Available online at www.sciencedirect.com

ScienceDirect

Vision Research 46 (2006) 4333–4345

Vision Research

www.elsevier.com/locate/visres

Visual causes versus correlates of attentional selection in dynamic scenes

Ran Carmi *, Laurent Itti

Neuroscience Program, University of Southern California, USA

Received 29 December 2005; received in revised form 22 July 2006

Abstract

What are the visual causes, rather than mere correlates, of attentional selection and how do they compare to each other during natural vision? To address these questions, we first strung together semantically unrelated dynamic scenes into MTV-style video clips, and performed eye tracking experiments with human observers. We then quantified predictions of saccade target selection based on seven bottom-up models, including intensity variance, orientation contrast, intensity contrast, color contrast, flicker contrast, motion contrast, and integrated saliency. On average, all tested models predicted saccade target selection well above chance. Dynamic models were particularly predictive of saccades that were most likely bottom-up driven-initiated shortly after scene onsets, leading to maximal inter-observer similarity. Static models showed mixed results in these circumstances, with intensity variance and orientation contrast featuring particularly weak prediction accuracy (lower than their own average, and approximately 4 times lower than dynamic models). These results indicate that dynamic visual cues play a dominant causal role in attracting attention. In comparison, some static visual cues play a weaker causal role, while other static cues are not causal at all, and may instead reflect top-down causes.

Keywords: Attention; Eye movements; Natural vision; Natural scenes; Modeling

1. Introduction

Orienting to salient visual cues, such as color or motion contrasts, provides a fast heuristic for focusing limited neurocomputational resources on behaviorally relevant sensory inputs. Converging evidence from neurophysiological (Fecteau, Bell, & Munoz, 2004; Gottlieb, Kusunoki, & Goldberg, 1998), psychophysical (Folk, Remington, & Johnston, 1992; Jonides & Yantis, 1988) and developmental (Atkinson & Braddick, 2003; Finlay & Ivinskis, 1984) studies indicates that dynamic stimuli are particularly effective in attracting human attention. Nonetheless, most computational studies of saliency[1] effects (the impact of bottom-up influences on attentional selection) examined visual correlates of fixations in the context of static scenes (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Mannan, Ruddock, & Wooding, 1997; Oliva, Torralba, Castelhano, & Henderson, 2003; Parkhurst, Law, & Niebur, 2002; Parkhurst & Niebur, 2003; Peters, Iyer, Itti, & Koch, 2005; Reinagel & Zador, 1999; Tatler, Baddeley, & Gilchrist, 2005; Torralba, 2003). Such studies provided valuable accounts of saliency effects, but the scalability of their conclusions to the dynamic real world remains an open question. Furthermore, the focus on correlations provides limited insight into causal mechanisms of attentional selection. For example: top-down guided orienting towards objects that have luminance-defined contours may lead to non-causal correlations between local edges and fixation locations.

Psychophysicists solve the potential confound between bottom-up and top-down causes by constructing multi-element search arrays, and measuring the extent to which task-irrelevant bottom-up cues, such as color or motion singletons, reduce search efficiency (Abrams & Christ, 2005; Folk et al., 1992; Franconeri, Hollingworth, &

* Corresponding author.
E-mail address: rancarmi@gmail.com (R. Carmi).

[1] Unless otherwise specified, we use the term “saliency” to refer to any bottom-up measure of conspicuity. The term “integrated saliency” refers to a particular bottom-up model that combines different visual contrasts into a unified saliency measure (see Section 2.5).

0042-6989/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2006.08.019

Simons, 2005; Hillstrom & Yantis, 1994; Jonides & Yantis, 1988; Theeuwes, 1994; Yantis & Egeth, 1999). Such studies have been instrumental in identifying strong bottom-up influences that capture attention involuntarily in the presence of competing top-down influences. However, the focus on experimental conditions that discourage observers from paying attention to salient stimuli may underestimate the impact of bottom-up cues in real world environments. Moreover, the costs relative in reaction time incurred by different visual cues provide, at best, indirect estimates of relative impact on attentional selection.

In this study, we quantified saliency effects in the context of complex dynamic scenes by measuring the prediction accuracy of seven bottom-up models of attentional selection. To minimizes potential top-down confounds without sacrificing real world relevance (ecological validity), we generated MTV-style video clips by stringing together semantically-unrelated clip snippets (clippets). The abrupt transitions (jump cuts) between clippets were deliberately designed to maximize semantic unrelatedness each MTV-style clip contained at most one clippet from a given continuous clip, and no attempt was made to conceal the cuts.

We measured saliency effects for different saccade populations, and particularly focused on subsets of saccades that were most likely to be bottom-up driven, such as saccades initiated shortly after jump cuts, leading to maximal inter-observer similarity (minimal variability). The rationale for our methodology is based on previous reports of a trade-off between bottom-up and top-down influences (Henderson & Hollingworth, 1999; Hernandez-Peon, Scherrer, & Jouvet, 1956; James, 1890). This trade-off implies that attentional selections should depend most heavily on bottom-up influences in circumstances that are least likely to involve top-down influences.

The results show that certain static cues, including luminance variance and orientation contrast, are the least predictive of attentional selection in exactly those circumstances in which the impact of bottom-up cues is expected to be the strongest. In the same circumstances, other visual cues, including intensity contrast, color contrast, and to a greater extent flicker contrast, motion contrast, and integrated saliency are the most predictive of attentional selection. In the discussion, we propose novel hypotheses and related future studies that could further elucidate mechanisms of attentional selection in realistic environments.

2. Methods

2.1. Participants

Eight human observers (3 women and 5 men), 23- to 32-years-old, provided written informed consent, and were compensated for their time ($12/h). All observers were healthy, had normal or corrected-to-normal vision, and were naïve as to the purpose of the experiment.

2.2. Stimuli

Fifty video clips (30 Hz, 640×480 pixels/frame, 4.5–30 s long, mean ± SD: 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 strung together by jump cuts into 50 scene-shuffled (MTV-style) clips (see Fig. 1 and Supp. Videos S1–S4). The range of clippet lengths was chosen such that observers would have enough time to perform several saccades within each clippet. The clippet lengths were randomized within the chosen range to minimize the ability of observers to anticipate the exact timing of jump cuts.

2.3. Experimental design

Observers 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 ($28° \times 21°$ 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". Observers 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 observers 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, in which instructionless free viewing sometimes led to observers disengaging from the display and looking around the room. A previous study found no task-related effects compared to free viewing observers who did not disengage from the display (Itti, 2005).

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), which tracks the pupil and corneal reflection. Calibration and saccade extraction procedures are described elsewhere (Itti, 2005). In this experiment, the calibration accuracy was $0.66° \pm 0.46°$ (mean ± SD), and a total of 10221 saccades were extracted from the raw eye-position data. Thirty-four saccades (0.3%) either started or ended outside of the display bounds, and were thus excluded from the data analysis, which was based on the remaining 10187 saccades.

2.5. Bottom-up attention-priority maps

Two-dimensional attention-priority, or saliency, maps (40×30 pixels/frame) were generated based on seven computational models: intensity variance (squared RMS contrast), integrated saliency, and individual saliency components (contrasts in color, intensity, orientation, flicker, and motion).

The intensity variance map was computed per input frame (30 Hz) based on the variance of pixel intensities in independent image patches:

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

where p refers to an individual image patch, m and n are its the width and in pixels (16×16, subtending $0.7° \times 0.7°$), I is the intensity of an image pixel, and I_p is the mean intensity of the patch. This model is used here, because it was previously proposed as a measure of perceptual contrast in natural images (Bex & Makous, 2002), and particularly as a visual correlate of fixation locations (Parkhurst & Niebur, 2003; Reinagel & Zador, 1999).

The other bottom-up maps were each computed by a series of non-linear integrations of center-surround differences across several scales (and feature dimensions, in the case of the integrated saliency model). Maps were initially computed at the input frame rate (30 Hz), fed into a two-dimensional

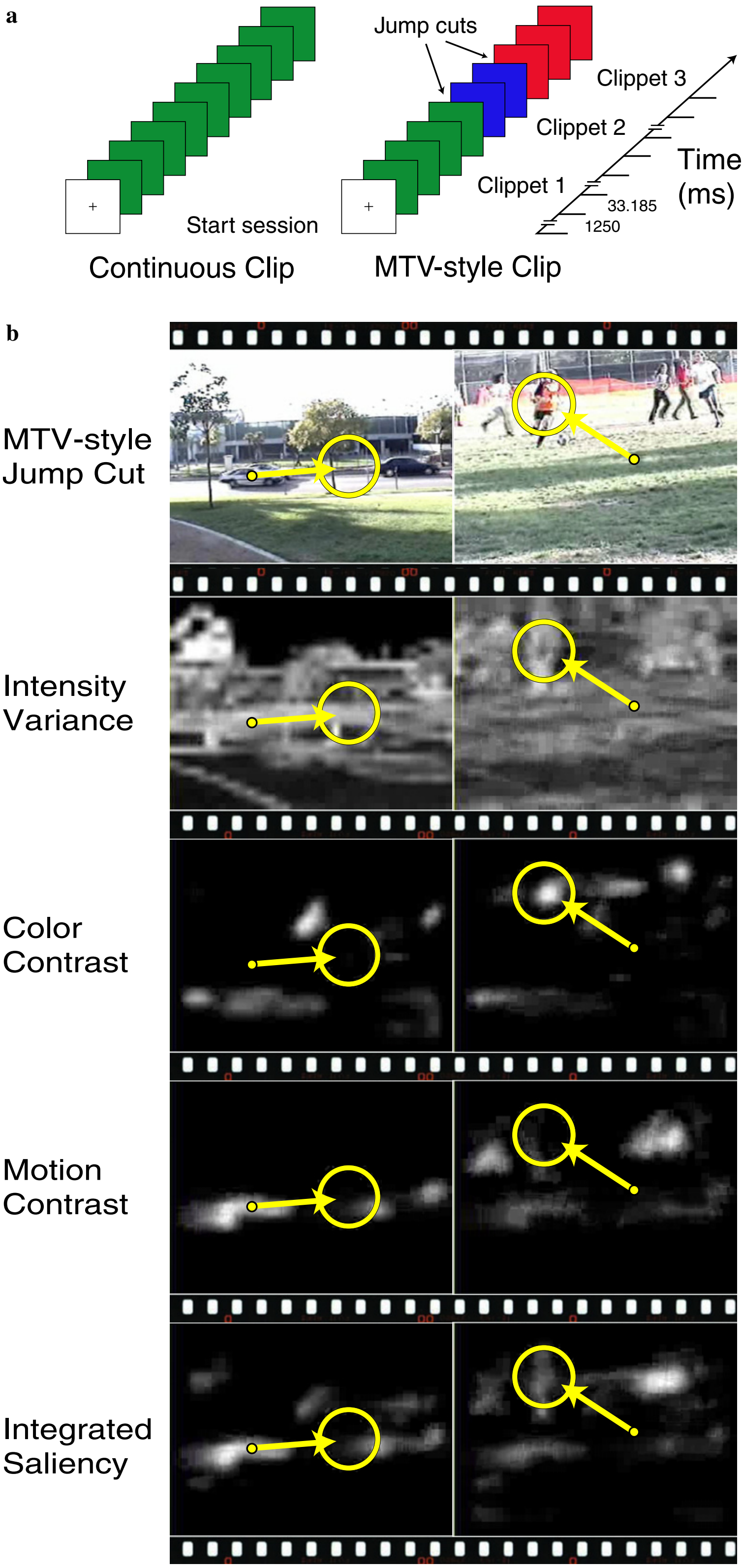

Fig. 1. MTV-style clips and attention-priority maps. (a) Schematic of the MTV-style scene shuffling manipulation. Each colored square depicts a video frame. Color changes indicate jump cuts—abrupt transitions between semantically unrelated clippets. (b) Two consecutive saccades from an MTV-style clip (#11, participant MC, $\Delta t = 298.7$ ms) that straddle a jump cut. Light-colored (yellow) markers depict the instantaneous eye-positions prior to saccade initiation (discs), the saccade trajectories (arrows), and the saccade targets (rings). Uppermost filmstrips depict the instantaneous input frames at the time of saccade initiation. Lower filmstrips depict the corresponding attention-priority maps based on the intensity variance, color contrast, motion contrast, and integrated saliency model (Supp. Videos S1–S4, respectively).

layer of leaky integrator neurons that provide temporal smoothing at 10 kHz, and eventually downsampled to the eye tracker sampling rate (240 Hz). These computations are described extensively elsewhere (Itti, 2005; Itti & Koch, 2000). An earlier version of this saliency model was published as part of a larger framework for simulating attention shifts (Itti & 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, but they are outside the scope of the current investigation, which aims to characterize saliency effects per se. The particular scale of attention-priority maps was chosen such that local measurements (16 × 16 pixels, 0.7° × 0.7°) corresponded to the largest effect size reported for visual correlates of attentional selection in the context of static images (Parkhurst & Niebur, 2003). All simulations were run on a Linux-based computer cluster (total run time for analyzing all the video clips using all the models: 792 processor hours). The software that was used to generate attention-priority maps is freely available for academic research, and can be downloaded from: http://ilab.usc.edu/toolkit.

2.6. Bottom-up prediction of single saccades

Normalized prediction for all human 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. Measurements were taken at the end of the fixation period prior to saccade initiation, as defined by the last eye-position sample during the preceding fixation. The timing of these measurements is based on the assumption that bottom-up influences are mostly accrued during the preceding fixation (Caspi, Beutter, & Eckstein, 2004; Parkhurst et al., 2002). We did not explicitly take into account the 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 & Koch, 2000). We also did not try to optimize the sampling latency, and instead used subjective observations to verify that the saliency of newly appearing targets reaches its peak value in close proximity to the initiation of human saccades towards these targets. Supplementary Supp. Video S5 demonstrates that the particular latency we chose agrees well with the timing of human selections in the context of a synthetic test clip. Whatever the optimal latency is, sampling attention-priority maps prior to saccade target selection is important for establishing causation rather than mere correlation.

We compensated for potential inaccuracies in human saccade targeting and the eye-tracking apparatus by sampling the maximal local value in an aperture around each saccade target ($r = 3.15°$). The aperture size was chosen rather arbitrarily to be on the scale of the parafovea. It should be noted that any choice of aperture size involves a trade-off between false positives and false negatives. For example, if a saccade is initiated towards and lands on non-salient text that happens to be located next to more salient stimuli, then too big of an aperture would lead to a false positive. In contrast, if a saccade is initiated towards a salient moving target but misses it slightly, then too small of an aperture would lead to a false negative. We did not try to optimize model performance by systematically varying the aperture size. In any case, the baseline measures (see next section) provide saccade-by-saccade safeguards against any biases that may be introduced by the particular choice of aperture size.

2.7. Baseline sampling

To quantify and compare the agreement between human attentional selection and different attention-priority maps (see next section), we utilized two types of baseline measures: one based on a uniform distribution of potential targets and the other based on a distribution of human-fixated locations. Baseline measures are important because they minimize potential artifacts due to the distribution of saliency values, which may vary substantially across different attention-priority maps as a function of the underlying model and the instantaneous input. To calculate the baseline, attention-priority maps were sampled at a randomly selected location concurrently with the initiation of each human saccade. Other than the randomness of the location, the sampling procedure for these so-called random saccades was identical to the one described above for human saccades. Baseline measures reward sparse maps with high target selectivity at the expense of dense maps with low target selectivity. For example, in the absence of a baseline, models could achieve high hit rates and prediction accuracy by generating uniform attention-priority maps (every sample will be a hit). With a baseline the hit rates of human and random saccades will be identical in the case of a uniform attention-priority map, reflecting its low prediction accuracy prediction accuracy.

It has been proposed that baseline sampling should be based on a distribution of human-fixated locations rather than a uniform distribution (Parkhurst & Niebur, 2003; Tatler et al., 2005). This proposal is motivated by reports of centrally-biased distributions of human fixations (Itti, 2005; Parkhurst & Niebur, 2003; Tatler et al., 2005), coupled with the assumption that such biases are caused by motor constraints or top-down influences rather than bottom-up influences. If this assumption is valid, then sampling baseline targets from a uniform distribution of locations may lead to artifactual results, especially when measuring saliency effects as a function of viewing time (and assuming that viewing sessions begin with a central fixation cross and involve centrally-biased distribution of saliency values, as was the case in most related studies performed to date). Whether or not this was an issue in previous studies, it should not be a concern in this study, because the temporal analyses presented here are aligned to jump cuts, which are not preceded by a predetermined fixation cross (central or otherwise). Furthermore, it is unclear whether the use of a human-fixated baseline is justified even in the context in which it was initially proposed. If bottom-up influences play a causal role in determining the fixational center bias, then using a human-fixated baseline would underestimate the magnitude of saliency effects, potentially leading to an even bigger artifact than the one it aims to remove. The causes of fixational center bias and their relative impact are not well understood, so the rationale for preferring a human-fixated baseline over a (simpler) uniform baseline seems tenuous at best. Nevertheless, to remove any doubts from the minds of readers about the potential dependence of the results presented here on the baseline type, we computed the key results using both the uniform baseline and human fixated baseline (see Fig. 5).

2.8. Performance metrics for quantifying the agreement between human attentional selection and attention-priority maps

2.8.1. DOH metric

The difference of histograms (DOH) metric quantifies the human tendency to initiate saccades towards salient targets by measuring the rightward shift of the human saccade histogram relative to the baseline saccade histogram:

$$\mathrm{DOH} = (1/\mathrm{DOH}_I) \times \sum_{i=1}^{n} W_i \times (H_i - R_i) \qquad (2)$$

where H_i and R_i are the fractions of human and baseline 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 reflects the assumption that deviations from the baseline in high saliency bins are more likely to reflect signal than noise, and should be thus weighted more strongly than similar deviations in low saliency bins. We used a linear weighting scheme because of its simplicity, but other monotonic functions could serve the same purpose.

DOH values are expressed as percentages of DOH_I, which reflects the ideal rightward shift of the human saccade histogram relative to the baseline saccade histogram:

$$\mathrm{DOH}_I = (W_n - W_1) \times (1 - p) = 0.8633 \qquad (3)$$

Theoretically, the largest possible saliency difference between human and baseline targets would occur if human and baseline saccades always land on the maximal and minimal saliency values, respectively. However, even if assuming an ideal model that always generates a single saliency value at saccade targets, and 0 elsewhere (see Fig. 2), a certain fraction of baseline saccades would land on the maximal salinecy value by chance, with approximate probability:

$$p = N_a/N_m = 0.0408 \tag{4}$$

where $N_a = 49$ is the number of pixels in an aperture around the saccade target ($r = 3.15°$, defined by 9 adjacent rows consisting of 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, where $W_m = 40$ is the map width, and$H_m = 30$ is the map height.

In the ideal scenario, the human histogram (saccade probability as a function of saliency at saccade target) will only contain saccades in the highest bin (90–100% of the max saliency), while the baseline histogram will have $1-p$ saccades in the lowest bin (0–10% of the max saliency), and p saccades in the highest bin. In comparison, the null scenario occurs when a model is unpredictive of attentional selection, in which case human and baseline saccades would be just as likely to hit salient targets, leading to a complete overlap between human and baseline histograms. To summarize, the expected range of DOH values is between 0 (chance) and 100 (ideal). Models that are worse predictors than chance would lead to negative DOH values.

It is interesting to note that the DOH values reported here provide a conservative estimate for the relative contribution of bottom-up versus top-down influences on attentional selection. Given that different observers do not always look at the same place simultaneously, even the ideal attention-priority map should sometimes contain more than one potential candidate. Consequently, the probability of baseline saccades landing on valid attention candidates would be higher than reported here, leading to a lower DOH upper bound. More realistic estimates could potentially be computed by taking into account the actual extent of inter-observer similarity. Depending on the metric used to quantify inter-observer similarity, a potential downside of this approach would be that it would make the upper bound dependent on the number of observers considered. The conclusions of this study are independent of the upper bound because they only rely on differences in bottom-up impact across conditions that share the same upper bound. We included the upper bound in the metric definition, because it makes the metric values intuitively more meaningful. Moreover, computing a realistic upper bound would be critical for any attempt to quantify the relative contribution of bottom-up versus top-down influences, which is an exciting follow-up question that is outside the scope of this study.

2.8.2. Percentile metric

The percentile metric is defined as:

$$P = (1/N) \times \sum_{i=1}^{N} p_i \tag{5}$$

where N is the number of human saccades, and p_i is the percentile of the sampled value of the attention-priority map at a human saccade target prior to saccade initiation. Percentiles were calculated by generating 100 baseline samples for each human saccade, and counting the number of baseline samples whose value was smaller than or equal to the human sample. This metric is similar to the ROC metric proposed in a previous study (Tatler et al., 2005), but is more appropriate in the context of dynamic stimuli that involve ever changing attention-priority maps. The ROC metric is useful in the context of a static attention-priority map that involves two stable distributions of fixated and non-fixated locations. In our data, the distribution of saliencies at non-fixated locations is unique for each human saccade, and the discriminability between that distribution and the saliency at the saccade target is equiv-

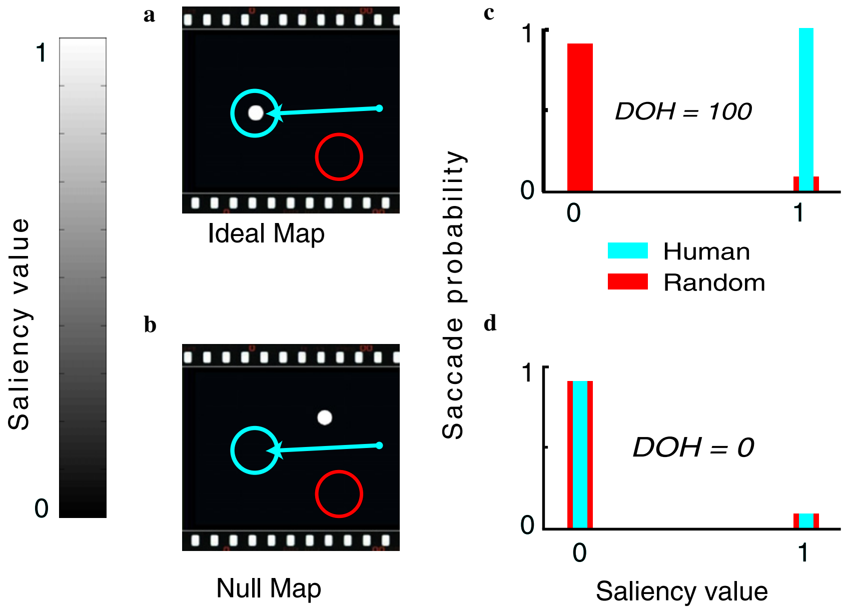

Fig. 2. Ideal and null predictions of attentional selection. (a) Ideal attention-priority map prior to saccade initiation containing a single positive value at the human saccade target, and zero elsewhere. Light-colored (cyan) markers depict the instantaneous eye-position (disc), the saccade trajectory (arrow), and the saccade target (ring). A random target is depicted by a dark-colored (red) ring (b) Same as a, but showing a null attention-priority map. Any map that contains positive values at random locations would qualify as a null map, but in this case only a single location is selected randomly and set to a positive value. (c) Saccade probability as a function of saliency value at saccade target based on the ideal map, which leads to the largest possible shift to the right of the human saccade histogram (light cyan) relative to the random saccade histogram (dark red). (d) Same as c, but based on the null map. Human and random saccades are equally likely to land on positive values, leading to identical histograms that are perfectly aligned. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

alent to the percentile of the human sample. Similar to ROC, the expected range of percentile values is from 50 (chance) to 100% (best possible prediction accuracy).

2.9. Pros and cons of different performance metrics

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 & Niebur, 2003; Reinagel & Zador, 1999; Tatler et al., 2005; Torralba, 2003), including: linearity, meaningful upper bound, priority weighting, directionality, and sensitivity to high-order statistics. The strongest 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 particular context of quantifying the agreement between human attentional selection and attention-priority maps. For example: both KL-divergence and DOH estimate the overall dissimilarity between two probability density functions—the saliency at human fixated vs. random locations. In contrast the DOH, the KL-divergence metric is nonlinear (metric values for different conditions or models cannot be compared as interval variables), has an infinite upper bound, contains no saliency-based weighting to boost the signal-to-noise ratio, and is bi-directional (no distinction between instances in which models are more versus less predictive than chance). In comparison, the ROC metric (Tatler et al., 2005) estimates the overall discriminability between two probability density functions (saliency at fixated vs. non-fixated locations). Relative disadvantages of the ROC metric are the lack of saliency-based weighting, and its smaller range of possible values (this range is probably even smaller than it appears, considering that the upper bound could only be reached if the underlying distributions are linearly separable). Furthermore, the ROC metric is most useful for static rather than dynamic conditions, as described in Section 2.8.2. The percentile metric is similar to ROC, but is computed on a saccade-by-saccade basis, which makes it equally applicable to both static and dynamic conditions. A relative advantage of the percentile metric compared DOH is its simplicity, but similar to the other metrics considered, it contains no saliency-based weighting (although such weighting could be added easily when computing the average metric value across saccades).

2.10. Advantages of jump cuts over clip onsets as temporal anchor point for measuring saliency effects as a function of viewing time

(1) Contrary to clip onsets, the exact timing of jump cuts is neither controlled by participants nor exactly predictable. Consequently, jump cuts are less sensitive to potential top-down artifacts and provide a cleaner dissociation of bottom-up and top-down influences.
(2) The fact that jump cuts occur during natural visual exploration minimizes center bias artifacts, which may arise due to a combination of factors, as described in Section 2.7. Several previous studies attempted to correct these potential artifacts post hoc during the analysis stage (Parkhurst & Niebur, 2003; Reinagel & Zador, 1999; Tatler et al., 2005). The relative advantage of jump cuts over clip onsets in this context is that they are not preceded by a predetermined fixation location (central or otherwise). Consequently, jump cuts minimize potential artifacts in measuring saliency effects without making unwarranted assumptions about the underlying causes of center bias.
(3) In our experiment, observers were exposed to more jump cuts than clip onsets (by an order of magnitude). Correspondingly, there are many more saccades available for analysis after jump cuts versus clip onsets, leading to relatively higher signal to noise ratio when measuring saliency effects as a function of viewing time.

3. Results

3.1. Average saliency effects based on all saccades

In realistic viewing conditions, overt attentional selections (saccades) are strongly coupled with covert attentional selections (Findlay, 2004; Kustov & Robinson, 1996; Sheinberg & Logothetis, 2001; Sperling & Weichselgartner, 1995). This coupling provides the rationale for studying attentional selection using saccade-based measures, as is done in this study. Fig. 1b shows examples of the instantaneous input, corresponding attention-priority (saliency) maps, and two consecutive saccades that straddle an MTV-style jump cut. For each saccade and attention-priority map, we sampled the map value at the saccade target and simultaneously at a random target (see Sections 2.6 and 2.7). Fig. 3 shows the overall human and random saccade histograms (saccade probability as a function of saliency at the saccade target) for representative models. The random saccade histograms reflect the probability density function of saliency values, while the human saccade histograms show the extent to which human selection of attention targets is biased towards salient locations. Figs. 1 and 3 demonstrate that different models generate different attention-priority maps for the same input, in terms of both the location and density of saliency values. For example: the intensity variance model generates the densest maps, with only 2% of random saccades landing on the lowest possible saliency value (0–10% of the max), while the motion contrast model generates the sparsest maps, with approximately 50% of random saccades landing on the lowest possible saliency value. The average prediction accuracy of all the tested bottom-up models was significantly higher than chance (DOH $= 0$, $z \gg 1.96$, $p \ll 0.01$). The most predictive model— integrated saliency—was on average 1.7 times more predictive than the least predictive model—intensity variance ($t(10185) = 21.8406$, $p \ll 0.01$).

3.2. "Bottom-up" labeling of saccades

The average prediction accuracy reported in Fig. 3 is suggestive of the relative impact of different visual cues on attentional selection, but these results may in fact be misleading because they are based on all saccades, including those that were not determined by bottom-up influences. To test the relative impact of bottom-up influences, it is informative to focus on bottom-up driven saccades. Unfortunately, we do not know how to unambiguously label particular saccades performed during visual exploration of real world scenes as "top-down guided" or "bottom-up driven". In fact, if attentional selections are determined by continuous interactions between bottom-up and top-down influences, then such unambiguous labeling of saccades is an ill-posed problem.

That said, it is possible to identify special circumstances in which humans are particularly sensitive to bottom-up influences. For example, saccades that are initiated shortly

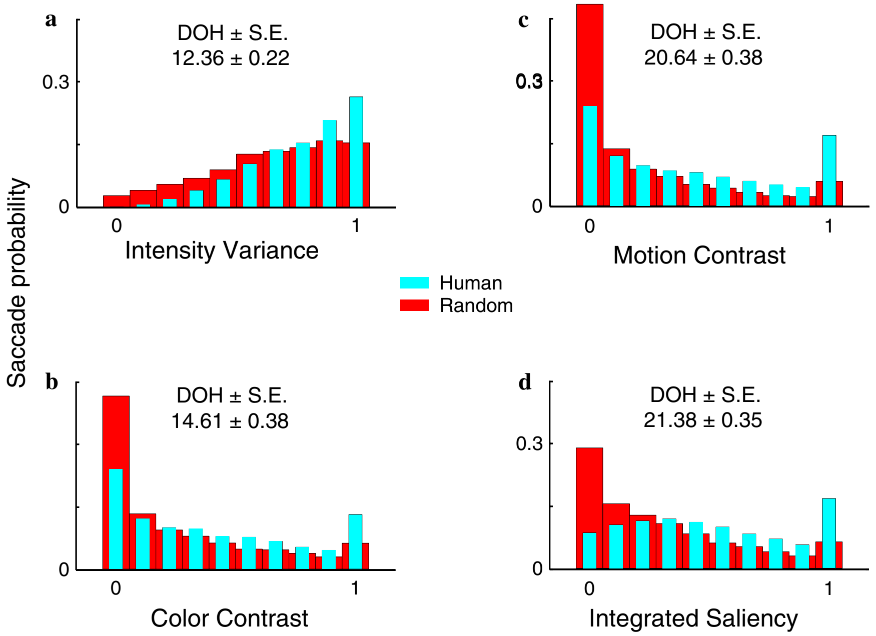

Fig. 3. Saccade histograms and the average prediction accuracy of representative bottom-up models. Numbers above histograms show the prediction accuracy of each model based on the DOH metric (see Section 2.8.1). (a) Intensity Variance, (b) Color Contrast, (c) Motion Contrast and (d) Integrated Saliency.

after exposure to novel scenes may be more bottom-up driven than later saccades, given that bottom-up influences are faster acting than top-down influences (Henderson, 2003; Wolfe, Alvarez, & Horowitz, 2000). The existing evidence for this hypothesis is mixed: one study found relatively stronger saliency effects early after stimulus onset than later on (Parkhurst et al., 2002) but a more recent study found no interaction between saliency effects and viewing time (Tatler et al., 2005). Another special circumstance that may indicate "bottom-up driven" saccades is when observers look at the same location simultaneously. The rationale is that top-down influences depend on prior knowledge and specific expectations that may not be the same for different observers, and lead them to look at different locations at the same time. In contrast, bottom-up influences depend more exclusively on the instantaneous stimulus content, which is physically identical for different observers, and thus more likely to simultaneously attract their attention to the same location. In other words, saccades that lead to relatively high inter-observer similarity are more likely to have been driven by bottom-up versus top-down influences (Mannan et al., 1997). Alternatively, differences in the level of inter-observer variability may only reflect changes in the similarity between top-down influences affecting different observers (top-down divergence), without involving changes in the impact of bottom-up influences (Tatler et al., 2005).

3.3. Saliency effects as a function of viewing time

To examine the potential interactions between saliency effects and viewing time, we quantified the accuracy of different bottom-up models in predicting attentional selection as a function of time and saccade index between adjacent jump cuts. Both analyses led to the same pattern of results, so to conserve space and facilitate direct comparisons with previous studies that examined this issue (Parkhurst et al., 2002; Tatler et al., 2005), we show only the saccade index analysis (see Fig. 4). Section 2.10 describes the methodological advantages of

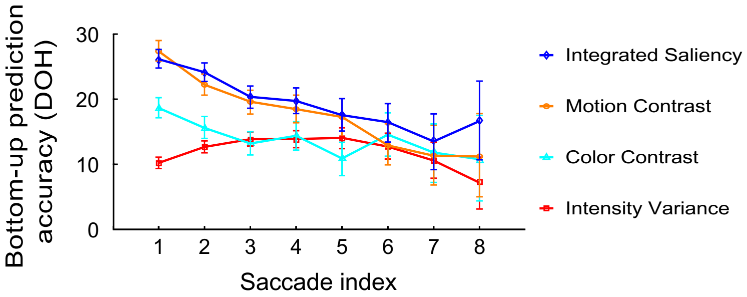

Fig. 4. Saliency effects as a function of saccade index between adjacent jump cuts. Saccades were pooled over all participants and clippets. Prediction accuracy was quantified using the DOH metric (see Section 2.8.1). Error bars depict standard errors based on 1000 bootstrap subsamples (Efron & Tibshirani, 1993).

aligning the temporal analysis of saliency effects to jump cuts instead of clip onsets.

Fig. 4 demonstrates that the integrated saliency model is 2.6 times better than the intensity variance model in predicting attentional selection ($t(10185) = 18.1212$, $p \ll 0.01$), when the analysis is based on the first saccades after jump cuts. It also shows that the prediction accuracy of the motion contrast and integrated saliency models peaks immediately after jump cuts, followed by slow decreases across seven consecutive saccades. Similarly, the prediction accuracy of the color contrast model decreases over time, but only across the first 3–5 saccades. The prediction accuracy of the intensity variance model shows the opposite initial trend—it starts low and increases slowly across the first 4–5 saccades.

Two previous studies argued that relying on a uniform distribution of locations for baseline sampling may introduce artifactual saliency effects (Parkhurst & Niebur, 2003; Tatler et al., 2005). To avoid such artifacts, the authors proposed that baseline sampling should rely instead on a distribution of human-fixated locations. Section 2.7 explains in detail why we believe that using a uniform distribution of locations is more justified in general, and particularly in the context of this study. However, to remove any doubts from the minds of readers about the potential dependence of the results presented here on the baseline type, we re-analyzed saliency effects as a function of viewing time using both uniform and human fixated baselines. To examine whether the obtained results strongly depend on the newly proposed DOH metric, we also used a percentile-based metric (see Section 2.8.2).

Similar to Figs. 4 and 5 show the measured saliency effects as a function of viewing time for the best and worst bottom-up predictors (see Fig. 3), but using different metrics and baseline types. As in Fig. 4, the prediction accuracy of the integrated saliency model starts high and becomes lower over time, while the prediction accuracy of the intensity variance model starts low and becomes higher over time. These trends are not affected by either the metric type or the baseline type. Moreover, the prediction accuracy of the integrated saliency model for the first saccades after jump cuts is significantly higher than the corresponding prediction accuracy of the intensity variance model, regardless of the metric type or baseline type. There are also dissimilarities compared to Fig. 4. For example, both the metric type and baseline type modulate the magnitude of the differences in prediction accuracy between models. The biggest differences in prediction accuracy between the intensity variance and integrated saliency models were measured by the DOH metric using the uniform baseline, while the smallest differences were measured by the percentile metric using the human-fixated baseline. Another noticeable trend is that the baseline type differentially affects the prediction accuracy of different models. Specifically, the prediction accuracy of the intensity variance model is not significantly modulated by the baseline type,

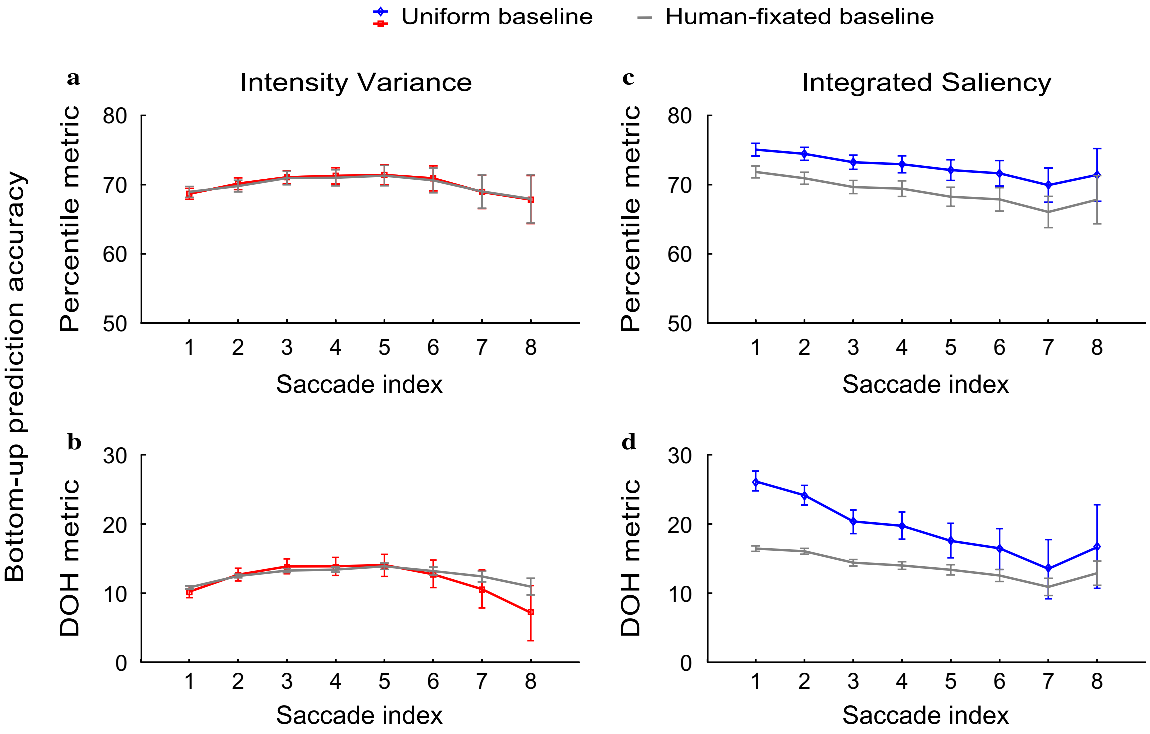

Fig. 5. Saliency effects as a function viewing time: effects of baseline type and metric type. (a) Similar to Fig. 4, but focusing on the Intensity Variance model (least predictive bottom-up model, see Fig. 3). Prediction accuracy was quantified by the percentile metric (see Section 2.8.2) using either the uniform baseline or the human-fixated baseline (see Section 2.7). (b) Same as a, but using the DOH metric (see Section 2.8.1). (c) Same as a, but focusing on the Integrated Saliency model (most predictive bottom-up model, see Fig. 3). (d) Same as b, but focusing on the Integrated Saliency model.

while the prediction accuracy of the integrated saliency model is significantly lower when the human-fixated baseline is used. This trend provides further evidence for the causal role of integrated saliency, but not of intensity variance, in determining attentional selection, since the human-fixated baseline is expected to underestimate causal saliency effects (see Section 2.7). In summary, Fig. 5 demonstrates that the key results presented in Fig. 4 do not depend on either the metric type or the baseline type. As described in Sections 2.7 and 2.10, there are compelling reasons to prefer the DOH metric with the uniform baseline over the available alternatives, so this is the metric of choice in the following analyses.

3.4. Saliency effects as a function of inter-observer variability

The second heuristic that we used to label saccades as “bottom-up driven” relied on identifying circumstances in which there was relatively high similarity (low variability) in attentional selection between different observers. To measure inter-observer variability, we fit a rectangle around the instantaneous gaze positions of different observers (at the end of each saccade made by each observer). The area of the bounding rectangle divided by the total display area reflects the extent to which observers look at the same location simultaneously. The main advantages of this metric are its simplicity and intuitiveness (0 indicates maximal similarity—observers look at the same location simultaneously, and 100 indicates maximal variability—different observers look at different corners of the display at exactly the same time). A potential disadvantage of this metric is that its values may be misleading in certain instances, for example: the area of the bounding rectangle will be zero if different observers are perfectly aligned horizontally or vertically, even though they may actually be looking at different locations along a line. In actuality, the eye-tracking data that we collected contained no such instances. To be on the safe side, we also quantified inter-observer variability based on the mean squared distance between the gaze positions of different observers. The pattern of results did not change as a function of the metric used, so to conserve space we only show the results based on the intuitively more appealing area metric.

Fig. 6a shows saliency effects as a function of inter-observer variability based on all the available saccades. It demonstrates that the integrated saliency model is 2.5 times better than the intensity variance model in predicting attentional selection ($t(10185) = 14.0763$, $p \ll 0.01$), when the analysis is based on saccades that led to minimal inter-observer variability (bounding rectangle area <1% of the total display area). Fig. 6a also demonstrates that saliency effects generally decrease as a function of inter-observer variability, although the intensity variance model shows a U-shaped pattern. Finally, Fig. 6b shows the accuracy of different bottom-up models in predicting attentional selection as a function of inter-observer variability, but based on the fastest first saccades (initiated within 250 ms after jump cuts). The first data point in Fig. 6b demonstrates that the integrated saliency model is 3.6 times better than the intensity variance model in predicting attentional selection ($t(10185) = 10.1349$, $p \ll 0.01$), when the analysis is based on saccades that are most likely to have been driven by bottom-up influences (initiated shortly after jump cuts, and leading to minimal inter-observer variability).

To summarize, Fig. 7 plots the prediction accuracy for all the tested models in two conditions: “All” saccades and “bottom-up” saccades (corresponding to the first data point in Fig. 6b). It demonstrates that the prediction accuracy of dynamic models (flicker contrast, motion contrast, and integrated saliency) is twice higher for “bottom-up” saccades compared to “All” saccades. The intensity contrast and color contrast models show a more moderate rel-

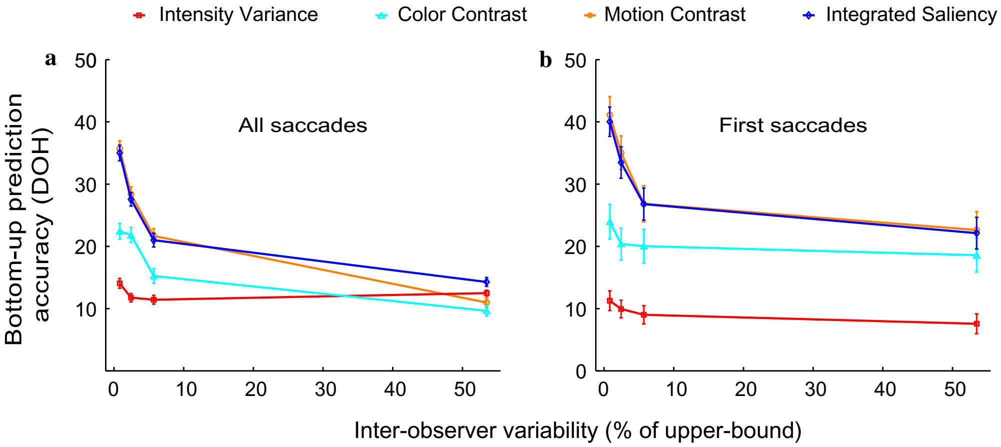

Fig. 6. Saliency effects as a function of inter-observer variability (the area of the smallest rectangle bounding the instantaneous eye-positions of different observers, divided by the display area). (a) Based on all the available saccades. Bin boundaries are the same as in b. (b) Based on the fastest first saccades (initiated within the initial 250 ms after jump cuts). To maximize the reliability of DOH values, saccades were grouped into quartiles that have the following bin boundaries (% of the display area): (0–0.81), (0.81–2.44), (2.44–5.70), and (5.70–53.46).

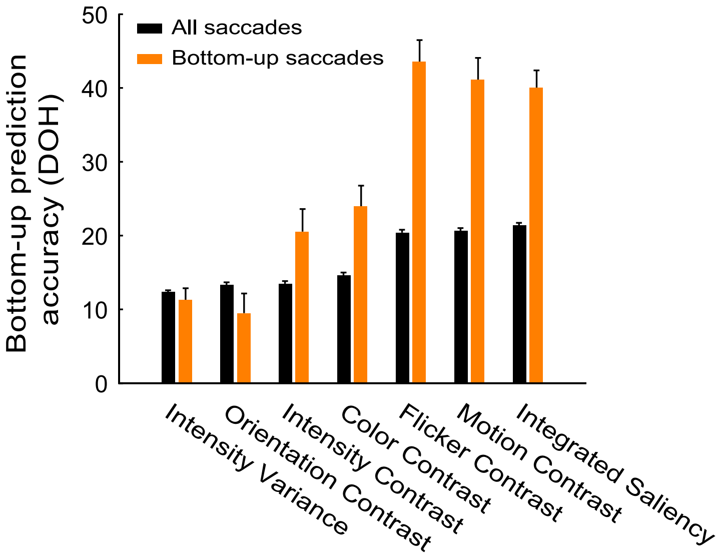

Fig. 7. Saliency effects for "All" versus "Bottom-up" saccades. The prediction accuracy for the "All" condition was quantified as shown in Fig. 3 and described in the text. The prediction accuracy for the "bottom-up" condition is based on the first data point in Fig. 6b, reflecting a subset of saccades that were initiated shortly after jump cuts, and led to minimal inter-observer similarity.

ative increase in prediction accuracy for "bottom-up" saccades, while the intensity variance and orientation contrast models show the opposite trend.

4. Discussion

4.1. Bottom-up causes versus correlates of attentional selection

The main contribution of this study is a quantitative classification of visual causes versus mere correlates of attentional selection in realistic viewing conditions. This dissociation was based on a simple notion, namely that causal bottom-up models should be increasingly more predictive of saccades that are more strongly driven by bottom-up influences. Our basic approach was to label particular saccade groups as more or less "bottom-up driven" based on two different heuristics, and then examine how the patterns of prediction accuracy change as a function of model type and "bottom-up" label. The results show that bottom-up models that are based on intensity contrast, color contrast—and to a greater extent—flicker contrast, motion contrast, and integrated saliency, show the highest prediction accuracy for "bottom-up" saccades (see Fig. 7). The reversed pattern of results—particularly low prediction accuracy for "bottom-up" saccades—was observed for other computational models, including intensity variance and orientation contrast. Assuming a trade-off between bottom-up and top-down influences (Henderson & Hollingworth, 1999; Hernandez-Peon et al., 1956; James, 1890), this result is indicative of top-down causality. For example, in the real world, there are likely to be significant correlations between certain visual features, such as local edges, and certain top-down influences, such as objects of interest that contain luminance-defined contours. Top-down guided saccades towards objects of interest may thus lead to significant yet non-causal correlations between local edge detectors and human attentional selection. But if local edges are correlated with object contours, then why not use them as a bottom-up shortcut to select behaviorally relevant information? The answer may lie in the relatively low magnitude of correlations between local edges and object contours, which may lead to unacceptably high rate of false positives. Specifically, natural scenes often contain textures that are replete with local edges, and it would be maladaptive to initiate saccades towards such edges, especially if other visual cues, such as motion contrasts, are more strongly correlated with behaviorally relevant information.

4.2. Static versus dynamic bottom-up models

An other dissociation that emerges in Fig. 7 involves static models with relatively low prediction accuracy versus dynamic models with relatively high prediction accuracy. This dynamic superiority may reflect an adaptation for detecting dynamic real world events that are critical for survival, such as the approach of predators or the fleeing of prey. Another evolutionary pressure for increased sensitivity to dynamic versus static visual cues may have been caused by biological camouflage which typically involves seamless blending into the background in terms of static features, such as shape and color (Curio, 1976),[2] are another evolutionary reason to be particularly sensitive to

[2] Nature also contains examples of dynamic camouflage, as employed by dragonflies during territorial aerial manoeuvres (Mizutani, Chahl, & Srinivasan, 2003), but these are relatively rare.

dynamic versus static visual cues. Among static bottom-up models, we found a small advantage in prediction accuracy to color contrast over intensity contrast. This result may reflect an evolutionary adaptation to detecting color contrasts, which may be particularly useful when searching for fruits embedded in foliage (Regan et al., 2001).

4.3. Interactions between bottom-up and top-down influences

The results of this study, particularly Figs. 4 and 5, corroborate an earlier report of saliency effects as a function of viewing time (Parkhurst et al., 2002), but are inconsistent with a more recent study of the same issue (Tatler et al., 2005). It is difficult to pinpoint the exact cause for these contradictory results, because several parameters are different across the relevant studies, including the stimuli, the subjects, the model type, and the metric type. Among these parameters, the model type seems to be the most likely culprit of the contradictory results. Given the variability in the pattern of results between the different static models in this study alone, it is not surprising that previous studies that utilized different static saliency models led to mixed results.

The jump cuts used in this study provide a unique opportunity to examine competitive interactions between older top-down influences and newer bottom-up influences. Immediately after a jump cut, there is likely to be a maximal deviation between top-down influences based on the pre-cut clippet and bottom-up influences based on the post-cut clippet. If older top-down influences were still active shortly after jump cuts, then the prediction accuracy of bottom-up models would have been at its lowest at that point in time. As far the new attention-priority maps are concerned, humans would be selecting targets at random with practically the same accuracy as the human-fixated baseline. Contrary to this hypothetical scenario, Figs. 4 and 5 show that for most of the bottom-up models tested, the prediction accuracy was at its highest shortly after jump cuts. This result demonstrates that there was little to no spill over of top-down influences across jump cuts.

Visual inspection of the video clips indicates that observers sometimes saccade towards faces and text shortly after jump cuts, potentially reflecting the impact of fast top-down influences. As a caveat, we noticed that faces often stand out in color contrast maps, whereas text sometimes stands out in intensity contrast maps (or motion contrast maps in the case of tickers). The extent to which preferential looking at faces or text is driven by bottom-up versus top-down influences is an open question. A related question is what do we mean exactly by "bottom-up" and "top-down"? If evolution or development equips us with dedicated face detectors, would it be justified to consider faces per se to be bottom-up influences? From a neural perspective, the answer would be yes if it could be shown that a face detector operates successfully without receiving any descending inputs (i.e., no information from upstream internal representations). In other words, the labels "bottom-up" and "top-down" cannot be separated from the underlying neural circuits. In this context, learning can be thought of as a process that progressively reshapes local neural circuits such that they become more bottom-up driven and less top-down guided.

4.4. Realism of stimuli used in studies of attentional selection

The stimulus set used in this study is substantially larger and more realistic than the collections of static images (Itti & Koch, 2000; Krieger et al., 2000; Mannan et al., 1997; Oliva et al., 2003; Parkhurst et al., 2002; Parkhurst & Niebur, 2003; Peters et al., 2005; Reinagel & Zador, 1999; Tatler et al., 2005; Torralba, 2003) and synthetic search arrays (Abrams & Christ, 2005; Folk et al., 1992; Franconeri et al., 2005; Hillstrom & Yantis, 1994; Jonides & Yantis, 1988; Theeuwes, 1994; Yantis & Egeth, 1999) that were used in previous studies for characterizing the impact of bottom-up influences on attentional selection.

The MTV-style clips used in this study are realistic or "natural" in the sense that very similar stimuli are encountered frequently by human observers in everyday life, such as when watching television or movies. Furthermore, while the real world (other than in television and film) seems to be continuous most of the time, human retinas are constantly exposed to an MTV-style version of the world due to saccadic eye movements. A striking demonstration of this phenomenon was recently shown at the ETRA conference (Wagner et al., 2006).

Nonetheless visual exploration of either continuous or MTV-style video clips does not capture the full complexity of sensory stimulation experienced in real world environments, which often involve three dimensions, a wide field of view, multi-sensory stimulation, and egomotion. The realism of laboratory stimuli could be further increased in several ways, such as by collecting or generating video clips that lack center bias. The main advantage of studying centrally-unbiased stimuli is that they would provide a better approximation of the selection challenge faced by human observers in the real world, where objects of interest could be located 360° around an observer at any given point in time. Another improvement to the realism of laboratory stimuli may be achieved by projecting video clips on a wall instead of displaying them on a computer monitor. This technique could be used to increase the experimental field of view without increasing the pixel resolution of the underlying stimuli. More expensive means to achieve the same or better increase in realism could involve head mounted displays.

4.5. Saliency modeling

The key elements that distinguish the most predictive bottom-up model used here (integrated saliency) from the available alternatives (Krieger et al., 2000; Mannan et al., 1997; Oliva et al., 2003; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999; Tatler et al., 2005; Torralba, 2003) are its neural grounding, inclusion of static and dynamic visual features, and non-linear spatial interactions.

Important elements that are poorly modeled in the current version of the integrated saliency model include differential sensitivities of foveal versus peripheral detectors, and interactions between foveal processing and scene understanding. These missing elements may act in opposite directions, so attempts to add one without the other (Itti & Koch, 2000; Parkhurst et al., 2002) may decrease rather than increase the realism of the models. For example, the uniform spatial resolution of computational saliency maps is likely to overestimate the saliency of non-fixated targets compared to biological saliency maps, which are based on a variable spatial resolution of photoreceptors and visual neurons (Connolly & Van Essen, 1984; Curcio, Sloan, Packer, Hendrickson, & Kalina, 1987). On the other hand, the lack of computational inhibition-of-return (Klein, 2000) is likely to underestimate the saliency of non-attended targets. Inhibition-of-return may in fact be a misnomer that refers to inhibitory top-down mechanisms that become active even before attention is withdrawn from the target. According to this hypothesis, fixated targets become relatively less salient as a function of fixation time due to diminishing informational gains. As a consequence, the relative saliency of peripheral stimuli increases, lowering the threshold of initiating a new saccade to the periphery. An interesting developmental implication of this hypothesis is that "sticky fixation" (Hood, Atkinson, & Braddick, 1998)—the special difficulty that infants have to disengage from fixated targets—may be attributable to a perceptual immaturity (slow information uptake) rather than an oculomotor immaturity. Adding an "inhibition-of-target" component to saliency models would be important for making them more predictive of the exact timing of saccades.

Acknowledgments

Supported by grants from NSF, NEI, NIMA, and the Zumberge Research and Innovation fund. Computation for the work described in this paper was supported by the University of Southern California Center for High Performance Computing and Communications (www.usc.edu/hpcc). We thank Rob Peters for useful discussions about metrics, Bruno Olshausen for his helpful editorial comments, and two anonymous reviewers for their stimulating comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.visres.2006.08.019.

References

Abrams, R. A., & Christ, S. E. (2005). The onset of receding motion captures attention: comment on Franconeri and Simons (2003). *Perception and Psychophysics, 67*(2), 219–223.

Atkinson, J., & Braddick, O. (2003). Neurobiological models of normal and abnormal visual development. In *The cognitive neuroscience of development* (pp. 43–72). Hove, East Sussex; New York: Psychology Press.

Bex, P. J., & 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*(6), 1096–1106.

Caspi, A., Beutter, B. R., & 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*(35), 13086–13090.

Connolly, M., & 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 Comparitive Neurology, 226*(4), 544–564.

Curcio, C. A., Sloan, K. R., Jr., Packer, O., Hendrickson, A. E., & Kalina, R. E. (1987). Distribution of cones in human and monkey retina: individual variability and radial asymmetry. *Science, 236*(4801), 579–582.

Curio, E. (1976). *The ethology of predation. Zoophysiology and ecology* (Vol. 7). Berlin, New York: Springer-Verlag (pp. x, 250 p.).

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap. Monographs on Statistics and Applied Probability* (Vol. 57). New York: Chapman & Hall (p. 436).

Fecteau, J. H., Bell, A. H., & Munoz, D. P. (2004). Neural correlates of the automatic and goal-driven biases in orienting spatial attention. *Journal of Neurophysiology, 92*(3), 1728–1737.

Findlay, J. M. (2004). Eye scanning and visual search. In F. Ferreira (Ed.), *The interface of language, vision, and action: Eye movements and the visual world*. UK: Psychology Press.

Finlay, D., & 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*(1), 29–36.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 1030–1044.

Franconeri, S. L., Hollingworth, A., & Simons, D. J. (2005). Do new objects capture attention? *Psychological Science, 16*(4), 275–281.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature, 391*(6666), 481–484.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498–504.

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

Hernandez-Peon, R., Scherrer, H., & Jouvet, M. (1956). Modification of electric activity in cochlear nucleus during attention in unanesthetized cats. *Science, 123*(3191), 331–332.

Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception and Psychophysics, 55*(4), 399–411.

Hood, B., Atkinson, J., & Braddick, O. (1998). Selection-for-action and the development of orienting and visual attention. In J. Richards (Ed.), *Cognitive neuroscience of attention: A developmental perspective*. New Jersey: Lawrence Erlbaum.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*(6), 1093–1123.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10-12), 1489–1506.

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

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception and Psychophysics, 43*(4), 346–354.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*(4), 138–147.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision, 13*(2–3), 201–214.

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature, 384*(6604), 74–77.

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

Mizutani, A., Chahl, J. S., & Srinivasan, M. V. (2003). Motion camouflage in dragonflies. *Nature, 423*(6940), 604.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. *International Conference on Image Processing, 1*, 253–256.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16*(2), 125–154.

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

Regan, B. C., Julliot, C., Simmen, B., Vienot, F., Charles-Dominique, P., & 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*(1407), 229–283.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10*(4), 341–350.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *Journal of Neuroscience, 21*(4), 1340–1350.

Sperling, G., & Weichselgartner, E. (1995). Episodic theory of the dynamics of spatial attention. *Psychological Review, 102*(3), 503–532.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research, 45*(5), 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*(4), 799–806.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A. Optics, Image Science, and Vision, 20*(7), 1407–1418.

Wagner, P., Bartl, K., Gunthner, W., Schneider, E., Brandt, T., & Ulbrich, H. (2006). A pivotable head mounted camera system that is aligned by three-dimensional eye movements. In *Proceedings of the 2006 symposium on eye tracking research & applications* (pp. 117–124). San Diego, California: ACM Press.

Wolfe, J. M., Alvarez, G. A., & Horowitz, T. S. (2000). Attention is fast but volition is slow. *Nature, 406*(6797), 691.

Yantis, S., & Egeth, H. E. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 661–676.