The Role of Memory in Guiding Attention during Natural Vision

Ran Carmi Neuroscience Program, Los Angeles, CA, USA

Laurent Itti Neuroscience Program, Los Angeles, CA, USA

What is the time frame in which perceptual memory guides attention? Current estimates range from a few hundred milliseconds to several seconds, minutes, or even days. Here we answer this question by establishing the time course of attentional selection in realistic viewing conditions. First, we transformed continuous video clips into MTV-style video clips by stringing together continuous clip segments using abrupt transitions (jump cuts). We then asked participants to visually explore either continuous or MTV-style clips, and recorded their saccades as objective behavioral indicators of attentional selections. The utilization of perceptual memory was estimated across viewing conditions and over time by quantifying the agreement between human attentional selections and predictions made by a neurally-grounded computational model. In the critical condition, jump cuts led to sharp declines in the impact of perceptual memory on attentional selection, followed by monotonic increases in memory utilization across 7 consecutive saccades and 2.5 seconds. These results demonstrate that perceptual memory traces play an important role in guiding attention across several saccades during natural vision. We propose novel hypotheses and experiments using hybrid natural-artificial stimuli to further elucidate neurocomputational mechanisms of attentional selection.

Keywords: Attention, Eye Movements, Memory, Modeling, Natural Vision, Natural Scenes

Introduction

Paying attention to the right thing at the right time underlies the ability of humans and other animals to learn, perceive, and interact with their environment. Such attentional selections are determined by interactions between memory-free and memory-dependent influences (James 1890, Hernandez-Peon et al. 1956, Henderson and Hollingworth 1999). Memory-free (bottom-up) influences are stimulus-centric factors, such as visual onsets (Jonides and Yantis 1988, Gottlieb et al. 1998), which can lead to automatic, or reactive, selection of attention targets. Memory-dependent (top-down) influences refer to behavioral goals and expectations (Yarbus 1967), which can guide attention proactively based on prior knowledge. Among potential top-down influences, there is a consensus that sensory snapshots are overwritten within a single fixation (up to a few hundred milliseconds), while semantic information can be accumulated and utilized for guiding attention across many fixations. In contrast, the utilization of perceptual memory, especially involving locations of previously fixated targets, has been debated extensively (Maljkovic and Nakayama 1996, Levin and Simons 1997, Chun and Jiang 1998, Horowitz and Wolfe 1998, Melcher and Kowler 2001, Hollingworth and Henderson 2002, Rensink 2002, Hayhoe et al. 2003, Najemnik and Geisler 2005).

Two competing theories about the accumulation and utilization of perceptual memory have emerged. According to the “world as an outside memory” (WOM) theory (O'Regan 1992, Rensink 2000), humans rely on the continuity of the world to access external information on demand, leading to conscious perceptions that are seemingly rich and continuous, in the absence of perceptual memory that persists across several fixations. More recent studies have shown that perceptual information can be accumulated over time during inspection of static scenes (Melcher 2001, Hollingworth and Henderson 2002). Nevertheless, other results have indicated that the WOM theory may still hold in the particular context of attentional selection (Horowitz and Wolfe 1998, Melcher and Kowler 2001, Najemnik and Geisler 2005). One suggestion is that attention targets are selected at random from an instantaneous pool of candidates (Horowitz and Wolfe 1998). Another study argued for a semi-random process, in which attentional selections are determined by proximity weighting of potential candidates relative to the instantaneous fixation location (Melcher and Kowler 2001). According to a third suggestion, humans select targets based on a continuously updated posterior probability map. While the proposed map relied on perfect memory, a temporal analysis of its predictive power showed that is mostly attributable to information accumulated during the last 1-2 fixations (Najemnik and Geisler 2005). The common thread to all these proposals is that location-specific memories are rarely utilized for guiding attention across more than 1-2 saccades.

In contrast, the "implicit memory" (IM) theory (Land and Furneaux 1997, Chun and Nakayama 2000) postulates that previously attended targets typically trigger perceptual memory traces that are routinely utilized across several fixations. Such memory traces may often be hidden from con-

scious introspection, leading to the common failure of human observers to explicitly report changes in scenes (Rensink 2002), even when their eye-movement patterns are clearly affected by such changes (Hollingworth and Henderson 2002). Proponents of the IM theory argue that location-specific memories are critical for facilitating target detection, especially in the context of visually-guided actions that require motor planning in spatial coordinates (Land and Lee 1994, Land and McLeod 2000, Hayhoe et al. 2003).

A general criticism of the traditional psychophysical approach for studying memory utilization is the questionable relevance of related conclusions to everyday life (Land and Furneaux 1997, Hayhoe et al. 2003). The crux of the argument is that the honorable tradition of designing highly controlled experiments often backfires by leading to oversimplifications and misrepresentations of the challenges posed by natural stimuli and tasks. In comparison, studies of attentional selection in realistic environments (Land and Lee 1994, Shinoda et al. 2001, Hayhoe et al. 2003) have so far focused on highly practiced behaviors, which may also reflect atypical interactions between perceptual memory and attention. Moreover, while such studies provide valuable accounts of eye-movement patterns in everyday life activities, they have so far been confined to descriptive analyses, which are inherently limited in their explanatory and predictive powers (Viviani 1990).

The main goal of the present study is to bridge the existing gap between real world relevance, hypothesis-driven experimentation, and predictive power. Specifically, we used computational tools to establish the time frame in which perceptual memory guides attention during visual exploration of novel dynamic scenes. The rationale for focusing on this particular behavior is that humans seem to spend a lot of time in everyday life visually exploring new people or environments, such as while watching television and films, without necessarily being engaged in highly practiced goal-oriented behaviors. Furthermore, visual exploration is the main drive for attentional selections during the first year of life starting from 3 months postnatally (Atkinson and Braddick 2003), indicating that it may be critical for learning regularities in one's environment.

To manipulate the potential impact of perceptual memory on attentional selection, we converted continuous video clips into MTV-style video clips (see Figure 1). The only criterion for choosing the original video clips was diversity maximization in an attempt to capture the stimulus complexity and heterogeneity that humans encounter in real world environments. The rationale for the MTV-style manipulation is that frequent jump cuts repeatedly undermine the utility of perceptual memory.

The WOM theory (O'Regan 1992, Rensink 2000) predicts that MTV-style jump cuts will have little to no effect on the utilization of perceptual memory, which is constantly being overwritten on a fixation-by-fixation basis even during continuous viewing. On the other hand, if perceptual memory is utilized across several fixations, as predicted by the IM theory (Land and Furneaux 1997, Chun and Nakayama 2000), then jump cuts are expected to repeatedly modulate the extent to which perceptual memory guides attention. Specifically, each jump cut should invalidate perceptual memory traces that were triggered by the preceding clip segment (clippet), followed by increasing accumulation and utilization of such traces based on the following clippet.

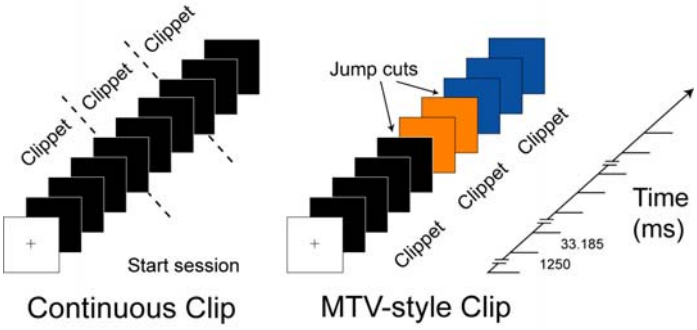

Figure 1. Schematics of continuous and MTV-style clips. Colored squares depict video frames. Different colors represent semantically unrelated clip segments (clippets).

The results show strong trans-saccadic utilization of perceptual memory across up to 7 saccades and 2.5 seconds. These findings support and quantitatively elaborate the IM theory (Land and Furneaux 1997, Chun and Nakayama 2000), but are inconsistent with the WOM theory (O'Regan 1992, Rensink 2000). We also observed a trend of delayed amnesia that was not predicted by either of these theories. To explain this trend, we propose novel testable hypotheses and related experiments that would require further integration of computational, behavioral, and neurophysiological techniques.

Methods

Participants

16 paid participants (6 women and 10 men), 23- to 32-years old, provided written informed consent, and were compensated for their time ($12/h). All participants were healthy, had normal or corrected-to-normal vision, and were naïve as to the purpose of the experiment.

Stimuli

50 video clips (30 Hz, 640x80 pixels/frame, 4.5-30 s, mean ± s.d.: 21.83 ± 8.41 s, no audio) from 12 heterogeneous sources, including indoor/outdoor, daytime/nighttime scenes shot at various locations in Los Angeles, video games, television newscasts, interviews, commercials, and sporting events. These continuous clips were randomly cut every 1-3 s (2.09 ± 0.57 s) into 523 clip segments (clippets), which were re-assembled into 50 MTV-style clips (see the left panel of Movie 1 and Movie 2 for examples). The uni-

form distribution of clippet lengths minimizes the ability to anticipate the exact timing of jump cuts.

Movie 1. Example #1 of an MTV-style clip, instantaneous eye positions of a human observer during visual exploration, and the corresponding attention-priority maps based on the saliency model.
- Left Panel: mtvclip03.
- Middle Panel: same as left panel, plus the instantaneous eye position of a human observer (MC), depicted by a small cyan square. Each time MC initiates a saccade, the square color changes to magenta, and a yellow ring is drawn around the saccade target.
- Right Panel: dynamic attention-priority maps generated by the saliency model based on mtvclip03. Each time MC initiates a saccade, a white ring is drawn around the saccade target.

Movie 2. Same as Movie 1, but for mtvclip04, participant JR.

Continuous and MTV-style clips were matched in length, and each MTV-style clip contained at most one clippet from a given continuous clip. As such, the MTV-style clips provide a unique opportunity to test conflicting predictions made by theories of memory and attention, while the continuous clips serve as an important control with a highly similar contribution of bottom-up influences and semantic factors.

The MTV-style manipulation was inspired by the cinematic practice of using jump cuts to compress time while preserving perceptual and semantic continuity (Hochberg 1986, Anderson 1996). The critical difference is that the MTV-style clips used here were deliberately designed to maximize the perceptual and semantic unrelatedness between adjacent clippets. Even the shortest clippets in our MTV-style clips (1 second long) are sufficiently long for human observers to recognize the depicted scene, and perform several saccades.

Experimental design

Participants were divided randomly into 2 groups of 8 participants each. One group inspected continuous clips, and the other group inspected MTV-style clips. All participants sat with their chin supported before 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: "try to 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 rather help the experimenters evaluate their general understanding of what they had watched. The purpose of the task was to allow participants to 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 tests with free viewing, in which observers tended to disengage from the display over time. A potential concern with this particular task is that it may bias observers to track motion in the display, which may lead to an unrealistically high agreement with a bottom-up model that includes motion cues. While this seems possible, we note that a previous study that relied on either free viewing or the task used here found no differences in the level of agreement with the saliency model (Itti 2005). Nevertheless, it would be interesting to test if and how the results would change when using different instructions, such as: "Try to understand the scenes..." or "Hit a button when you see a famous person..."

We relied on 2 separate groups of subjects (1 per viewing condition), because the point of this study was to measure how perceptual memory formed during the viewing session affected attentional selection. A single-group design, with or without randomized order of presentation, would have been vulnerable to asymmetrical priming effects (retention of scene gist and layout is likely to be more persistent during continuous versus MTV-style viewing conditions). A potential problem with our 2-group design is that differences between the groups may contaminate comparisons across viewing conditions. While this confound seems unlikely given the random assignment of subjects into groups, it remains a possibility. In any case, comparisons over time within a single viewing condition are not sensitive to inter-group differences.

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, see Figure 2). Calibration and saccade extraction procedures have been described elsewhere (Itti 2005). In short, the raw eye-position data was segmented into saccade, blink and fixation/smooth-pursuit periods. 8812 and 10187 saccades (>2°) were extracted from the raw eye-position data of the continuous and MTV-style groups, respectively. These totals exclude saccades that either started or ended outside of the display bounds (65 and 34 saccades from the continuous and MTV-style groups, respectively). The saccade amplitude criterion (>2°) minimizes the rate of false positives in saccade extraction, and also focuses the analysis on actual attention shifts

rather than mere gaze adjustments during inspection of the same object.

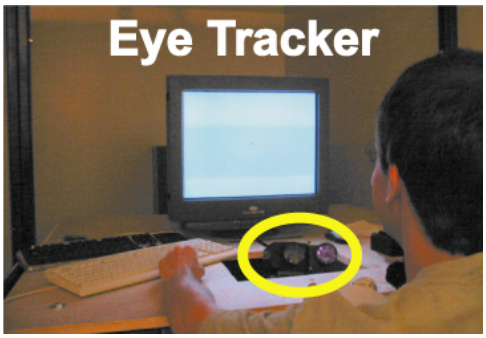

Figure 2. A human observer during a calibration session.

Attention-priority maps

Attention-priority maps were computed based on a saliency model (see Figure 3). Multi-scale feature pyramids were generated at the input frame rate (30Hz), and converted through a series of computations into proto-saliency maps (1 per input frame). Each proto-saliency map was fed into a two-dimensional layer of leaky integrator neurons that accounted for sensory persistence and provided temporal smoothing at 10 kHz.

This saliency model is used here as a probe for the availability of strictly bottom-up influences on attentional selection. It should be noted that an earlier version of this model was published as part of a larger framework for simulating attention shifts (Itti and Koch 2000), which included winner-take-all and inhibition-of-return. These model components were excluded here, because they do not reflect strictly bottom-up influences, and hence would have complicated the interpretation of the results. Moreover, it is unclear whether the still image-oriented inhibition-of-return that was used in previous versions of the saliency model is justified in the context of dynamic scenes. All simulations were run on a Linux-based computer cluster containing 1,830 dual-processor nodes, which is capable of performing 10.75 trillion calculations per second (http://www.usc.edu/hpcc). We used only a fraction of the available processing power in this study (90 nodes), which allowed us to compress months of processing time into a few hours. To obviate the need to store or explicitly downsample attention-priority maps, we sampled instantaneous maps as they were being generated based on the procedure described below (the sampling rate was bound by the eye-tracker frequency, or 240 Hz).

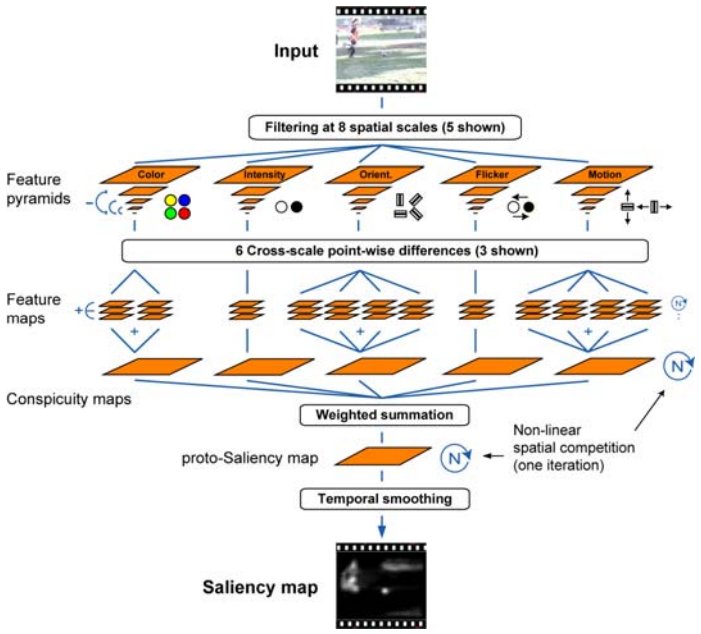

Figure 3. Schematic of the saliency model, which takes a video stream as input, detects stimuli that stand out in either space or time, and outputs a corresponding video stream of attention-priority maps. Saliency computations are based on a series of non-linear integrations of contrast detections across several scales and feature dimensions, including color, intensity, orientation, flicker and motion. These computations are motivated by neurophysiological (Frost and Nakayama 1983, Sillito et al. 1995, Gottlieb et al. 1998), psychophysical (Treisman and Gelade 1980, Polat and Sagi 1994), and theoretical (Koch and Ullman 1985) studies. The mathematical formulation of this model is described elsewhere (Itti and Koch 2000).

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 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, which affect individual selections in natural conditions, are mostly accrued during the preceding fixation (Parkhurst et al. 2002, Caspi et al. 2004). We did 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). 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.

We compensated for potential inaccuracies in human saccade targeting and the eye-tracking apparatus by sam-

pling the maximal local value in an aperture around each saccade target. The size of the aperture (r=3.15°) was chosen rather arbitrarily – we did not optimize the performance of the model by trying different aperture sizes. This issue seems peripheral for this study because sampling inaccuracies are not expected to differ consistently across the conditions tested here. Furthermore, the baseline measurements described below correct for sampling inaccuracies on a saccade-by-saccade basis.

For each human saccade, we simulated a concurrent random selection based on a spatially uniform distribution of potential targets. To appreciate the importance of using such an adaptive baseline, it is helpful to consider a hypothetical model that achieves optimal hit rates by generating uniform attention-priority maps. In the absence of a baseline, such a non-informative model will be deemed maximally predictive because human saccades will always target the maximal value in the attention-priority map. With the random baseline in place, uniform attention-priority maps would have no predictive power, because human and random saccades will achieve equal hit rates. Generally speaking, the random baseline ensures that both hit rate and target selectivity are taken into account when measuring model performance.

Several authors proposed that baseline samples should be drawn from a distribution of human selected locations rather than a uniform distribution (Parkhurst and Niebur 2003, Tatler et al. 2005). This proposal is motivated by the fact that observers tend to fixate centrally, which is further compounded by the central fixation cross that commonly precedes stimulus onsets. The rationale for using a non-uniform baseline is based on the assumption that the human tendency to fixate centrally arises from motor biases, such as a tendency to make short saccades or look towards the center of the display. If this assumption is valid, then differences between the saliency at human saccade targets versus random samples from a uniform distribution may not reflect bona-fide saliency effects. Instead, they may be entirely attributable to methodological limitations (the central fixation cross) and motor biases. Alternatively, the tendency of human observers to fixate centrally may be attributable to the centrally-biased distribution of saliency and objects of interests in photography-based stimuli (Tatler et al. 2005). If this assumption is valid, then baseline sampling from a distribution of human saccade targets would underestimate the actual magnitude of saliency effects.

The debate over which baseline should be used to measure saliency effects is peripheral to this study, as exemplified by the results using the ASH metric, which does not depend on any baseline-related assumptions. We chose to address this issue in this paper because of its important implications for developing generic metrics that quantify the agreement between human attentional selection and attention-priority maps (Mannan et al. 1997, Reinagel and Zador 1999, Krieger et al. 2000, Parkhurst et al. 2002, Oliva et al. 2003, Parkhurst and Niebur 2003, Torralba 2003, Itti 2005, Peters et al. 2005, Tatler et al. 2005).

The ASH metric

The average saliency at human saccade targets (ASH) metric is defined as:

$$ASH = (1/N) \times \sum_{i=1}^{N} S_i \quad (1)$$

where N is the number of human saccades, and S_i is the saliency at the human saccade target prior to saccade initiation (at the end of the preceding fixation).

The ASH metric computes hit rate while ignoring target selectivity. As such, it is limited in its applicability to comparisons (across viewing conditions or over time), which involve equivalent attention-priority maps in terms of the distribution and density of values.

The DOH metric

The difference of histograms (DOH) metric 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:

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

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 reflects the assumption that deviations from the baseline in mode-defined "high-priority" bins are more informative (should be given stronger weights) than similar deviations in "low-priority" bins. The rationale for this assumption is that model-defined priority should be indicative of functional importance, for example: if the tested model assigns higher value to location X compared to location Y then humans should be more likely to select location X rather than location Y. 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 random saccade histogram:

$$DOH_I = (W_n - W_1) \times (1-p) = 0.8633 \quad (3)$$

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 map values, respectively. However, even if assuming an ideal model that always generates a single value at saccade targets, and 0 elsewhere (see Figure 4), a certain fraction of random saccades would land on the maximal map value by chance, with probability:

$$p = N_a / N_m = 0.0408 \quad (4)$$

where $N_a = 49$ is the number of pixels in an aperture around the saccade target (r=3.15°, approximated 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.

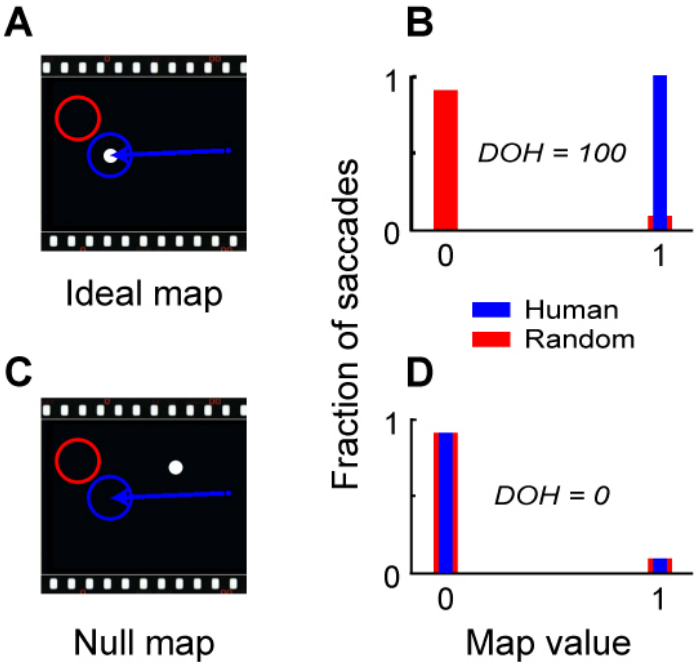

Figure 4. Hypothetical scenarios of predicting attentional selection.
(A) An ideal attention-priority map prior to saccade initiation, which contains a single positive value at the saccade target, and zero elsewhere. Blue markers depict the eye position of a human observer prior to saccade initiation (filled circle), the saccade trajectory (arrow), and the saccade target (ring). The red ring depicts a concurrent random target.
(B) Saccade distribution as a function of map values at saccade targets. 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. In this particular example, only a single random location is set to a non-zero value.
(D) same as B, but for the null scenario. Human and random saccades are just as likely to land on positive values, leading to identical histograms and no rightward shift of the human histogram relative to the random histogram.

In the ideal scenario, the human histogram (saccade distribution as a function of map value) 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. In comparison, the null scenario occurs when a model is unpredictive of attentional selection, in which case human and random saccades would be just as likely to hit high-priority targets, leading to a complete overlap between human and random histograms. To summarize, the expected range of DOH values expressed in percentages 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 (Mannan et al. 1997, Reinagel and Zador 1999, Krieger et al. 2000, Parkhurst et al. 2002, Oliva et al. 2003, Parkhurst and Niebur 2003, Torralba 2003, Itti 2005, Tatler et al. 2005), 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 (saliency at human fixated vs. random locations), but KL-divergence suffers from the following relative disadvantages: non-linearity (metric values should not be compared as interval variables across different groups of saccades or models), infinite upper-bound, no priority-based weighting, and bi-directionality (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). Similar to KL-divergence, it has no priority-based weighting. Moreover, the ROC metric provides a misleadingly high upper-bound due to its implicit assumption of linear discriminability.

Given that the inter-observer agreement in attentional selection is imperfect, 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 normalizing factor (DOH upper bound). Hence, the DOH values reported here represent conservative estimates of bottom-up impact. More realistic estimates of the practical upper bound could be based on inter-observer agreement, although such estimates would necessarily involve additional assumptions. For example, there is no consensus on how inter-observer agreement should be quantified in dynamic viewing conditions. Our conclusions are independent of the upper bound because they only depend on estimating differences in bottom-up impact across conditions and over time. We included the upper bound in the metric definition because it makes its values intuitively more meaningful as conservative estimates for the relative impact of bottom-up versus top-down influences on attentional selection.

Results

The conclusions of this paper are based on the assumption that the more humans act proactively, the less likely they are to act reactively, and vice versa. In other words, we assume that there is a functional trade-off between bottom-up and top-down influences (James 1890, Hernandez-Peon et al. 1956, Henderson and Hollingworth 1999). It should be noted, however, that bottom-up influences may sometimes overlap rather than compete with top-down influences. For example, motion contrast signals generated by an approaching predator may attract one's attention in a bottom-up manner, which may then help to recognize the predator, and trigger top-down attention guidance towards the exact same location in the visual field. This example notwithstanding, the fact that top-down mechanisms evolved at all suggests that they often provide a unique contribution compared to the evolutionary more conserved bottom-up influences (Land 1999).

The trade-off assumption allows us to make quantitative inferences about the impact of memory on attention guidance, even though we do not have at our disposal computational models of memory-driven attention in realistic viewing conditions. Specifically, we infer modulations in the impact of top-down influences by quantifying modulations in the impact of bottom-up influences. To this end, we set out to establish the availability of potential bottom-up influences using a computational saliency model (Figure 2) that has no memory other than sensory persistence (Itti and Koch 2000, Itti 2005). The key design features that distinguish this particular saliency model from the available alternatives (Mannan et al. 1997, Reinagel and Zador 1999, Krieger et al. 2000, Oliva et al. 2003, Parkhurst and Niebur 2003, Torralba 2003, Tatler et al. 2005) are its detection of dynamic signals, spatial interactions between local detectors, and neural grounding. While the results presented here are not tied in any way to the implementation details of a particular saliency model, they do depend on accurate estimation of bottom-up influences. In this context, the detection of dynamic signals is particularly critical, because previous studies have shown that stimulus changes are the strongest bottom-up influences on attentional selection in dynamic environments (Jonides and Yantis 1988, Gottlieb et al. 1998, Findlay and Walker 1999, Itti 2005).

By highlighting conspicuous locations in the display, saliency maps reflect the potential availability of bottom-up influences, but not their actual impact on attentional selection. To measure the bottom-up impact, we developed two metrics that can quantify the agreement between human attentional selection and attention-priority maps. The first metric - "Average Saliency at Human targets" (ASH) - simply computes the average saliency at human saccade targets. The utility of ASH is limited because it is highly sensitive to the baseline distribution of saliency values. For example, conditions that involve different models and/or set of stimuli may lead to very different distributions of saliency values and incomparable ASH values. The second metric – "Difference of Histograms" (DOH) – is a more powerful metric whose values are comparable across conditions that may have different baseline distributions of saliency values. In the following sections, we focus on data analyses using the relatively more generic DOH metric. The results of corresponding analyses based on the ASH metric are described in the text and shown in Figure 10.

Average bottom-up impact on attentional selection

Figure 5 and Figure 6 show two pairs of saccades that straddle jump cuts. Eye position markers and the output of the saliency model depict the basic measurements that we took in order to quantify the impact of bottom-up influences on attentional selection (see Methods).

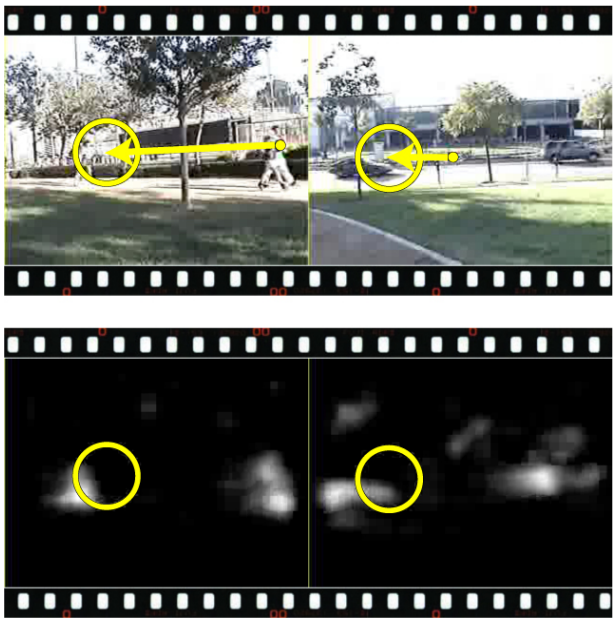

Figure 5. Saccades that straddle an MTV-style jump cut (mtvclip03, participant MC). The temporal offset between the saccade initiation times was 1261.0 ms, during which MC smoothly pursued the people jogging to the right, followed by a short saccade (<2°) towards the bottom-right (the camera was stationary in both clippets, see Movie 1). Superimposed markers (yellow) depict the eye-position prior to saccade initiation (filled circle), saccade trajectory (arrow), and saccade target (ring). Upper filmstrips depict the instantaneous input frames at the time of saccade initiation (the last eye tracker sample from the preceding fixation). Lower filmstrips depict the corresponding attention-priority maps based on the saliency model (see Figure 2).

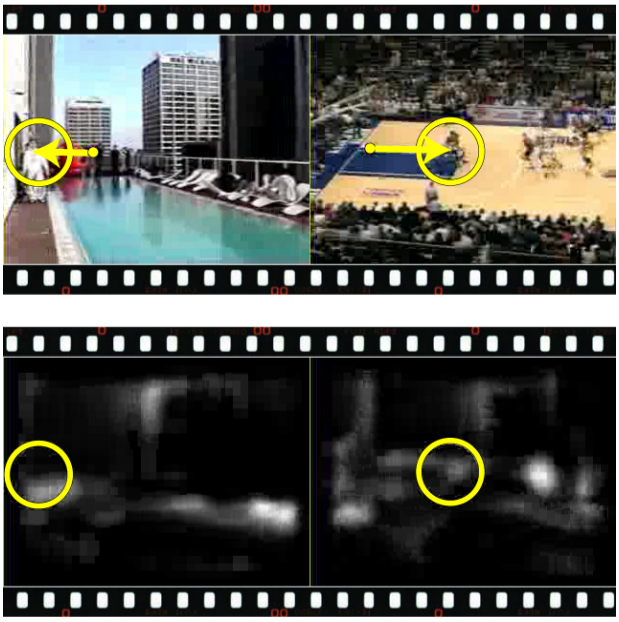

Figure 6. Same as Figure 5, but for mtvclip04, participant JR. The temporal offset between the saccade initiation times was 431.4 ms, during which JR smoothly pursued the fixated person who was moving to the right due to a leftward camera pan (see Movie 2).

Figure 7A shows the distribution of saccades as a function of normalized saliency in the continuous viewing condition.

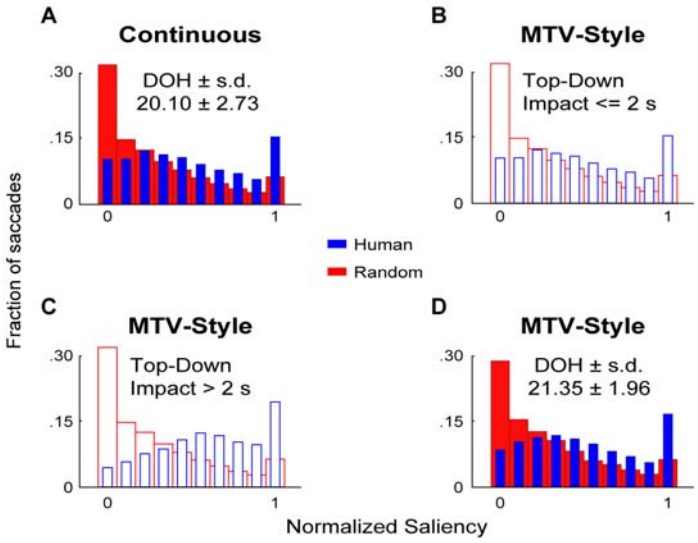

Figure 7. Saliency-based saccade distributions, and average bottom-up impact on attentional selection.
(A) The distribution of saccades, pooled over all participants and clips from the continuous group, as a function of normalized saliency at saccade target. Red and blue vertical bars represent the random and human saccade histograms, respectively. Numbers above histograms depict the bottom-up impact based on the saliency model (DOH ± the inter-participant s.d.).
(B) Hypothetical saccade histograms for the MTV-style group, assuming that the impact of top-down influences on attentional selection fades within 2 s during inspection of continuous clips.
(C) Same as B, but assuming that the impact of top-down influences on attentional selection persists for more than 2 s during inspection of continuous clips.
(D) Same as A, but for the MTV-style group.

It demonstrates that approximately 30% of the random saccades targeted the lowest possible saliency (0-10% of the max), while 15% of the human saccades targeted the highest possible saliency (90-100% of the max), with the remaining saccades targeting intermediate saliency values. The random saccade histogram reflects the actual distribution of saliency values, while the human saccade histogram shows the distribution of saliency values sampled by human saccades. The DOH metric measures the human tendency to visit salient locations by quantifying the rightward shift of the human saccade histogram relative to the random saccade histogram.

Assuming equivalent random histograms across viewing conditions, three scenarios are possible for the human saccade histogram in the MTV-style viewing condition:

1) The human histogram might be shifted to the left compared to its continuous counterpart. This would indicate that human observers were less likely to select targets based on bottom-up influences in the MTV-style condition, even though the potential availability of valid top-down influences was limited compared to the continuous condition. One possible interpretation would be that the rapid succession of novel scenes in the MTV-style condition induced observers to select targets more randomly (i.e., the trade-off assumption is invalid). Such a result would not be informative for distinguishing between the memory utilization theories addressed here.

2) The MTV-style manipulation might have no effect on the relative shift of the human saccade histogram (see Figure 7B). This would indicate that top-down influences during continuous viewing are rarely utilized beyond 2 s (the average length of persistent context in MTV-style clips), leading to the same overall bottom-up impact in both viewing conditions. This scenario is consistent with either the WOM theory or a transient version of the IM theory, in which perceptual memory is utilized across fixations, but rarely beyond 2 s.

3) The MTV-style human histogram might be shifted to the right compared to its continuous counterpart (see Figure 7C). This result would indicate that top-down influences that guided attention beyond 2 s during continuous viewing were replaced by bottom-up influences during MTV-style viewing. This scenario is inconsistent with the WOM theory, and instead supports a more persistent version of the IM theory.

Figure 7D shows the actual saccade histograms in the MTV-style condition. The random saccade histogram mirrors its continuous counterpart, while the human saccade histogram is slightly shifted to the right compared to the

continuous condition. However, these small differences across viewing conditions are not statistically significant ($t[14]=1.05$, $p>0.20$, based on the inter-observer s.d.'s shown in Figure 8). We obtained similar results using the simpler ASH metric, which does not depend on the random baseline: the average saliency at human saccade targets was 49.15 ± 2.16 in the continuous condition and 50.89 ± 1.79 in the MTV-style condition ($t[14]=0.62$, $p>0.20$).

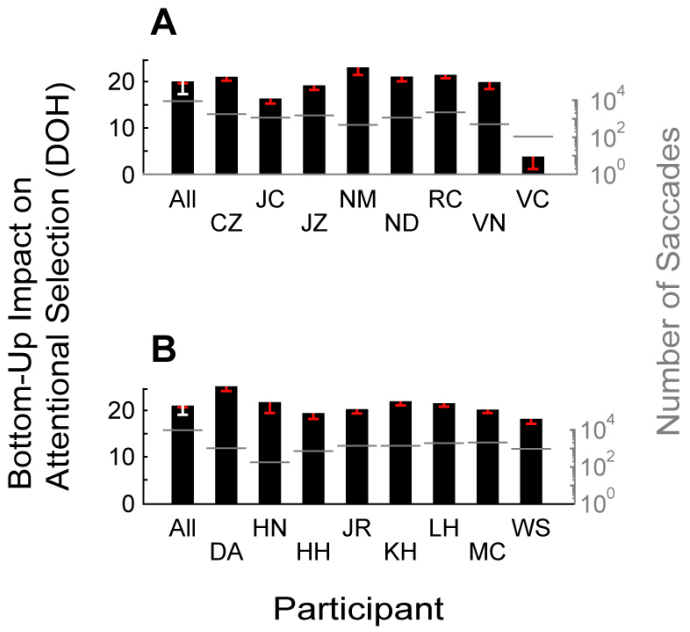

Figure 8. Average bottom-up impact on attentional selection per participant.
(A) Continuous group. Left ordinate depicts the DOH values (black bars). Red error bars depict the bootstrap s.d. based on 1000 subsamples (Efron and Tibshirani 1993). White error bars depict the weighted inter-participant s.d. based on the number of saccades per participant, shown by the right ordinate (gray horizontal bars, log scale).
(B) Same as A, but for the MTV-style group.

These results indicate that top-down influences were not utilized beyond 2 s in either of the viewing conditions. As such, they are consistent with either the WOM theory (O'Regan 1992, Rensink 2000), which claims that perceptual memory is not utilized across more than 1-2 fixations, or a transient version of the IM theory (Land and Furneaux 1997, Chun and Nakayama 2000), in which perceptual memory is utilized across several fixations, but rarely beyond 2 s. It is also possible that perceptual memory is utilized beyond 2 s, but the memory utilization probe is too coarse to detect existing differences between the two viewing conditions.

To address these alternative interpretations in the context of a positive effect, we analyzed the bottom-up impact on attentional selection at a finer time scale than 2 s, as described in the next section.

Time course of bottom-up impact on attentional selection

To reveal memory effects over time, it is informative to align this temporal analysis to anchor points in which memory is invalid, which are followed by potential memory accumulation and utilization. Theoretically, both clip onsets and jump cuts may provide such anchor points, but in practice jump cuts are superior for several reasons:

(1) Prior to clip onsets, observers fixate the center of a blank display. In contrast, jump cuts occur while observers are actively engaged in visual exploration. As a result, jump cuts provide a unique opportunity to examine the time course of competitive interactions between old and new attention-priority maps. The WOM theory predicts that old representations would be overwritten within 1-2 fixations, while a strong version of the IM theory predicts that memory effects due to the preceding clippet would persist across jump cuts. Alternatively, memory utilization may be contingent on the availability of persistent context. According to this weaker version of the IM theory, jump cuts would lead to rapid amnesia, followed by increasing impact of memory over time as persistent context builds up within a clippet.

(2) The deliberate initiation of clip onsets by observers may introduce unknown anticipatory artifacts. In comparison, jump cuts occur irregularly without soliciting input from the observer.

(3) The initial central fixation may conspire with other factors and produce an artifactual peak in bottom-up impact following clip onsets (see the "Prediction of single saccades" subsection in the Methods for an in-depth description of this issue). This potential artifact is avoided by analyzing the time course of bottom-up impact following jump cuts, which are not preceded by a predetermined fixation location (central or otherwise).

(4) Each participant is exposed to approximately 10 times more jump cuts than clip onsets, leading to the initiation of a correspondingly larger number of saccades following jump cuts vs. clip onsets. This order of magnitude difference in the number of saccades greatly improves the signal-to-noise ratio and statistical confidence of the results.

As described below, competing theories of memory utilization predict different patterns of memory utilization following jump cuts (see Figure 9A):

(1) Scenario #1 (black plot): The WOM theory (O'Regan 1992, Rensink 2000) holds that perceptual memory is being overwritten every 1-2 fixations even during continuous viewing. If this is true, jump cuts are expected to have little to no effect on the balance between bottom-up and top-down influences.

(2) Scenario #2 (blue plot): the attention system relies heavily on perceptual memory, but this reliance is contingent on the availability of persistent context. This version of the IM theory (Land and Furneaux 1997, Chun and Nakayama 2000) predicts that internal representations elicited by clippet X (before the jump cut) will be swiftly replaced by

novel internal representations triggered by visual inputs from clippet X+1 (after the jump cut). In this case, the impact of bottom-up influences is expected to peak early on, because they are faster acting than top-down influences (Wolfe et al. 2000, Henderson 2003). After the scene gist and layout are recognized, top-down influences are expected to kick in and gradually replace bottom-up influences.

(3) Scenario #3 (orange plot): internal representations based on clippet X interfere with the accumulation and utilization of novel representations based on clippet X+1. This version of the IM theory predicts that bottom-up impact will drop immediately after jump cuts, because at that point in time observers will be selecting attention targets based on irrelevant attention-priority maps shaped by clippet X (before the jump cut). Assuming that reliance on such maps will be transient, the same temporal pattern to the one described in scenario #2 is expected to emerge after a short delay.

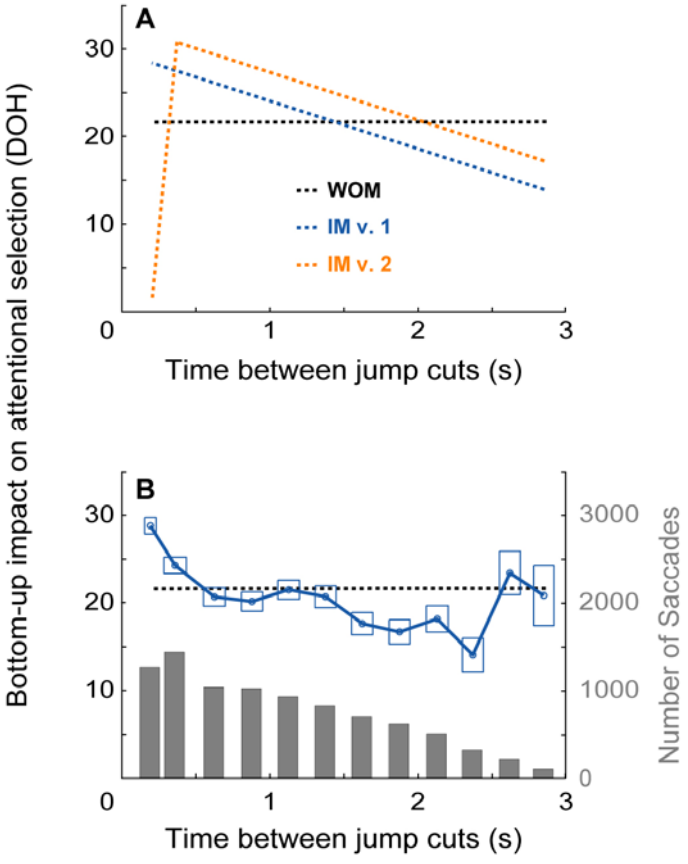

Figure 9. Time course of bottom-up impact on attentional selection.

(**A**) Hypothetical scenarios: 1. memory plays no role in guiding attention across fixations (WOM, black plot); 2. memory plays an important role in guiding attention across fixations, but only when semantically persistent context is available (IM v. 1, blue plot); 3. same as 2., but memory persists across jump cuts before it is overwritten (IM v. 2, orange plot).

(**B**) The actual bottom-up impact as a function of time between jump cuts (blue circles). Saccades were pooled over all participants and clippets, starting from the first jump cut after clip onset onwards, and binned into consecutive 250 ms intervals. The horizontal displacement of each data point was determined by calculating the average delay in saccade initiation relative to the preceding jump cut. The vertical sides of error boxes depict the bootstrap s.d. of DOH values based on 1000 subsamples (Efron and Tibshirani 1993). The horizontal sides of error boxes depict the s.d. of the average delay in saccade initiation relative to the preceding jump cut. Black bars depict the available number of saccades in each temporal bin.

Figure 9B shows the actual time course analysis as a function of time between MTV-style jump cuts. We measured bottom-up impact as a function of time by grouping saccades into consecutive 250 ms intervals between adjacent jump cuts. To account for the dead time prior to saccade initiation (Caspi et al. 2004), during which visual inputs from clippet X+1 could not possibly affect saccade targeting, the first data point in this analysis only includes saccades that were initiated at least 80 ms after jump cuts (1264/1537 saccades in the first temporal bin). To confirm that the results are not strongly affected by the dead time parameter, we performed additional analyses using other values (0,150 ms), which led to non-significant differences. Bottom-up impact on attentional selection peaked during the initial 250 ms after jump cuts, as depicted by the first data point (DOH = 28.81 ± 0.93). This result demonstrates that the visual system adapts rapidly to changing conditions, and is inconsistent with the notion of memory utilization across jump cuts (scenario #3). The following data points show monotonic decreases in bottom-up impact for up to 2.5 s after jump cuts, and thus rule out scenario #1 by indicating corresponding increases in the impact of competing top-down influences. This inference is further supported by the fact that observers tended to make more saccades faster in the first 0.5 s compared to the second 0.5 second following jump cuts.

To confirm that these results are not artifacts of the DOH metric or the temporal binning procedure, we performed additional analyses of bottom-up impact following jump cuts based on the simpler ASH metric (see Figure 10A) and saccade index (see Figure 10B). Regardless of the type of temporal analysis that was performed, a consistent decrease in bottom-up impact following jump cuts was evident across 7 consecutive saccades for up to 2.5 s.

Interestingly, Figures 9 and 10 also show late increases in bottom-up impact occurring 2.5 s after jump cuts (this trend is only significant in Figures 9B and 10B, which may be attributable to the finer temporal resolution of these figures compared to Figure 10A). The late increases in bottom-up impact cannot be explained by either of the memory utilization theories addressed here. Possible interpretations of this surprising result are discussed below in the section: "Attention and scene understanding".

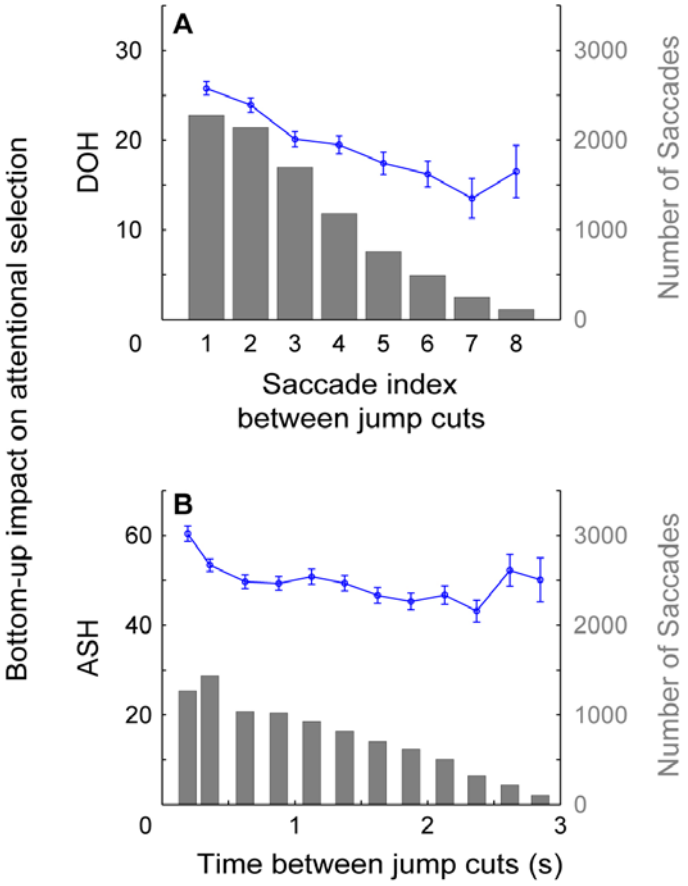

Figure 10. Time course of bottom-up impact on attentional selection.

(**A**) Similar to Figure 9B, but based on saccade index between jump cuts. This analysis demonstrates that the trend of weakening bottom-up impact over time is not an artifact of the temporal binning procedure.

(**B**) Similar to Figure 9B, but based on the ASH metric, which reflects the average saliency at human saccade targets. Error bars depict the standard error based on the number of saccades. This analysis demonstrates that the trend of weakening bottom-up impact over time is independent of the random saccade baseline that is used to compute DOH values.

Discussion

The results of this study support and elaborate the IM theory (Land and Furneaux 1997, Chun and Nakayama 2000) by establishing quantitatively the time frame in which perceptual memory guides attention during natural vision. Specifically, perceptual memory was shown to be utilized across 7 consecutive saccades and 2.5 s, even in the absence of visually-guided actions (see Figures 9 and 10). The peak in bottom-up impact that occurred immediately after jump cuts demonstrates that memory utilization is strongly contingent on the availability of semantically persistent context. Taken together, these results are inconsistent with the WOM theory (O'Regan 1992, Rensink 2000), other than in extreme circumstances, such as immediately after jump cuts, when it would be maladaptive to rely on perceptual memory for guiding attention. Figure 9 also shows evidence of delayed perceptual amnesia occurring 2.5 s after exposure to novel scenes. This surprising trend may reflect novelty or anticipatory effects, as discussed below in the section: "Attention and scene understanding".

Memory and attention

The existing controversy about the time frame in which perceptual memory guides attention may be attributable in large part to the difficulty of generalizing conclusions from studies performed in highly artificial laboratory conditions. For example, the challenges faced by the attention system while searching for small sinusoidal gratings embedded in static noise backgrounds (Najemnik and Geisler 2005) may be quite different from challenges encountered during a visual search and discrimination task involving sparse arrays of simple shapes (Maljkovic and Nakayama 1994). Alternatively, the plethora of estimates may reflect a real world flexibility of the attention system, which can automatically make pragmatic choices between relying on vision versus memory, depending on which source of information is more likely to improve performance (Oliva et al. 2004). In the following sections, we discuss two qualifications for this intuitively appealing proposal:

(1) Contrary to many laboratory conditions, real world environments are typically too complex and demanding to allow for exclusive reliance on either vision or memory to select particular targets. If a savannah monkey is being chased by a lion, it better run for its life while simultaneously examining the path ahead, keeping track of the lion, and looking for alternative escape routes. In such dynamic circumstances that involve several different agents, obstacles, and distractors, as well as a large field of view, survival depends on efficient allocation of limited visual and mental resources. In this example, the chased monkey would likely benefit from retaining accurate internal representations of pertinent information, such as the lion's location, speed, and direction, while ignoring irrelevant information, such as the lion's color and texture. In other circumstances, such as while searching for fruits embedded in foliage, the relative importance of colors and textures may increase compared to motion signals, which may be irrelevant (leaves blowing in the wind). The important point here is not the type of perceptual information that may be retained in different circumstances, but rather the complexity of real world challenges, which often necessitate the involvement of both vision (or other forms of sensation) and memory.

(2) Vision and memory are not interchangeable sources of information. For example: while watching players taking jump shots in a basketball game, our participants sometimes made saccades towards the hoop, even before the ball left the player's hands (i.e., before the ball's trajectory could have been analyzed based on its visual motion). It appears that such attentional selections depended on simultaneous

integration of several bottom-up and top-down influences, including the movement of the player, prior knowledge of what typically happens to balls when players take jump shots, and the exact location of the hoop. The conclusion that vision and memory can substitute for each other depending on their instantaneous utility may only apply to artificial laboratory stimuli that often undermine the utility of prior world knowledge.

Natural versus artificial approaches to studying vision

In addition to establishing the time frame in which perceptual memory guides attention, the more general contribution of this study is in the context of the growing debate about the usefulness of natural versus artificial stimuli and tasks to study biological vision (Felsen and Dan 2005, Rust and Movshon 2005). This debate has so far focused on low-level vision and the response properties of single neurons, and here we re-examine it in the context of high-level vision.

Proponents of the artificial approach argue that it is the best way to perform hypothesis-driven experiments, and contend that it is sufficient for characterizing neural mechanisms employed in everyday life. Furthermore, they criticize the natural approach as being too difficult and also unhelpful for generating novel hypotheses. In contrast, proponents of the natural approach point to the slow progress in understanding neural computation at the single neuron level and even more so at the system level. This state of affairs may be attributable to several factors:

1) The inherent complexity of biological brains.

2) The difficult technical challenges of collecting network-wide neural data with sufficient spatial and temporal resolutions.

3) The relevance critique: highly artificial stimuli and tasks may lead to results that are unrepresentative of how brains operate in the real world.

As for the first two factors, there is little one could do other than develop new technologies, and perhaps also shift more resources to studying simpler organisms, such as invertebrates, insects, and rodents. For those who prefer to focus on humans, the methodology presented in this paper shows that natural stimuli and tasks can be used to do hypothesis-driven research. The "Attention and scene understanding" section below also discusses several novel hypotheses that arise from this study.

The relevance question may also be raised in the context of this study, given that jump cuts are highly disruptive, and may induce unnatural visual behavior. The general answer to this criticism is that one should be willing to sacrifice some realism to gain explanatory and predictive powers. The key question is whether the sacrifice undermines the real world relevance of the study, which seems unlikely in this case for several reasons:

1) Jump cuts are ubiquitous in motion pictures, even though people are often not aware of their occurrence (Hochberg 1986, Anderson 1996). 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), human perception does not appear to be adversely affected by jump cuts. On the contrary, many people (especially younger ones...) seem to be particularly attracted to MTV-style stimuli, perhaps because of the higher information content and associated excitement compared to continuous stimuli.

2) Saccades make continuous real world stimuli appear MTV-like on the retinas of human observers. This striking phenomenon was recently demonstrated by movie clips recorded using a head-mounted camera that was moved in real time according to three-dimensional eye movements (Wagner et al. 2006).

The take-home message is that hypothesis-driven studies of high-level vision can and should rely on much more realistic stimuli than dots, bars, gratings, and plaids, the bread and butter of traditional psychophysics.

Neural Implications

The MTV-style manipulation provides a controlled stimulus-based technique for repeatedly inducing amnesia, followed by cumulative reliance on top-down influences. This technique could be useful for determining the functional connectivity between brain areas that control attentional selection. Specifically, recurring surges in activation after jump cuts would highlight brain areas that are mainly involved with bottom-up processing, while increased activation over time between jump cuts would highlight top-down areas. It remains to be seen whether the currently available neuroimaging technology is powerful enough for this purpose, but recent advances in fMRI (Hasson et al. 2004) and EEG (Michel et al. 2004) seem promising. Similarly, the repeated disruption of top-down signals triggered by jump cuts might also be useful for characterizing the response properties of individual neurons. By dissociating bottom-up from top-down input sources, MTV-style stimuli could help remove some of the confounds that plague traditional approaches for characterizing neuronal receptive fields (Lehky and Sejnowski 1988, Olshausen and Field 2005).

Attention and scene understanding

Variants of the MTV-style manipulation could also be instrumental for studying interactions between attention and scene understanding. It is often assumed that attention plays a minor role in scene understanding because people can recognize the gist of static scenes very rapidly without making any attention shifts (Henderson 2003). However, accurate perception of dynamic scenes, which pose unique and more complex perceptual challenges compared to static scenes, may require well-coordinated attentional selections. Suggestive evidence for this proposal was provided by a study of people diagnosed with autism, which demon-

strated strong correlations between anomalous attentional selections and difficulties in comprehending social interactions (Klin et al. 2002).

In this context, the late increases in bottom-up impact (see Figure 9B) may provide fertile ground for future research. First, one should consider the possibility that this surprising trend may simply reflect an artifact due to insufficient data. Starting from the first second onwards, later temporal bins in Figure 9B are based on relatively fewer clippets and contain fewer saccades than earlier temporal bins. For example: the last two temporal bins in Figure 9B contain 219 and 104 saccades, compared to 1264 and 1436 saccades in the first two temporal bins. Consequently, the reliability of DOH values decreases over time, as reflected by the relatively larger standard errors in later temporal bins. While it would be important to replicate this result using more data, the >300 saccades in the last two temporal bins seem to provide sufficient statistical power when using the bootstrapping technique. The rest of this section describes two alternative explanations for this result:

1) One hypothesis is that participants learned to anticipate the occurrence of jump cuts over the course of the experiment by estimating some sort of a hazard function. Indeed, it has been shown that humans can exploit knowledge of elapsed time to anticipate sensory events (Eagleman et al. 2005). Such anticipations may have prompted observers to shift to “bottom-up mode” when they perceived a high likelihood of an impending jump cut, thus minimizing the frequency of anomalous memory-driven selections.

2) Alternatively, the late increases in bottom-up impact may reflect novelty effects arising from the typical rate of change in natural stimuli. According to this hypothesis, observers shifted to “bottom-up mode” because of novel bottom-up influences that are likely to indicate important new events or changes to previously attended targets.

Theoretically, both novelty and anticipatory effects may conspire in biasing observers to shift to “bottom-up mode”. To test the relative contributions of novelty versus anticipatory effects, future studies could manipulate the rate of stimulus changes, such as by extending the range of clippet lengths from 1-3 s to 3-5 s. If the late increases in bottom-up impact would not be affected by this manipulation, then the novelty hypothesis would be supported. In contrast, the anticipation hypothesis predicts that the bottom-up increases reported here would be delayed further when clippets are longer (e.g., occur 4.5 s instead of 2.5 s following jump cuts). Further testing of these hypotheses could be done by shortening clippets (e.g., to 1-2 s). In this case, the anticipation hypothesis predicts that the late increases in bottom-up impact would still occur shortly before jump cuts (i.e., sooner than reported here). On the other hand, the novelty hypothesis predicts that shorter clippets might not allow for enough stimulus changes to accumulate, thus eliminating the late increases in bottom-up impact. A potential complication of shortening clippets is that the behavior of participants may become idiosyncratic if the length of persistent context becomes too short for allowing natural visual exploration.

Other experiments could manipulate the frame rate within clippets while preserving the same distribution of clippet lengths across conditions. According to the anticipation hypothesis, both static and dynamic scenes should lead to the same anticipatory effect, because it is the impending jump cut rather than the stimulus content, which prompts observers to shift to "bottom-up mode". However, if the late increases in bottom-up impact reflect novelty effects, then such increases are only expected to occur during inspection of MTV-style clips composed of dynamic scenes. The rationale behind this prediction is that static scenes contain a constant baseline contribution of bottom-up influences, and thus lack the stream of novelty that characterizes dynamic scenes.

The irregular timing of jump cuts combined with the swiftness of human perception (Henderson 2003) reduce the utility of anticipatory shifts to "bottom-up mode". It seems doubtful that the benefits of such anticipations would be large enough to offset the potential costs, including the need to continuously estimate the likelihood of an impending jump cut and the suboptimal selection of targets before the jump cut. We thus propose that the following chain of events accounts for the time course of attentional selection revealed by this study (see Figure 9B):

(1) Once a novel scene is experienced but before scene recognition (e.g., immediately following jump cuts), the impact of bottom-up influences on attentional selection is most pronounced. During this short period of time, top-down influences are unreliable, and bottom-up influences represent the best guess of where pertinent information is located. The fact that observers tended to make more saccades faster in the first 0.5 s after jump cuts than later on further supports this interpretation, given that bottom-up influences are known to be faster acting than top-down influences (Wolfe et al. 2000, Henderson 2003)

(2) After the scene is recognized, prior knowledge and perceptual memory become increasingly more reliable, leading to increases in the impact of top-down influences on attentional selection, coupled with corresponding decreases in the impact of bottom-up influences.

(3) As time goes by between jump cuts, objects and people move around compared to their initial positions, prompting observers to increase again their relative sensitivity to bottom-up influences.

This chain of events may repeat itself during continuous viewing conditions, leading to oscillatory changes in the balance between bottom-up and top-down influences on attentional selection. Future experiments with a larger number of participants would be needed in order to detect such oscillatory patterns in continuous viewing conditions.

Acknowledgments

The work described in this manuscript was funded by grants from NSF, HFSP, NIMA, and the Zumberge Research and Innovation fund. Computation for the work described in this manuscript was supported by the University of Southern California Center for High Performance Computing and Communications (www.usc.edu/hpcc). We thank A. Almor, I. Bargad, M. Baudry, I. Biederman, M. Cordey, G. Karmi, C. Koch, Z. L. Lu, C. v. d. Malsburg, V. Navalpakkam, B. W. Mel, R. Peters, W. Soussoue, R. Sorek, and B. Tjan, for helpful comments on earlier versions of this manuscript. We also thank Dan Simons and two other anonymous reviewers for their helpful comments.

References

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

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

Caspi, A., B. R. Beutter, and M. P. Eckstein. (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.

Chun, M. M., and Y. Jiang. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. Cognitive Psychology **36**:28-71.

Chun, M. M., and K. Nakayama. (2000). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. Visual Cognition **7**:65-81.

Eagleman, D. M., P. U. Tse, D. Buonomano, P. Janssen, A. C. Nobre, and A. O. Holcombe. (2005). Time and the brain: how subjective time relates to neural time. J Neurosci **25**:10369-10371.

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

Felsen, G., and Y. Dan. (2005). A natural approach to studying vision. Nat Neurosci **8**:1643-1646.

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

Frost, B. J., and K. Nakayama. (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. Lawrence Erlbaum Associates, Hillsdale, N.J.

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

Hasson, U., Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. (2004). Intersubject synchronization of cortical activity during natural vision. Science **303**:1634-1640.

Hayhoe, M. M., A. Shrivastava, R. Mruczek, and J. B. Pelz. (2003). Visual memory and motor planning in a natural task. Journal of Vision **3**:49-63.

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

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

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

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

Hollingworth, A., and J. M. Henderson. (2002). Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception and Performance **28**:113-136.

Horowitz, T. S., and J. M. Wolfe. (1998). Visual search has no memory. Nature **394**:575-577.

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

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

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

Klin, A., W. Jones, R. Schultz, F. Volkmar, and D. Cohen. (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 S. Ullman. (1985). Shifts in Selective Visual-Attention - Towards the Underlying Neural Circuitry. Human Neurobiology **4**:219-227.

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

Land, M. F. (1999). Motion and vision: why animals move their eyes. Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology **185**:341-352.

Land, M. F., and S. Furneaux. (1997). The knowledge base of the oculomotor system. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **352**:1231-1239.

Land, M. F., and D. N. Lee. (1994). Where we look when we steer. Nature **369**:742-744.

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

Lehky, S. R., and T. J. Sejnowski. (1988). Network model of shape-from-shading: neural function arises from both receptive and projective fields. Nature **333**:452-454.

Levin, D. T., and D. J. Simons. (1997). Failure to detect changes to attended objects in motion pictures. Psychonomic Bulletin & Review **4**:501–506.

Maljkovic, V., and K. Nakayama. (1994). Priming of pop-out: I. Role of features. Memory and Cognition **22**:657-672.

Maljkovic, V., and K. Nakayama. (1996). Priming of pop-out: II. The role of position. Perception & Psychophysics **58**:977-991.

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

Melcher, D. (2001). Persistence of visual memory for scenes. Nature **412**:401.

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

Michel, C. M., M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. Grave de Peralta. (2004). EEG source imaging. Clinical Neurophysiology **115**:2195-2222.

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

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

Oliva, A., J. M. Wolfe, and H. C. Arsenio. (2004). Panoramic search: the interaction of memory and vision in search through a familiar scene. Journal of Experimental Psychology: Human Perception and Performance **30**:1132-1146.

Olshausen, B. A., and D. J. Field. (2005). How close are we to understanding V1? Neural Computation **17**:1665-1699.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. Canadian Journal of Psychology **46**:461-488.

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

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

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

Polat, U., and D. Sagi. (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.

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

Rensink, R. A. (2000). Seeing, sensing, and scrutinizing. Vision Research **40**:1469-1487.

Rensink, R. A. (2002). Change detection. Annual Review of Psychology **53**:245-277.

Rust, N. C., and J. A. Movshon. (2005). In praise of artifice. Nat Neurosci **8**:1647-1650.

Shinoda, H., M. M. Hayhoe, and A. Shrivastava. (2001). What controls attention in natural environments? Vision Research **41**:3535-3545.

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

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

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

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 G. Gelade. (1980). Feature-Integration Theory of Attention. Cognitive Psychology **12**:97-136.

Viviani, P. (1990). Eye movements in visual search: cognitive, perceptual and motor control aspects. Reviews of Oculomotor Research **4**:353-393.

Wagner, P., K. Bartl, W. Gunthner, E. Schneider, T. Brandt, and H. Ulbrich. (2006). A pivotable head mounted camera system that is aligned by three-dimensional eye movements. Pages 117-124 *in* Proceedings of the 2006 symposium on Eye Tracking Research & Applications. ACM Press, San Diego, California.

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

Yarbus, A. L. (1967). Eye movements and vision. Plenum Press, New York.