Cerebral Cortex, 2017; 1–15

doi: 10.1093/cercor/bhw399
Original Article

ORIGINAL ARTICLE

Differential Contribution of Low- and High-level Image Content to Eye Movements in Monkeys and Humans

Niklas Wilming[1,2,3,4,9], Tim C. Kietzmann[1,5], Megan Jutras[2,3,9], Cheng Xue[6], Stefan Treue[6,7,8], Elizabeth A. Buffalo[2,3,9] and Peter König[1,4]

[1]Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany, [2]Department of Physiology and Biophysics, University of Washington, Seattle, WA 98195, USA, [3]Yerkes National Primate Research Center, Atlanta, GA 30329, USA, [4]Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [5]Medical Research Council, Cognition and Brain Sciences Unit, Cambridge CB2 7EF, UK, [6]Cognitive Neuroscience Laboratory, German Primate Center - Leibniz-Institute for Primate Research, Goettingen, Germany, [7]Faculty of Biology and Psychology, Goettingen University, Goettingen, Germany, [8]Leibniz-ScienceCampus Primate Cognition, Goettingen, Germany and [9]Washington National Primate Research Center, Seattle, WA 09195, USA

Address correspondence to Niklas Wilming. Email: nwilming@uke.de

Abstract

Oculomotor selection exerts a fundamental impact on our experience of the environment. To better understand the underlying principles, researchers typically rely on behavioral data from humans, and electrophysiological recordings in macaque monkeys. This approach rests on the assumption that the same selection processes are at play in both species. To test this assumption, we compared the viewing behavior of 106 humans and 11 macaques in an unconstrained free-viewing task. Our data-driven clustering analyses revealed distinct human and macaque clusters, indicating species-specific selection strategies. Yet, cross-species predictions were found to be above chance, indicating some level of shared behavior. Analyses relying on computational models of visual saliency indicate that such cross-species commonalities in free viewing are largely due to similar low-level selection mechanisms, with only a small contribution by shared higher level selection mechanisms and with consistent viewing behavior of monkeys being a subset of the consistent viewing behavior of humans.

Key words: human macaque comparison, low-level salience, oculomotor control, overt visual attention

Introduction

Eye movements are an essential aspect of our everyday behavior, because the direction of gaze determines what parts of our visual environment are processed with high-accuracy foveal vision. The importance of eye movements is reflected in their ubiquity (saccades occur at a rate of ca. 3–5 Hz) and in viewing strategies that are specifically tailored toward behavior (Land and Hayhoe 2001; Land and Tatler 2001; Sullivan et al. 2012; Johnson et al. 2014). Understanding the underlying cortical saccade target selection process is therefore fundamental for understanding vision and human cognition at a larger scale (Petersen and Posner 2012).

The processes underlying such overt visual selection have traditionally been approached by behavioral measurements, mostly performed on humans, and by electrophysiology, performed in macaque monkeys, which are the most prominent

© The Author 2017. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

model system for studying attentional selection (Bisley 2011). This approach rests on the fundamental assumption that common neural mechanisms are at play in both species. Only then is the link of (monkey) neuronal mechanisms to (human) behavior valid. To verify this assumption, it is crucial to investigate whether the overall behavioral phenomenon to be understood, overt visual attention, is comparable in human and macaque. Here, we addressed this issue by comparing patterns of eye movements recorded from 11 monkeys and 106 human observers, while they were performing a task that comes natural to both species: free viewing. Free viewing has the advantage that it does not require explicit instructions or training. Furthermore, monkeys do not need to be externally rewarded because they are intrinsically motivated to freely explore visual scenes. Thus, free viewing can be completed without instructions, training, or explicit reward and it therefore remains undefined which parts of the stimuli should be attended. While tasks that require training can result in comparable behavior, they potentially mask the natural modus operandi of overt visual selection. Consequently, if free-viewing behavior is similar across humans and monkeys, it is because both species have intrinsically chosen a selection strategy that emphasizes the same locations, not because the task dictates which locations promise success. Free viewing therefore provides an unbiased view of the natural selection processes of overt attention in macaques and humans.

To compare viewing behavior across species, we followed a 2-staged approach. We first compared cross-species similarity in fixation locations. Using data-driven agglomerative clustering, we found that the 2 species form distinct clusters of viewing behavior, indicating species-specific selection strategies. Despite these differences, cross-species predictions were clearly above chance, indicating shared behavior. Following these observations, we tested in how far these differences and similarities in viewing behavior can be understood in terms of different explanatory dimensions, commonly assumed to jointly contribute to the guidance of eye movements. Distinctions are typically made between stimulus-dependent, context-dependent, and geometrical factors. Stimulus-dependent influences are, for example, the saliency conveyed by low-level images features (Itti and Koch 2001; Parkhurst et al. 2002), objects (Einhäuser et al. 2008; Nuthmann and Henderson 2010), and stimulus interpretation (Kietzmann et al. 2011). Context-dependent aspects include the task (Castelhano et al. 2009; Betz et al. 2010) and scene context (Torralba et al. 2006; Kietzmann and König 2015). Geometrical aspects include oculomotor biases, like the center bias of fixations (Tatler and Vincent 2009) and saccadic momentum (Smith and Henderson 2011; Wilming et al. 2013). All of these aspects interact in the selection process and consistently make strong contributions to the guidance of eye movements (Kollmorgen et al. 2010). These well-established dimensions therefore provide a good starting point to understand the observed similarities and dissimilarities across species. However, while low-level stimulus features are a well-controlled and well-studied explanatory dimension, higher level factors are less clearly defined in the context of free viewing on natural scenes, which comprises the current data set. We therefore initiate our investigation by comparing the relative contribution of low-level stimulus features and subsequently test any other residual, presumably higher level, factors across both species.

To estimate the relative contribution of these different factors, we first estimated the consistency of viewing behavior within humans and monkeys. The consistency within a species measures the similarity of viewing behavior across many observers and thereby forms an upper bound for the similarity of fixation selection strategies (Wilming et al. 2011). The reliability of such consistency estimates depends on the number of observers (Wilming et al. 2011). In particular, small groups tend to underestimate the consistency within a group of observers and consistency estimates approach an asymptotic level as the group size increases. In this study, we compare 11 monkey and 106 human observers and, to our knowledge, our data set is the first to reach this asymptotic level. This analysis revealed an overall reduced consistency in macaques compared with humans. We then decomposed the respective upper bound into a stimulus-driven part and residual viewing behavior that must be driven by the remaining explanatory dimensions. These analyses revealed that the predictive power of low-level features is comparable across species. This implies that low-level features can explain large parts of the consistent macaque viewing behavior, but provide comparably limited predictive power in humans. However, the absolute impact of different low-level feature dimensions exhibits large similarities across species, suggesting that similar low-level selection mechanisms are at play in both macaques and humans. Following this observation, we tested whether commonalities across species can be observed beyond these, presumably shared, low-level mechanisms. We found that a joint model, combining low-level saliency and cross-species predictions, that is, data from humans to predict macaques and data from macaques to predict humans, only yields marginally better prediction accuracy than the low-level model alone. Thus, while our data suggest that human and macaque share common low-level selection mechanisms, other, potentially higher level effects only generalize to a small degree across species.

Materials and Methods

Participants

Eye movements were recorded from 11 rhesus monkeys (Macaca mulatta, 8 male). Recordings were carried out across 3 different locations. Data from 4 monkeys were recorded at the Yerkes National Primate Research Center (YNPRC) in Atlanta, USA, in accordance with National Institute of Health guidelines and protocols were approved by the Emory University Institutional Animal Care and Use Committee. Data from 3 additional monkeys were recorded at the Washington National Primate Research Center (WNPRC) in Seattle, USA, in accordance with National Institute of Health guidelines and protocols were approved by the Washington University Institutional Animal Care and Use Committee. Data from 4 additional monkeys were recorded at the German Primate Center (DPZ) in Goettingen, Germany, in accordance with European Directive 2010/63/EU, corresponding German animal welfare law and institutional guidelines. The animals were group-housed with other macaque monkeys. The facility provides the animals with an enriched environment (including a multitude of toys and wooden structures, natural as well as artificial light, exceeding the size requirements of the European regulations, including access to outdoor space; Calapai et al. 2016). All procedures were approved by the appropriate regional government office (Niedersaechsisches Landesamt fuer Verbraucherschutz und Lebensmittelsicherheit, LAVES). Eye-movement recordings from humans came from 2 previous studies that used the same stimuli and comparable tasks. We analyzed data from 106 observers, 58 from Açik et al. (2010) and 48 from Onat et al. (2014). Açik et al. recruited participants from different age ranges (18 children with mean age 7.6 years, 6 female; 23 university students with mean age 22.1, 11 female; 17 older adults with mean age 80.6, 10 female). Onat et al. recruited 48 students (mean age 23.1 years, 25 male). The majority of participants were therefore recruited from the general

student population at the University of Osnabrück. All participants (main and control experiments) gave written informed consent and all experimental procedures for eye-movement recordings from humans were in compliance with guidelines described in the Declaration of Helsinki and approved by the ethics committee of the University of Osnabrück.

Stimuli

Stimuli consisted of 192 images from 3 different categories (64 images in each category). "Natural" scenes were taken from the "McGill Calibrated Color Image Database" and depict mainly bushes, flowers, and similar outdoor scenes. "Urban" scenes depicted urban and manmade scenes taken around Zürich, Switzerland. "Fractal" images were taken from Elena's Fractal Gallery, Maria's Fractal Explorer Gallery, and Chaotic N-Space Network available online, and depicted computer-generated fractals. Figure 1A shows example stimuli from all categories. Please see Açik et al. (2010) for more details.

Apparatus

Recordings at the Yerkes National Primate Center and the Washington national primate center were carried out with an ISCAN infrared eye-tracking system while each monkey sat in a dimly illuminated room. Monkeys were head fixed during recordings. Stimuli were presented on a CRT Monitor with a resolution of 800 × 600 pixels and a refresh rate of 120 Hz. The viewing distance was 60 cm. Recordings at the German Primate Center were carried out in similar conditions but an SR-Research EyeLink 1000 was used for recording of eye movements. The viewing distance was 57 cm and stimuli were presented on a TFT screen (60 Hz, 1920 × 1080 pixels). The size of the images in degrees of visual angle was matched between all 3 setups (33.3° × 25°).

Human eye movements were recorded with an EyeLink 1000 system (Açik et al. 2010) or an EyeLink II system (Onat et al. 2014). Human eye-movement recordings were carried out at the University of Osnabrück, Germany. Onat et al. presented stimuli on a CRT Monitor with a resolution of 1280 × 960 pixels and a refresh rate of 85 Hz. The viewing distance was 80 cm.

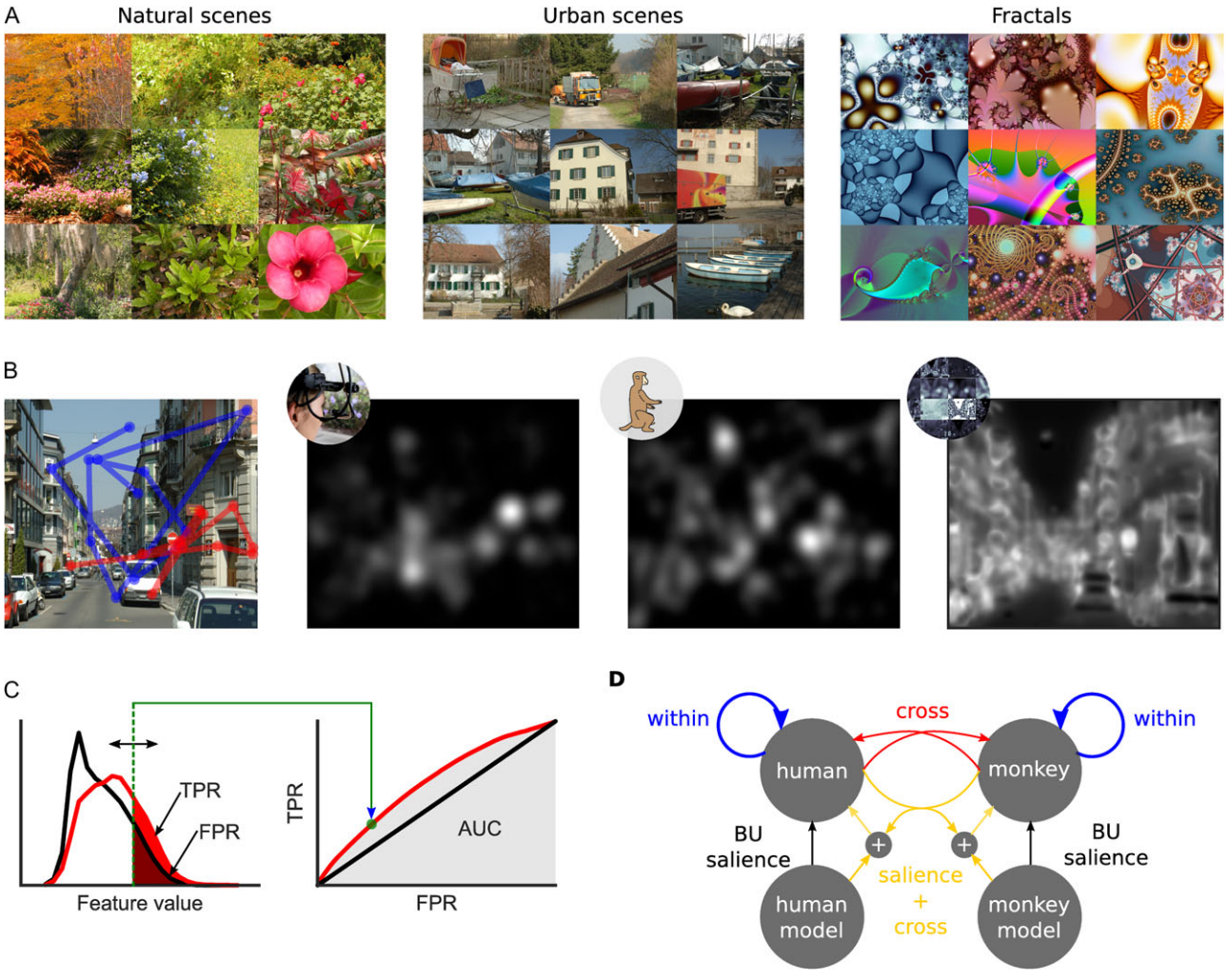

Figure 1. Study overview. (*A*) Nine example images from the categories natural scenes, urban scenes, and fractal scenes. (*B*) One example stimulus with one monkey (blue) and one human (red) eye-movement trace. The next 3 plots show the density of human fixations on the example image, the density of monkey fixations, and a predicted saliency map for the example stimulus. (*C*) The computation of AUC values. Left: Feature values at fixated locations (red) and non-fixated control locations (black) are classified as fixated or not fixated by a simple threshold (green dotted line). Moving the threshold and plotting the false alarm rate (FPR) against the hit rate (TPR) generates a receiver operating characteristic (ROC) curve which is shown on the right. The area under this curve (AUC) is a measure of classification quality. (*D*) Different predictors and comparisons in this study.

Açik et al. used a 60-Hz TFT screen with the same resolution and a viewing distance of 65 cm. The size of the images in degrees of visual angle was 35° × 26° (Açik et al. 2010) and 30° × 22° (Onat et al. 2014) for the human recordings.

Procedure and Task

Monkeys performed a free-viewing task and were not explicitly rewarded for image viewing. Images were shown until a total looking time inside the image of 10 s had accumulated. Monkeys at the Washington and Yerkes National Primate Center carried out a color change task between free-viewing trials. In this task, the monkey was required to hold a touch bar and maintain fixation on a small rectangle (0.3°) that appeared at various locations on the screen. The rectangle changed color from gray to an equiluminant yellow at a randomly chosen time between 500 and 1100 ms. Upon release of the touch bar within 500 ms after the color change, a drop of blended chow was delivered as reward (Jutras et al. 2009; Jutras and Buffalo 2010). Monkeys at the German Primate Center carried out a fixation control task that required them to saccade to a point on the screen and were rewarded for maintaining fixation for 1.25 s. Data from the control trials were not included in subsequent analyses.

Macaque recordings were carried out on 3 consecutive days. This kept sessions short enough for monkeys to attend to all images without losing interest. On the first 2 days, 66 randomly sampled images were shown twice and on the last day, 60 images were shown twice. The order of presentation was the same for all monkeys. Due to a technical error, the data from 1 day from one monkey was discarded. To increase the amount of available data, and to potentially compare effects of memory later on, 2 monkeys repeated the experiment after 4 weeks.

Human observers were instructed to "freely view" the same images for 6 s (Onat et al. 2014). In contrast, Açik et al. (2010) showed images for 5 s and instructed participants to "study the image carefully". After each image, participants were then shown a 3.2° image patch and had to judge if it was taken from the image presented just before. We consider the 2 tasks to be comparable since the patch recognition task does not require special viewing strategies. In particular, patch locations were drawn uniformly from the entire image and patches are large and easily identifiable such that freely inspecting the image allows successful completion of the task. This was also reflected by the high task performance of the participants (85% across all age ranges). Both studies therefore used similar instructions, and the data were pooled accordingly. All analyses were performed on the first 5 s of image viewing.

Data Pre-processing

Saccade detection for humans was based on 3 measures: eye movement of at least 0.10°, with a velocity of at least 30°/s, and an acceleration of at least 8000°/s^2. After saccade onset, minimal saccade velocity was 25°/s. Saccade detection for monkeys was carried out similarly but we additionally required that each saccade lasted at least 21 ms and traveled at least 0.35° of visual angle. This was necessary to compensate for the lower sampling rate of the ISCANN system (240 Hz vs. 500 Hz and 1 kHz). Samples in between 2 saccades were labeled as fixations.

Monkey eye-tracking data recorded with the additional color change task were calibrated in 2 steps. Before each recording session, monkeys carried out a block of color change trials. Since the color change was subtle, monkeys had to fixate the rectangle in order to detect the color change. We manually adjusted the gain and offset of the eye tracker until fixations were on the color change rectangle. To improve the manual calibration after the recording, we used the color change trials between picture presentations. We fitted a 2D affine transformation (least-squares fit) between average eye position after onset of the color change rectangle and the position of the rectangle in visual space. This took care of translations and skew in the monkey eye-tracking data. Monkeys from the German Primate Center were calibrated using a 12-point calibration grid before the task. Human eye tracking was calibrated with a 12-point grid before the experiment started.

Since the stimulus presentation time was different between experiments (5, 6, and 10 s), we only used the first 5 s of image viewing for subsequent analysis. We rescaled all eye-tracking data to the stimulus size used for monkeys in Atlanta and Seattle (800 × 600 pixel).

Performance Measure: Computation of AUC Values

This study investigated how well different factors predicted fixation locations of humans and monkeys. Specifically, we were interested in the predictive power of bottom-up-salience, within-species consistency and cross-species consistency and fixation densities of individual observers. These factors were quantified by "predictors" (described in detail below) that assign a score to every location in an image, which scales with the predicted likelihood of fixating this location. To assess the quality of each predictor, we evaluated if fixated locations ("actuals") received higher predictor scores than non-fixated control locations ("controls"). We computed the area under the receiver operating characteristic (ROC) curve (AUC), separating feature values at actual and control locations, as our performance measure.

The AUC is computed by classifying actual fixation locations and control locations as fixated or non-fixated based on the respective score at actual and control locations based on a simple threshold. Varying this threshold generated ROC curves for each predictor and the AUC is computed as the area under this ROC curve. The area sums to 1.0, if the classification is perfect, that is, the distributions of score values at actual and control locations are perfectly separated. A value of 0.5 indicates a classification at a chance level. Perfect misclassification results in an area under the curve of zero. To account for the center bias of fixations, control locations were drawn from the spatial bias of each observer (Tatler et al. 2005; Tatler 2007; Tatler and Vincent 2009). That is, control data were taken from the same subject on all other stimuli of the same category. Each predictor was evaluated for every observer and averaged over all stimuli within a category. Finally, we here aim at understanding the factors contributing to the consistent viewing behavior in each species. We will therefore express the predictive power of individual predictors relative to the within-species consistency, which serves as an upper bound. Since an AUC of 0.5 implies chance level performance, we subtract 0.5 from both AUC values before computing the ratio.

The within- and cross-species predictions consist of fixation densities that are generated by smoothing all fixations that form a predictor (e.g. all fixations on an image of one species for the cross-species predictor) with a Gaussian filter of FWHM = 2° and subsequently normalizing the 2D map to unit volume.

Consistency Between Individual Observers and Hierarchical Clustering

To investigate the similarity of viewing behavior of all pairs of observers, we computed AUC that indicate how well fixations from observer A on one stimulus predict the fixations from another observer B on the same stimulus. For each stimulus, we computed a fixation density map from fixations of observer A and computed how well the density values separate actual and control locations from observer B. Averaging across stimuli within a category yielded 3 similarity matrices with 117 × 117 AUC values each. These matrices are not symmetric since observer A and B have different control locations that affect the AUC value. We symmetrized the matrices by computing the average AUC of A predicting B and B predicting A, as required for the subsequent hierarchical clustering.

In a second step, we applied hierarchical agglomerative clustering with Ward's minimum variance criterion to each of the 3 category specific similarity matrices (Ward 1963). Agglomerative hierarchical clustering starts with individual observers and repeatedly merges them into clusters. In each iteration, clusters from the previous iteration (starting with individual observers) are merged such that the total within cluster variance increases as little as possible.

For technical reasons, we could not include all human observers in the clustering. Açik et al. (2010) balanced stimulus presentation across observers, such that pairs of observers saw the entire data set. This implies that these pairs cannot predict each other since they did not see the same images. Agglomerative hierarchical clustering cannot easily deal with such missing values unless they are imputed, and the imputation strategy used can affect the clustering. To circumvent this ambiguity, we decided to use only a subset of the 117 × 117 similarity matrices, excluding observers that had missing values. Notably, however, the main results reported here hold true even if the additional observers are included, independent of the imputation strategy.

Cross- and Within-Species Consistency

How well fixation locations from one species predicted locations from the other was quantified with a cross-species predictor. Computing a fixation density with all fixations from one species on the image in question generated the cross-species prediction for a specific image. This yielded a score for each location in a visual stimulus, which was subsequently used to compute the cross-species AUC. The within-species prediction was similar, but used fixations from the own species (without the subject currently being evaluated). This within-species consistency forms an effective upper bound for the predictability of viewing behavior (Wilming et al. 2011).

The within-species consistency AUC is a biased estimator that tends to produce smaller values with fewer observers. To allow meaningful comparisons between humans and monkeys, we subsampled our human data set to fewer observers (see Statistical Comparisons).

Feature-Fixation Correlations

To compare the influence of image features on viewing behavior, we computed how well different features predicted fixation locations. In total, we computed 16 different features on 3 spatial scales. First, we computed luminance, blue-yellow, red-green, and saturation channels of all stimuli. A first group of features represents a smoothed (Gaussian filter, $\sigma \in$ {8 pixel, 16 pixel, 32 pixel}) version of these simple features. A second group of features is computed by computing contrast in a Gaussian circular aperture ($\sigma \in$ {8 pixel, 16 pixel, 32 pixel}) on the luminance, blue-yellow, red-green, and saturation channels. Contrast was computed according to the following formula:

$$C = (G(X^2, \sigma) - G(X, \sigma)^2)^{1/2},$$

where G convolves the input X with a Gaussian kernel with standard deviation σ. The third group represents the second-level contrast maps (contrast of the contrast maps, $\sigma \in$ {38 pixel, 76 pixel, 151 pixel}), which describe texture contrast. A fourth group consists of an edge detection filter (Sobel filter) and the intrinsic dimensionality 0–2 of local image patches (cf. Onat et al. 2014). Intrinsic dimensionality describes how well a local patch is described by an edge, corner, or surface. Each feature was computed on 3 different spatial scales. This was achieved by down sampling of the original stimulus with a Gaussian pyramid up to 2 times. Each step of the pyramid halves the length and the width of the stimulus yielding the high (no down sampling), medium (once), and low (twice) spatial scale. We then computed the AUC for predicting fixation locations from each feature. This yielded a vector of 48 AUC values for each observer and category.

We then compared AUC values across monkey and human observers to investigate to what extent different features indicate similar predictive power, and therefore suggest similar selection mechanisms. AUC values <0.5 indicate that a feature is anti-predictive of fixation locations, that is, would be predictive if feature maps were multiplied with −1. This has an important implication for comparing features between species. The AUC value of features that are more predictive, for example, in monkeys, will be closer to one compared with the human feature AUC values. Those features that are more anti-predictive in monkeys will be closer to zero. If features are more effective in one species, independent of whether they are predictive or anti-predictive, feature AUCs will fall on a line that pivots around 0.5. In a regression that predicts monkey feature AUC values based on human feature AUC values such an increase in effectiveness would be demonstrated with a slope that is different from one. Regression coefficients larger than one indicate that features effective in humans are more effective in monkeys; coefficients smaller than one indicate that humans are more driven by bottom-up features.

We therefore repeatedly (N = 1000) regressed the pattern of average monkey feature-fixation AUCs onto an average of feature-fixation AUC vectors from 11 randomly sampled human observers. Subsampling AUC values from human observers allowed us to partially estimate the variance introduced by only having 11 monkey observers. Finally, we thus tested whether regression coefficients were different from one with a 2-sided t-test and computed the average variance explained by the regression models.

Salience Model

The details of the salience model have been described previously (Wilming et al. 2013). Briefly, the bottom-up salience model consists of a weighted linear combination of the set of 48 features described above. Weights were obtained by applying a logistic regression, predicting whether an observation (vector of feature values) was taken from a fixation or from a control location. Control locations were sampled from the spatial bias of fixations, that is, data from the same subject recorded on other

stimuli. The respective feature weights were obtained for each species, observer and stimulus category separately. To evaluate the performance of the saliency model, we performed a leave-one-out cross-validation on the level of observers. This ensured that data from the observer to be predicted were never used during weight estimation. Thus, weights were estimated with data from all other subjects, which allowed us to focus on bottom-up influences that are shared across observers in each species.

Statistical Comparisons

Because we had different numbers of observers across species, we needed to correct for potential statistical differences introduced by this imbalance. We accomplished this in 2 ways. First, we bootstrapped 95% confidence intervals (CIs) for human AUC values by repeatedly selecting 11 random human observers (with replacement) and averaging their AUC values. This estimated the distribution of average AUC values to be expected had only 11 human observers participated in the experiment. Mean monkey AUC values falling outside of the 95% CI for humans were interpreted as support for the hypothesis that moneys and humans show a true species difference. For visualization purposes, we also bootstrapped 95% CIs for monkey observers by repeatedly sampling 11 monkey observers with replacement. Furthermore, estimates of within-species consistency, measured via AUC, exhibit a dependency on the number of observers used for prediction: fewer observers produce on average smaller AUC values than larger group sizes (Wilming et al. 2011). When comparing humans with monkeys, any difference in within-species consistency is therefore potentially due to different sample sizes. To account for this possibility, we estimated the effect of smaller group sizes in our human data. We sampled groups of observers with $N = 11$ observers 500 times. In each group, we estimated the consistency between human observers. We were also interested in the development of the performance of within-humans consistency over the number of observers used for computing the prediction. We therefore additionally subsampled each group and predicted each observer in the group with different numbers of observers from the same group. We predicted each observer in a group with 1, 2, 3, …, 10 randomly sampled observers. These subsamplings yielded 500 estimates for each number of predicting observers, which we used to determine intervals that contained 95% of the samples around the estimated mean. Because the estimated mean appeared to follow an exponential function across the number of observers used for prediction, we also fitted an exponential function to the data. This was done to aid data visualization.

Combined Cross-Species and Salience Model and Model Comparison

We found that the bottom-up saliency prediction and the cross-species prediction gave comparable AUC values for predicting fixation locations of both species. We were therefore interested in evaluating whether the cross-species prediction and the bottom-up saliency model explained unique or shared aspects of the viewing behavior. To investigate this issue, we estimated the predictive power of a combined model, including both predictors in a logistic regression, again classifying image locations as fixated or non-fixated. In detail, to classify fixation and control locations of one human observer, we trained a saliency model on all other human observers and used all monkey fixation data to compute the cross-species prediction. The resulting model therefore contained 2 predictors: bottom-up saliency and cross-species predictions. Conversely, to classify monkey fixation locations, we used the data of all other monkeys to train the saliency model together with all human fixation data as cross-species prediction (for a schematically depiction of the approach, see Fig. 1*D*). We predicted each individual observer using a leave-one-observer-out cross-validation. For each logistic regression computed, we standardized the mean and standard deviation of the 2 predictors.

This analysis allowed us to assess whether the combination of bottom-up salience and the cross-species prediction improves over using either predictor alone. To ensure that a larger number of free parameters in the combined model did not provide a predictive advantage, despite reporting cross-validated test performance, we fitted an additional model in which we combined 2 bottom-up salience models: one trained on the same-species data and a second one trained on the cross-species data. This joint model has the same number of free parameters as the original joint model (combining bottom-up saliency and cross-species predictions) and therefore allowed us to verify that any observed improvements are due to additionally explained fixation locations, rather than being of pure technical nature. Our rationale for combining the 2 bottom-up saliency models was that the cross-species bottom-up salience model can only capture those aspects of the cross-species prediction that are driven by bottom-up salience. Any improvements of the cross-species + salience model over such a cross-salience + within-salience model must therefore be due to factors captured in the cross-species prediction that are not due to bottom-up influences.

Control Experiment: Influence of Head Restraint

We conducted a control experiment to investigate whether the difference in the usage of head restraints across species may explain the results observed. Since the results of this experiment are, strictly speaking, not necessary to evaluate the main results we report them here. The influence of head restraints was accomplished by comparing human viewing behavior with and without head restraints (head-fixed vs. head-free condition). Participants freely viewed all images from the urban and fractals category and additionally carried out a guided viewing task, which required them to fixate a dot on the screen that changed position as soon as the eyes landed on its location.

Participants, Apparatus, and Stimuli

We recruited 19 participants (12 female, 7 male, mean age 24, ranging from 18 to 41). Participants watched all images from the urban and fractals category. Stimuli were displayed on a BenQ XL2420T with a resolution of 1920 × 1080 (53.7° × 30.2°). Viewing distance was 60 cm. Gaze position was tracked with a remote SR-Research EyeLink 1000 (SR Research Ltd., Ottawa, ON, Canada) with a sampling rate of 500 Hz. During the head-fixed condition, participants bit into a mouth guard that was individually fitted to their teeth. The mouth guard was then attached to a chin rest that was used as an additional restraint. During the head-free condition, chin rest and mouth guard were removed and participants were only instructed to sit still.

Procedure and Task

During the experiment, participants viewed all images in 2 blocks of 64 images that corresponded to the 2 different conditions. Stimuli, condition, and category order were randomized but images from one category were kept as sub-blocks within each condition block. Stimuli and condition order were counterbalanced across subjects such that pairs of participants saw all

128 images. Images were shown for 6 s and were preceded by a cross that had to be fixated before a trial started. After participants viewed all images in a condition block they carried out 10 trials of a guided viewing task. A dot appeared on the screen, which changed position as soon as the participants gaze location was within 2.2° of the dot. Participants had to "chase" the dot, which changed position 200 times. This allowed us to collect data from 2000 saccades per subject per restraint condition. Subjects were instructed to view images freely. The eye tracker was calibrated before each condition block with a 13-point calibration grid (validation error below 0.55°). The entire experiment took about 60–70 min with briefing, fitting of the mouth guard, experiment, breaks, and debriefing.

Analysis and Results

We analyzed the timing of fixations, the distribution of fixation locations, and the saccadic main sequence in both restraint conditions. Fixation durations were analyzed by computing a cumulative histogram for each participant and restraint condition (Supplementary Fig. 1A). Plotting these against each other showed almost perfectly straight lines along the diagonal. To statistically corroborate this, we fitted a regression on a per subject basis and computed the variance explained by the regression line. The line fits had an average slope of 1.01 (SD = 0.05) and provided an exceedingly good fit ($r^2 > 0.99$). Systematic deviations caused by condition differences would curve the relationship between cumulative fixation durations and would therefore not be well described by a linear relationship. The distribution of fixation locations was analyzed by computing within-condition and across-condition AUC prediction scores in analogy to the within-species and cross-species AUC scores. Specifically, we computed how well participants in the head-free and head-fixed condition predicted each other and how well participants from the head-fixed condition predicted the head-free condition and vice versa. We used leave-one-out cross-validation for all AUC scores. This yielded 19 AUC scores per comparison, which we compared by plotting their cumulative histograms. Additionally, we computed paired *t*-tests which did not reject the null hypothesis of no difference between conditions (Supplementary Fig. 1B and C; urbans: within <-> within: $P > 0.4$, cross <-> within head fixed: $P > 0.19$, cross <-> within head free: $P > 0.84$; fractals: within <-> within: $P > 0.21$, cross <-> within head fixed: $P > 0.14$, cross <-> within head free: $P > 0.38$). We also compared the saccadic main sequence, that is, saccade amplitude versus peak velocity, between head-fixed and head-free conditions. The main sequence in both conditions was highly similar within subject (Supplementary Fig. 1D, $r^2 > 0.98$). In summary, fixing the head in a central position in front of the screen did not change viewing behavior in an appreciable way in this experiment.

Results

We began our comparisons of viewing behavior between species by comparing how well the viewing behavior of individual observers can be predicted by fixation locations from another observer. Each observer in the study, monkey and human, viewed 64 images from 3 different categories (natural, urban, and fractal scenes; example images in Fig. 1A). For each image, we computed the density of fixations across space from one observer (see Fig. 1B for 2 example densities) and used this fixation density as a predictor for fixation locations of another observer. Prediction accuracy was quantified by computing the AUC, separating fixation density values at fixation locations of the predicted observer from density values at fixation locations on other images viewed by the same observer ("control locations"; see Fig. 1C). The choice of control locations accounts for the influence of the center bias in fixation selection (Tatler et al. 2005; Tatler and Vincent 2009). AUC values around 0.5 indicate chance prediction performance, whereas values close to 1 indicate perfect separation of actual and control fixations. Thus, if the viewing behavior of one observer is similar to the viewing behavior of another, high AUC values can be expected. We computed AUC values for all pairs of observers, yielding 3 similarity matrices (Fig. 2A), one for each stimulus category, that show how well each observer predicted each other observer. We then applied hierarchical agglomerative clustering to these similarity data to test for groups of participants with similar viewing behavior and rearranged the similarity matrices accordingly (see Materials and Methods for more details). This data-driven approach resulted in humans and monkeys sorted into different top-level clusters for natural and urban stimuli (Fig. 2B). The cluster structure for fractals is similar, but 3 human observers are part of the top-level monkey cluster and are then separated at the next level. In all 3 categories, monkeys and humans predicted each other less compared with the same species predicting itself (unpaired *t*-tests comparing within-species vs. cross-species, all Ps < 0.05 with Bonferroni–Holm correction for 6 comparisons; average AUC values; humans = H; monkeys = M; naturals: H<->M = 0.539, H<->H = 0.579, M<->M = 0.615; urbans: H<->M = 0.589, H<->H = 0.699, M<->M = 0.626; fractals: H<->M = 0.612, H<->H = 0.658, M<->M = 0.622). Concentrating only on the monkey clusters, that is, the right branch of the top-level cluster revealed that monkeys were split into 2 groups in all stimulus categories. Interestingly, the same monkeys are assigned to the 2 groups, with the exception of the natural stimuli where one monkey switches clusters. The 2 top-level monkey clusters separate monkeys from the US laboratories (identical rearing conditions; cluster 1) from the macaques at the DPZ, who are, however, joined by 2 other US monkeys (cluster 2) on naturals and urbans and one other monkey for fractals. Differences between monkeys might therefore be in part explained by housing and rearing conditions at the different recording sites. Yet, comparably small numbers of monkeys in the respective clusters do not lend themselves to reliable statistical inference and make it difficult to distinguish natural variation from systematic influences. We therefore do not explore distinctions between monkey clusters any further in this manuscript. Overall, our findings suggest that viewing behavior of monkeys and humans is only comparable to a limited extent, and that viewing behavior across monkeys is not homogeneous.

The clustering analysis raised the question of which aspects of viewing behavior between humans and monkeys are comparable and which are not. To address this question, we investigated how well viewing behavior of both species can be predicted by factors that are known to drive fixation selection strategies in humans. Figure 1D shows a summary of all comparisons performed.

We started by quantifying the similarity of viewing behavior across observers within each species individually (Fig. 1D, blue line). In contrast to the cluster analysis, this "within-species consistency" was computed by predicting individual observers from the data of all other observers. Again using AUC, this allowed us to estimate similarities in viewing behavior within a group, not just behavior shared between pairs of observers (Wilming et al. 2011). At the same time, the within-species consistency is the best-known predictor of individual viewing behavior in humans (Onat et al. 2014). Intuitively, the consistency

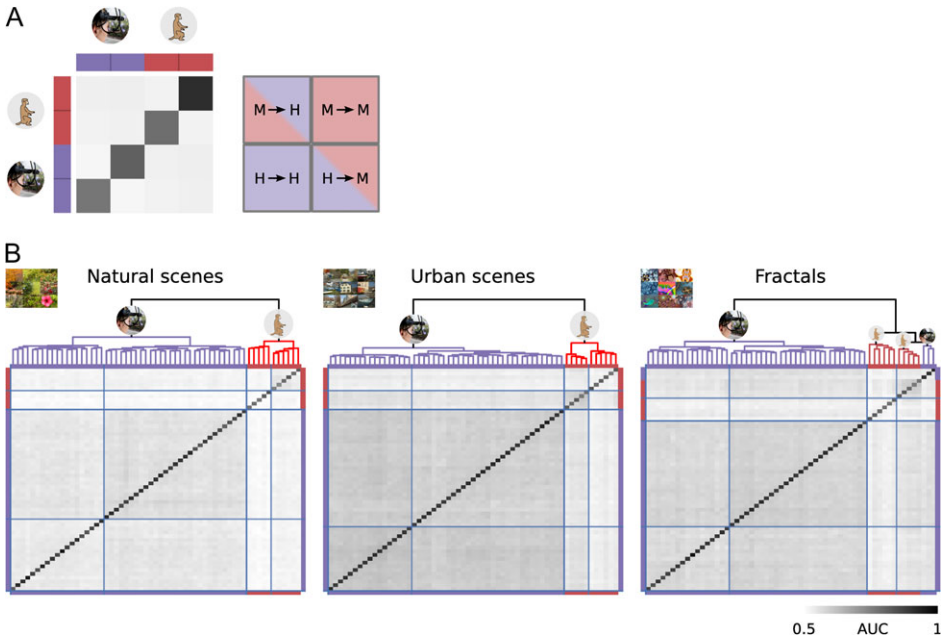

Figure 2. Similarity of viewing behavior between all observers for different categories of stimuli. (*A*) Schematic of a 4 × 4 similarity matrix to visualize who predicts who in the full matrix below, for example, H->M depicts areas where human observers predict monkey observers. (*B*) Full similarity matrix constructed from AUC values between pairs of observers (left = natural scenes, center = urban scenes, right = fractals). The intensity of individual points encode how well an observer predicts another. The species is encoded by different colors on the side of each matrix (purple = humans; red = monkeys). Rows and columns are sorted according to the results of hierarchical clustering. The dendrogram on the top shows the cluster structure, links are colored according to their species composition (purple only humans, red only monkeys, black mixed). Monkeys and humans are sorted into different clusters by the hierarchical clustering algorithm.

is high when all individuals in a group share the same viewing behavior and thus select the same fixation locations. Applied to monkeys and humans, respectively, this analysis revealed that the consistency between humans was higher than between monkeys ($<H>_{naturals}$ = 0.658, $<H>_{urbans}$ = 0.767, $<H>_{fractals}$ = 0.727, $<M>_{naturals}$ = 0.598, $<M>_{urbans}$ = 0.680, $<M>_{fractals}$ = 0.664; Fig. 3 top panels, blue symbols, and bars; Table 1 summarizes AUC values numerically; unpaired *t*-tests $P < 0.05$ with Bonferroni–Holm correction). To exclude the possibility that the estimate of the within-monkey consistency was lower simply because we had less observers available to estimate shared viewing behavior, we estimated the within-human consistency also with fewer observers (Fig. 3, lower right plot in each panel). The difference between within-human and within-monkey consistency persisted even when the number of predicting participants was kept identical in both species, that is, using only 10 human observers for the prediction (Fig. 3, lower right plot in each panel). In particular using groups of 11 humans where each observer is predicted by the remaining 10 yielded consistency estimates that were very close to those when 105 observer predicted 1 observer (mean difference in AUC: 0.002, 0.0001, 0.01). These results demonstrate that for equal sample sizes intra-species predictability is higher for humans than monkeys.

In a next step, we computed how well viewing behavior of one species predicts that of the other species, that is, cross-species consistency (Fig. 1*D*, red lines). For all observers and images, we computed how well the fixation density of all fixations of the respective other species on the stimulus separated actual and control fixation locations. We then averaged across stimuli and members of the predicted species. This analysis quantifies the amount of shared viewing behavior between species. We found that the cross-species prediction was below the within-species consistency for both species and all categories (Fig. 3 top panels, red symbols and bars; $<M->H>_{naturals}$ = 0.579, $<M->H>_{urbans}$ = 0.666, $<M->H>_{fractals}$ = 0.672, $<H->M>_{naturals}$ = 0.551, $<H->M>_{urbans}$ = 0.626, $<H->M>_{fractals}$ = 0.614; paired *t*-tests all *P*s < 0.05 with Bonferroni–Holm correction for 6 tests). In total, the cross-species consistency (M->H) reached 50%, 62%, and 76% of the within-human consistency and 51%, 70%, and 70% of the within-monkey consistency for the 3 stimulus classes, respectively (we subtract the chance level of 0.5 from each AUC value before computing ratios to avoid artificial inflation of these scores). This observation holds not only on average, but for all individual observers, that is, the cross-species is smaller than the within-species consistency for all 11 monkey and all 106 human observers.

Interestingly, we observed that monkeys predicted each other as well as they predicted humans, that is, the cross-species (M->H) score is comparable to the within-monkey score (unpaired *t*-tests all $P > 0.05$ with Bonferroni–Holm correction for 3 tests; bootstrapped CIs are overlapping; mean differences of AUC values are −0.019, −0.014, and 0.008 w.r.t to naturals, urbans, and fractals). This is a first indication that the consistent viewing behavior of monkeys is a subset of the consistent viewing behavior of humans. In combination, the cross-species scores and the extent to which monkeys predict humans suggests that both species are driven by similar factors. Furthermore, the larger consistency of human viewing behavior signals the presence of additional factors that are not present in monkeys. The presence of such factors would also explain why the cross-species score is not symmetric.

To better understand similarities and differences across species, we made use of a bottom-up salience model to test in how far low-level factors could explain the effects observed. We computed a set of 42 different visual features organized on 3 different spatial scales (see Materials and Methods). The resulting saliency model consists of a weighted sum of these features.

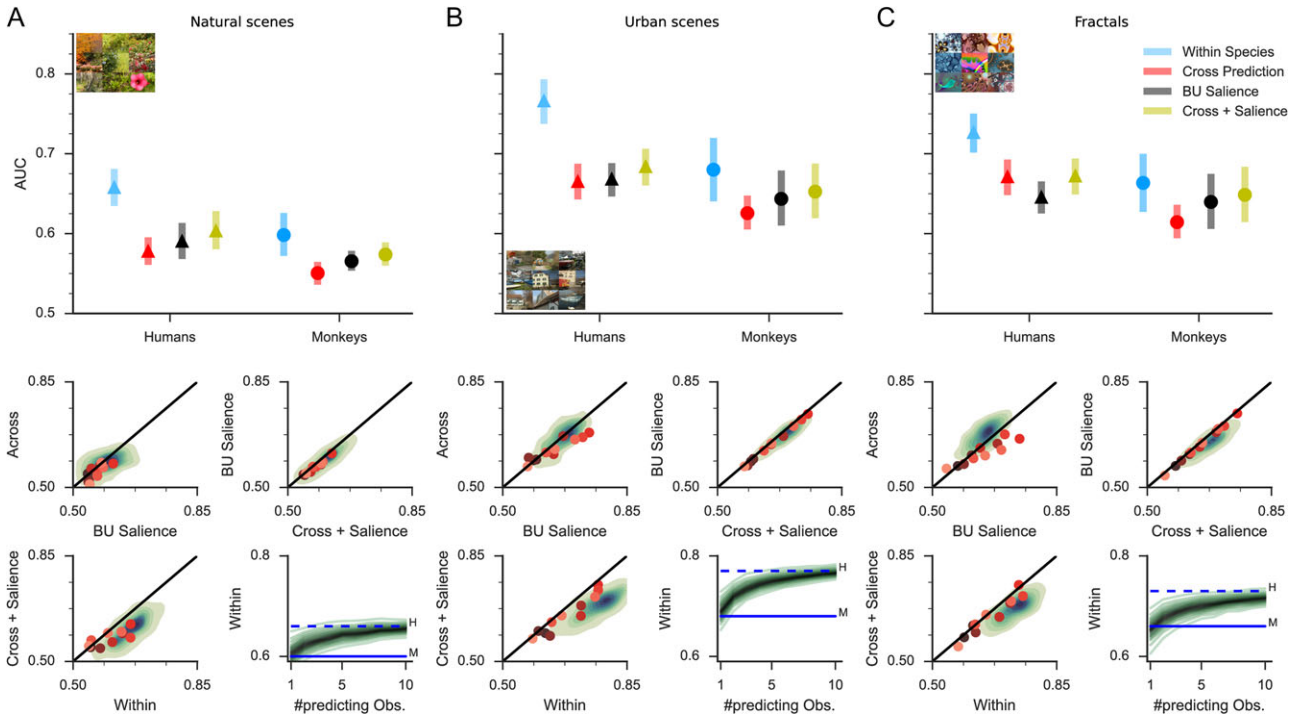

Figure 3. Predicting monkey and human fixation locations with different predictors. Columns show results for natural, urban, and fractal scenes (from left to right). The bar plots on the top show mean value and 95% CIs for individual predictors. CIs are computed by repeatedly sampling 11 observers to allow better comparison between human and monkey data. Three scatter plots at the bottom show individual comparisons. Red dots show individual monkeys (monkeys are indexed by hue). Green contours show a density estimate of the distribution of human AUC; each shade increases the contained amount of observers by 10%. Bottom right plots in each panel show within-human estimates when the number of predicting observers is subsampled. Dashed blue lines indicate our adjusted estimate for the within-human consistency had only 11 observers participated.

Table 1 Summary of AUC scores and percentages relative to the within-species consistency of different predictors

Category	Within		Cross		Salience		Salience + cross	
	Human	Monkey	Human	Monkey	Human	Monkey	Human	Monkey
Natural	0.658	0.598	0.579	0.551	0.591	0.565	0.60	0.574
			50%	51%	57%	67%	66%	75%
Urban	0.767	0.680	0.666	0.626	0.669	0.644	0.69	0.653
			62%	70%	63%	80%	69%	85%
Fractal	0.727	0.664	0.672	0.614	0.646	0.640	0.67	0.649
			76%	70%	64%	85%	76%	91%

We chose this type of model because it allowed us to estimate weights for each observer and category independently with a logistic regression that optimally separated actual and control fixations based on their predicted empirical saliency values (rightmost panel in Fig. 1B shows one example saliency map). Furthermore, our model comprised "pure" bottom-up features that do not exploit semantic aspects of the stimuli (e.g. faces, objects, or text). Whereas more complex models can indeed improve predictive performance (as indicated by the success of deep network models, e.g. Kümmerer et al. 2014), they mix lower and higher level features in their prediction. This runs counter to the goal of the current work, which was to separate lower and higher level factors. Our model is therefore designed to act as a probe into bottom-up influences on viewing behavior, not to maximize prediction accuracy. To ensure comparability to the within and cross-species similarity, we use a leave-one-observer-out cross-validation scheme to fit the saliency model. That is, predictions for a specific observer are independent from the data of that observer. The salience model therefore must predict behavior of an observer it has not encountered before, which implies that the salience models can only capitalize on viewing behavior that is consistent between observers. The within-species consistency therefore acts as an upper limit on the performance of the salience model (Wilming et al. 2011). Once fitted, the saliency model was applied in the same manner as the within- and cross-species predictions (Fig. 1D, black line). Correspondingly, the model prediction accuracy was computed based on AUC values. Figure 3 shows the results of the bottom-up saliency model. Black triangles and circles show the respective mean performance for humans and monkeys ($\langle H \rangle_{naturals} = 0.591$, $\langle H \rangle_{urbans} = 0.669$, $\langle H \rangle_{fractals} = 0.646$, $\langle M \rangle_{naturals} = 0.565$, $\langle M \rangle_{urbans} = 0.644$, $\langle M \rangle_{fractals} = 0.640$). CIs for human and monkey bottom-up salience AUCs are largely overlapping (Fig. 3, black bars, unpaired *t*-tests all $P > 0.05$ with Bonferroni–Holm correction). To further investigate the similarity of the bottom-up salience model in both species, we compared the predictive

power of individual features that enter the salience model by computing feature-specific AUC values. We found that the resulting feature-fixation AUCs were highly correlated across species on urban scenes and fractals and to some extent on natural scenes (Fig. 4). A linear regression between feature AUC values of humans and monkeys showed that the pattern of human feature-fixation AUC values explained 94% and 97% of the variance between the monkey feature-fixation AUC values for urbans and fractals and 27% on natural scenes. To estimate the distribution of regression slopes, we recalculated the linear regression on randomly sampled subsets of 11 human and monkey observers (sampling with replacement, see Materials and Methods for more details). The distributions of slopes, across bootstrapping samples, on urban and fractal scenes were approximately centered on 1 for urban and fractal scenes but 1 was not contained in the 95% interval for natural scenes (average slopes are: 0.52, 0.90, and 0.97 for naturals, urbans, and fractal scenes, respectively). These results show that feature-fixation AUCs were almost identical for urban and fractal scenes and thus the models were nearly identical for these stimulus categories. The finding of smaller feature AUC values on naturals is in line with the observation that, in general, all AUC values are lower on natural scenes compared with urban and fractal scenes. Yet, the pattern of results within this category is highly similar to the other 2 categories (see Fig. 3A). This suggests that the observed differences are not related to species differences but rather to an overall category difference. In summary, the results of the bottom-up saliency models and similarities in feature AUCs suggest that viewing behavior that is shared between species is to a large extent driven by bottom-up salience.

Interestingly, bottom-up salience reached 67%, 80%, and 85% of the within-monkey consistency AUC scores, suggesting that bottom-up salience explains a large part of the consistent viewing behavior between monkeys. In humans, however, we observed only 57%, 63%, and 64%, indicating that human viewing behavior is strongly guided by factors beyond low-level features (unpaired *t*-tests, all $P < 0.05$ with Bonferroni–Holm correction for 3 tests). These numbers imply that, across categories, 23% of the consistent viewing behavior in monkeys cannot be explained by bottom-up salience while the gap is 39% for humans. This demonstrates that although the bottom-up saliency models of both species are similar and explain viewing behavior to a comparable degree, the gap to the respective within-species consistency leaves a large fraction of human viewing behavior to be explained.

With this in mind, it is interesting to know whether there are other factors contributing to successful cross-species predictions, besides a presumably shared bottom-up selection mechanisms. Despite the fact that bottom-up salience AUC values were larger than the cross-species prediction for the majority of monkeys (8/11, 8/11, 10/11 w.r.t. naturals, urbans, and fractals), both factors might explain different aspects of viewing behavior. To explicitly investigate this possibility, we tested one additional linear model that combined bottom-up saliency and the cross-species prediction (Fig. 1*D*, yellow lines). We reasoned that performance of the combined model could only improve over the individual predictors, if both predictors explain, at least in part, different aspects of the consistent viewing behavior. The combined model used a logistic regression to obtain optimal weights for z-scored bottom-up salience values (with the weighting of the individual features optimized

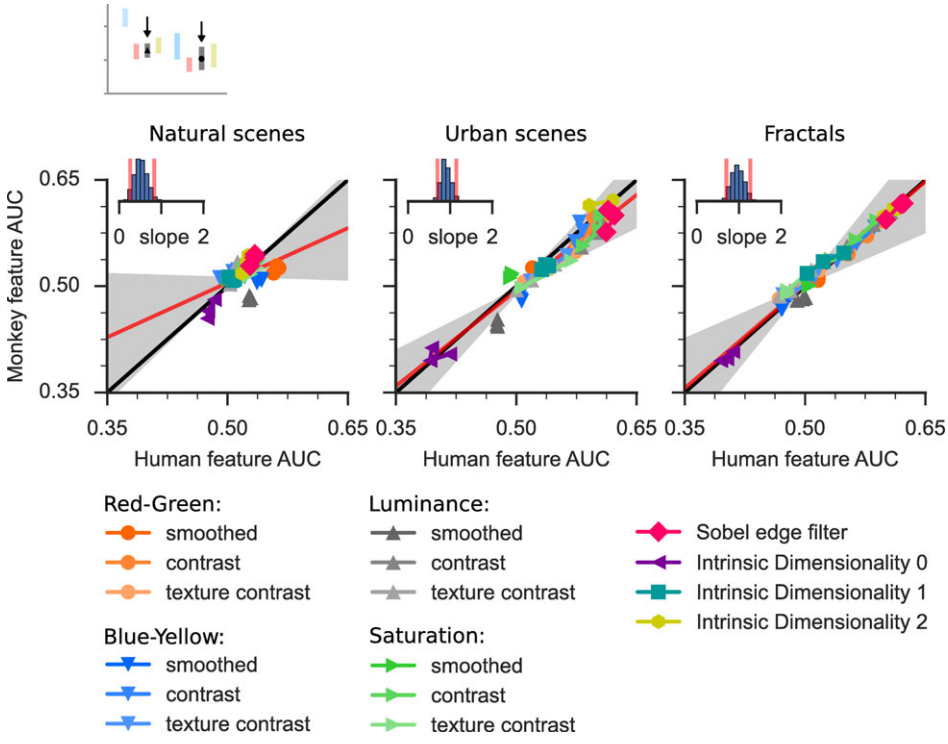

Figure 4. Comparison of bottom-up salience across species. Panels show feature-fixation AUC values for humans and monkeys (see Materials and Methods for feature definitions). Red lines show the best fit linear model that explains monkey feature-fixation AUCs based on human feature-fixation AUCs. The shaded blue area contains 95% of all bootstrapped regression lines created by repeatedly subsampling human and monkey observers with replacement. Small insets show the bootstrapped distribution of slopes of the linear regression. The area between the red bars contains 95% of bootstrapped slopes. Columns show different categories (naturals, urbans, and fractals).

as before, that is, by leave-one-observer-out cross-validation), and z-scored cross-species predictions. Figure 5 plots the bottom-up salience values against cross-species predictions for fixated locations and non-fixated locations. Model predictions can best be understood in terms of a classifier that compares the respective bottom-up salience and cross-species values of a test fixation against a given threshold (here zero for both models). Fixations with values exceeding the threshold are classified as "fixated", all others as "non-fixated" (thresholds shown in Fig. 5 as dashed black lines; red coloring indicates more actuals than control fixations, blue the reverse). As can be seen in Fig. 5, a model combining low-level saliency and cross-species predictions can only improve over the 2 individual models when a combination of both models, that is, a slanted decision function, classifies more points correctly (i.e. better separates red from blue points). We found that the combined model exhibited only small, but consistent, improvements over bottom-up salience and the cross-species prediction alone, for all monkeys on all categories (paired t-tests all $P < 0.05$ with Bonferroni–Holm correction for 6 tests; differences of combined-salience AUC values, naturals: 0.008, urbans: 0.009, fractals: 0.009). The facts that the decision function of the combined model is slanted (Fig. 5B, green lines), and that the combined model improves over its constituent models, together imply that bottom-up salience and the cross-species prediction both significantly contribute to the prediction. Yet, it has to be considered that the dynamic range, that is, the difference between the bottom-up saliency model and the intra-species predictability is rather small. As a consequence, the performance of the combined model was very close to the within-monkey estimate (naturals: 0.574 vs. 0.598; urbans: 0.653 vs. 0.68, fractals: 0.649 vs. 0.664; 75%, 85%, and 91%, respectively). For humans, the combined model showed marginal improvements over either bottom-up salience or the cross-species prediction (naturals: 0.01, paired t-tests against bottom-up salience $P < 0.05$; urbans: 0.02, paired t-tests against bottom-up salience $P < 0.05$; fractals: 0.001, paired t-tests against bottom-up salience $P < 0.05$; Bonferroni–Holm correction for 6 tests) but was still far away from the within-human estimate (naturals: 0.60 vs. 0.66, urbans: 0.69 vs. 0.77, fractals: 0.67 vs. 0.73; 66%, 69%, 76% relative to the within-human estimate). For both species, these results demonstrate that shared factors, beyond bottom-up selection, are very limited. Rather, shared viewing behavior is predominantly driven by low-level selection.

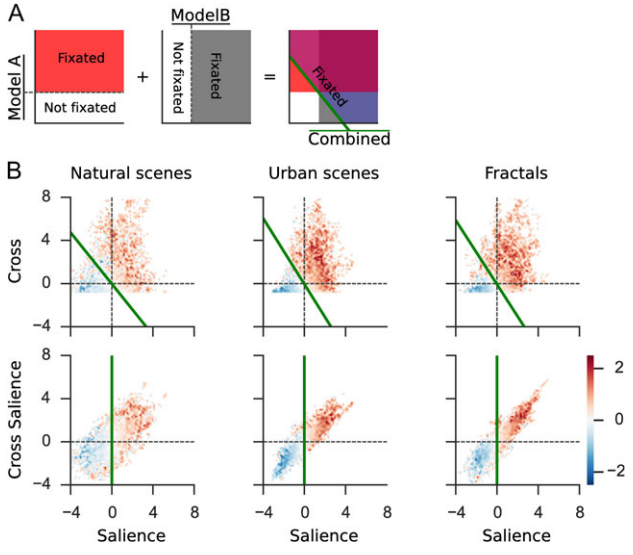

Figure 5. Combination of different models to predict monkey fixation locations. (*A*) Cartoon drawing to exemplify how 2 models form a combined model. Model A and B both project data into a space in which samples are classified as fixated or non-fixated based on their position w.r.t a threshold (dashed lines). The combined model weights and sums both models to potentially improve its prediction. The slope of the resulting decision function (green line) depicts the relative impact of the 2 combined models. (*B*) Performance of the combined model on different stimulus categories (naturals, urbans, fractals). Color encodes the log likelihood of observing fixations with a specific combination of predictor values in the data set (e.g. log(P(fix | cross prediction, within salience)/P(control | cross prediction, within salience)) in the top row and log(P(fix | cross-salience, within salience)/P(control | cross-salience, within salience)) in the bottom row), that is, redish colors imply more actual fixations and blueish colors non-fixated locations. Top row: Each panel shows the seperating decision function of a within-species bottom-up model and the cross-species prediction (vertical and horizontal dashed lines, respectively). Diagonal green lines show the decision function of the model that combines within-species saliency and the cross-species prediction. The bottom row shows decision functions for within-species salience and cross-species salience models. Here, the decision function (green line) for a model that combines within-species and cross-species salience falls onto the function for the within-species salience model, that is, the combined model utilizes exclusively the species-specific saliency model, neglecting the respective model derived from the other species.

Although very little improvement was observed with the combined model, with respect to the question whether or not monkeys are a good model system for human fixation behavior, it is nevertheless interesting to investigate why the combined cross-species + salience predicts monkey fixations better than bottom-up salience alone. In particular, we wanted to test whether the human-cross-species prediction contains explanatory power beyond its bottom-up salience component. As a reference, we compared the combined model (saliency + cross-species) with a model that combines 2 saliency models: a cross-species (trained on all human observers) and a within-species model (trained on leave-one-observer-out data from monkey observers), as described earlier. We found that the combined-salience model did not use the cross-species salience values for its prediction. This implies that the combined-salience model is identical to the one obtained by using within-monkey salience only (the decision function of the combined-salience model falls onto the decision function of the within-monkey salience model in Fig. 5; put differently, the weight assigned to the cross-species bottom-up saliency model by the logistic regression is close to zero). This suggests that the little improvement of the combination of the cross-species prediction and bottom-up salience over bottom-up salience alone is due to factors that are contained in human viewing behavior but not in our bottom-up salience model.

Taken together, bottom-up salience is able to account for most of the consistent viewing behavior between monkeys, while human viewing behavior makes only a small contribution when predicting monkey behavior. For humans, too, the addition of monkey data adds little to improve predictive performance beyond a pure bottom-up model. Moreover, while consistent macaque viewing behavior is well understood in terms of low-level salience, human viewing behavior is more complex. Thus, our data indicate that both species share similar low-level selection mechanisms, whereas similarities beyond low-level are very limited.

Discussion

The present data revealed that humans and monkeys form distinct clusters of viewing behavior, where predictions within each species are significantly better than predictions across

species. Within species, we found that patterns of human fixation locations are more consistent than those of monkeys, even when we account for the low number of monkey observers. This implies that humans and monkeys do not use identical selection strategies in a free-viewing task; a task that requires no training for monkeys or humans. Despite these overall differences, prediction accuracies across species are significantly above-chance level. This implies that monkeys and humans do share some degree of fixation selection strategies. To investigate in what aspects the selection strategies of these species overlap we compared the predictive power of a set of low-level stimulus features. This revealed remarkable similarities across species. First, the overall predictive power of low-level saliency models is comparable. Additionally, the predictive power of individual low-level features was strongly correlated between species, indicating that similar low-level selection mechanisms are at work in both species. Following this, we compared whether the fixation data of one species adds explanatory power beyond such low-level factors for the respective other species. Here we found only small improvements in predictive power, indicating that, at least in the current data set, low-level salience is the most important shared component in the guidance of eye movements across species. Taken together, our findings show that free viewing of pictures produces consistent behavior in monkeys that can be related to human behavior in a meaningful way. The large influence of bottom-up salience in both monkeys and humans, the large number of data points that can be acquired in a short period of time and the fact that no training is required, make free viewing in monkeys very well suited to the study of the neural basis of low-level stimulus-driven oculomotor control. Yet, our results reveal important limitations, as viewing commonalities were limited to low-level selection, excluding shared higher level selection mechanisms. They, therefore, have large implications for future electrophysiological studies in macaques and behavioral comparisons across species.

Although the current study is the largest to date, it is not the first to compare viewing behavior in macaques and humans. Consistent with our findings, other groups, too, reported above-chance predictions of fixation points by salience models in both humans and monkeys for videos clips (Berg et al. 2009) and gray scale images (Einhäuser et al. 2006). However, there is disagreement in how far the effects are qualitatively comparable in humans and monkeys. Einhäuser et al. (2006) argue that both species are equally driven by bottom-up salience, while Berg et al. (2009) report that bottom-up salience is more predictive for human eye-movement behavior. Using a considerably larger set of observers and colored stimuli from 3 different categories, we do not find consistent differences across species that would argue for a true species differences. It should be noted that the low-level AUC values reported previously (Einhäuser et al. 2006; Berg et al. 2009) indicate only a small influence of stimulus features (AUC $\leq$0.59) in monkeys during free viewing. This is in line with the current monkey data for natural scenes. However, looking at urbans and fractals, we observe larger AUC values for a bottom-up salience model (AUC = 0.65 and 0.64). Bottom-up salience could therefore be more prominent in guiding monkey eye movements than previously thought, dependent on the stimulus category used. Together with the large correlations in the feature AUCs across species, this data suggest comparable low-level selection mechanisms in macaques and humans. Moreover, while we only observe very small similarities beyond low-level stimulus features in the current free-viewing paradigm, it is possible that other experimental settings or categories reveal larger consistencies. For instance, similarities between monkeys and humans exist during the viewing of faces (Guo et al. 2003, 2009; Ghazanfar et al. 2006; Shepherd et al. 2010) and scenes containing simple social interactions (McFarland et al. 2013; Solyst and Buffalo 2014). It should be noted, however, that these studies did not estimate the contribution of low-level saliency. It therefore remains unclear in how far the shared viewing behavior observed was driven by shared low-level selection mechanisms.

The interpretation of our results depends on how similar recording conditions were between species. Naturally, there are a few differences between humans and monkeys that we were not able to control. These include the fact that monkeys needed to be head fixed during recordings, that monkeys had less exposure to the kinds of stimuli that were presented, and that monkeys received reward during calibration trials and, more generally, underwent extensive training regimes for other tasks. We will discuss these issues in turn in the next paragraphs.

First, all of the monkeys were head fixed during the recordings. The human participants, however, were only instructed not to move their head, but were otherwise not constrained. To rule out the possibility that a head restraint interferes with fixation selection, we conducted a control experiment in which participants viewed urban scenes and fractals, with and without a head restraint (custom molded bite-bar and chin rest). We analyzed 3 aspects of fixation behavior: fixation durations, selected fixation locations, and the saccadic main sequence. Fixation durations were almost identical between conditions (linear fits had an average slope of 1.01 with SD = 0.05, $r^2 > 0.99$, Supplementary Fig. 1). We compared fixated locations across conditions by computing within-condition and across-condition AUC prediction scores, in analogy to the within-species and cross-species AUC scores. The resulting AUC values were very similar (Supplementary Fig. 1B and C; $P > 0.21$ for all comparisons), again indicating no systematic differences across conditions. Finally, no significant differences were observed between the saccadic main sequence between species in both conditions (within subject $r^2 > 0.98$). In summary, fixing the head in a central position in front of the screen did not change the overall pattern of viewing behavior. We therefore conclude that the difference in head restraint between species is not relevant for the interpretation of our results.

Second, compared with humans, monkeys have limited prior exposure to the kind of stimuli shown in this study (urban and natural stimulus categories). Thus, differences in specific associations, memories, and emotions, triggered for humans but not for macaques, might explain why we find consistent viewing behavior beyond bottom-up salience in humans but not in monkeys. What speaks against this possibility is the fact that we observed a very similar pattern of results across all categories, including fractal stimuli, for which both species have little to no prior exposure. Accordingly, our results with fractal stimuli corroborate our conclusions even for previously unknown stimulus categories.

Third, to achieve successful calibration, monkeys needed either to detect a color change of a rectangle located at random screen locations or fixate a cross at random screen locations. More generally, monkeys were previously trained in a wide variety of tasks that associated specific actions with rewards (DPZ monkeys: classification of 2D and 3D random dot stimuli, grasping tasks, and fixation task; YNPRC and WNPRC: delayed match to sample, change detection, visual search, and covert attention task). Is it therefore conceivable that monkeys searched the free-viewing stimuli for reward? We believe that

this is unlikely for several reasons. First, monkey fixation durations on free-viewing images are much shorter than the minimum fixation requirements during reward trials (color change task: 500–1100 ms required, average fixation duration: 225 ± 153 ms; fixation task: 1250 ms required, average fixation duration: 257 ± 214 ms). This shows that the monkeys did not simply transfer the temporal properties of reward providing actions to the free-viewing trials. Second, if operant conditioning engendered specific consistent viewing strategies, we should have observed elevated within-monkey AUC scores. In contrast, within-monkey AUC scores did not substantially exceed the scores for bottom-up saliency. Finally, speaking specifically against an influence of interleaved calibration, the fixation epochs during calibration trials are in close analogy to drift correction procedures that preceded stimulus presentation in the human experiments. Forced fixation epochs were therefore present in both human and monkey experiments. Hence, we believe that neither training for other tasks nor calibration procedures undermine the interpretation of our results.

Our results suggest an effective, and comparable saliency model in both species. However, whether or not bottom-up salience has a causal influence on eye movements is a matter of ongoing debate (Li 2002; Einhäuser and König 2003; Mazer and Gallant 2003; Henderson et al. 2007; Einhäuser et al. 2008; Arcizet et al. 2011; Schütz et al. 2011; Betz et al. 2013). In many cases, salience models have been used successfully to predict eye-movement targets during free viewing of images (Itti and Baldi 2005; Kienzle et al. 2007; Zhang et al. 2008; Bruce and Tsotsos 2009; Hwang et al. 2009; Judd et al. 2009; Zhao and Koch 2011). Furthermore, recent reports provided evidence for the existence of a functional saliency map in the human brain (Bogler et al. 2011; Ossandón et al. 2012). In summary, there is good evidence for the existence of saliency-like selection mechanisms in the brain. However, even if saliency turns out to be a correlate, for example, of object detection (Einhäuser et al. 2008; Nuthmann and Henderson 2010), rather than a true selection mechanism, the current finding of above-chance cross-species predictions, and comparable low-level AUC values still indicate shared mechanisms in the guidance of overt visual attention.

The results that we have observed are on a purely behavioral level, and it therefore remains an open question as to how these map onto the neural level. On the one hand, it is known that large homologies exist between human and monkey early visual (Orban et al. 2004) and higher level ventral visual areas (Kriegeskorte et al. 2008; Kornblith et al. 2013; Yovel and Freiwald 2013; Cichy et al. 2014). It is therefore tempting to hypothesize that the observed similarities in viewing behavior are the result of similar neural processing of visual space. On the other hand, marked differences exist across species. The temporal lobe is much larger in humans (Rilling and Seligman 2002) and at the same time, fewer and less clear homologous brain structures exist in dorsal areas (Orban et al. 2004). For example, human LIP possesses more retinotopically organized areas than monkey LIP (Patel et al. 2010). And, on a larger scale, oculomotor control in the human brain appears to be more lateralized while the monkey brain shows more contralateral specificity (Kagan et al. 2010; Oleksiak et al. 2011). It is therefore possible that the human brain possesses different mechanisms to drive eye-movement behavior. In this context, our behavioral findings also resonate well with the absence of a ventral attention network in macaque monkeys but overlap of the dorsal attention network in both species during a paradigm that required detection of target images in a rapidly presented stream of images (Patel et al. 2015). Many areas in the dorsal attention network are retinotopically organized and the network likely contains a priority map for guiding eye movements (Bisley 2011), making it a plausible candidate for stimulus-driven control of eye movements. Thus, while the current set of results suggests similar low-level selection mechanisms, our analyses additionally suggest that homologies w.r.t. higher level selection of fixation locations are limited and cannot be assumed a priori for electrophysiological studies on macaques. Comparisons of the neural basis of higher level oculomotor selection mechanisms across species should therefore be validated by behavioral comparisons—ideally with tasks that require the same amount of training for both species. Finally, we would like to emphasize that our results do not preclude similarities between species during other cognitive tasks (e.g. Wisconsin Card Sorting Task (Nakahara et al. 2002) or covert attention tasks (Caspari et al. 2015)).

Supplementary Material

Supplementary material are available at *Cerebral Cortex* online.

Funding

The authors gratefully acknowledge the funding of EU Grants ERC-2010-AdG-269716 “Multisense” and H2020-FeT-PROaCT-2014 641321—“socSMCs”. This work was supported by National Institute of Mental Health Grants MH080007 and MH093807 (to E.A.B.) and the National Institutes of Health, ORIP-0D010425. C.X. and S.T. were supported by the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center 889 “Cellular mechanisms of sensory processing” and the Research Unit 1847 “The physiology of distributed computing underlying higher brain functions in non-human primates” and a DFG Postdoctoral Fellowship to T.C. Kietzmann.

Notes

We would like to thank Alexandra Vormberg for assistance with data collection for the head restraint experiment. *Conflict of Interest*: None declared.

References

Açik A, Sarwary A, Schultze-Kraft R, Onat S, König P. 2010. Developmental changes in natural viewing behavior: bottom-up and top-down differences between children, young adults and older adults. Front Psychol. 1:207.

Arcizet F, Mirpour K, Bisley JW. 2011. A pure salience response in posterior parietal cortex. Cereb Cortex. 21:2498–2506.

Berg DJ, Boehnke SE, Marino RA, Munoz DP, Itti L. 2009. Free viewing of dynamic stimuli by humans and monkeys. J Vis. 9:1–15.

Betz T, Kietzmann T, Wilming N, König P. 2010. Investigating task-dependent top-down effects on overt visual attention. J Vis. 10:1–14.

Betz T, Wilming N, Bogler C, Haynes J-D, König P. 2013. Dissociation between saliency signals and activity in early visual cortex. J Vis. 13:1–12.

Bisley JW. 2011. The neural basis of visual attention. J Physiol. 589:49–57.

Bogler C, Bode S, Haynes J-D. 2011. Decoding successive computational stages of saliency processing. Curr Biol. 1–5.

Bruce NDB, Tsotsos J. 2009. Saliency, attention, and visual search: an information theoretic approach. J Vis. 9:1–24.

Calapai A, Berger M, Niessing M, Heisig K, Brockhausen R, Treue S, Gail A. 2016. A cage-based training, cognitive testing and enrichment system optimized for rhesus macaques in neuroscience research. Behav Res Methods. doi:10.3758/s13428-016-0707-3.

Caspari XN, Janssens T, Mantini XD, Vandenberghe XR, Vanduffel XW. 2015. Covert shifts of spatial attention in the macaque monkey. J Neurosci. 35:7695–7714.

Castelhano M, Mack M, Henderson J. 2009. Viewing task influences eye movement control during active scene perception. J Vis. 9:1–15.

Cichy RM, Pantazis D, Oliva A. 2014. Resolving human object recognition in space and time. Nat Neurosci. 17:1–10.

Einhäuser W, König P. 2003. Does luminance-contrast contribute to a saliency map for overt visual attention? Eur J Neurosci. 17:1089–1097.

Einhäuser W, Kruse W, Hoffmann K-P, König P. 2006. Differences of monkey and human overt attention under natural conditions. Vision Res. 46:1194–1209.

Einhäuser W, Spain M, Perona P. 2008. Objects predict fixations better than early saliency. J Vis. 8:1–26.

Ghazanfar AA, Nielsen K, Logothetis NK. 2006. Eye movements of monkey observers viewing vocalizing conspecifics. Cognition. 101:515–529.

Guo K, Meints K, Hall C, Hall S, Mills D. 2009. Left gaze bias in humans, rhesus monkeys and domestic dogs. Anim Cogn. 12:409–418.

Guo K, Robertson RG, Mahmoodi S, Tadmor Y, Young MP. 2003. How do monkeys view faces? A study of eye movements. Exp brain Res. 150:363–374.

Henderson J, Brockmole J, Castelhano M, Mack M. 2007. Visual saliency does not account for eye movements during visual search in real-world scenes. In: Roger PG Van Gompel, Martin H. Fischer, Wayne S. Murray, Robin L. Hill, editors. Eye movements: A Window on Mind and Brain. Amsterdam: Elsevier Science. p. 537–562.

Hwang AD, Higgins EC, Pomplun M. 2009. A model of top-down attentional control during visual search in complex scenes. J Vis. 9:1–18.

Itti L, Baldi P. 2005. Bayesian surprise attracts human attention. Vision Res. 49:1295–1306.

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

Johnson L, Sullivan B, Hayhoe M, Ballard D. 2014. Predicting human visuomotor behaviour in a driving task. Phil Trans R Soc B. 369:20130044.

Judd T, Ehinger K, Durand F, Torralba A. 2009. Learning to predict where humans look. Comput Vis. 2106–2113.

Jutras MJ, Buffalo EA. 2010. Recognition memory signals in the macaque hippocampus. Proc Natl Acad Sci USA. 107:401–406.

Jutras MJ, Fries P, Buffalo EA. 2009. Gamma-band synchronization in the macaque hippocampus and memory formation. J Neurosci. 29:12521–12531.

Kagan I, Iyer A, Lindner A, Andersen RA. 2010. Space representation for eye movements is more contralateral in monkeys than in humans. Proc Natl Acad Sci USA. 107:7933–7938.

Kienzle W, Wichmann F, Scholkopf B. 2007. A nonparametric approach to bottom-up visual saliency. Adv Neural Inf Process Syst. 19:689–696.

Kietzmann T, Geuter S, König P. 2011. Overt visual attention as a causal factor of perceptual awareness. PLoS One. 6(7):e22614.

Kietzmann T, König P. 2015. Effects of contextual information and stimulus ambiguity on overt visual sampling behavior. Vision Res. 110:76–86.

Kollmorgen S, Nortmann N, Schröder S, König P. 2010. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. PLoS Comput Biol. 6: e1000791.

Kornblith S, Cheng X, Ohayon S, Tsao DY. 2013. A network for scene processing in the macaque temporal lobe. Neuron. 79: 766–781.

Kriegeskorte N, Mur M, Ruff D, Kiani R. 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 60:1126–1141.

Kümmerer M, Theis L, Bethge M. 2014. Deep gaze I: boosting saliency prediction with feature maps trained on imageNet. arXiv. 1411:1045.

Land MF, Hayhoe M. 2001. In what ways do eye movements contribute to everyday activities? Vision Res. 41:3559–3565.

Land MF, Tatler BW. 2001. Steering with the head: the visual strategy of a racing driver. Curr Biol. 11:1215–1220.

Li Z. 2002. A saliency map in primary visual cortex. Trends Cogn Sci. 6:9–16.

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

McFarland R, Roebuck H, Yan Y, Majolo B, Li W, Guo K. 2013. Social interactions through the eyes of macaques and humans. PLoS One. 8:e56437.

Nakahara K, Hayashi T, Konishi S, Miyashita Y. 2002. Functional MRI of macaque monkeys performing a cognitive set-shifting task. Science. 295:1532–1536.

Nuthmann A, Henderson J. 2010. Object-based attentional selection in scene viewing. J Vis. 10:1–19.

Oleksiak A, Postma A, van der Ham IJ, Klink PC, van Wezel RJ. 2011. A review of lateralization of spatial functioning in nonhuman primates. Brain Res Rev. 67:56–72.

Onat S, Açık A, Schumann F, König P. 2014. The contributions of image content and behavioral relevancy to overt attention. PLoS One. 9:e93254.

Orban G a, Van Essen D, Vanduffel W. 2004. Comparative mapping of higher visual areas in monkeys and humans. Trends Cogn Sci. 8:315–324.

Ossandón JP, Onat S, Cazzoli D, Nyffeler T, Müri R, König P. 2012. Unmasking the contribution of low-level features to the guidance of attention. Neuropsychologia. 50: 3478–3487.

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

Patel GH, Shulman GL, Baker JT, Akbudak E, Snyder AZ, Snyder LH, Corbetta M. 2010. Topographic organization of macaque area LIP. Proc Natl Acad Sci USA. 107:4728–4733.

Patel GH, Yang D, Jamerson EC, Snyder LH, Corbetta M, Ferrera VP. 2015. Functional evolution of new and expanded attention networks in humans. Proc Natl Acad Sci. 112: E5377–E5377.

Petersen SE, Posner MI. 2012. The attention system of the human brain: 20 years after. Annu Rev Neurosci. 35:73–89.

Rilling JK, Seligman RA. 2002. A quantitative morphometric comparative analysis of the primate temporal lobe. J Hum Evol. 42:505–533.

Schütz AC, Braun DI, Gegenfurtner KR. 2011. Eye movements and perception: a selective review. J Vis. 11:1–30.

Shepherd S V, Steckenfinger SA, Hasson U, Ghazanfar AA. 2010. Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. Curr Biol. 20: 649–656.

Smith TJ, Henderson JM. 2011. Does oculomotor inhibition of return influence fixation probability during scene search? Atten Percept Psychophys. 73:2384–2398.

Solyst JA, Buffalo EA. 2014. Social relevance drives viewing behavior independent of low-level salience in rhesus macaques. Front Neurosci. 8:1–13.

Sullivan B, Johnson L, Rothkopf C, Ballard D, Hayhoe M. 2012. The role of uncertainty and reward on eye movements in a virtual driving task. J Vis. 12:1–17.

Tatler B. 2007. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. J Vis. 7: 1–17.

Tatler BW, Baddeley RJ, Gilchrist ID. 2005. Visual correlates of fixation selection: effects of scale and time. Vision Res. 45: 643–659.

Tatler BW, Vincent BT. 2009. The prominence of behavioural biases in eye guidance. Vis cogn. 17:1029–1054.

Torralba A, Oliva A, Castelhano MS, Henderson JM. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev. 113:766–786.

Ward JH. 1963. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 58:236–244.

Wilming N, Betz T, Kietzmann TC, König P. 2011. Measures and limits of models of fixation selection. PLoS One. 6:e24038.

Wilming N, Harst S, Schmidt N, König P. 2013. Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. PLoS Comput Biol. 9:e1002871.

Yovel G, Freiwald WA. 2013. Face recognition systems in monkey and human: are they the same thing? F1000Prime Rep. 5:10.

Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW. 2008. SUN: a Bayesian framework for saliency using natural statistics. J Vis. 8:32.

Zhao Q, Koch C. 2011. Learning a saliency map using fixated locations in natural scenes. J Vis. 11:1–15.