Causal Saliency Effects During Natural Vision

Ran Carmi
Neuroscience Program
University of Southern California
Los Angeles, CA 90089
carmi@usc.edu

Laurent Itti
Neuroscience Program
University of Southern California
Los Angeles, CA 90089
itti@usc.edu

Abstract

Salient stimuli, such as color or motion contrasts, attract human attention, thus providing a fast heuristic for focusing limited neural resources on behaviorally relevant sensory inputs. Here we address the following questions: What types of saliency attract attention and how do they compare to each other during natural vision? We asked human participants to inspect scene-shuffled video clips, tracked their instantaneous eye-position, and quantified how well a battery of computational saliency models predicted overt attentional selections (saccades). Saliency effects were measured as a function of total viewing time, proximity to abrupt scene transitions (jump cuts), and inter-participant consistency. All saliency models predicted overall attentional selection well above chance, with dynamic models being equally predictive to each other, and up to 3.6 times more predictive than static models. The prediction accuracy of all dynamic models was twice higher than their average for saccades that were initiated immediately after jump cuts, and led to maximal inter-participant consistency. Static models showed mixed results in these circumstances, with some models having weaker prediction accuracy than their average. These results demonstrate that dynamic visual cues play a dominant causal role in attracting attention, while static visual cues correlate with attentional selection mostly due to top-down causes.

Keywords: Attention; Eye-movements; Natural Vision; Computational Modeling; Saliency

1. Introduction

Attentional selections are the result of complex interactions between bottom-up and top-down influences [Findlay and Walker 1999; Henderson 2003; Hernandez-Peon et al. 1956; James 1890]. Among bottom-up influences, dynamic stimuli are very effective in attracting human attention, as indicated by converging evidence from neurophysiological [Fecteau et al. 2004; Gottlieb et al. 1998], psychophysical [Folk et al. 1992; Jonides and Yantis 1988] and developmental [Atkinson and Braddick 2003; Finlay and Ivinskis 1984] studies. Nevertheless, the common computational approach for studying the impact of bottom-up influences on attentional selection (henceforth, saliency effects) is to analyze the visual correlates of fixation selections during inspection of still images [Itti and Koch 2000; Krieger et al. 2000; Mannan et al. 1997; Oliva et al. 2003; Parkhurst et al. 2002; Parkhurst and Niebur 2003; Peters et al. 2005; Reinagel and Zador 1999; Tatler et al. 2005; Torralba 2003]. Such studies provide valuable accounts of saliency effects during visual exploration of static scenes, but the scalability of their conclusions to more dynamic environments is an open question. Furthermore, the typical focus of these computational accounts on correlation rather than causation weakens their explanatory and predictive powers.

A general obstacle to characterizing bottom-up influences is that any observed effects (or lack thereof) may actually reflect top-down effects [Yantis and Egeth 1999]. Thus, the prevailing psychophysical approach to characterizing the types of stimuli that attract attention is to identify task-irrelevant bottom-up cues that increase reaction time when participants search for synthetic targets embedded in multi-element arrays [Abrams and Christ 2005; Folk et al. 1992; Franconeri et al. 2005; Hillstrom and Yantis 1994; Jonides and Yantis 1988; Theeuwes 1994; Yantis and Egeth 1999]. Such studies are instrumental for identifying strong bottom-up influences that capture attention involuntarily in the context of competing top-down influences. However, the focus on experimental conditions that discourage participants from paying attention to salient stimuli may underestimate saliency effects in real world environments, which are likely to involve stronger correlations between saliency and behavioral relevance. More generally, it is unclear whether psychophysical stimuli, which are typically highly simplified, lead to the same behavioral patterns as real world stimuli.

To reduce the potential interference from top-down influences without sacrificing real world relevance, we generated scene-shuffled clips that contain jump cuts every couple of seconds. Such jump cuts repeatedly break any expectations that observers may have formed based on the recent input history (top-down influences). Consequently, they temporarily bias observers to select targets based on the instantaneous input (bottom-up influences). While jump cuts may not be common in the natural world, they are nonetheless ubiquitous in motion pictures, even though people are often not aware of their occurrence [Anderson 1996; Hochberg 1986]. The use of jump cuts was pioneered by Jean-Luc Godard in his 1960 movie *Breathless*, and later popularized by MTV in the 1980s [Thompson and Bordwell 2003]. Contrary to earlier predictions [Gibson 1979/1986], a continuous perceptual experience is rarely disrupted by jump cuts. Moreover, humans seem to be particularly drawn to stimuli containing jump cuts. A possible explanation is that such rapidly changing stimuli lead to faster information uptake than continuous stimuli, which may become boring once all the pertinent information is extracted. Be that as it may, the important point in this context is that humans seem to behave naturally during visual exploration of MTV-style stimuli.

We measured saliency effects for different saccade populations defined by their likelihood of being bottom-up driven, such as saccades initiated shortly after jump cuts. The rationale for this focus is based on the trade-off between bottom-up and top-down influences [Henderson and Hollingworth 1999; Hernandez-Peon

ETRA 2006, San Diego, California, 27–29 March 2006.
© 2006 ACM 1-59593-305-0/06/0003 $5.00

et al. 1956; James 1890], which suggests that the magnitude of saliency effects should be highest when top-down effects are minimized. In contrast, visual correlates of attentional selection that reflect top-down causes are expected to show the opposite trend, namely: relatively low correlations when top-down effects are minimized. Indeed, our results show that the correlations between human attentional selection and certain visual cues, including intensity variance and orientation contrast, reflect top-down causes. Other visual cues, including intensity contrast, color contrast, and to a greater extent: flicker contrast, motion contrast, and integrated saliency, are shown to causally attract attention.

2. Methods

2.1 Participants

Paid participants (3 women and 5 men), 23- to 32-years old, provided written informed consent. All participants were healthy, had normal or corrected-to-normal vision, and were naïve as to the purpose of the experiment.

2.2 Stimuli

50 video clips (30 Hz, 640x480 pixels/frame, 4.5-30 seconds long, mean ± s.d.: 21.83 ± 8.41 s, no audio) from 12 heterogeneous sources, including indoor/outdoor daytime/nighttime scenes, video games, television programs, commercials, and sporting events. These continuous clips were cut every 1-3 s (2.09 ± 0.57 s) into 523 clip snippets (clippets), which were re-assembled into 50 scene-shuffled (MTV-style) clips (Fig. 1a and Video S3 in Supplemental Material). The range of clippet lengths was chosen based on previous results showing that inter-participants consistency in attentional selection diverges significantly within this time frame [Mannan et al. 1997]. We thus hypothesized that the relative impact of bottom-up influences may also change significantly within this time frame. The clippet lengths were randomized within the chosen range to minimize the ability of participants to anticipate the exact timing of jump cuts. Continuous and MTV-style clips were matched in length, and each MTV-style clip contained at most one clippet from a given continuous clip. This MTV-style manipulation was inspired by the cinematic practice of introducing jump cuts to compress time while preserving semantic continuity [Anderson 1996; Hochberg 1986; Thompson and Bordwell 2003]. The critical difference is that our MTV-style clips were deliberately designed to maximize semantic unrelatedness between adjacent scenes depicted in different clippets, and no attempt was made to hide the scene transitions.

2.3 Experimental design

Participants inspected MTV-style video clips while sitting with their chin supported in front of a 22" color monitor (60 Hz refresh rate) at a viewing distance of 80 cm (28° x 21° usable field-of-view). Their task was: "follow the main actors and actions, and expect to be asked general questions after the eye-tracking session is over". Participants were told that the questions will not pertain to small details, such as specific small objects, or the content of text messages, but would instead help the experimenters evaluate their general understanding of what they had watched. The purpose of the task was to let participants engage in natural visual exploration, while encouraging them to pay close attention to the display throughout the viewing session. The motivation for providing a task came from preliminary testing, which revealed that instructionless free viewing sometimes led to idiosyncratic patterns of eye movements over time as observers lost interest and disengaged from the display.

2.4 Data acquisition and processing

Instantaneous position of the right eye was recorded using an infrared-video-based eye tracker (ISCAN RK-464, 240 Hz, <1° spatial error), which tracks the pupil and corneal reflection. Calibration and saccade extraction procedures have been described elsewhere [Itti 2005]. A total of 10221 saccades were extracted from the raw eye-position data. 34 saccades (0.3% of the total number) either started or ended outside of the display bounds, and were thus excluded from the data analysis (see below), which was based on the remaining 10187 saccades.

2.5 Bottom-up attention-priority maps

Instantaneous 2D attention-priority maps (240Hz, 40x30 pixels/frame) based on 7 computational models: intensity variance[1], integrated saliency, and individual saliency components (contrasts in color, intensity, orientation, flicker, and motion) were generated using a Linux-based computer cluster (total run time: 792 processor hours).

The intensity variance map was computed based on the variance of pixel intensities in an image patch:

$$C_p = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(I(i,j) - \bar{I}_p \right)^2$$

Where p refers to an image patch, $m=16$, $n=16$ are the width and height, respectively, of the patch in pixels (corresponding to 0.7°x0.7° in our display), I is the intensity of a pixel, and $\bar{I}_p$ is the mean intensity of the patch. The motivation for using this particular model comes from studies that demonstrated its correlation with perceptual contrast in natural images [Bex and Makous 2002], particularly in the context of attentional selection [Parkhurst and Niebur 2003; Reinagel and Zador 1999]. The particular scale of attention-priority maps was chosen such that local measurements (0.7°x0.7°) corresponded to the largest effect size reported for visual correlates of attentional selection [Parkhurst and Niebur 2003].

Saliency maps were computed based on a series of nonlinear integrations of center-surround differences across several scales and feature channels. Maps for individual saliency components were generated by consecutive runs of the integrated saliency model, in which all feature channels but one were inactivated. The computations in this model have already been described extensively elsewhere [Itti 2005; Itti and Koch 2000]. They are motivated by neurophysiological [Bisley and Goldberg 2003; Frost and Nakayama 1983; Gottlieb et al. 1998; Sillito et al. 1995], psychophysical [Polat and Sagi 1994; Treisman and Gelade 1980], and computational [Koch and Ullman 1985] studies. An earlier version of the integrated saliency model was published as part of a larger framework for simulating attention shifts [Itti and Koch 2000], which also included winner-take-all and inhibition-of-return. These operations may be useful for an upstream saccade generation module that integrates bottom-up and top-down influences. As such, they are outside the scope of

[1] The square root of intensity variance is also known as RMS contrast.

the current investigation that relies on attention-priority maps as probes for the potential availability of bottom-up influences.

2.6 Bottom-up prediction of single saccades

For each human saccade, we generated a concurrent random saccade by sampling its target from a spatially uniform distribution of possible locations. Such random saccades ensure that both hit rate and target specificity are taken into account when evaluating the prediction accuracy of attention-priority maps. In the absence of a specificity criterion, models that generate uniform attention-priority maps will achieve optimal hit rates and will be deemed maximally predictive. With the random baseline, such useless models would be considered minimally predictive, because both human and random saccades will have exactly the same hit rate. An important advantage of calculating the random baseline per saccade is that it eliminates potential artifacts due to varying distributions of saliency values in different attention-priority maps (e.g., across scenes or over time).

The role of random saccades is to reflect the chance levels of landing on different saliency values. Saccadic tendencies, such as biases for making short saccades [Melcher and Kowler 2001], were not included in either the model or the baseline. The rationale for this decision is that saccadic tendencies may reflect bottom-up influences, such as due to centrally-biased distribution of saliency values [Parkhurst and Niebur 2003; Reinagel and Zador 1999; Tatler et al. 2005]. In this case, the underlying cause for saccadic tendencies should already be included implicitly in bottom-up models, and including them explicitly in the baseline as well would lead to underestimating the actual magnitude of bottom-up influences. Alternatively, saccadic tendencies may reflect motor constraints that are independent of the actual stimulus content or related internal representations. In this case, saliency comparisons between human and random saccades would be contaminated by motor constraints that may lead to misestimating the actual magnitude of saliency effects. Unfortunately, the extent to which saccadic tendencies are influenced by saliency distribution versus motor constraints during natural vision is unknown, so introducing them explicitly would involve unwarranted assumptions about their origin. In any case, the analyses presented here do not depend on estimating the actual magnitude of bottom-up versus top-down influences during natural vision, which is a fascinating research question in its own right. Rather, we only compare the prediction accuracy of different bottom-up models in identical conditions, or of the same model across different conditions. For such comparisons, saccadic tendencies, regardless of their origin, are not expected to bias the results in any consistent way.

Normalized prediction for all saccades was calculated by sampling the attention-priority map at the saccade target, and dividing that local value by the global maximal value in the instantaneous attention-priority map. Local sampling was done by calculating the maximal local value in an aperture around each saccade target, thus compensating for potential inaccuracies in human saccade targeting and the eye-tracking apparatus. The fixed aperture size (r=3.15°) minimizes false negatives and false positives based on our subjective evaluation of where several randomly chosen human saccades were actually targeted. We did not try to optimize the aperture size (e.g., based on cortical magnification and saccade length), under the assumption that any errors in the human saccade sampling would be offset by corresponding errors in the random saccade sampling. To establish causal rather than correlational effects, measurements were taken at the end of the fixation period prior to saccade initiation, as defined by a standard saccade extraction procedure [Itti 2005]. The timing of such measurements does not explicitly take into account known sensory-motor delays in saccade execution [Caspi et al. 2004], because such delays are already included in the internal dynamics of the saliency model [Itti and Koch 2000]. The rationale for choosing this particular timing was two-fold: first, it is independent of fixation duration; second, attentional selection during natural vision is likely to be influenced most strongly by visual information accrued during the preceding fixation [Caspi et al. 2004; Najemnik and Geisler 2005; Parkhurst et al. 2002]. Optimizing the timing of saliency sampling for individual saccades is not crucial here, because the focus of the current investigation is on characterizing differences in saliency effects between groups of saccades.

2.7 Ideal prediction of attentional selection

Theoretically, the largest possible difference between model responses at human vs. random saccade targets would occur if human and random saccades always land on the maximal and minimal model response, respectively. However, even if assuming an ideal model that always generates a single response at saccade target, and 0 everywhere else (Fig. 2a), a certain fraction of random saccades would land on the maximal model response by chance. The probability of chance hits is given by $p = N_t / N_m = 0.0408$, where $N_t = 49$ is the number of pixels in an aperture around the saccade target (approximated by 9 adjacent rows: 1,5,7,7,9,7,7,5,1 pixels), and $N_m = W_m \times H_m = 1200$ is the number of pixels in the attention-priority map (40x30).

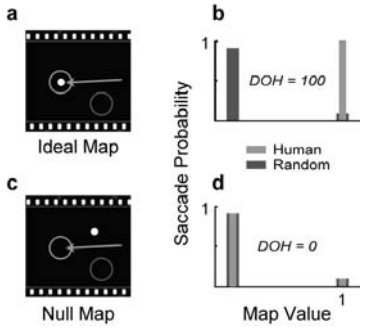

Figure 2. Hypothetical predictions of attentional selection. (*a*) An ideal attention-priority map prior to the initiation of a human saccade. It contains a positive value at the saccade target, and zero elsewhere. Eye-position prior to saccade initiation (filled circle), saccade trajectory (arrow), and saccade target (ring) are depicted in light gray. Dark gray ring depicts a random saccade target. (*b*) Saccade probability as a function of map values at saccade targets for the ideal scenario. The ideal scenario leads to the maximal rightward shift of the human histogram relative to the random histogram. (*c*) A null attention-priority map prior to saccade initiation. Any map that contains positive values at random locations would qualify as a null map, but in this case, only one random location is set to a non-zero value to facilitate direct comparisons to the ideal scenario. (*d*) Same as *b*, but for the null scenario. Human and random saccades are just as likely to land on positive values, leading to no rightward shift of the human histogram relative to the random histogram.

In the ideal scenario, the human histogram (saccade probability as a function of model response) will only contain saccades in the highest bin (90-100% of the max response), while the random histogram will have 1-*p* saccades in the lowest bin (0-10% of the max response), and *p* saccades in the highest bin (Fig. 2b). In comparison, the null scenario occurs when a model is unpredictive of attentional selection (Fig. 2c), in which case human and random saccades would be just as likely to target high priority targets, leading to no rightward shift between the human and random histograms (Fig. 2d).

2.8 DOH metric

The difference of histograms (DOH) metric outputs a scalar that quantifies the human tendency to initiate saccades towards high priority targets by measuring the rightward shift of the human saccade histogram relative to the random saccade histogram:

$$DOH = (1/DOH_I) \times \sum_{i=1}^{n} W_i \times (H_i - R_i)$$

Where H_i and R_i are the fractions of human and random saccades, respectively, which fall in bin i with boundaries $(i-1)/n$, i/n , where n=10 is the number of bins, and $W_i = (i-0.5)/n$ is the mid-value of bin i.

The weighting vector ensures that deviations from the baseline for high priority candidates, as defined by the attention-priority map, receive higher weights than corresponding deviations for low priority candidates. This is important because high priority candidates are more likely to attract attention than low priority candidates, and thus are more informative for measuring prediction accuracy.

DOH values are expressed as percentages of DOH_I , which reflects the ideal rightward shift of the human saccade histogram relative to the random saccade histogram (Fig. 2b):

$$DOH_I = (W_n - W_1) \times (1-p) = 0.8633$$

Hence, the expected range of DOH values is from 0 (chance) to 100 (ideal). Models that are worse predictors than chance would lead to negative DOH values.

The DOH metric has several advantages compared to previously suggested metrics [Itti 2005; Krieger et al. 2000; Mannan et al. 1997; Oliva et al. 2003; Parkhurst et al. 2002; Parkhurst and Niebur 2003; Reinagel and Zador 1999; Tatler et al. 2005; Torralba 2003], namely: linearity, meaningful upper bound, intuitiveness, priority weighting, directionality, and sensitivity to high-order statistics. The most advanced alternatives to DOH are KL-divergence [Itti 2005] and ROC analysis [Tatler et al. 2005]. The main advantage of the KL-divergence and ROC metrics relative to the DOH metric is their grounding in information theory and signal detection theory, respectively. However, both of these metrics are inferior to DOH in the specific context of measuring saliency effects. For example: both KL-divergence and DOH estimate the overall dissimilarity between different probability density functions, but KL-divergence suffers from the following relative disadvantages: non-linearity (i.e., metric values cannot be compared as interval variables), infinite upper-bound, no priority weighting, and bi-directionality (i.e., no distinction between instances in which models are more predictive versus less predictive than chance). In contrast, the ROC metric [Tatler et al. 2005] estimates the overall discriminability between different probability density functions. As such, it does not include any priority weighting. Furthermore, it introduces unwarranted assumptions about linear discriminability, which are not required when using dissimilarity-based metrics, such as KL-divergence or DOH.

It is important to realize that the DOH values reported here provide a conservative estimate for the overall impact of bottom-up versus top-down influences on attentional selection. This is because inter-participant consistency in attentional selection is imperfect, indicating that even the ideal attention-priority map should sometimes contain more than one potential candidate. Consequently, the probability of random saccades landing on valid attention candidates would be higher than reported here, leading to a lower DOH upper bound. Our conclusions are immune to this fact because they only depend on differences in prediction accuracy. The normalizing factor is interesting as a first step towards quantifying the relative contribution of bottom-up versus top-down influences, which is outside the scope of the current investigation.

3. Results

3.1 Overall saliency effects

We compared the accuracy of different bottom-up models in predicting attentional selection, which is strongly coupled with saccade target selection during natural vision [Findlay 2004; Kustov and Robinson 1996; Sheinberg and Logothetis 2001; Sperling and Weichselgartner 1995]. Table 1 shows the overall accuracy of different bottom-up models in predicting attentional selection.

Table 1. Overall accuracy of different bottom-up models in predicting attentional selection in the MTV-style experiment. Models are rank ordered by DOH from low to high. Significance values for pairs of adjacent DOH values are based on 2-tail t-tests

	DOH Mean	DOH SE*	t[10185], p value
Intensity Variance	12.3602	0.22155	-
Orientation Contrast	13.3311	0.33725	3.10, <0.005**
Intensity Contrast	13.4708	0.35572	0.29, >0.2
Color Contrast	14.6101	0.38036	2.26, <0.05*
Intensity Transient	20.3957	0.38582	10.75, <0.0001**
Motion Contrast	20.6382	0.37502	0.44, >0.2
Integrated Saliency	21.3771	0.34837	1.39, <0.2

* based on 1000 bootstrap subsamples [Efron and Tibshirani 1993]

DOH values clearly dissociate into 2 main groups of static versus dynamic models. To conserve space, we only show detailed analyses for 2 representative models from each group: intensity variance, color contrast, motion contrast, and integrated saliency.

Fig. 1b (see color plate) shows examples of human saccades and model responses. Fig. 3 shows the overall human and random saccade histograms (saccade probability as a function of model response). The random saccade histograms reflect the probability density function of model responses, while the human saccade

histograms demonstrate how much human selection of attention targets is biased towards locations with high model responses. As Figs. 1b and 3 demonstrate, different models generate diverse attention-priority maps, in terms of location and density of candidate targets. For example: the intensity variance model generates the densest maps, with only 2% of random saccades landing on minimal priority candidates (0-10% of the max). In contrast, the motion contrast model generates the sparsest maps, with 50% of random saccades landing on the lowest possible model response bin.

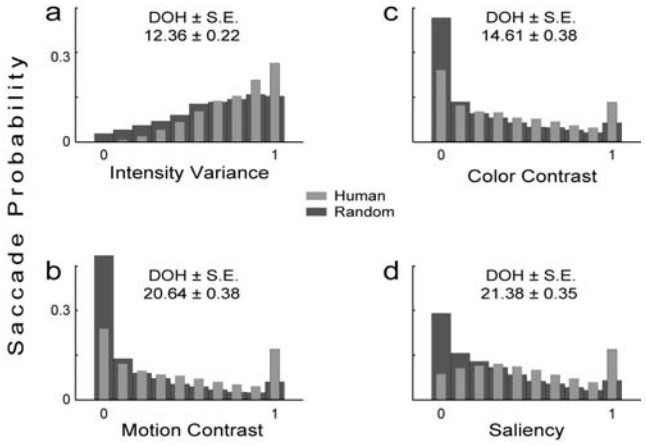

Figure 3: Overall bottom-up predictions of attentional selection. Different panels show the random and human saccade probabilities for different models as a function of map value at saccade target prior to saccade initiation. Dark gray and light gray vertical bars represent the random and human saccade histograms, respectively. Numbers above histograms show the DOH values.

In the text, we particularly focus on differences in prediction accuracy between intensity variance and integrated saliency. Intensity variance was chosen as a representative static model because it has been shown to correlate with detectability and attentional selection in natural images [Bex and Makous 2002; Parkhurst and Niebur 2003; Reinagel and Zador 1999]. Integrated saliency was chosen as the representative dynamic model because of its relatively broad applicability to both static and dynamic stimuli (the other dynamic models - intensity transient and motion contrast - require a minimum of 2 consecutive frames to produce an output). This choice did not matter here, because none of the circumstances we examined led to any statistically significant differences in prediction accuracy between different dynamic models. Fig. 3 and table 1 demonstrate that all the tested bottom-up models were significantly better than chance (DOH=0) in predicting attentional selection ($z>1.96$, $p<0.01$), with integrated saliency being 1.7 times more predictive than intensity variance ($t[10185]=21.8406$, $p<<0.01$).

3.2 Saliency effects as function of viewing time and inter-participant consistency

While establishing the overall superiority of dynamic over static features in attracting attention, Fig. 3 may in fact underestimate this superiority, because not all saccades are equally informative indicators of saliency effects. Specifically, top-down guided saccades are generally uninformative for estimating the relative impact of different types of saliency on attentional selection, while bottom-up driven saccades are particularly informative for this purpose. Unfortunately, tools to unambiguously label particular saccades performed during visual exploration of real world scenes as "top-down guided" or "bottom-up driven" are currently unavailable. Nevertheless, the two following heuristics may be used to approximate such labeling: First, bottom-up influences are faster acting than top-down influences [Henderson 2003; Wolfe et al. 2000]. Hence, saccades that occur shortly after exposure to novel scenes may be more bottom-up driven than later saccades. The evidence for this hypothesis is mixed: one study [Parkhurst et al. 2002] found relatively stronger saliency effects early after stimulus onset than later on, but another study found no interaction between saliency effects and viewing time [Tatler et al. 2005]. Second, top-down influences depend critically on prior knowledge and specific expectations that may be quite different for different participants, for example: Native English speakers tend to read from left to right, while native Hebrew speakers tend to read from right to left. In contrast, bottom-up influences depend more exclusively on the instantaneous stimulus content, which is physically identical for different participants. Thus, saccades that lead to increased inter-participant consistency in attentional selection may have been driven more strongly by bottom-up rather than top-down influences [Mannan et al. 1997]. Alternatively, changes in inter-participant consistency may simply reflect cross-participant divergence in top-down influences [Tatler et al. 2005], in which case no interaction is expected between inter-participant consistency in attentional selection and saliency effects.

To test for potential interactions between viewing time and saliency effects, we examined the accuracy of different bottom-up models in predicting attentional selection as a function of saccade index between adjacent jump cuts (Fig. 4a, see color plate). We also performed a similar analysis by actual time (based on consecutive 250 ms bins), which led to practically identical results. To facilitate direct comparisons to previous studies that examined the same issue [Parkhurst et al. 2002; Tatler et al. 2005], we only show the saccade index plot. As explained below, there are at least two methodological advantages to focusing the analysis on proximity to jump cuts rather than clip onsets, which may also cause surges in saliency effects [Henderson 2003; Parkhurst et al. 2002] (but see [Tatler et al. 2005]). One such advantage is the proper elimination of central bias artifacts, which may arise due to a combination of factors [Parkhurst et al. 2002; Parkhurst and Niebur 2003; Reinagel and Zador 1999; Tatler et al. 2005] [2]: First, the distribution of saliency values and objects of interest in photographs is often spatially biased towards the center of the display. Second, viewing sessions traditionally begin with a central fixation cross. Third, humans tend to make short saccades.

[2] Center bias and previous approaches for dealing with it have led to misestimating saliency effects. Overestimation may occur when saliency effects are measured at the beginning of a viewing session that was preceded by a central fixation cross [Itti 2005; Parkhurst et al. 2002]. Underestimation may occur either by simply ignoring the earliest and strongest saliency effects [Reinagel and Zador 1999], or by introducing fixation-dependent biases into the baseline [Parkhurst and Niebur 2003; Tatler et al. 2005]. The problem with the latter approach is the unwarranted assumption that the observed tendency to fixate towards the center of the display is mainly driven by motor biases rather than the actual distribution of saliency, which is indeed centrally biased in photography-based stimuli [Parkhurst et al. 2002; Reinagel and Zador 1999; Tatler et al. 2005].

Acting together, these factors may introduce an artifactual peak in saliency effects immediately after clip onsets, simply because participants were artificially induced to fixate more salient stimuli at the beginning of viewing sessions. We avoided this potential artifact by measuring saliency effects after jump cuts, which are not preceded by an artificial central fixation. Another advantage of analyzing saliency effects following jump cuts rather than clip onsets is signal to noise ratio (SNR): Each participant is exposed to approximately 10 times more jump cuts than clip onsets. Consequently, participants perform approximately 10 times more saccades following jump cuts compared to clip onsets, leading to relatively higher SNR for measuring saliency effects. Fig. 4a demonstrates that integrated saliency was 2.7 times better than intensity variance in predicting attentional selection ($t[10185]=18.1212$, $p<<0.01$), when the analysis was based on the first saccade initiated after jump cuts. It also indicates that the impact of motion contrast and integrated saliency peaks immediately after jump cuts, followed by slow decreases. Color contrast displays a similar pattern, but only for up to 5 saccades, while intensity variance displays a bell-shaped trend. We also performed the same analysis for clip onsets, which mirrored the initial trends of the jump cuts analysis (from the 1 s and saccade 4 onwards, saliency effects appeared constant in the clip onsets analysis, probably due to artifactual masking introduced by the asynchrony of jump cuts across clips). In summary, we observed decreased saliency effects with viewing time for all tested models, except the intensity variance model.

To test for potential interactions between inter-participant consistency and saliency effects, we examined the accuracy of different bottom-up models in predicting attentional selection as a function of inter-participant consistency. Fig. 4b demonstrates that integrated saliency was 2.5 times better than intensity variance in predicting attentional selection ($t[10185]=14.0763$, $p<<0.01$), when the analysis was based on saccades that brought the eye-position of a given participant closest to the instantaneous eye-position of other participants (area of bounding rectangle: 0°-4.8°). Fig. 4b further demonstrates a positive relationship between inter-participant consistency and saliency effects for all tested models, except the intensity variance model. Similarly, Fig. 4c shows the accuracy of different bottom-up models in predicting attentional selection as a function of inter-participant consistency, but only for the fastest 1st saccades (initiated within 250 ms after jump cuts). Finally, table 2 shows values of the first data point in Fig. 4c.

Table 2. Same as table 1, but based on a subset of saccades that were initiated within 250 ms after jump cuts and led to the highest inter-participant consistency

	DOH Mean	DOH SE*	t[10185], p value
Orientation Contrast	9.4893	2.6632	-
Intensity Variance	11.2693	1.584	0.47, >0.2
Intensity Contrast	20.5364	3.0515	4.14, <0.0001**
Color Contrast	23.975	2.7901	0.80, >0.2
Integrated Saliency	40.0302	2.3546	4.07, <0.0001**
Motion Contrast	41.1012	2.976	0.32, >0.2
Intensity Transient	43.5548	2.933	0.58, >0.2

* based on 1000 bootstrap subsamples [Efron and Tibshirani 1993]

It demonstrates that integrated saliency was 3.6 times better than intensity variance in predicting attentional selection ($t[10185]=10.1349$, $p<<0.01$), when the analysis was based most exclusively on bottom-up driven saccades.

4. Discussion

For the first time, our study establishes causal links between saliency and attentional selection during natural vision. This predictive power is attributable to the MTV-style manipulation, and the particular timing in which saliency effects were measured (i.e., prior to saccade initiation). The observed superiority of dynamic over static saliency in attracting human attention likely reflects an evolutionary adaptation to real world environments: important events in everyday life, such as the approach of predators or mates, may be detected more rapidly and selectively based on dynamic rather than static features. Moreover, biological camouflage strategies typically involve seamless blending into the background in terms of static features, such as shape and color [Curio 1976][3]. In such circumstances, attentional selection mechanisms based on static features are rendered useless, and organisms must rely on dynamic features for rapid detection of behaviorally relevant information, such as the location of predators or prey. Among bottom-up models based on static features, we found superior prediction accuracy for color versus intensity. This result may reflect an evolutionary adaptation for detecting color contrasts, such as when searching for colorful fruits embedded in foliage [Regan et al. 2001]. The following subsections address the methodological innovations that made this study possible, and discuss further implications of the results in light of previous studies, as well as promising future directions:

4.1 Stimuli

The stimulus set used here consists of 50 video clips from 12 different sources [Itti 2005], and is substantially larger and richer compared to the collections of still images [Itti and Koch 2000; Krieger et al. 2000; Mannan et al. 1997; Oliva et al. 2003; Parkhurst et al. 2002; Parkhurst and Niebur 2003; Peters et al. 2005; Reinagel and Zador 1999; Tatler et al. 2005; Torralba 2003] and synthetic search arrays [Abrams and Christ 2005; Folk et al. 1992; Franconeri et al. 2005; Hillstrom and Yantis 1994; Jonides and Yantis 1988; Theeuwes 1994; Yantis and Egeth 1999] used in previous studies.

It should be noted that visual exploration of video clips does not capture the full complexity of sensory stimulation experienced in real world environments, which often involve wider fields of view, multi-sensory stimulation, and egomotion. Unfortunately, the computational tools at our disposal are not powerful enough to handle unconstrained real world stimuli. We think that the use of motion pictures as stimuli strikes a reasonable balance between real world relevance and computational power.

4.2 Saliency modeling

Neural grounding, spatial interactions between local detectors, and detection of dynamic saliency are critical elements that distinguish the integrated saliency model and its components from

[3] Examples of dynamic camouflage, as employed by dragonflies during territorial aerial manoeuvres [Mizutani et al. 2003], can also be found in nature, but they are rare compared to static camouflage.

the available alternatives [Krieger et al. 2000; Mannan et al. 1997; Oliva et al. 2003; Parkhurst and Niebur 2003; Reinagel and Zador 1999; Tatler et al. 2005; Torralba 2003]. Extracting dynamic features from natural time varying stimuli requires a computational leap compared to extracting static features from still images. Our results demonstrate that this methodological advance led to dramatically improved model accuracy in predicting human attentional selection.

The most predictive bottom-up model used here (integrated saliency) still contains several limitations, and here we focus on two of them: First, it fails to account for the variable spatial resolution of the primate visual system [Connolly and Van Essen 1984; Curcio et al. 1987], or for potential differences in saliency processing between the fovea and periphery. Such sensory asymmetries and processing asymmetries may act in opposite directions, assuming that peripheral saliency is more informative than foveal saliency for making new attentional selections. If so, introducing one without the other [Parkhurst et al. 2002] would lead to less realistic modeling of saliency effects than introducing neither. Future accounts of processing asymmetries could pave the way towards realistic integration of both sensory and processing asymmetries into saliency models. Another limitation of the integrated saliency model is that it includes no excitatory spatial interactions, which can lead to perceptual grouping and strongly attract attention [Driver et al. 1992]. Nevertheless, the integrated saliency model does include inhibitory spatial interactions, and is thus not strictly local, contrary to most other saliency models.

An important step towards integrating bottom-up and top-down influences would be the addition of visual short-term memory to existing saliency models. The persistence of accrued sensory information during natural vision, as well as the nature of interactions between old and new visual inputs, are important open questions in this context. Additional task-independent top-down effects may be introduced by weighted cue combination based on the relative impact of individual cues on attentional selection, as revealed here. Such a scheme would simulate long-term learning of average cue reliability [Jacobs 2002].

4.3 DOH metric

In addition to comparing the performance of competing models in different conditions, as we did here, future studies could utilize the DOH metric for measuring model performance in the context of different activities or people, such as novices versus experts [Land and McLeod 2000], or people with autism versus control subjects [Klin et al. 2002]. The DOH metric can be used to measure the performance of any model that generates attention-priority maps, regardless of its underlying computations (bottom-up, top-down, or both).

4.4 Bottom-up saccade labeling

Here we promoted the idea that the relative impact of bottom-up influences on attentional selection should be measured when top-down influences are as little involved as possible (without losing realism). This approach is diametrically opposed to the psychophysical practice of characterizing bottom-up cues based on their ability to involuntarily capture attention in the context of a competing top-down task [Abrams and Christ 2005; Folk et al. 1992; Franconeri et al. 2005; Hillstrom and Yantis 1994; Jonides and Yantis 1988; Theeuwes 1994; Yantis and Egeth 1999]. As rationale, we note that humans spend a lot of time in everyday life visually exploring other people or new environments, without necessarily being engaged in highly demanding goal-oriented behaviors. In such less constrained circumstances, perception and action may critically depend on saliency effects, regardless of whether the selected stimuli meet the laboratory criterion for bottom-up attention capture. Indeed, our data demonstrate that certain bottom-up influences, such as color contrasts, play an important role in attracting attention during natural vision, even though they do not lead to involuntary attention capture in the lab [Folk et al. 1992; Jonides and Yantis 1988]. We thus hypothesize that purely bottom-up or purely top-down selections are rare in real world environments. A key open question is the extent of cooperativeness versus competitiveness between bottom-up and top-down influences in realistic environments.

Acknowledgements

Supported by grants from NSF, NEI, NIMA, and the Zumberge Research and Innovation fund. Computation for the work described in this letter was supported by the University of Southern California Center for High Performance Computing and Communications (*www.usc.edu/hpcc*).

References

Abrams, R. A., and Christ, S. E. (2005). The onset of receding motion captures attention: comment on Franconeri and Simons (2003). *Perception & Psychophysics 67*, 219-223.

Anderson, J. D. (1996). *The reality of illusion: an ecological approach to cognitive film theory* (Carbondale, Southern Illinois University Press).

Atkinson, J., and Braddick, O. (2003). *Neurobiological Models of Normal and Abnormal Visual Development* (Hove, East Sussex ; New York, Psychology Press).

Bex, P. J., and Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *Journal of the Optical Society of America A. Optics, Image Science, and Vision 19*, 1096-1106.

Bisley, J. W., and Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science 299*, 81-86.

Caspi, A., Beutter, B. R., and Eckstein, M. P. (2004). The time course of visual information accrual guiding eye movement decisions. *Proceedings of the National Academy of Sciences of the United States of America 101*, 13086-13090.

Connolly, M., and Van Essen, D. (1984). The representation of the visual field in parvicellular and magnocellular layers of the lateral geniculate nucleus in the macaque monkey. *Journal of Comparative Neurology 226*, 544-564.

Curcio, C. A., Sloan, K. R., Jr., Packer, O., Hendrickson, A. E., and Kalina, R. E. (1987). Distribution of cones in human and monkey retina: individual variability and radial asymmetry. *Science 236*, 579-582.

Curio, E. (1976). *The ethology of predation* (Berlin; New York, Springer-Verlag).

Driver, J., Mcleod, P., and Dienes, Z. (1992). Motion coherence and conjunction search: implications for guided search theory. *Perception & Psychophysics 51*, 79-85.

Efron, B., and Tibshirani, R. (1993). *An introduction to the bootstrap* (New York, Chapman & Hall).

Fecteau, J. H., Bell, A. H., and Munoz, D. P. (2004). Neural correlates of the automatic and goal-driven biases in orienting spatial attention. *Journal of Neurophysiology 92*, 1728-1737.

Findlay, J. M. (2004). *Eye scanning and visual search*. In The Interface of Language, Vision, and Action: Eye Movements and the Visual World, J. M. Henderson, and F. Ferreira, eds. (UK, Psychology Press).

Findlay, J. M., and Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences 22*, 661-674.

Finlay, D., and Ivinskis, A. (1984). Cardiac and Visual Responses to Moving Stimuli Presented Either Successively or Simultaneously to the Central and Peripheral Visual-Fields in 4-Month-Old Infants. *Developmental Psychology 20*, 29-36.

Folk, C. L., Remington, R. W., and Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance 18*, 1030-1044.

Franconeri, S. L., Hollingworth, A., and Simons, D. J. (2005). Do new objects capture attention? *Psychological Science 16*, 275-281.

Frost, B. J., and Nakayama, K. (1983). Single Visual Neurons Code Opposing Motion Independent of Direction. *Science 220*, 744-745.

Gibson, J. J. (1979/1986). *The ecological approach to visual perception* (Hillsdale, N.J., Lawrence Erlbaum Associates).

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

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences 7*, 498-504.

Henderson, J. M., and Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology 50*, 243-271.

Hernandez-Peon, R., Scherrer, H., and Jouvet, M. (1956). Modification of electric activity in cochlear nucleus during attention in unanesthetized cats. *Science 123*, 331-332.

Hillstrom, A. P., and Yantis, S. (1994). Visual motion and attentional capture. *Perception and Psychophysics 55*, 399-411.

Hochberg, J. E. (1986). *Representation of motion and space in video and cinematic displays*. In Handbook of perception and human performance: Vol. 1. Sensory processes and perception, K. R. Boff, R. Kaufman, and J. P. Thomas, eds. (New York, Wiley), pp. 22-21 to 22-64.

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

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40*, 1489-1506.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences 6*, 345-350.

James, W. (1890). *Principles of Psychology* (Oxford, England, Henry Holt).

Jonides, J., and Yantis, S. (1988). Uniqueness of Abrupt Visual Onset in Capturing Attention. *Perception and Psychophysics 43*, 346-354.

Klin, A., Jones, W., Schultz, R., Volkmar, F., and Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry 59*, 809-816.

Koch, C., and 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., and 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., and Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature 384*, 74-77.

Land, M. F., and Mcleod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience 3*, 1340-1345.

Mannan, S. K., Ruddock, K. H., and Wooding, D. S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception 26*, 1059-1072.

Melcher, D., and Kowler, E. (2001). Visual scene memory and the guidance of saccadic eye movements. *Vision Research 41*, 3597-3611.

Mizutani, A., Chahl, J. S., and Srinivasan, M. V. (2003). Motion camouflage in dragonflies. *Nature 423*, 604-604.

Najemnik, J., and Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature 434*, 387-391.

Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. (2003). *Top-down control of visual attention in object detection*. Paper presented at: International Conference on Image Processing.

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research 42*, 107-123.

Parkhurst, D. J., and Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision 16*, 125-154.

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

Polat, U., and Sagi, D. (1994). Spatial interactions in human vision: from near to far via experience-dependent cascades of connections. *Proceedings of the National Academy of Sciences of the United States of America 91*, 1206-1209.

Regan, B. C., Julliot, C., Simmen, B., Vienot, F., Charles-Dominique, P., and Mollon, J. D. (2001). Fruits, foliage and the evolution of primate colour vision. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 356*, 229-283.

Reinagel, P., and Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network 10*, 341-350.

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

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., and Davis, J. (1995). Visual Cortical Mechanisms Detecting Focal Orientation Discontinuities. *Nature 378*, 492-496.

Sperling, G., and Weichselgartner, E. (1995). Episodic Theory of the Dynamics of Spatial Attention. *Psychological Review 102*, 503-532.

Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research 45*, 643-659.

Theeuwes, J. (1994). Stimulus-Driven Capture and Attentional Set - Selective Search for Color and Visual Abrupt Onsets. *Journal of Experimental Psychology: Human Perception and Performance 20*, 799-806.

Thompson, K., and Bordwell, D. (2003). *Film history : an introduction*, 2nd edn (Boston, McGraw-Hill).

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A. Optics, Image Science, and Vision 20*, 1407-1418.

Treisman, A. M., and Gelade, G. (1980). Feature-Integration Theory of Attention. *Cognitive Psychology 12*, 97-136.

Wolfe, J. M., Alvarez, G. A., and Horowitz, T. S. (2000). Attention is fast but volition is slow. *Nature 406*, 691.

Yantis, S., and Egeth, H. E. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance 25*, 661-676.

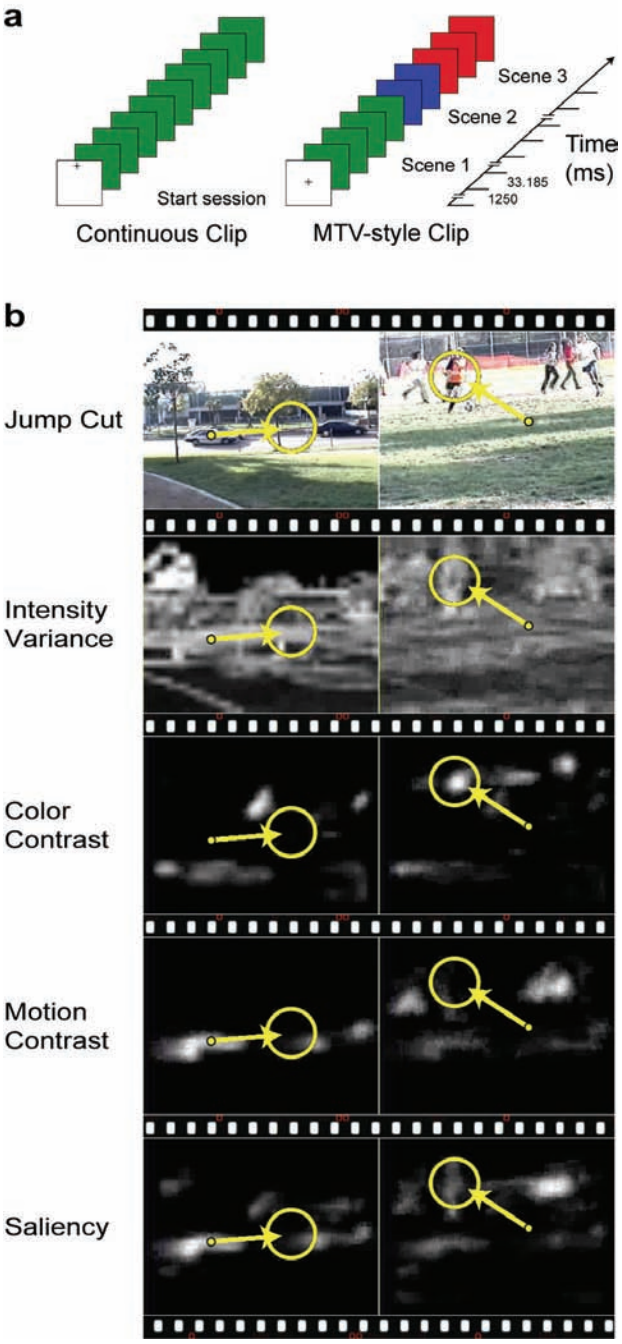

Figure 1: Jump cuts and attention-priority maps. (*a*) Schematic of the MTV-style scene shuffling. Each colored square depicts a video frame. Color changes indicate abrupt transitions between semantically unrelated scenes. (*b*) Two consecutive saccades from an MTV-style clip (#11, participant MC) that straddle a jump cut. Yellow rings depict saccade targets. Filled yellow circles mark eye-positions prior to saccade initiation. Yellow arrows show saccade trajectories. Uppermost filmstrips depict the instantaneous input frames at the time of saccade initiation. Lower filmstrips depict the corresponding intensity variance, color contrast, motion contrast, and saliency maps.

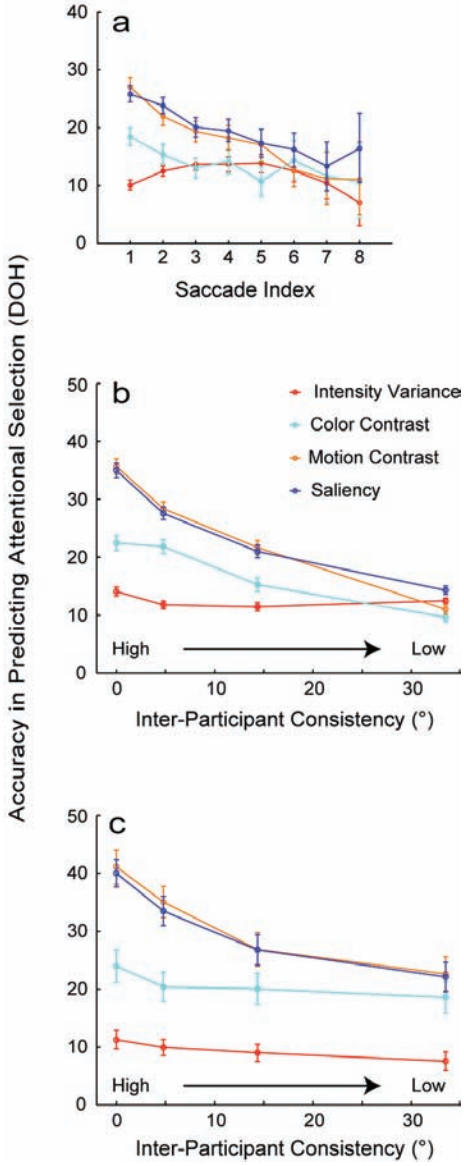

Figure 4: Saliency effects as a function of viewing time (saccade index) and inter-participant consistency. (*a*) Saliency effects as a function of saccade index between adjacent jump cuts, based on pooling saccades across all participants and jump cuts. Error bars depict the S.E. for 1000 bootstrap subsamples.
(*b*) Saliency effects as a function of inter-participant consistency, which is measured by the area of the smallest rectangle bounding the instantaneous eye-positions of different participants. DOH values were computed for the same areas as in c, based on all available saccades. (*c*) Saliency effects as a function of inter-participant consistency, but only for saccades initiated within 250 ms after jump cuts. To maximize the reliability of DOH values, saccades were divided into quartiles, resulting in the following area bins: [0°-4.8°], (4.8°-14.4°]; (14.4°-33.5°]; (33.5°-314.3°].