42$, t-test). The relatively rapid decay observed in humans for the relation of bottom–up features to fixation is consistent with the view that bottom–up features—if at all—are relevant for humans only during the first few fixations. However, it is important to note that we observe such a dependence on fixation number only for high modification levels, which potentially affect the global appearance of the stimulus as natural. In conclusion, there is a slight tendency for humans to fixate points of higher IK-saliency earlier than monkeys. Since this effect is only present for high peak modification levels, this may be attributed to saliency arising from modifications deviating from the general global natural appearance.

For S_{mod} the 3-factor ANOVA reveals no 3-factor interaction ($p = 0.70$). There is a strong dependence on species and modification ($p < 10^{-4}$ for the factors species and modification). As there is significant interaction between peak modification level and fixation number ($p = 0.003$), we again analyze the different peak modification levels separately (Fig. 7, bottom). We find a species difference for all but one peak modification level ($\alpha = -0.6$: $p = 0.01$; $\alpha = -0.4$: $p = 0.12$; $\alpha = -0.2$: $p = 0.0009$; $\alpha = +0.2$: $p = 0.0004$; $\alpha = +0.4$: $p = 0.0005$; $\alpha = +0.6$: $p = 0.001$; $\alpha = +0.8$: $p = 0.001$; $\alpha = +1.0$: $p < 10^{-4}$). In all cases where the species difference is significant, monkeys are more susceptible to the modifications than humans over almost all fixation numbers (64/70 data-points, Fig. 7, bottom). Only for the strongest modification ($\alpha = +1.0$) there is a significant effect of fixation number ($p = 0.0001$), which is—however—not monotonic over time; and at no peak modification level, there is any interaction between species and fixation number ($p > 0.27$ for each α). These data demonstrate a strong difference between humans and monkeys with respect to the influence of S_{mod}, when comparing the data fixation by fixation. It is important to note the difference to the analysis of Section 3.6. The present analysis is not confounded by the fact that monkeys have a shorter fixation duration and thus make on average more fixations on each given stimulus, which might include larger numbers of less salient targets solely on the basis of this fact. Still—as long as there is no effect of fixation number itself—the effect of contrast modifications is most prominent in the natural range, which is consistent with the aforementioned analysis. In conclusion, the fixation-by-fixation analysis demonstrates that monkeys are more susceptible than humans to an imposed local low-level feature, such as contrast modification.

3.9. Model independent measure

Although all features under investigation showed a significant relation to overt attention, ROC analysis revealed that the prediction performance of these features and models was nevertheless poor. Hence, we apply the normalized distance (see Section 2.4.9) as a measure, which measures, independent of model-assumptions and features, how well the fixation pattern of one subject predicts the fixations of another subject. By construction, the normalized distance takes larger values the better one subject predicts

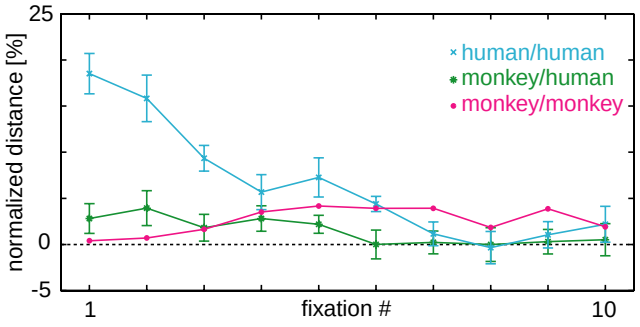

Fig. 8. Normalized distance. Normalized distances (% smaller than chance level) for increasing fixation numbers. Higher values correspond to better prediction. Mean over pair-wise comparisons in humans ($n = 9$, cyan), monkeys ($n = 1$, magenta) and between species ($n = 14$). Errorbars denote standard errors across pairs of subjects.

the other subject. A normalized distance of 0% implies that there is no image-specific consistency between the two subjects compared. Averaged across all 9 inter-human pair-wise comparisons, the normalized distance for the first human fixation is at 19%. This value drops monotonically towards 0% for the first 7 fixations (Fig. 8, cyan line), with the first 6 fixations are significantly larger than 0% ($p < 10^{-4}$; $p = 2 \times 10^{-4}$; $p = 1 \times 10^{-4}$; $p = 0.02$; $p = 0.009$; $p < 10^{-4}$, t-test for 1st to 6th fixations). More importantly, for the first 3 fixations the normalized distances are significantly larger across all humans than the respective normalized distances for monkeys ($p < 10^{-4}$, t-test for each of the first 3 fixations; Fig. 8, magenta) and than the mean inter-species distances ($p < 0.002$, t-test; Fig. 8, green) Furthermore, for the first 3 fixations even the lowest value of any of the inter-human comparisons exceeds the normalized distance between the two monkeys. This implies that for the first 3 fixations any human predicts the fixation of any other human better than one monkey predicts the other monkey.

To check for the possibility that monkeys look at the same items, but just in different order, we for each pair of subjects measure the minimum difference between any two fixations on the same stimulus. Since in this case the minimum over all control fixations would underestimate the true random value, we for each image use the distance to a randomly chosen different image as baseline. Computing the analogous measure to normalized distance yields 2.7% for the comparison of the two monkeys and $19.6 \pm 9.9\%$ on average for the human comparisons, which is significantly larger ($p = 9 \times 10^{-4}$, t-test). The intra-species value for humans is also significantly larger than the inter-species comparison ($9.2 \pm 6.0\%$; $p = 0.005$, t-test), which rules out that just one of the monkeys is behaving differently from all other subjects. Furthermore, again any pair of humans predicts each other better than the two monkeys (minimum intra-humans: 7.2%). These data demonstrate that—irrespective of any specific features—humans have a higher tendency to look at similar items than monkeys. In conjunction with the finding that intrinsic low-level features (luminance-contrast, texture-contrast, and IK-saliency) are related to human and monkey overt attention to a similar degree, this result provides further evidence that humans more than monkeys rely on high-level features or on cognitive scene interpretation.

4. Discussion

In the present study, we show that intrinsic low-level features such as luminance-contrast, texture-contrast, and saliency—as defined by a model on the basis of luminance and orientation differences—are related to overt attention in humans and monkeys to a similar degree. However, an imposed low-level feature that has no spatially fixed relation to higher order items, affects humans and monkeys differently. This difference is most prominent if the modification does not introduce deviations from the global context of the stimulus being natural. In addition, humans show a higher consistency as to which items they direct their attention to. Taken together, these data suggest that humans and monkeys might employ different processing strategies under natural conditions: while monkeys predominantly direct their attention according to low-level features, humans are more driven by high-level scene interpretation.

Since its original formulation in 1985 the saliency map model for visual attention (Koch & Ullman, 1985) has undergone various modifications. The basic scheme, however, has remained unchanged (see Itti & Koch, 2001, for review). The stimulus is analysed in distinct feature domains. Maps of differences in each feature are generated at different spatial scales and the resulting maps are added linearly. Selection of the next attended location in the saliency map is made according to a winner-takes-all scheme, which penalizes previously attended locations to suppress a return to the previously visited location (“inhibition-of-return”).

Several recent studies have compared human eye-movements to the prediction of saliency-map models (Itti, 2005; Parkhurst et al., 2002; Peters, Iyer, Itti, & Koch, 2005; Tatler et al., 2005). While all these studies find that the prediction performance of their respective saliency map models is significantly above chance, the reported effects—when taking only luminance and orientation into account—are small. Using the ROC area as measure, Tatler et al. (2005) find a maximum of 63% for contrast, which is decreased for lower spatial frequencies, thus on average (given that low frequencies are predominant in natural scenes) well compatible with our result observed on unmodified images (54 and 53% for monkeys and humans). When taking only luminance or orientation channel into account, Itti (2005) finds—using a different metric—an effect of about 10% above random fixations. In this case the random baseline is uniformly sampled over the image area, leading to central bias being a potential confound that could lower this number (see discussion in Tatler et al., 2005). Finally Peters et al. (2005), using so-called

"normalized scanpath saliency" find that on average IK-saliency (or "BSM" in their terms) at fixations is 0.69 standard-deviations above the mean IK-saliency in each image. For outdoor grayscale images alone their value is 0.64. While this value well exceeds the value for uniformly sampled random fixations, a control similar to our baseline yields a value of 0.39 (Peters, personal communication). Finally, the ROC area analysis on the Peters et al. (2005) data yields 68%, but a control with randomly shuffled images also yields 63% on their data (personal communication)—unlike in our case, where the control is always close to 0.5. This is in line with the 4% points above chance we observe. Given the difference in exact protocol and image material, the data of Peters et al. (2005) are consequently also compatible with the present data on humans. The results of all these studies on human overt attention—though not always using directly comparable measures—are in general in agreement with the size of the effects we observe in humans and monkeys. In general, the prediction of such static models and features is—though significantly above chance—rather poor as compared to models for dynamic scenes (Itti, 2005).

The aforementioned data confirm that saliency-map models predict human fixations in natural scenes above chance. However, only few studies have systematically assessed the impact of individual features in overt attentional behavior without specific model assumptions. In humans, a correlative effect of luminance-contrast and fixation had first been described by Reinagel and Zador (1999) and was later confirmed in several other studies (Einhäuser & König, 2003; Krieger et al., 2000; Parkhurst & Niebur, 2003). Using a larger set of images and thus presenting each individual image fewer times than in our earlier study, we confirm the correlative effect of luminance-contrast with overt attention in humans. We also use a larger presentation size and find that the effect is already visible without restricting the analysis to a certain frequency range. This facilitates interpretation of the data for which the stimulus was modified in the contrast domain. Most importantly, we demonstrate that the effect is also present in monkeys and that it is of similar size to that observed in humans, at least when considering luminance-contrast.

To account for results from a previous study from our laboratory using human subjects alone (Einhäuser & König, 2003), Parkhurst and Niebur (2004) suggested an extension of the saliency map model, which makes distinct predictions for the effects of luminance-contrast and texture-contrast. Using some general assumptions on the relative scale of first and second order effects, this model predicts eye-tracking and psychophysical data. The model predicts that texture-contrast is approximately 10-times more important than luminance-contrast in attracting human overt attention. Here, we use a definition of texture-contrast that does not require specific model assumptions and is a canonical generalization of the luminance-contrast definition. With respect to human subjects we confirm an interaction between texture-contrast and overt attention. In addition, we find an interaction between texture-contrast and overt attention in monkeys. The interaction is not significantly different to the corresponding interaction seen in the human results.

In terms of understanding the mechanisms underlying overt attention, this description is phenomenological and therefore incomplete. The model cannot exclude the possibility that the observed relations do not arise from some unknown image property in natural scenes that is correlated with both luminance-contrast and overt attention. In an earlier study of human overt attention using the same modification paradigm described here, we found evidence for such a higher order image property (Einhäuser & König, 2003). Texture-contrast may account partly for this observation (Parkhurst & Niebur, 2004). However, our present findings suggest that the saliency of texture-contrast (S_{TC}) does not depend on the modification level, while the effect of modifications (S_{mod}) does. Thus, this demonstrates the need to consider an additional higher order property.

The main goal of the present study was to compare overt attention in humans and monkeys viewing natural scenes. While we find no difference in the phenomenology of overt attention on unmodified images, contrast modifications appeared to produce different effects in the two species. The conceptual advantage of measuring the effect of such modifications as compared to features inherent in natural scenes, is that the feature "contrast-modification" has no fixed correlation to any localized high level feature. Hence, inter-species differences of contrast-modification cannot be attributed to such a local high-order bottom–up feature. By construction, contrast-modification correlates to low-level features such as luminance-contrast and texture-contrast. As we observe no inter-species difference for those features, it is unlikely that their relation to contrast-modification accounts for the species difference. Finally, modifications introduce a local deviation from the fact that the scene is natural. This may well account for monkeys and humans attending to the modifications, especially if they are strong. However, it cannot fully explain the inter-species difference, which is most prominent for contrast modifications within the range of luminance-contrasts observed in natural images.

There are a variety of possible explanations for the observed inter-species difference. A straight-forward explanation might suggest that monkeys who perform tasks involving fixations for several hours a day actively search for a fixation spot in the natural image and thus have a lower threshold for finding the modified regions. This interpretation, however, is unlikely for several reasons. First, neither monkey had been trained to perform search tasks or tasks involving natural scenes. Second, monkeys are rewarded irrespective of their eye-movements. Third, if this interpretation would hold, the difference should also be equally strong for modifications inside and outside the range of natural contrast fluctuations.

An alternative interpretation might suggest that monkeys and humans perceive images of natural scenes differently. One possibility would be that monkeys are not capable of interpreting a photograph as a representation of the real world. However, this conclusion is not consistent with the results of recordings from object-sensitive neurons in macaque inferotemporal cortex (Booth & Rolls, 1998). They were not able to identify differences in the responses to real objects and image of the same objects. In view of these results, there is little evidence that processing of natural images and real scenes is different in the monkey visual cortex. Furthermore, the speculation would invalidate all monkey experiments involving images of natural scenes as being representative of natural conditions.

It is entirely plausible, however, that the two species differ in the degree to which they perceive images as "natural." In this view humans would assign a higher order interpretation to the stimulus. top–down signals would then guide their attention shifts based on the contrast-distribution usually consistent with such stimuli more than bottom–up signals from the actual stimulus. Only when the modifications were sufficiently large to appear unnatural would bottom–up signals from the stimulus drive overt attentional behavior. If monkeys lack similar knowledge on the statistical relations in natural scenes—potentially because they grew up in a cage environment—they are more driven by the bottom–up signal and hence more susceptible to the modifications. Such a view is supported by Parkhurst and Niebur (2003), who point out that the type of image (natural vs. artificial, outdoor vs. indoor) has great impact on human overt visual behaviors. It should be emphasized that the proposed transition from a dominant bottom–up mode to a dominant top–down mode may be gradual. In this view, small modifications would suffice for monkeys to be dominated by bottom–up cues, while the threshold for humans to shift away from a top–down dominated mode is higher.

To achieve comparable behavioral relevance for both species we exclusively used outdoor scenes containing few nameable objects. As in any study comparing monkey and human behavior, we still cannot exclude, however, the possibility that the two species implicitly assign a different interpretation to the task and that this in turn affects the relative importance of bottom–up and top–down signals.

While further experiments are needed to test these hypotheses, our main finding is unaffected; rhesus monkeys employ apparently different strategies to guide overt visual attention when viewing natural scenes to those employed by humans. The conclusion implied by our result on the applicability of monkey studies to human observers is twofold: for studies that only rely on correlative effects of local low-level features to attention, our data assure that these effects are similar in both humans and monkeys. In particular, saliency map models, which use local contrasts, are likely to predict monkey attention as well—or as badly—as they predict human attention. This provides some justification to test predictions of such computational concepts—originally developed for human attention—in monkey models. On the other hand several studies compare the processing of natural scenes to noise stimuli that are similar in their local statistical structure, but do not have a globally natural appearance (e.g., Guo et al., 2003; Rainer, Augath, Trinath, & Logothetis, 2002). When relating the results of such studies to human processing (which none of the studies themselves explicitly implies, though), one has to take into account the species differences in weighing local features against high-level scene interpretation. Irrespective of the underlying cause of this difference, our results hence emphasize the care that should be taken when relating studies performed in the monkey to human perception.

Acknowledgments

We thank Honda RI Europe (W.E.), Volkswagen Foundation, (W.K., and K.P.H., grant "Plasticity of Spatial Cognition"), Deutsche Forschungsgemeinschaft (KPH, SFB 509 "Neurovision," TP B2) and the Swiss National Science Foundation (W.E., project-no: PBEZ2–107367; PK, Grant No. 31-61415.01) for their financial support. We are grateful to D. Walther for supplying a stand-alone matlab-implementation of the saliency-map model and to A. Horstmann and Dr. W. Lindner for their technical assistance in carrying out the experiments.

References

Astafiev, S. V., Shulman, G. L., Stanley, C. M., Snyder, A. Z., Van Essen, D. C., & Corbetta, M. (2003). Functional organization of human intraparietal and frontal cortex for attending, looking, and pointing. *The Journal of Neuroscience, 23*, 4689–4699.

Booth, M. C., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex, 8*, 510–523.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17*, 1089–1097.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9*, 303–308.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature, 391*, 481–484.

Guo, K., Robertson, R. G., Mahmoodi, S., Tadmor, Y., & Young, M. P. (2003). How do monkeys view faces? A study of eye movements. *Experimental Brain Research, 150*, 363–374.

Horwitz, G. D., & Newsome, W. T. (1999). Separate signals for target selection and movement specification in the superior colliculus. *Science, 284*, 1158–1161.

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 modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194–203.

Itti, L., & Baldi, P. (2005) A principled approach to detecting surprising events in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 631–637.

Itti, L. (2005) Quantitative modeling of perceptual salience at human eye position. Visual cognition, in press.

Judge, S. J., Richmond, B. J., & Chu, F. C. (1980). Implantation of magnetic search coils for measurement of eye position: An improved method. *Vision Research, 20*, 535–538.

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.

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature, 384*, 74–77.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*, 9–16.

Mazer, J. A., & Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search: Evidence for a ventral stream salience map. *Neuron, 40*, 1241–1250.

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.

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., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16*, 125–154.

Parkhurst, D., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience, 19*, 783–789.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Peters, R. J., Iyer, A., Itti, L, & Koch, C. (2005). Components of bottom–up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42.

Rainer, G., Augath, M., Trinath, T., & Logothetis, N. K. (2002). The effect of image scrambling on visual cortical BOLD activity in the anesthetized monkey. *Neuroimage, 16*, 607–616.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems, 10*, 341–350.

Robinson, D. L., & Petersen, S. E. (1992). The pulvinar and visual salience. *Trends in Neuroscience, 15*, 127–132.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *The Journal of Neuroscience, 21*, 1340–1350.

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.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology, 12*, 97–136.

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.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Van Essen, D. C., Drury, H. A., Joshi, S., & Miller, M. I. (1998). Functional and structural mapping of human cerebral cortex: Solutions are in the surfaces. *Proceedings of the National Academy of Sciences of the United States of America, 95*, 788–795.