Available online at www.sciencedirect.com
ScienceDirect
Vision Research 46 (2006) 4333–4345
Vision Research
www.elsevier.com/locate/visres

Visual causes versus correlates of attentional selection in dynamic scenes

Ran Carmi *, Laurent Itti

Neuroscience Program, University of Southern California, USA

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

Abstract

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

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

1. Introduction

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

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

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

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

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

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

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

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

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

2. Methods

2.1. Participants

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

2.2. Stimuli

Fifty video clips (30 Hz, 640×480 pixels/frame, 4.5–30 s long, mean ± SD: 21.83 ± 8.41 s, no audio) from 12 heterogeneous sources, including indoor/outdoor daytime/nighttime scenes, video games, television programs, commercials, and sporting events. These continuous clips were cut every 1–3 s (2.09 ± 0.57 s) into 523 clip snippets (clippets), which were strung together by jump cuts into 50 scene-shuffled (MTV-style) clips (see Fig. 1 and Supp. Videos S1–S4). The range of clippet lengths was chosen such that observers would have enough time to perform several saccades within each clippet. The clippet lengths were randomized within the chosen range to minimize the ability of observers to anticipate the exact timing of jump cuts.

2.3. Experimental design

Observers inspected MTV-style video clips while sitting with their chin supported in front of a 22″ color monitor (60 Hz refresh rate) at a viewing distance of 80 cm ($28° \times 21°$ usable field of view). Their task was: "follow the main actors and actions, and expect to be asked general questions after the eye-tracking session is over". Observers were told that the questions will not pertain to small details, such as specific small objects, or the content of text messages, but would instead help the experimenters evaluate their general understanding of what they had watched. The purpose of the task was to let observers engage in natural visual exploration, while encouraging them to pay close attention to the display throughout the viewing session. The motivation for providing a task came from preliminary testing, in which instructionless free viewing sometimes led to observers disengaging from the display and looking around the room. A previous study found no task-related effects compared to free viewing observers who did not disengage from the display (Itti, 2005).

2.4. Data acquisition and processing

Instantaneous position of the right eye was recorded using an infrared-video-based eye tracker (ISCAN RK-464, 240 Hz), which tracks the pupil and corneal reflection. Calibration and saccade extraction procedures are described elsewhere (Itti, 2005). In this experiment, the calibration accuracy was $0.66° \pm 0.46°$ (mean ± SD), and a total of 10221 saccades were extracted from the raw eye-position data. Thirty-four saccades (0.3%) either started or ended outside of the display bounds, and were thus excluded from the data analysis, which was based on the remaining 10187 saccades.

2.5. Bottom-up attention-priority maps

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

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

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

where p refers to an individual image patch, m and n are its the width and in pixels (16×16, subtending $0.7° \times 0.7°$), I is the intensity of an image pixel, and $\bar{I}_p$ 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 a visual correlate of fixation locations (Parkhurst & Niebur, 2003; Reinagel & Zador, 1999).

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