ATTENTION, MOVIE CUTS, AND NATURAL VISION:
A FUNCTIONAL PERSPECTIVE

by

Ran Carmi

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(NEUROSCIENCE)

May 2007

Copyright 2007
Ran Carmi
DEDICATION

To my family and friends who supported me lovingly while I pursued this research.
# TABLE OF CONTENTS

**DEDICATION** ............................................................................................................. ii

**TABLE OF CONTENTS** ............................................................................................ iii

**LIST OF FIGURES** ..................................................................................................... v

**ABSTRACT** ............................................................................................................... vii

**ABSTRACT** ............................................................................................................... vii

1. **GENERAL INTRODUCTION** ............................................................................ 1

2. **THE IMPACT OF MOVIE CUTS ON THE HUMAN VISUAL SYSTEM** ...... 4
   2.1. Abstract ........................................................................................................ 4
   2.2. Introduction .................................................................................................. 5
   2.3. Methods ....................................................................................................... 9
       2.3.1. Participants ........................................................................................... 9
       2.3.2. Stimuli .................................................................................................. 9
       2.3.3. Experimental design ........................................................................... 9
       2.3.4. Data acquisition and processing ......................................................... 10
       2.3.5. Inter-observer similarity/variability ................................................... 12
   2.4. Results ........................................................................................................ 14
   2.5. Discussion .................................................................................................. 22
       2.5.1. The impact of movie cuts on human gaze behavior ......................... 22
       2.5.2. The perceptual seamlessness of movie cuts ....................................... 27

3. **VISUAL CAUSES VS. CORRELATES OF ATTENTIONAL SELECTION**. 33
   3.1. Abstract ...................................................................................................... 33
   3.2. Introduction ................................................................................................ 34
   3.3. Methods ...................................................................................................... 37
       3.3.1. Participants ........................................................................................... 37
       3.3.2. Stimuli .................................................................................................. 37
       3.3.3. Experimental design ........................................................................... 40
       3.3.4. Data acquisition and processing ......................................................... 40
       3.3.5. Bottom-up attention-priority maps ..................................................... 41
       3.3.6. Bottom-up prediction of single saccades ........................................... 42
       3.3.7. Baseline sampling .............................................................................. 43
       3.3.8. Performance metrics .......................................................................... 45
       3.3.9. Advantages of jump cuts over clip onsets as temporal anchor points 52
   3.4. Results ........................................................................................................ 53
       3.4.1. Average saliency effects based on all saccades ................................. 53
       3.4.2. "Bottom-up" labeling of saccades ....................................................... 54
3.4.3. Saliency effects as a function of viewing time ................................... 56
3.4.4. Saliency effects as a function of inter-observer variability ................ 60
3.5. Discussion ............................................................................................... 63
  3.5.1. Bottom-up causes versus correlates of attentional selection .......... 63
  3.5.2. Static versus dynamic bottom-up models ....................................... 64
  3.5.3. Interactions between bottom-up and top-down influences ............. 65
  3.5.4. Realism of stimuli used in studies of attentional selection .......... 67
  3.5.5. Saliency modeling ........................................................................... 68
3.6. Acknowledgements .................................................................................. 69

4. THE ROLE OF PERCEPTUAL MEMORY IN GUIDING ATTENTION ....... 70
  4.1. Abstract ................................................................................................. 70
  4.2. Introduction .......................................................................................... 71
  4.3. Methods ............................................................................................... 76
    4.3.1. Participants .................................................................................... 76
    4.3.2. Stimuli ......................................................................................... 76
    4.3.3. Experimental design ..................................................................... 77
    4.3.4. Data acquisition and processing ............................................... 79
    4.3.5. Attention-priority maps ............................................................... 80
    4.3.6. Prediction of single saccades ..................................................... 82
    4.3.7. The ASH metric ........................................................................ 85
    4.3.8. The DOH metric ........................................................................ 86
  4.4. Results ................................................................................................. 91
    4.4.1. Average bottom-up impact on attentional selection ................. 93
    4.4.2. Time course of bottom-up impact on attentional selection ....... 99
  4.5. Discussion ........................................................................................... 105
    4.5.1. Memory and attention ............................................................... 106
    4.5.2. Natural versus artificial approaches to studying vision .......... 108
    4.5.3. Neural Implications .................................................................. 111
    4.5.4. Attention and scene understanding ........................................... 111
  4.6. Acknowledgments ................................................................................. 116

5. SUMMARY .................................................................................................. 117

ALPHABETIZED BIBLIOGRAPHY ................................................................... 121
LIST OF FIGURES

Figure 2.1. Continuous and MTV-style clips ............................................................ 14
Figure 2.2. Spatial distributions of gaze positions .................................................... 15
Figure 2.3. The extent of inter-observer similarity .................................................. 16
Figure 2.4. Inter-observer similarity/variability as a function of viewing time
from clip onsets .................................................................................................. 17
Figure 2.5. Inter-observer similarity/variability as a function of viewing time
between MTV-style cuts ................................................................................. 19
Figure 2.6. Spatial distributions of gaze positions as function of viewing time
between MTV-style cuts ................................................................................ 20
Figure 2.7. Spatiotemporal dependence between gaze positions ......................... 21
Figure 3.1. MTV-style clips and attention-priority maps ....................................... 38
Figure 3.2. Ideal and null predictions of attentional selection .............................. 47
Figure 3.3. Saccade histograms and the average prediction accuracy of
representative bottom-up models .................................................................. 54
Figure 3.4. Saliency effects as a function of saccade index between adjacent
jump cuts ......................................................................................................... 57
Figure 3.5. Saliency effects as a function viewing time, baseline type and
metric type ....................................................................................................... 58
Figure 3.6. Saliency effects as a function of inter-observer variability .............. 61
Figure 3.7. Saliency effects for "All" versus "Bottom-up" saccades ....................... 62
Figure 4.3. Schematic of the saliency model ......................................................... 82
Figure 4.4. Hypothetical scenarios of predicting attentional selection ................ 88
Figure 4.5. Saccades that straddle an MTV-style jump cut (mtvclip03,
participant MC) .............................................................................................. 93
Figure 4.6. Same as Figure 4.5, but for mtvclip04, participant JR ..................... 94
Figure 4.7. Saliency-based saccade distributions, and average bottom-up impact
on attentional selection .................................................................................. 95
Figure 4.8. Average bottom-up impact on attentional selection per participant...... 98

Figure 4.9. Time course of bottom-up impact on attentional selection.................. 102

Figure 4.10. Time course of bottom-up impact on attentional selection................. 104
ABSTRACT

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. Attentional selection enables biological organisms with limited neural resources to extract the most pertinent information from the barrage of sensory inputs that they normally experience in the real world. The goal of the research described in this dissertation is to characterize the output, input, and associated computational transformations that underlie attentional selection during natural vision. Such functional understanding of attentional selection would help build more intelligent machines that could behave autonomously in real world environments. It could also lead to better diagnostic tools and treatments for medical conditions, such as Autism, ADHD, and Parkinson Disease, which are characterized by anomalous patterns of attentional selection.

The following key questions are addressed: What do movies reveal about the nature of sensory inputs to the human visual system? Which sensory inputs causally attract attention and how do they compare to each other? How is attentional selection affected by interactions between past and present sensory inputs? To answer these questions, the eyes of human observers were tracked non-intrusively as they watched either continuous or MTV-style (discontinuous) video clips. The main conclusions of this dissertation are based on a series of quantitative analyses, in which human gaze behavior and the prediction accuracy of related computational models were compared across viewing conditions, space, and time.
The results indicate that sensory inputs are naturally discontinuous, and that the human visual system can keep the mind informed of the most pertinent information at every point in time without requiring either visual or mental continuity. It is also shown that dynamic visual correlates of attentional selection (e.g., motion contrasts) play a dominant causal role in attracting attention during natural vision. In comparison, some static visual correlates (e.g., color contrasts) play a relatively weaker causal role, whereas others (e.g., orientation contrast) are noncausal correlates, potentially reflecting top-down causes. Lastly, it is demonstrated that perceptual memory is utilized for guiding attention across several gaze shifts when persistent visual context is available.
1. GENERAL INTRODUCTION

The overarching goal of the research described in this dissertation is to reveal the functional mechanisms of attentional selection during natural vision. The term "functional" refers to system-level understanding of how attention works. By "natural vision" I mean: visual behavior that people engage in during everyday life. Gaining deeper knowledge of the functional mechanisms that control human attention would facilitate the construction of intelligent machines that could behave autonomously in real world environments. It may also help to better understand, diagnose, and treat medical conditions, such as Autism, ADHD, and Parkinson Disease, which are characterized by anomalous patterns of attentional selection.

This dissertation is composed of three chapters, each of which is based on a stand-alone manuscript (Carmi & Itti, 2006a, Carmi & Itti, 2006b, Carmi & Itti, 2006c). The relevant literature background is surveyed comprehensively in the respective introduction sections, so this general introduction focuses on the key questions addressed in each chapter, and outlines the basic methodological approach that was taken to answer these questions.

The research described in Chapter 2: "THE IMPACT OF MOVIE CUTS ON THE HUMAN VISUAL SYSTEM" is motivated by the observation that movie cuts give rise to physically unnatural visual discontinuities that are often imperceptible. This peculiar mismatch between physics and psychology has been explained in terms of either visual or mental continuity across movie cuts, but the related evidence is either anecdotal or inconsistent. The following questions are asked: what is the
impact of movie cuts on patterns of human gaze behavior, and what do these patterns reveal about the perception of movies in particular and visual perception in general? Chapter 2 also discusses the role of movie editing techniques and consistent narrative in blurring the sensory, perceptual, and cognitive boundaries between movies and real world environments.

The research described in Chapter 3: "VISUAL CAUSES VS. CORRELATES OF ATTENTIONAL SELECTION" delves deeper into the computational mechanisms of attentional selection. It addresses two basic questions: what are the visual causes, rather than mere correlates, of attentional selection and how do they compare to each other during natural vision? Chapter 3 also proposes new experiments and modeling enhancements that could further improve the ecological validity (real world relevance) of attention studies without sacrificing their explanatory and predictive powers.

The research described in Chapter 4: "THE ROLE OF PERCEPTUAL MEMORY IN GUIDING ATTENTION" addresses the time frame in which perceptual memory guides attention, a controversial question among visual psychologists. Chapter 4 also revisits an ongoing debate about the pros and cons of the artificial approach versus the natural approach to studying biological vision. It argues that studies of high-level vision that rely exclusively on the artificial approach are particularly sensitive to the pitfalls of ecological invalidity. Lastly, chapter 3 proposes a new set of hypothesis-driven experiments and a hybrid natural-artificial
approach to study interactions between attentional selection and scene understanding.

To answer these questions, the following experimental manipulation was performed: 50 continuous video clips were cut into 523 clip snippets (clippets), which were shuffled and restrung together into 50 MTV-style clips. This MTV-style manipulation is revealing because of its likely effects on the mental representations of viewers. For example, the abrupt transitions (jump cuts) between clippets are expected to invalidate mental representations that were formed based on the preceding clippet, and lead to the formation of novel mental representations based on the following clippet. To examine these effects, two groups of human observers were asked in two separate experiments to inspect either continuous or MTV-style clips, and their eyes were tracked with a non-intrusive eye tracker. This dissertation describes a series of quantitative analyses that compared human gaze behavior and the prediction accuracy of related computational models across viewing conditions, space, and time.
2. THE IMPACT OF MOVIE CUTS ON THE HUMAN VISUAL SYSTEM

2.1. Abstract

Movie cuts give rise to physically unnatural visual discontinuities that are often imperceptible. This peculiar mismatch between physics and psychology has been explained in terms of either visual or mental continuity, but the related evidence is either anecdotal or inconsistent. Here we address two questions: what is the impact of movie cuts on patterns of human gaze behavior, and what do these patterns reveal about the perceptual seamlessness of movie cuts? We tracked the eyes of human observers as they watched either continuous or MTV-style (discontinuous) video clips that were generated by stringing together visually and semantically unrelated clip snippets. Inter-observer similarity and other parameters of gaze behavior were quantified and compared across viewing conditions, space and time. In both the continuous and MTV-style viewing conditions, the gaze positions of 4 human observers were clustered within less than 5% of the display area for more than 50% of the viewing time. On average, the MTV-style manipulation led to increased inter-observer similarity and stronger center bias in the distribution of gaze positions. MTV-style cuts repeatedly led to increases in the extent of inter-observer similarity lasting for 0.5 s, followed by monotonic decreases for the next 1.5 s. These results demonstrate that the human visual system can adapt rapidly to an extremely discontinuous stream of sensory data. The implication is that the perceptual seamlessness of movie cuts reflects a natural adaptation to visual discontinuities that is independent of either visual or mental continuity.
2.2. Introduction

Movies\(^1\) are physically quite different from the real world in both their spatial and temporal properties (Anderson, 1996, Cutting, 2005, Gibson, 1979/1986, Hochberg, 1986). In space, they are typically displayed on two-dimensional screens that span a much smaller field-of-view than the 360 degrees afforded by the three-dimensional real world. In time, movies are composed of sequences of discrete frames, whereas the real world is continuous. For example, during one second of film observers are exposed to a rapid succession of 24 frames, each of which is flashed 2 or 3 times (for a total display frequency of either 48Hz or 72Hz). The brief periods of physical darkness between frames are imperceptible because these display frequencies are sufficiently higher than the flicker fusion threshold of human observers - the minimal frequency at which an intermittent light source becomes indistinguishable from a steady light source (Seitz et al., 2005). In comparison, movie cuts typically occur at temporal frequencies of less than 0.5 Hz (much lower than the flicker fusion threshold of human observers), but nonetheless they often go unnoticed (Hochberg & Brooks, 1996).

The craft of making movie cuts imperceptible has been a hallmark of Hollywood-style moviemaking for over 50 years, yet vision scientists and film theorists alike have paid little attention to the question of why movie cuts actually work (Anderson, 1996). The late J. J. Gibson was one of the few pioneers who theorized about visual perception of movie cuts (Gibson, 1979/1986). He proposed

---

\(^1\) We use the term "movie" in its general sense: "a series of pictures displayed on a screen in rapid succession". This usage encompasses all the known mediums for displaying dynamic stimuli, including film, television, and computer-generated animations.
that the most seamless cuts are those that retain key visual invariances across successive shots, such as the relative position or orientation of key objects. According to this "visual continuity" hypothesis, transitions between successive non-overlapping shots, such as fades, wipes, and dissolves, are very important because they signal to viewers the passage of time or the change of location between such shots. Consistently with this notion, Gibson also argued that jump cuts - abrupt cuts between non-overlapping shots - are "ill-founded", because they break visual continuity without any warning. These arguments notwithstanding, the rise of Music Television (MTV) in the 1980's popularized the use of jump cuts, which are much more prevalent today than they were at the time the "visual continuity" hypothesis was proposed back in 1979 (Anderson, 1996, Cutting, 2005, Thompson & Bordwell, 2003). Furthermore, fades, wipes, and dissolves are less popular today than they used to be (Cutting, 2005). It thus appears that the evolution of moviemaking in the last couple of decades is largely inconsistent with the "visual continuity" hypothesis.

An alternative theoretical account was provided by J. Hochberg (Hochberg, 1986) who proposed that the perceptual seamlessness of movie cuts is attributable to the sparseness and continuity of mental representations. According to this "mental continuity" hypothesis, viewers maintain highly incomplete (sparse) mental representations that are not disrupted by movie cuts. Hochberg argued that the visual discontinuities across jump cuts go unnoticed because they are not accompanied by disruptions to mental representations. Presumably, mental continuity is maintained because the post-cut shot (side B) answers the key question raised in the minds of
viewers by the pre-cut shot (side A). Indeed, prescriptive manuals of film editing have included such guidelines for decades (Reisz & Millar, 1968). For example, an actor looking off-screen in side A raises the question: "what is he looking at? A jump cut to a dog in side B may be perceptually seamless because it answers the question raised by side A. Audio can also help to maintain mental continuity in the face of visual discontinuities, for example: the cut to the dog in side B may be made even more seamless by adding a bark sound to side A. MTV is probably the best example of how consistent audio helps to make jump cuts perceptually seamless (Cutting, 2005). The crux of the "mental continuity" hypothesis is that viewers can cope with large visual discontinuities as long as the narrative remains consistent across shots.

Here we pursued two goals - one exploratory and one hypothesis-driven. The exploratory goal was to study the impact of movie cuts on human gaze behavior. To the best of our knowledge, such exploration is yet to be performed, even though it may have important implications for visual perception in general, and the perception of movies in particular. The hypothesis-driven goal was to experimentally test the validity of existing hypotheses about the perceptual seamlessness of movie cuts. To achieve these goals, we first collected 50 continuous video clips from heterogeneous sources. These clips were cut into 523 clip snippets (clippets), which were shuffled and restrung together into 50 MTV-style clips. The rationale for this MTV-style manipulation is as follows: if movie cuts are perceptually seamless due to visual or mental continuity, then our randomly generated MTV-style cuts should be anything
but perceptually seamless. Consequently, viewers of MTV-style clips are likely to become confused shortly after MTV-style cuts, and display more erratic gaze behavior. To test this prediction, we tracked the eyes of human observers as they watched either continuous or MTV-style clips, and analyzed patterns of gaze behavior over space and time.

The main findings of this study are inconsistent with both the visual and mental continuity hypotheses: The gaze behavior of different human observers became more rather than less synchronized during inspection of MTV-style versus continuous clips. Furthermore, inter-observer similarity increased shortly after MTV-style cuts, exactly when viewers should have been the most confused according to the visual and mental continuity hypotheses. To account for these results, we propose that the physical dissimilarities between real world stimuli and movies disappear at the sensory level. According to this "natural adaptability" hypothesis, visual perception is always based on discontinuous sensations due to the nature of human eye-movements, regardless of whether the underlying stimulus is continuous or not. Presumably, the human visual system evolved the ability to keep the mind informed of the most pertinent information at every point in time without requiring either visual or mental continuity. In the discussion section, we propose new experiments that could be performed to test this "natural adaptability" hypothesis. We also discuss the role of maintaining a consistent narrative across shots and the effects of editing techniques, which often seem to be misinterpreted.
2.3. Methods

2.3.1. 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.

2.3.2. 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 snippets (clippets), which were scrambled and re-assembled into 50 MTV-style clips. Continuous and MTV-style clips were matched in length, and each MTV-style clip contained at most one clippet from a given continuous clip.

2.3.3. 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
sometimes looked around the room instead of at the display monitor.

A 2-group design (1 per viewing condition) was preferred over a single-group
design of subjects, because we wanted to minimize priming effects. A potential
problem with this 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.

2.3.4. 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. The calibration procedures are only summarized here, because they were
already described in detail elsewhere (Itti, 2005). First, a 5-point calibration was used
once at the beginning of each viewing session. This initial calibration provided a
sanity check during the experiment: it was used by the ISCAN software to
superimpose in real time the instantaneous gaze position of a human observer on a copy of the observer video clip that was displayed on a monitor at the experimenter station. Second, each block of 5 video clips was preceded by a 9-point calibration session that was used to map eye tracker coordinates to screen coordinates offline (mean ± s.d. of the calibration accuracy: 0.66° ± 0.46°). Each calibration trace was filtered for blinks, then automatically segmented into two fixation periods (corresponding to the central cross and calibration point that flickered in succession), or discarded if the segmentation failed quality control criteria (enforcing minimal fixation duration and maximal variability in gaze positions during each fixation). Eye traces from the entire block of 5 video clips that followed each 9-point calibration session were discarded if more than 3 calibration points failed the quality control criteria mentioned above. Eye traces based on individual video clips were also discarded if they failed additional quality control criteria (excessive eye-blinks, loss of tracking due to head motion or excessive wetting of the eye, loss of corneal reflection due to excessive squinting). In theory, assuming that all quality control criteria were met, each group of subjects could have yielded up to 400 eye traces (50 clips, 8 observers). In practice, the actual yield was 235 and 285 eye traces from the continuous and MTV-style groups, respectively. To ensure fair comparisons across groups while maximizing the amount of data used, the analyses presented here are based on 200 eye traces from each group (50 clips, 4 observers). It is important compare the same number of observers across groups because metrics of inter-observer similarity (see next section) depend on the number of observers used. For
example, a single observer has the maximal possible similarity with itself, and inter-observer similarity generally decreases (variability increases) as more observers are considered.

2.3.5. Inter-observer similarity/variability

Inter-observer similarity/variability was quantified using 2 metrics:

1. % coverage, as measured by the area of the rectangle bounding the concurrent gaze positions of different observers, divided by the display area:

\[
IOV_A = 100 \times \frac{(\max(\bar{x}) - \min(\bar{x})) \times (\max(\bar{y}) - \min(\bar{y}))}{A_{\text{max}}}
\]

where \( \bar{x}, \bar{y} \) are vectors of gaze positions, and \( A_{\text{max}} = W \times H \) is the area of the display with width \( W = 640 \) and height \( H = 480 \).

The lower bound of this area metric is 0, reflecting a circumstance in which different observers look simultaneously at exactly the same location (maximal similarity and minimal variability). The upper bound is 100, reflecting a circumstance in which different observers look simultaneously at different corners of the display.

2. Mean Euclidean distance between the concurrent gaze positions of different observers (in degrees of visual angle):

\[
IOV_D = F_{dpp} \times \frac{2 \times \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{n \times (n - 1)}
\]
where \( F_{dpp} = 1/23 \) is a factor that converts between pixels and degrees, \( n = 4 \) is the number of gaze positions (1 per observer), and \( x_i, y_i \) are the coordinates of gaze position \( i \).

The circumstances in which the lower and upper bounds would be reached are identical to what is described above for the area metric, but the corresponding values for this distance metric are 0 and TBD, respectively.

The values of the area metric are more intuitively meaningful than those of the distance metric, which is why the former is preferred. A potential disadvantage of the area metric compared to the distance metric is that only the former would fail in certain extreme circumstances, specifically: when different observers are perfectly aligned either horizontally or vertically. In these circumstances, the area of the bounding rectangle would be 0, and the area metric would erroneously report that there is maximal similarity between observers. In practice, such circumstances never arose in either of the viewing conditions tested in this study.
2.4. Results

Figure 2.1 shows frame examples from the continuous and MTV-style viewing conditions with superimposed gaze positions of 4 human observers. Figure 2.1A shows 4 frames from a continuous clip, whereas Figure 2.1B shows 4 corresponding frames from an MTV-style clip. As reflected in the gaze positions superimposed on the first frame of each clip, observers were looking at a central fixation cross before clip onsets. Each of the other 3 frame examples in the MTV-style condition was displayed 0.5 s after the most recent cut, and all observers initiated at least one rapid gaze shift during that period.

Figure 2.1. Continuous and MTV-style clips.

(A) Sample frames from a continuous clip (see Movie 1). Bright yellow numbers at the bottom left of each frame depict the frame display time (format: hh:mm:ss.msec). Colored discs depict the instantaneous gaze positions of 4 human observers. Certain frames contain more than one disc of the same color, marking instances in which the eye was in motion (the eye tracker operates at 240Hz, 8 times faster than the frame rate, so up to 8 discs may be visible).
(B) Same as A, but for an MTV-style clip that shares the same first second with the continuous clip shown in A (see Movie 2). Black digits at the top left of each frame depict the clippet index.

To address the impact of movie cuts on the human visual system we performed several spatial and spatiotemporal analyses. Figure 2.2 shows the overall spatial distributions of gaze positions in the continuous and MTV-style viewing conditions. It demonstrates that both viewing conditions led to centrally-biased distributions of gaze positions, but that the MTV-style condition was characterized by a relatively stronger center-bias.

Figure 2.2. Spatial distributions of gaze positions.

(A) Continuous viewing condition. The abscissa and ordinate of the distribution map correspond to the screen width and height, respectively. Each pixel in the distribution map subtends 0.7° X 0.7°, which is equivalent to the mean calibration accuracy. Pixels values sum to 100 and reflect the percentage of gaze positions falling in each pixel (see horizontal color bar).

(B) Same as A, but in the MTV-style viewing condition.

Figure 2.3 shows spatiotemporal correlations between different observers, as measured by the extent of inter-observer similarity (expressed by its reciprocal -variability, see section 2.5). A value of 0 means that observers look at the same position simultaneously (maximal similarity, minimal variability), whereas a value of
100 means that observers are looking at different corners of the display simultaneously (minimal similarity, maximal variability).

Figure 2.3. The extent of inter-observer similarity.

(A) % of time steps as a function of inter-observer variability (divided into 20 equally sized bins) in the continuous viewing condition.

(B) Same as A, but in the MTV-style viewing condition.

The histograms in each viewing condition are based on 3 data sets:

1. Ordered human data. For each clip, the 4 associated eye traces (1 per observer) were traversed sequentially, and inter-observer variability was computed in each time step based on the synchronized gaze positions of different observers.

2. Shuffled human data. For each clip, inter-observer variability was computed the same number of times as in 1, but the 4 gaze positions in each time step were selected at random (1 from each of the 4 associated eye traces).

3. Random data. Inter-observer variability was computed the same number of times as in 1 and 2, but the 4 gaze positions in each time step were sampled at random from a uniform distribution of screen coordinates.

Figure 2.3 shows that the extent of inter-observer variability in both viewing conditions was much lower than expected by chance, especially if the correct temporal order of gaze positions is taken into account. For example, human observers in the continuous condition spent approximately 55% of the viewing time
looking at an area smaller than 5% of the display area (Figure 2.3A, left-most black bar). This trend was even stronger in the MTV-style condition, in which 65% of the viewing time was spent looking at an area of the same size (Figure 2.3B, left-most black bar). Another interesting trend in Figure 2.3 is the larger overlap between the shuffled human histogram and the ordered human histogram versus the random histogram.

To further examine the role of temporal factors, we quantified changes in inter-observer similarity as a function of viewing time from clip onsets, as shown in Figure 2.4. The strong correspondence between Figure 2.4A (based on the distance metric) and Figure 2.4B (based on the area metric) demonstrates that the results do not depend on the type of metric that was used to quantify inter-observer variability. Figure 2.4 also shows that inter-observer variability was consistently higher in the continuous condition compared to the MTV-style condition, consistently with the results presented in Figure 2.3. In both viewing conditions, inter-observer variability increased sharply as a function of time for the first 2-2.5 seconds from clip onsets.

![Figure 2.4](image)

**Figure 2.4.** Inter-observer similarity/variability as a function of viewing time from clip onsets.
(A) Dark blue and light orange plots show inter-observer variability as function of time from clip onset in the continuous and MTV-style conditions, respectively. For each clip, inter-observer variability was quantified for consecutive 250 ms time slots. The plots depict the mean inter-observer variability in each time slot (averaged across clips). Error bars depict the standard errors. Dashed lines depict the mean inter-observer variability across time slots. Dotted red line depicts the mean inter-observer variability expected by chance based on 4 random observers (35.88%, the mean of the random histogram shown in Figure 2.3).

(B) Same as A, but using a distance metric (see section 2.5). The mean inter-observer variability expected by chance is 12.74°.

Figure 2.4 shows that after the initial sharp increase, inter-observer variability continued to increase at a much slower pace in the continuous condition, and hovered around the mean in the MTV-style condition. In both conditions, inter-observer variability remained much lower than the chance baseline (dotted lines) throughout the viewing session. The large fluctuations during the first 10-15 seconds in the MTV-style condition indicate a temporal pattern that may be masked later on due to the asynchrony of MTV-style cuts across different clips (see section 2.2). To further examine this issue, we quantified changes in inter-observer similarity as a function of viewing time between MTV-style cuts. Figure 2.5 shows that inter-observer variability decreased shortly after MTV-style cuts, followed by gradual monotonic increases (black plot). This temporal pattern disappeared when the temporal sequence of gaze positions was shuffled (light cyan plot).
Figure 2.5. Inter-observer similarity/variability as a function of viewing time between MTV-style cuts.

For each cliplet (excluding the first cliplet of each clip), inter-observer variability was quantified for consecutive 250 ms temporal bins. Black and cyan plots are based on the ordered and shuffled human data, respectively (see Figure 2.3), and show the mean inter-observer variability over time (averaged across cliplets). Error bars depict the standard errors. Dashed lines depict the mean inter-observer variability across temporal bins. Dotted red line depicts the mean inter-observer variability expected by chance (see Figure 2.4A).

We also examined the spatial distributions of gaze positions as a function of time between MTV-style cuts. Figure 2.6 shows that the center-bias in gaze positions increased shortly after MTV-style cuts (second panel), and slowly decreased thereafter.
Figure 2.6. Spatial distributions of gaze positions as function of viewing time between MTV-style cuts.

Each of the 12 panels corresponds to a time slot of 250 ms, as described for Figure 2.5. The abscissa and ordinate of each panel correspond to the screen width and height, respectively. Each pixel in each panel subtends 0.7° X 0.7°, which is equivalent to the mean calibration accuracy. Pixels values sum to 100 and reflect the percentage of gaze positions falling in each pixel (see horizontal color bar).

Lastly, Figure 2.7 shows the spatiotemporal dependence between gaze positions. It demonstrates that the mean Euclidean distance between gaze positions increased as a function of temporal offset. In both viewing conditions, the average distance between gaze positions was reached when their temporal offset was 1 second (dashed lines).
Figure 2.7. Spatiotemporal dependence between gaze positions.

1000 gaze positions were selected at random from each eye trace. For each selected gaze position, a second gaze position was selected from the same eye trace based on 1000 temporal offsets, ranging from 4.166 msec (two successive gaze positions) to 4.166 sec (two gaze positions separated in time by 1000 other gaze positions). The mean Euclidean distance between two selected gaze positions was computed for all possible temporal offsets. The continuous (dark blue) and MTV-style (light orange) plots are based on the mean Euclidean distance per temporal offset, pooled across all eye traces in each condition. Error bars show the standard errors. Dashed lines depict the mean Euclidean distance across all temporal offsets in each condition. The dotted red line depicts the mean Euclidean distance expected by chance - 12.74°. It was calculated based on 100000 pairs of independent gaze positions sampled at random from a uniform distribution of screen coordinates.
2.5. Discussion

We examined spatiotemporal patterns of human gaze behavior in the context of continuous and MTV-style viewing conditions. The implications of the results in the context of understanding why people look where they do are discussed in section 4.1. The second goal of this study was to investigate why movie cuts are perceptually seamless, even though they often involve large visual discontinuities. The implications of the answer go far beyond the perception of movies, as discussed in section 4.2.

2.5.1. The impact of movie cuts on human gaze behavior

2.5.1.1. Spatial distribution of gaze positions

Both the continuous and MTV-style conditions were characterized by center bias in gaze positions (see Figure 2.2), echoing previous results from eye tracking studies that utilized static scenes (Parkhurst et al., 2002, Tatler et al., 2005). Center bias may arise from motor, sensory (bottom-up), or cognitive (top-down) causes. First, viewers were seated with their head supported by a chin-rest that was centered with respect to the display monitor. Second, viewers were asked to fixate at the center of the display before clip onsets. Both of these methodological factors may cause a motor bias to look at the center of the display. Third, most of the stimuli used in this study were video clips originally designed for human consumption, including ads, video games, televised sport events, etc. Such stimuli often contain objects of interest at or near the center of the frame, and this stimulus bias is likely to
cause sensory and cognitive biases to look at the center. The increased center bias in
the MTV-style condition is illuminating in this context, as discussed in section
2.5.1.4.

2.5.1.2. Overall inter-observer similarity

Figure 2.3 shows that eliminating temporal factors has a strong effect on the
extent of inter-observer similarity (black versus cyan histograms). This result is
consistent with previous accounts of inter-observer similarity in the context of
dynamic scenes (Stelmach et al., 1991, Tosi et al., 1997), and inconsistent with other
accounts based on static scenes (Mannan et al., 1997b, Tatler et al., 2005, Wooding,
2002). The results of this and previous studies indicate that dynamic stimuli lead to
a much stronger involvement of temporal factors in determining gaze position. It is
interesting to note that the impact of eliminating spatial factors has an even stronger
effect on the extent of inter-observer similarity (cyan versus red histograms in Figure
2.3). It is unclear whether this finding reflects a genuinely stronger involvement of
spatial versus temporal factors in determining gaze position, which is independent of
the motor and stimulus biases introduced by our methodology. Section 4.1.4
describes future studies that could resolve this issue by ensuring that the most
pertinent stimuli are not presented straight ahead from the perspective of viewers.
Figure 2.3 also shows that the MTV-style condition was characterized by stronger
inter-observer similarity, and the implication of this finding is discussed in section
2.5.1.4.
2.5.1.3. Changes in inter-observer similarity as a function of viewing time

Figure 2.4 shows a sharp increase in inter-observer variability during the first couple of seconds after clip onsets in both the continuous and MTV-style conditions. This result is consistent with previous accounts of changes in inter-observer similarity as a function of viewing time from the onset of static scenes (Mannan et al., 1997b, Tatler et al., 2005). The time frame of this increase is informative, but the trend itself is not surprising: the fixation cross before clip onsets practically ensures that inter-observer variability will increase as a function of viewing time from clip onsets. The novel aspect of our results is the temporal pattern that follows: inter-observer variability was consistently lower in the MTV-style condition relative to the continuous condition during the reminder of the viewing time. This finding demonstrates that the MTV-style manipulation led to strong increases in inter-observer similarity, as also evident in the histograms shown in Figure 2.3. Figure 2.4 also shows strong fluctuations in inter-observer variability during the first 10-15 seconds in the MTV-style condition. These fluctuations may be caused by MTV-style cuts, which occur asynchronously in different clips. This asynchrony may lead to progressively stronger masking effects as a function of viewing time when data is pooled across clips. To avoid these masking effects, we performed an additional temporal analysis using MTV-style cuts as anchor points, as shown in Figure 2.5.

Figure 2.5 shows that inter-observer variability repeatedly decreased shortly after MTV-style cuts, followed by gradual monotonic increases until the next MTV-style cut (black plot). Figure 2.5 also shows that the light cyan plot (human/shuffled
analysis) is flat and consistently higher than the black plot (human/ordered analysis), reiterating the important role of temporal factors in determining gaze position and inter-observer similarity. To examine the associated spatial factors, we also plotted the spatial distributions of gaze positions as a function of viewing time between MTV-style cuts, as shown in Figure 2.6. It demonstrates that increases/decreases in inter-observer similarity are coupled with increases/decreases in the extent of center-bias. This coupling could be caused by several factors:

1) The MTV-style condition increases the rate at which new stimuli become available at the center of gaze, making it less necessary or beneficial to engage in active visual exploration. A reduced drive to move their eyes is likely to lead viewers to spend more time looking at the center of the display, indirectly leading to higher inter-observer similarity.

2) The visual and mental discontinuities caused by MTV-style cuts may confuse viewers and lead them to "space out". In the absence of a compelling reason to look anywhere in particular, they may stare at the center of the display simply because it is located straight in front of their eyes. The implication is that the higher inter-observer similarity in the MTV-style condition is merely an artifact of motor biases.

3) Viewers in the MTV-style condition were under time pressure to make sense of a given scene before it changes to a completely different scene. Under these circumstances and considering the stimulus-related center bias, viewers may have
actively elected to focus their gaze at the center of the display, which often contained the most pertinent information about the observed scene.

To test hypothesis #1, we examined the spatiotemporal dependence between gaze positions, as shown in Figure 2.7. Unsurprisingly, the spatial distance between gaze positions decreased as function of the temporal offset between them. The interesting result is the lack of any significant differences in the shapes of these functions between the continuous and MTV-style conditions. This finding argues against the intuitively appealing notion that MTV-style stimuli inhibit active visual exploration. If this was the case, then the spatiotemporal dependence between gaze positions should have been stronger in the MTV-style condition. Alternatively, the MTV-style manipulation may cause two opposing trends: Increased stimulus dynamic at the center of gaze may indeed inhibit active visual exploration, but this inhibitory effect may be offset by excitatory effects due to increased stimulus dynamic outside the center of gaze. Future studies could resolve this uncertainty by employing gaze-contingent displays to selectively manipulate the extent of stimulus dynamics at the center of gaze versus the periphery. Until such evidence is presented, the simpler explanation for the observed lack of differences between the continuous and MTV-style conditions in the context of Figure 2.7 is that there were no differences. Either way, hypothesis #1 would be insufficient to explain the results shown in Figure 2.6 and Figure 2.7.

Conclusive evidence for either hypothesis #2 or hypothesis #3 is unavailable yet, but there are several reasons to prefer hypothesis #3. First, previously published
results show that the frequency of rapid gaze shifts and the prediction accuracy of a bottom-up model of attentional selection both increase shortly after MTV-style cuts (Carmi & Itti, 2006b). These findings suggest that MTV-style cuts lead to active stimulus-dependent selections of attention targets rather than passive staring at the center of the display. Anecdotally, we also noticed that viewers sometimes shift their gaze synchronously to a non central target shortly after MTV-style cuts, especially when post-cut clippets contain salient stimuli in non central locations (for example, see Figure 2.1B, clippet #2).

Future studies could further examine the causes behind why people look where they do by manipulating motor biases, stimulus biases, or both. For example, if the display monitor is positioned such that it is not aligned with the viewers' center of gaze, and they would still look at the center of the display shortly after MTV-style cuts, then this result would support stimulus-dependent causes. Alternatively, if only the stimulus-related center bias is removed, and viewers would still look at the center of the display after MTV-style cuts, then this finding would support motor-dependent causes.

2.5.2. The perceptual seamlessness of movie cuts

The existing hypotheses for why movie cuts often go unnoticed focus on continuity. The "visual continuity" hypothesis argues that key visual invariances are typically retained across cuts and thus allow viewers to make sense of successive shots through direct perception (Gibson, 1979/1986). In the case of cuts between non-overlapping shots, the "visual continuity" hypothesis claims that conventional
transitions, such as fades, wipes, and dissolves, are important for making such cuts perceptually acceptable. Since this hypothesis was proposed back in 1979, jump cuts were popularized by the rise of MTV in the 1980's, and their frequency has apparently increased (Anderson, 1996, Cutting, 2005). Jump cuts occur abruptly and retain no visual invariances between successive shots, so the substantial rise in their frequency seems sufficient to doom the "visual continuity" hypothesis (Hochberg, 1986). Concurrently, the use of fades, wipes, and dissolves has also declined in Hollywood-style moviemaking (Cutting, 2005), further weakening the plausibility of the "visual continuity" hypothesis.

An alternative explanation for why movie cuts often go unnoticed is provided by the "mental continuity" hypothesis (Hochberg, 1986, Hochberg & Brooks, 1996). The general claim is that mental continuity is retained in some form across cuts, even in the face of large visual discontinuities. As the proponents of this hypothesis admit, the details are a bit vague, because of our limited knowledge about the nature of mental representations. Nevertheless, it is possible to identify key conditions that must be met for the "mental continuity" hypothesis to hold:

1) Mental representations are sparse. If this was not the case, then large visual discontinuities would have led inevitably to mental discontinuities. "Change blindness" studies, which established the insensitivity of human observers to many types of changes across successive images or shots, are indeed supportive of mental sparseness (Rensink, 2002).²

² An important methodological detail that is not always apparent in "change blindness" demos is the obligatory presence of a mask, such as a uniformly gray frame, which is flashed briefly between the two key frames. In the absence of a mask, viewers
2) Mental representations are highly conserved across viewers. If this was not the case, then different people would have noticed different changes across cuts, making it practically impossible to ensure perceptual seamlessness for all viewers simultaneously. Experimentally, correlations between the change detection rates of different human observers are indicative of mental similarity. Further support for this notion has been provided by brain imaging and eye tracking studies, which indicate that the human visual systems of different observers "tick together" when they are exposed to the same movie (Hasson et al., 2004, Stelmach et al., 1991, Tosi et al., 1997).

3) Moviemakers can control and manipulate the viewers' focus of attention. This condition is motivated by the experimental finding that "change blindness" is contingent on manipulating objects or features that are outside the viewers' focus of attention (Rensink, 2002). The implication for mental continuity is that the viewers' attention must either remain focused on non-changing items across shots or be shifted predictably. The evidence for this notion is anecdotal in nature. For example, some moviemakers have argued that match-action cuts, which connect shots of the same action from different angles, are perceptually seamless exactly because they carry the eyes of viewers through the cut (Anderson, 1996). The validity of such claims is yet to be verified experimentally, but at the very least they show that moviemakers consciously hypothesize about the viewers' focus of attention. Hochberg & Brooks (Hochberg & Brooks, 1996) also demonstrated how a

---

are rarely blind to visual changes, even if they are made to semantically unimportant items. In movies, masking is typically achieved without introducing artificial stimuli, such as by ensuring that key objects are located in overlapping screen positions across successive shots (Hochberg, 1986).
common editing practice of manipulating the actors' gaze direction strongly affects how viewers interpret successive shots. On the other hand, there is also anecdotal evidence that is inconsistent with the "mental continuity" hypothesis. For example, to avoid artifactual motion signals or for artistic purposes, editors sometimes deliberately break continuity by inserting cutaways to shots of completely different scenes between two slightly different shots of the same scene (Hochberg, 1986).

This study provides the first experimental test - and refutation - of the "mental continuity" hypothesis. The most revealing results in this context are Figure 2.5 and Figure 2.6, which were already discussed in section 4.1.3. In short, these figures demonstrate that MTV-style cuts, which were deliberately designed to break visual and mental continuity, led to the exact opposite results of those predicted by the "mental continuity" hypothesis. Instead of confusing viewers and leading to erratic gaze behavior, NTV-style cuts led to increased synchronization between viewers and consistent changes in their gaze patterns.

To account for these results, we propose an alternative explanation for the perceptual seamlessness of movie cuts. According to this "natural adaptability" hypothesis, visual perception is always based on discontinuous sensations, regardless of whether the underlying stimulus is continuous or not. The similarities between visual sensations of real world stimuli and movies can be appreciated by considering the natural repertoire of eye movements. Almost all animals with good vision

---

3 Gaze direction is a strong cue for eliciting "joint attention" - synchronized focus of attention (in this case, between actors and viewers).
4 Cutaways and other editing techniques that eliminate artifactual motion signals are often referred to as "continuity editing" techniques (Anderson, 1996), seemingly reflecting a theoretical bias. Rather than maintain veridical continuity, as the term "continuity editing" suggests, the actual role of such techniques is to prevent anomalous perceptions of continuity.
employ stereotypical eye movement patterns (Land, 1999), including fixations (shots with a stable camera), abrupt gaze shifts (cuts), and smooth-pursuit (camera pans)\(^5\). These eye-movement patterns allowed our ancestors to function successfully in a complex dynamic world using limited neural resources, which preclude simultaneous and detailed processing of the entire visual field. The implication is that moviemakers have learned to mimic natural sensations that have been experienced by humans and their ancestors well before the invention of movies\(^6\). A similar idea was already proposed 40 years ago by the director John Huston who reportedly likened movie cuts to the visual experience caused by normal gaze shifts (Messaris, 1994). We argue that the actual biological equivalents of movie cuts are large gaze shifts, which are a relatively rare subtype of normal gaze shifts (Parkhurst et al., 2002). This distinction may be important for understanding why movie cuts typically occur at a slower rate (once every few seconds) than normal gaze shifts (a few times each second). A direct experimental support for the analogy between gaze shifts and movie cuts was recently provided by a unique demonstration of the actual visual stimulation falling on the retina of a human observer walking in a real world environment, which looks a lot like an MTV-style clip (Wagner et al., 2006).

The presumed convergence between real world stimuli and movies at the sensory level explains why movies are not perceived differently than the real world, but it does not provide a functional answer for the mystery of perceptual seamlessness. In fact, this convergence elevates the mystery of perceptual

\(^5\) Movie analogues of egomotion include dolly shots and zooming.

\(^6\) This implication contradicts a previous proposal that moviegoers have to learn a special kind of visual language in order to make sense of movies (Carroll, 1980).
seamlessness from a question that is specific to the perception of movies to a general mystery about real world visual perception. The functional precursor of the "natural adaptability" hypothesis was proposed a decade ago by J. D. Anderson in the context of why jump cuts between non-overlapping shots rarely lead to artifactual motion signals (Anderson, 1996). He argued that the visual system simply resets and discards the information from the previous shot when the low-level visual differences between successive shots are large enough. More generally, we argue that the human visual system has evolved the ability to rapidly shift neural resources between different attention targets without requiring either visual or mental continuity. This adaptability is necessary because continuity may often be absent or undesirable, either because of the natural discontinuity of sensations or due to rapid and pertinent changes, such as the sudden approach of a predator.

In the context of movies, the natural adaptability of the human visual system may be the key factor that renders movie cuts perceptually seamless (once artifactual motion signals are eliminated). Future studies could further test the "natural adaptability" hypothesis in several ways, such as by comparing professionally-made versus randomly-generated jump cuts in terms of their impact on human gaze behavior and comprehensibility. Lastly, it seems important to emphasize that the "natural adaptability" hypothesis aims to explain the perceptual seamlessness of individual cuts, and does not imply that maintaining a consistent narrative across shots is unimportant. On the contrary, a consistent narrative is likely to be critical for maintaining the attention of viewers over time frames of minutes or longer.
3. VISUAL CAUSES VS. CORRELATES OF ATTENTIONAL SELECTION

3.1. 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 (e.g., motion contrast) play a dominant causal role in attracting attention. In comparison, some static visual cues (e.g., color contrast) play a weaker causal role, while others (e.g., orientation contrast) are not causal at all, potentially reflecting top-down causes.
3.2. 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 et al., 2004, Gottlieb et al., 1998), psychophysical (Folk et al., 1992, Jonides & Yantis, 1988) and developmental (Atkinson & Braddick, 2003, Finlay & Ivinskis, 1984) studies indicate that dynamic stimuli are particularly effective in attracting human attention. Nonetheless, most computational studies of saliency effects (the impact of bottom-up influences on attentional selection) examined visual correlates of fixations in the context of static scenes (Krieger et al., 2000, Mannan et al., 1997a, Oliva et al., 2003, Parkhurst et al., 2002, Parkhurst & Niebur, 2003, Peters et al., 2005, Reinagel & Zador, 1999, Tatler et al., 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 et al., 2005, Hillstrom & Yantis, 1994, Jonides & Yantis, 1988, Theeuwes, 1994, Yantis & Egeth,
Such studies have been instrumental in identifying strong bottom-up influences that capture attention involuntarily in the context of competing top-down influences. However, the focus on experimental conditions that discourage observers from paying attention to salient stimuli may underestimate the impact of bottom-up cues in real world environments. Moreover, the relative costs in reaction time incurred by different visual cues provide, at best, indirect estimates of their 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 minimize potential top-down confounds without sacrificing ecological validity (real world relevance), 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 this methodology is based on previous reports of a trade-off between bottom-up and top-down influences (Henderson & Hollingworth, 1999, Hernandez-Peon et al., 1956, James, 1890). This trade-off implies that
attentional selections would 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 peak. 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.
3.3. Methods

3.3.1. Participants

8 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.

3.3.2. Stimuli

50 video clips (30 Hz, 640 x 480 pixels/frame, 4.5-30 seconds long, mean ± s.d.: 21.83 ± 8.41 s, no audio) from 12 heterogeneous sources, including indoor/outdoor daytime/nighttime scenes, video games, television programs, commercials, and sporting events. These continuous clips were cut every 1-3 s (2.09 ± 0.57 s) into 523 clip snippets (clippets), which were strung together by jump cuts into 50 scene-shuffled (MTV-style) clips (see Figure 3.1). The range of clippet lengths was chosen based on preliminary observations showing that it is sufficiently long to allow natural visual exploration. The clippet lengths were randomized to minimize the ability of observers to anticipate the exact timing of jump cuts.
Figure 3.1. 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 clips.

(b) Two consecutive saccades from an MTV-style clip (#11, participant MC, Δt=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 (see Section 3.3.5).
3.3.3. 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° x 21° usable field-of-view). Their task was: “follow the main actors and actions, and expect to be asked general questions after the eye-tracking session is over”. 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 resulted in observers disengaging from the display and looking around the room.

3.3.4. 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 were already described elsewhere (Itti, 2005). In this experiment, the calibration accuracy was 0.66° ± 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.
3.3.5. Bottom-up attention-priority maps

2D attention-priority maps (40x30 pixels/frame) were generated based on 7 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 (30Hz) based on the variance of pixel intensities in independent image patches:

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

(1)

Where \( p \) refers to an image patch, \( m, n \) are its width and height in pixels (16x16, corresponding to 0.7°x0.7° in our display), \( I \) is the intensity of an image pixel, and \( \bar{I}_p \) 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 an attention attractor (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 (30Hz), fed into a two-dimensional layer of leaky integrator neurons that provided temporal smoothing at 10 kHz, and eventually downsampled to the eye tracker sampling rate (240 Hz). These computations have already been 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 (0.7°x0.7°) 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: 792 processor hours).

3.3.6. 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. 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 ($r=3.15^\circ$). 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 section 3.3.7) provide saccade-by-saccade safeguards against any biases that may be introduced by the particular choice of aperture size.

3.3.7. 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 distributions 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, the maximal hit rate and prediction accuracy could be achieved by simply generating uniform attention-priority maps (every sample will be a hit). If a baseline is used, the hit rates of human and random saccades will be identical, reflecting low 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, particularly when saliency effects are measured as a function of viewing time from stimulus onsets that is preceded by a central fixation cross. Whether or not this has been 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 Figure 3.5).

3.3.8. Performance metrics

3.3.8.1. DOH metric

The difference of histograms (DOH) metric quantifies the human tendency (above chance) of initiating saccades towards salient targets:

$$\text{DOH} = (1/ \text{DOH}_f) \times \sum_{i=1}^{n} W_i \times (H_i - R_i)$$  \hspace{1cm} (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 $i$. 

45
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:

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

(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 Figure 3.2), a certain fraction of baseline saccades would land on the maximal saliency value by chance, with approximate probability:

$$p = N_a / N_m = 0.0408$$

(4)

where $N_a = 49$ is the number of pixels in an aperture around the saccade target ($r=3.15^\circ$, 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.
Figure 3.2. 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.
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.

3.3.8.2. Percentile metric

The percentile metric is defined as:

$$P = \left( \frac{1}{N} \right) \times \sum_{i=1}^{N} p_i$$  \hspace{1cm} (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 equivalent 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).

3.3.8.3. 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., 1997a, 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 non-linear (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 the Methods 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).

3.3.9. Advantages of jump cuts over clip onsets as temporal anchor points

(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 the Methods 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.
3.4. Results

3.4.1. 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). Figure 3.1b 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 3.3.6 and 3.3.7).

Figure 3.3 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. Figure 3.1 and Figure 3.3 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>>1.96, p<<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<<0.01).

Figure 3.3. 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 3.3.8.1).

(a) Intensity Variance. (b) Color Contrast. (c) Motion Contrast. (d) Integrated Saliency.

3.4.2. "Bottom-up" labeling of saccades

The average prediction accuracy reported in Figure 3.3 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 had little to do with bottom-up influences. To test the relative impact of bottom-up influences, it would be most 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.

Nonetheless, it is possible to identify special circumstances in which humans are particularly sensitive to bottom-up influences. For example, saccades that are initiated shortly 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 et al., 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.,
Alternatively, differences in the level of inter-observer variability may reflect top-down divergence, regardless of changes in the impact of bottom-up influences (Tatler et al., 2005).

3.4.3. 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 Figure 3.4). The Methods section 3.3.9 describes the methodological advantages of aligning the temporal analysis of saliency effects to jump cuts instead of clip onsets.

Figure 3.4 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<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 7 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.
Figure 3.4. Saliency effects as a function of saccade index between adjacent jump cuts.

Saccades were pooled across all participants and clippets. Prediction accuracy was quantified using the DOH metric (see Section 3.3.8.1). Error bars depict standard errors based on 1000 bootstrap subsamples (Efron & Tibshirani, 1993).

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 3.3.7 explains in detail why using a uniform distribution of locations may be more justified in the context of this study and in general. 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 DOH metric, we also utilized a percentile-based metric (see Section 3.3.8.2).
Figure 3.5 repeats the analyses shown in Figure 3.4, but using different metrics and baseline types. For brevity sake, Figure 3.5 focuses on the least and most predictive bottom-up models (see Figure 3.3).

![Saliency models](image)

**Figure 3.5.** Saliency effects as a function viewing time, baseline type and metric type.

(a) Similar to Figure 3.4, but focusing on the intensity variance model (least predictive bottom-up model, see Figure 3.3). Prediction accuracy was quantified by the percentile metric (see Section 3.3.8.2) using either the uniform baseline or the human-fixated baseline (see Section 3.3.7).

(b) Same as a, but using the DOH metric (see Section 3.3.8.1).

(c) Same as a, but focusing on the integrated saliency model (most predictive bottom-up model, see Figure 3.3).

(d) Same as b, but focusing on the integrated saliency model.

As in Figure 3.4, 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, shortly after
jump cuts, the prediction accuracy of the integrated saliency model is significantly higher than the corresponding prediction accuracy of the intensity variance model for all the tested, regardless of which metric type or baseline type is used. There are also dissimilarities compared to Figure 3.4. 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 with the uniform baseline, while the smallest differences were measured by the percentile metric with 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, 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 3.3.7). In summary, Figure 3.5 demonstrates that the key results presented in Figure 3.4 do not depend on either the metric type or the baseline type. As described in Sections 3.3.7 and 3.3.8, 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.
3.4.4. 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.

Figure 3.6a 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 ($t_{[10185]}=14.0763$, $p<<0.01$), when the analysis is based on saccades that led to minimal inter-observer variability (bounding rectangle area < 1% of the total display area). Figure 3.6a 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, Figure 3.6b 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 Figure 3.6b demonstrates that the integrated saliency model is 3.6 times better than the intensity variance model in predicting attentional selection ($t_{[10185]}=10.1349$, $p<<0.01$), 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).

Figure 3.6. Saliency effects as a function of inter-observer variability.

(a) Based on all the available saccades. Bin boundaries match those 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]; (5.70-53.46].

To summarize, Figure 3.7 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 Figure 3.6b). 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 relative increase in prediction accuracy for "Bottom-up" saccades, while the intensity variance and orientation contrast models show the opposite trend.

Figure 3.7. Saliency effects for "All" versus "Bottom-up" saccades.
The prediction accuracy for the "All" condition (black bars) was quantified as explained in Figure 3.3.

The prediction accuracy for the "Bottom-up" condition (light orange bars) is based on a subset of saccades that were initiated shortly after jump cuts, and led to maximal inter-observer similarity (equivalent to the first data point in Figure 3.6b).

3.5. Discussion

3.5.1. Bottom-up causes versus correlates of attentional selection

This study dissociated visual causes from mere correlates of attentional selection in realistic viewing conditions. This dissociation was accomplished based on the notion that causal bottom-up models should be most predictive of saccades that are most strongly driven by bottom-up influences. Specifically, we labeled particular saccade groups as more or less likely to be "bottom up driven", and measured the prediction accuracy of different bottom-up models as a function of this 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 Figure 3.7). 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.

3.5.2. Static versus dynamic bottom-up models

Another dissociation that emerges in Figure 3.7 is between static models with relatively low prediction accuracy and 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 present due to biological camouflage, which typically involves seamless blending into the background in terms of static features, such as shape and color (Curio, 1976)\(^7\). 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

\(^7\) Nature also shows examples of dynamic camouflage, as employed by dragonflies during territorial aerial manoeuvres (Mizutani et al., 2003), but these are relatively rare.
adaptation to detecting color contrasts, which may be particularly useful when searching for fruits embedded in foliage (Regan et al., 2001).

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

The results of this study, particularly Figure 3.4 and Figure 3.5, 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,
Figure 3.4 and Figure 3.5 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-acting 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 are open questions. Another 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 think of faces as 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 circuitry. 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.
3.5.4. Realism of stimuli used in studies of 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 movies. Furthermore, while the real world (apart from movies) 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. Head mounted displays could also achieve the same or better increase in realism, albeit more expensively.

3.5.5. 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., 1997a, 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 et al., 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 may 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 et al., 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 such an "inhibition-of-target" component to saliency models would be important for making them more predictive of the exact timing of saccades.

3.6. Acknowledgements

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 2 anonymous reviewers for their stimulating comments.
4. THE ROLE OF PERCEPTUAL MEMORY IN GUIDING ATTENTION

4.1. Abstract

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 plays 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.
4.2. 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 (Henderson & Hollingworth, 1999, Hernandez-Peon et al., 1956, James, 1890). Memory-free (bottom-up) influences are stimulus-centric factors, such as visual onsets (Gottlieb et al., 1998, Jonides & Yantis, 1988), 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 (Chun & Jiang, 1998, Hayhoe et al., 2003, Hollingworth & Henderson, 2002, Horowitz & Wolfe, 1998, Levin & Simons, 1997, Maljkovic & Nakayama, 1996, Melcher & Kowler, 2001, Najemnik & Geisler, 2005, Rensink, 2002).

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 (Hollingworth & Henderson, 2002, Melcher, 2001). Nevertheless, other results have indicated that the WOM theory may still hold in the particular context of attentional selection (Horowitz & Wolfe, 1998, Melcher & Kowler, 2001, Najemnik & Geisler, 2005). One suggestion is that attention targets are selected at random from an instantaneous pool of candidates (Horowitz & 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 & 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 a strong dominant effect for information accumulated during the last 1-2 fixations (Najemnik & 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 (Chun & Nakayama, 2000, Land & Furneaux, 1997) 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 conscious introspection, leading to the common failure of human observers to explicitly report changes in scenes (Rensink,
even when their eye-movement patterns are clearly affected by such changes (Hollingworth & 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 (Hayhoe et al., 2003, Land & Lee, 1994, Land & McLeod, 2000).

A general criticism of the traditional psychophysical approach for studying memory utilization is the questionable relevance of related conclusions to everyday life (Hayhoe et al., 2003, Land & Furneaux, 1997). 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 (Hayhoe et al., 2003, Land & Lee, 1994, Shinoda et al., 2001) 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 & 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 4.1). 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.

**Figure 4.1.** Continuous and MTV-style clips. Colored squares depict video frames. Different colors represent visually and semantically unrelated clip segments (clippets).
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 (Chun & Nakayama, 2000, Land & Furneaux, 1997), 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.

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 (Chun & Nakayama, 2000, Land & Furneaux, 1997), 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.
4.3. Methods

4.3.1. 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.

4.3.2. 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. The uniform distribution of clippet lengths minimizes the ability to anticipate the exact timing of jump cuts.

Continuous and MTV-style clips were matched in length, and each MTV-style clip contained at most one clippet from a given continuous clip. 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 matched 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.
(Anderson, 1996, Hochberg, 1986). 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 a few saccades. Incidentally, the term "jump cut" seems to mean different things to different people. We use it in the general sense of a cut that leads to a jump in continuity. A more specific meaning of "jump cut" is a cut between similar shots that leads to artifactual motion signals, such as a person or a face that appear to jump towards the camera.

4.3.3. 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
sometimes looked around the room instead of at the display monitor. 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.
4.3.4. Data acquisition and processing

Instantaneous position of the right eye was recorded using an infrared-video-based eye tracker (ISCAN RK-464, 240 Hz, see Figure 4.2). 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. The value of the saccade amplitude threshold was chosen somewhat arbitrarily (on the scale of the fovea), rather than based on the average visual angle of objects in our stimuli, which is unknown.
4.3.5. Attention-priority maps

Attention-priority maps were computed based on a saliency model (see Figure 4.3). 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 & 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 4.3. 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 & Nakayama, 1983, Gottlieb et al., 1998, Sillito et al., 1995), psychophysical (Polat & Sagi, 1994, Treisman & Gelade, 1980), and theoretical (Koch & Ullman, 1985) studies. The mathematical formulation of this model is described elsewhere (Itti & Koch, 2000).

4.3.6. 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 (Caspi et al., 2004, Parkhurst et al., 2002). 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 & 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 sampling the maximal local value in an aperture around each saccade target. The size of the aperture (r=3.15°) was chosen rather arbitrarily (on the scale of the parafovea) – we did not optimize the performance of the model by trying different aperture sizes. This issue seems peripheral, 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. The sampling procedure was identical to the one described above for human saccades. 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 & 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 (Itti, 2005, Krieger et al., 2000, Mannan et al., 1997a, Oliva et al.,

4.3.7. The ASH metric

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

$$ASH = \frac{1}{N} \sum_{i=1}^{N} S_i$$ (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.
4.3.8. 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:

\[
DOH = \left( \frac{1}{DOH_I} \right) \times \sum_{i=1}^{n} W_i \times (H_i - R_i)
\]  

(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:

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

(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 4.4), a certain fraction of random saccades would land on the maximal map value by chance, with probability:

\[ p = \frac{N_a}{N_m} = 0.0408 \]  \hspace{1cm} (4)

where \( N_a = 49 \) is the number of pixels in an aperture around the saccade target (\( r=3.15^\circ \), 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 4.4. 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 (Itti, 2005, Krieger et al., 2000, Mannan et al., 1997a, 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 (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.

4.4. 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 (Henderson & Hollingworth, 1999, Hernandez-Peon et al., 1956, James, 1890). 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 (see Figure 4.3) that has no memory other than
sensory persistence (Itti, 2005, Itti & Koch, 2000). The key design features that distinguish this particular saliency model from the available alternatives (Krieger et al., 2000, Mannan et al., 1997a, Oliva et al., 2003, Parkhurst & Niebur, 2003, Reinagel & Zador, 1999, Tatler et al., 2005, Torralba, 2003) 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 (Findlay & Walker, 1999, Gottlieb et al., 1998, Itti, 2005, Jonides & Yantis, 1988).

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 4.10.

4.4.1. Average bottom-up impact on attentional selection

Figure 4.5 shows a pair of saccades that straddle an MTV-style jump cut.

Figure 4.5. 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). 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 4.3).

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. Figure 4.6 shows another example of saccades that straddle an MTV-style jump cut.

Figure 4.6. Same as Figure 4.5, 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.
Figure 4.7A shows the distribution of saccades as a function of normalized saliency in the continuous viewing condition.

Figure 4.7. 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 4.7B). 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 4.7C). 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 4.7D 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 4.8). 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).
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 (Chun & Nakayama, 2000, Land & Furneaux, 1997), 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.

4.4.2. 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 4.9A):

(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 (Chun & Nakayama, 2000, Land & Furneaux, 1997) predicts that internal representations elicited by clipper X (before the jump cut) will be swiftly replaced by novel internal representations triggered by visual inputs from clipper 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 (Henderson, 2003, Wolfe et al., 2000). 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 clipper X interfere with the accumulation and utilization of novel representations based on clipper 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 clipper 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 4.9. 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 & 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 4.9B 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 cliplet 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 4.10A) and saccade index (see Figure 4.10B). 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.

Figure 4.10. Time course of bottom-up impact on attentional selection.
(A) Similar to Figure 4.9B, 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 4.9B, 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.

Interestingly, Figure 4.9 and Figure 4.10 also show late increases in bottom-up impact occurring 2.5 s after jump cuts (this trend is only significant in Figure 4.9B and Figure 4.10B, which may be attributable to the finer temporal resolution of these figures compared to Figure 4.10A). 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 Section 4.5.4: "Attention and scene understanding".

4.5. Discussion

The results of this study support and elaborate the IM theory (Chun & Nakayama, 2000, Land & Furneaux, 1997) 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 Figure 4.9 and Figure 4.10). 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 4.9 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 Section 4.5.4: "Attention and scene understanding".

4.5.1. 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 windowed sine-wave gratings embedded in static noise backgrounds (Najemnik & Geisler, 2005) may be quite different from challenges encountered during a visual search and discrimination task involving sparse arrays of simple shapes (Maljkovic & 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.

4.5.2. 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 & Dan, 2005, Rust & 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 conclusions. In this context, it is important to distinguish between general realism and real world relevance as it pertains to a particular research question. The jump cuts used here are informative for studying memory-driven attention exactly because they repeatedly perturb the balance between bottom-up and top-down influences. This perturbation is not entirely realistic in the general sense, given that real world visual stimuli do not seem to change as frequently and abruptly. That being said, there are compelling reasons to believe that the particular issue addressed here - the time frame in which perceptual memory guides attention between jump cuts - sufficiently resembles real world conditions:

1) Jump cuts are ubiquitous in motion pictures, even though people are often not aware of their occurrence (Anderson, 1996, Hochberg, 1986). The use of jump
cuts was pioneered by Jean-Luc Godard in his 1960 movie *Breathless*, and later popularized by MTV in the 1980s (Thompson & 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).

A potentially important distinction between the jump cuts used here and the ones people experience in the real world is the relatedness between stimuli across a jump cut. In fact, the craft of movie editing is based on manipulating visual and semantic relatedness across cuts, as well as the timing of cuts, in much finer detail than done here. This is a promising avenue for future research, as there is precious little scientific knowledge about why editing techniques work, or even how effective they are.

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, or plaids - the bread and butter of traditional psychophysics. To achieve sufficient control, researchers could either manipulate natural stimuli, or use software that can generate more realistic artificial stimuli. For example: video game engines can generate
highly complex dynamic scenes, while providing very specific knowledge of the stimulus content and properties.

4.5.3. 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 & Sejnowski, 1988, Olshausen & Field, 2005).

4.5.4. 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 demonstrated 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 4.9B) 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 4.9B are based on relatively fewer clips and contain fewer saccades than earlier temporal bins. For example: the last two temporal bins in Figure 4.9B 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 4.9B):
(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 (Henderson, 2003, Wolfe et al., 2000).

(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.
4.6. 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.
5. SUMMARY

This dissertation described a behavioral-computational approach to studying functional mechanisms of attentional selection during natural vision. To this end, the eyes of human observers were tracked non-intrusively as they watched either continuous or MTV-style (discontinuous) video clips. The main conclusions are based on a series of quantitative analyses, in which human gaze behavior and the prediction accuracy of related computational models were compared across viewing conditions, space, and time.

The research described in Chapter 2 was motivated by the observation that movie cuts give rise to physically unnatural visual discontinuities that are often imperceptible. This peculiar mismatch between physics and psychology has been explained in terms of either visual or mental continuity across movie cuts, but the related evidence is either anecdotal or inconsistent. The following questions were asked: what is the impact of movie cuts on patterns of human gaze behavior, and what do these patterns reveal about the perception of movies in particular and visual perception in general? To answer these questions, several model-free analyses were performed, comparing inter-observer similarity and other key parameters of gaze behavior across viewing conditions, space, and time. The results demonstrated that the human visual system adapts rapidly to changing conditions. Movie cuts that preserve neither visual nor semantic continuity across successive shots failed to confuse viewers, and instead synchronized their gaze behavior. To explain these results, it is argued that the physical dissimilarities between continuous real world
stimuli and discontinuous movies disappear at the sensory level due to the discontinuous nature of human eye-movements. According to a newly proposed "natural adaptability" hypothesis, the human visual system evolved the ability to keep the mind informed of the most pertinent information at every point in time without requiring either visual or mental continuity. Presumably, this adaptability has been necessary because continuity may often be absent or undesirable, either because of the natural discontinuity of sensations or due to rapid and pertinent real world changes, such as the sudden approach of predators. Chapter 2 also discussed the role of movie editing techniques and consistent narrative in blurring the sensory, perceptual, and cognitive boundaries between the human experience of movies and real world environments.

The research described in Chapter 3 delved deeper into the computational mechanisms of attentional selection. It addressed two basic questions: what are the visual causes, rather than mere correlates, of attentional selection and how do they compare to each other during natural vision? To answer these questions, a series of model-based analyses were performed, comparing the prediction accuracy of different models of attentional selection as a function of time and inter-observer similarity. The results demonstrated that dynamic visual correlates of attentional selection (e.g., motion contrasts) play a dominant causal role in attracting attention. In comparison, some static visual correlates (e.g., color contrasts) play a relatively weaker causal role, whereas others (e.g., orientation contrast) are non causal correlates, potentially reflecting top-down causes. Chapter 2 also proposed new
experiments and modeling enhancements that could further improve the ecological validity (real world relevance) of attention studies without sacrificing their explanatory and predictive powers.

The research described in Chapter 4 addressed the time frame in which perceptual memory guides attention, a controversial question among visual psychologists. To address this question, changes in the impact of top-down attention were measured across viewing conditions and over time by quantifying corresponding changes in the impact of bottom-up attention. The results demonstrated that perceptual memory is utilized for guiding attention across several gaze shifts when persistent visual context is available.

Chapter 4 also revisited an ongoing debate about the pros and cons of the artificial approach versus the natural approach to studying biological vision. It argued that studies of high-level vision that rely exclusively on the artificial approach are particularly sensitive to the pitfalls of ecological invalidity. Lastly, chapter 4 proposed a new set of hypothesis-driven experiments and a hybrid natural-artificial approach to study interactions between attentional selection and scene understanding.

The body of research described in this dissertation characterized key inputs, outputs, and associated computational transformations that underlie attentional selection during natural vision. It brings us closer to realizing the long term goals of this research, namely: to build intelligent machines that could behave autonomously in real world environments, and to better understand, diagnose, and treat medical
conditions, such as Autism, ADHD, and Parkinson's Disease, which are characterized by anomalous patterns of attentional selection.
ALPHABETIZED BIBLIOGRAPHY


