Computational mechanisms for gaze direction in interactive visual environments

Robert J. Peters*
Department of Computer Science
University of Southern California

Laurent Itti†
Departments of Computer Science, Neuroscience and Psychology
University of Southern California

Figure 1: Four game frames with the eye position of one human (small cyan square, arrows) gazing back and forth between Mario and his enemy in a video game console system. Computational image-processing heuristics drawn from biological vision could help such systems predict and adapt to the behavior of their human operators.

Abstract

Next-generation immersive virtual environments and video games will require virtual agents with human-like visual attention and gaze behaviors. A critical step is to devise efficient visual processing heuristics to select locations that would attract human gaze in complex dynamic environments. One promising approach to designing such heuristics draws on ideas from computational neuroscience. We compared several such heuristics with eye movement recordings from five observers playing video games, and found that heuristics which detect outliers from the global distribution of visual features were better predictors of human gaze than were purely local heuristics. Heuristics sensitive to dynamic events performed best overall. Further, heuristic prediction power differed more between games than between different human observers. Our findings suggest simple neurally-inspired algorithmic methods to predict where humans look while playing video games.

CR Categories: H.1.2 [Models and Principles]: User/Machine Systems—Software Psychology I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Perceptual Reasoning I.3.3 [Computer Graphics]: Picture/Image Generation—Viewing Algorithms

*e-mail: rjpeters@usc.edu; website: http://ilab.usc.edu/rjpeters/

†e-mail: itti@usc.edu; website: http://ilab.usc.edu/

Keywords: visual attention, active vision, eye movements, computational modeling, video games, immersive environments

1 Introduction

Increasingly, many interactive graphics environments are starting to share a paradigm in which one or more human operators participate in a virtual world along with a potentially large number of virtual agents. This paradigm applies to mass-market video games, as well as high-end simulators/trainers for flying or driving, and large-scale combat simulators. As with computer systems in general, the shared environment is often "seen" in different ways by human and virtual agents. The human participants sense the environment through visual and auditory displays; in contrast, the virtual agents typically have direct access to some higher-level representation (such as an object-oriented scene graph) that generates the rendered visual and auditory signals. Thus, although human and virtual agents are operating in the same environment, they are not seeing it through the same eyes.

The next generation of interactive computer graphics systems will require artificial agents that can more faithfully mimic their human counterparts, while also being more aware of their human collaborators and adversaries. This shift may be driven by market forces in the case of video games, by safety concerns in the case of flying or driving simulators, or by increasing security concerns in the case of combat simulators or civil police simulators. One approach to this goal is to incorporate knowledge of biological vision into interactive graphics, using empirical results from eye-tracking and computational modeling. This offers two specific benefits: (1) the system could better predict the behavior of its human operator, allowing it to tailor the visual display appropriately (for example, by placing important information near the likely focus of attention, or by avoiding distracting displays during important moments), and (2) the

system could better mimic human behavior, allowing the virtual peers of the human operator to behave more naturally.

This empirical approach carries two main challenges: one is the difficulty in constructing a natural visual stimulus in a laboratory setting, and the other is the difficulty in implementing a computational algorithm that mimics the observed gaze behaviors. Most studies to date have addressed these challenges in complementary but disjoint ways. For example, it has been shown that humans preferentially gaze towards regions with multiple superimposed orientations (corners or crosses) [Zetzsche et al. 1998; Privitera and Stark 2000], higher than average spatial contrast (variance of pixel intensities) [Reinagel and Zador 1999] and entropy [Privitera and Stark 2000], and higher than average "saliency" [Parkhurst et al. 2002; Peters et al. 2005] as computed by a biologically-inspired model of bottom-up attention [Itti et al. 1998], yet these results are typically obtained with static scenes (still images) and address only low-level, bottom-up image properties. Adding a temporal dimension brings a number of complications in trying to realistically predict gaze targets with a computational model.

On the other hand, other studies have used naturalistic interactive or immersive environments to give high-level accounts of gaze behavior in terms of objects, agents, "gist," and short-term memory [Yarbus 1967; Henderson and Hollingworth 1999; Rensink 2000; Land and Hayhoe 2001; Sodhi et al. 2002; Hayhoe et al. 2003; Bailenson and Yee 2005], for example, to describe how task-relevant information guides eye movements during sandwich-making [Land and Hayhoe 2001; Hayhoe et al. 2003] or how distractions such as setting the radio or answering a phone affect eye movements while driving [Sodhi et al. 2002]. These studies have provided important constraints regarding goal-oriented high-level vision, but it remains to be seen how such high-level stimulus properties can be computed from raw visual input (*i.e.*, a time-varying 2-D pixel array). That is, the finding that people tend to orient towards "the radio," "an adversary," "the bread," "the jelly," or other real-world entities, which have been labeled beforehand by the experimenter and which are of particular interest during a given task, needs to be translated into computer algorithms that can first locate and identify these entities from raw video data. In fact, in interactive environments with natural tasks, it is intuitively expected that volitional top-down influences may more strongly guide gaze than reflex-like bottom-up influences, such that it is unclear whether computationally tractable bottom-up image analysis heuristics would play a significant role at all.

We report here a step toward addressing both of these challenges at once: we recorded observers' eye movements while they were exposed to dynamic visual input, in an interactive video game playing task (Fig. 1), and compared the recorded gaze behavior with the predictions of nine heuristic computational metrics based on low-level image features (such as color, intensity, motion) and saliency. We tested these metrics for their ability to predict human gaze targets and found that, despite the presumptive strength of top-down influences, all of the metrics scored significantly above chance. Those metrics that were sensitive to dynamic events were especially strong gaze predictors. Furthermore, we found that the metrics' predictive ability differs dramatically depending on the game-play paradigm, more so than on the particular human observer. Such metrics could be used as a first step toward endowing virtual agents with simple yet realistic visual systems.

2 Methods

2.1 Eye-movement recordings

Five subjects (three male, two female) participated with informed consent, under a protocol approved by the Institutional Review Board of the University of Southern California. Subjects played four or five five-minute segments of standard Nintendo GameCube games, including Mario Kart, Wave Race, Super Mario Sunshine, Hulk, and Pac Man World. Stimuli were presented on a 22" computer monitor (LaCie Corp; 640×480 pixels, 75 Hz refresh, mean screen luminance 30 cd/m^2, room 4 cd/m^2); subjects were seated at a viewing distance of 80 cm (28°× 21° usable field-of-view) and rested on a chin-rest. To allow later analysis with our computational heuristics, the video game frames were displayed and simultaneously recorded on a Linux computer under `SCHED_FIFO` scheduling to ensure microsecond-accurate timing [Finney 2001]. Each subject's right eye position was recorded at 240Hz with a hardware-based eye-tracking system (ISCAN, Inc.). Each game segment was preceded by a calibration sequence, in which subjects fixated a series of points on an invisible 3×3 grid, to set up a correspondence between the eye camera and stimulus display coordinate systems [Stampe 1993]. After calibration, subjects fixated a central cross and pressed a key to start the game play. In post-processing, 8,449 saccades of amplitude 2° or more were extracted from 1,740,972 recorded eye position samples over 24 five-minute game-play sessions. The corresponding 216,000 recorded video frames (about 185 GB of raw pixel data) were processed on a cluster with 48 CPUs through the computational heuristics described next.

2.2 Computational heuristics

We tested nine heuristic metrics for predicting gaze targets. These fall into two broad groups, one in which only local features within small image patches are considered, and another in which global context is used to determine whether a local area is unusual and thus worthy of interest. This second group of heuristics can be further subdivided into those that assess static features (which treat each frame individually, with no temporal context), and others that assess dynamic features. Finally, several metrics can be combined together to form a new metric that is different from the sum of its parts; in particular we test two combined metrics that correspond to "saliency." Fig. 2a shows heuristic responses for several sample video frames.

2.2.1 Local heuristics

We evaluate two purely local heuristics that compute local variance [Reinagel and Zador 1999] and local entropy [Privitera and Stark 2000] in 16×16 image patches, which have been previously shown to correlate with eye movements over still images. Variance for an image patch k is computed as $V(k) = (\sum_{(x,y)}(I(x,y) - \overline{I}_k)^2)/(N-1)$ where (x,y) loop over the $N = 16 \times 16$ set P_k of pixel coordinates that define patch k, $I(x,y)$ is the image intensity at (x,y), and $\overline{I}_k$ is the mean intensity over patch k. Entropy is similarly computed as $E(k) = -\sum_{i \in G(k)} F_i(k) \log F_i(k)$ where $F_i(k)$ is the frequency of occurrence of gray level i within the 16×16 patch k of interest, and $G(k)$ is the set of all gray levels present in the patch.

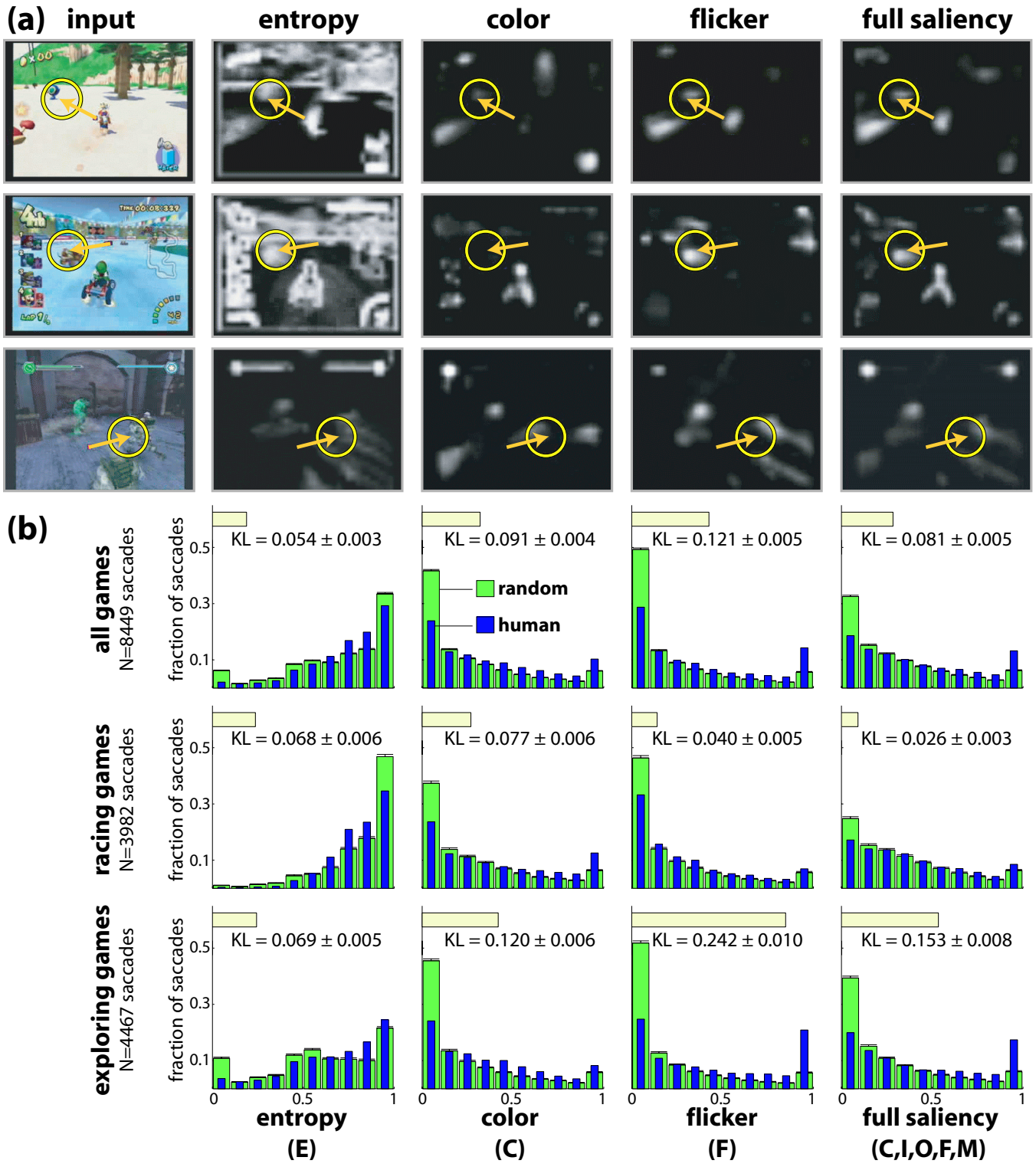

Figure 2: Scoring different metrics for their ability to predict human gaze targets during video-game playing. **(a)** The left column shows sample video frames from different sessions. In each frame, the subject is about to make a saccade from the point at the base of the arrow to the point at the tip of the arrow, surrounded by the circle. At the time of saccade onset, we sample the master maps generated by each metric (columns 2–5: entropy, color, flicker, and full saliency) at the location of the saccade target and at a number of uniformly random locations. **(b)** Across all saccades from 24 sessions with 5 observers, these histograms show how frequently locations with particular heuristic values are the targets of human saccades (narrow, dark bars), and of random saccades (light, wide bars). The Kullback-Leibler (KL) distance quantifies the dissimilarity between these distributions; higher KL values indicate that the metric is better able to distinguish human fixation locations from other locations. KL distances are indicated here by light horizontal bars atop each subpanel. The top row of subpanels shows heuristic scores across all game types, while the second and third rows show scores for racing games and exploration games, respectively. Only four of our nine metrics are shown here, but Fig. 3 summarizes the KL distances for all nine metrics.

2.2.2 Outlier-based heuristics

We also evaluate heuristics which respond to image outliers in visual feature dimensions known to play a role in human attention [Wolfe and Horowitz 2004]. Our software implementation of these heuristics has been previously described [Itti et al. 1998] and is freely available from the web (http://ilab.usc.edu/toolkit/).

Twelve neuronal features are implemented, sensitive to color contrast (red/green and blue/yellow, separately), temporal flicker (onset and offset of light intensity, combined), intensity contrast (light-on-dark and dark-on-light, combined), four orientations (0°, 45°, 90°, 135°), and four oriented motion energies (up, down, left, right). These features detect spatial outliers in image space, using a center-surround architecture inspired from biological receptive fields. Center and surround scales are obtained using dyadic pyramids with 9 scales (from scale 0, the original image, to scale 8, the image reduced by a factor 256). Center-surround differences are then computed as point-wise differences across pyramid scales, for combinations of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($\delta = \{3, 4\}$); thus, six feature maps are computed for each of the 12 features. Each feature map is additionally endowed with internal dynamics that provide a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition [Itti et al. 1998]. In this way, initially noisy feature maps can be reduced to sparse representations of only outlier locations which strongly stand out from their surroundings.

Here we combine the two color features into a single color channel, and similarly combine the four orientation features, and the four motion features. We hence evaluate five heuristics sensitive to outliers in the general dimensions of color, intensity, orientation, flicker, and motion, plus one that combines intensity, color, and orientation into a measure of static saliency, and one that adds motion and flicker to yield a measure of full static/dynamic saliency.

2.3 Scoring the heuristics

Each heuristic generates a topographic dynamic master response map, assigning a response value to every spatial location in each video frame, and a good master response map should highlight locations fixated by observers, more often than expected by chance. Thus, each heuristic is scored as follows: at the onset of each human saccade, we sample the heuristic's master map activity around the saccade's future endpoint, and around a uniformly random endpoint. We quantify differences between histograms of master map samples collected from human and random saccades using the Kullback-Leibler (KL) distance: heuristics which better predict human scanpaths exhibit higher KL distances, since observers typically gaze towards a non-uniform minority of regions with highest heuristic responses while avoiding the majority of regions with low heuristic responses (see Fig. 2b). The KL distance offers several advantages over simpler scoring schemes [Reinagel and Zador 1999; Parkhurst et al. 2002], including (1) being agnostic about the mechanism for selecting a saccade given the instantaneous distribution of heuristic responses over visual space, and (2) being invariant to reparameterizations, such that applying any continuous monotonic nonlinearity to master map values would not affect scoring.

3 Results

Fig. 3a shows results obtained overall. All metrics tested performed significantly above the chance-level KL score of zero (t-tests, $p < 10^{-100}$ or better). All KL scores significantly differed from one another, suggesting a strict ranking of all metrics (paired t-tests for equality of KL scores, $p < 10^{-100}$ or better, except for motion vs. flicker, $p < 10^{-7}$). Surprisingly, under our conditions of active game-play and with our video game stimuli, we found that motion alone (M) was the best predictor of human gaze, with flicker (F) a close second. Combining these with additional static features line color, intensity, and orientation to give the full saliency metric (C,I,O,F,M) actually impaired prediction performance. Orientation alone (O), intensity alone (I), and static saliency (C,I,O), indeed, were the poorest predictors, although still significantly above chance. Local heuristics computing variance (V) and entropy (E) scored respectably, but were surpassed by full saliency (C,I,O,F,M), color alone (C), flicker (F), and motion (M). Notably, color alone (C) scored better than the other static features, in contrast to results obtained previously with static images [Parkhurst et al. 2002], in which color was found to be the weakest gaze attractor.

We found two interesting dissociations within the scores obtained by the different heuristics. A split by game type (Fig. 3b) suggested an interesting dissociation between racing games (Mario Kart and Wave Race), and exploration games (Super Mario Sunshine, Hulk, Pac Man World), which involve a more open-ended storyline. Racing games involve primarily obstacle-avoidance while the player navigates a predetermined course. In contrast, exploration games impose looser restrictions on the player's navigation, but instead implicitly require the player to search the environment for objects or other agents with which to engage. All of the metrics we tested, except variance (V), scored better at predicting observers' gaze during exploration games than during racing games. In addition to this overall effect, we also found that the top-scoring heuristics were different for racing and exploring games. For the exploration games, the best heuristics were the dynamic outlier-based features motion (M) and flicker (F), followed closely by the static feature color (C). In contrast, for the racing games, the best heuristics were instead the local features entropy (E) and variance (V), again along with color (C).

In a second split, we found that there was less variability in heuristic performance across subjects than across games (Fig. 3c), especially so for the overall best metrics motion (M) and flicker (F). This is an important finding for two reasons. First, it suggests that heuristics could be further tailored to particular immersive environments, for example by finding the optimal weighted combination of all metrics for a given game type. Second, it may be reasonable to avoid such tailoring for individual human subjects, achieving an important savings within the limited resources of an interactive graphics system.

4 Discussion

In this study we have quantitatively evaluated nine simple computational metrics for their ability to predict where people look while playing video games. We found that all metrics score significantly above chance, hence representing easily computable shortcuts to the targets selected by human gaze. Although the predictive ability of these simple metrics is clearly limited, it is encouraging that they work at all. Our study provides direct experimental evidence that bottom-up

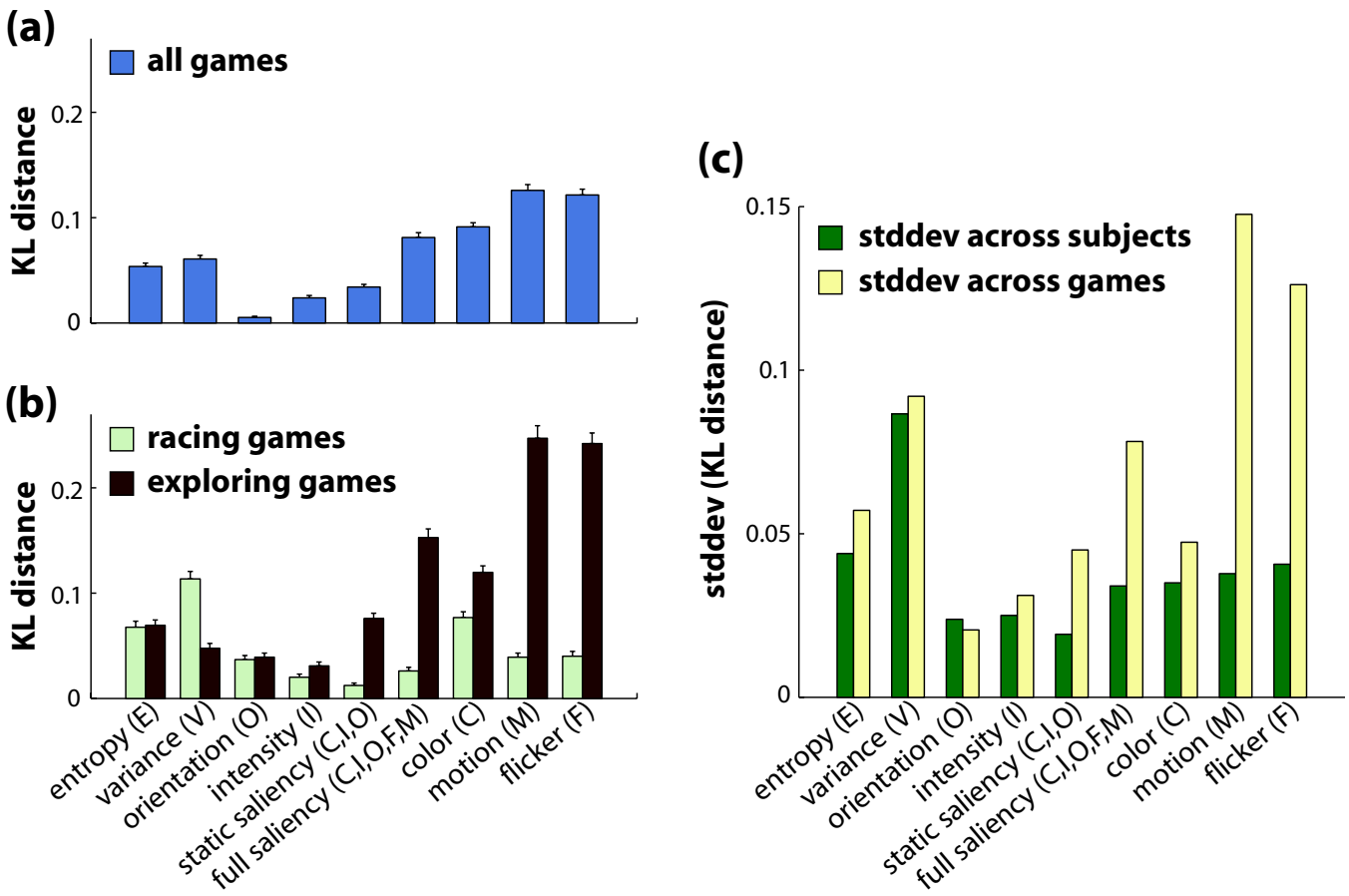

Figure 3: Nine heuristics were scored by the *KL* distance (section 2.3) to quantify how well the heuristic discriminates human gaze targets from random locations. **(a)** Across all game sessions, the dynamic features motion (M) and flicker (F) score better than static features color (C), intensity (I) and orientation (O), and local metrics entropy (E) and variance (V). **(b)** When the sessions were split by game paradigm, all metrics except variance were better predictors of human gaze in exploration games than an racing games. **(c)** Sessions were grouped either by subject or by game, then the average *KL* score was computed within each group, and finally the standard deviation was computed across groups. For all metrics except orientation, there was higher variability due to game than due to the subject.

image analysis can predict a non-negligible fraction of human gaze targets, even in situations where top-down influences are expected to dominate, driving eye movements based on task demands rather than on visual inputs. Our finding that the heuristics tested here significantly correlate with actual gaze position suggests that these heuristics could directly be used to endow virtual agents with gaze behaviors that would approximate human behavior. These results provide empirical validation that previously proposed visual systems for virtual agents indeed yield outputs that correlate with human behavior [Terzopoulos and Rabie 1997; Itti et al. 2003; Peters and O'Sullivan 2003]. Preliminary unpublished results suggest that the heuristics are better at predicting gaze when observers passively view video games than when they interactively play the games; thus, one direction for future study is to better understand this difference — is it due to bottom-up influences that are simply missed by our heuristics, or is it due to other (top-down, task-related) influences that are simply outside the scope of bottom-up heuristics?

A virtual agent could use such metrics to automatically compute where to look next in any virtual, real, or mixed environment. Although an analysis of a computer-graphics scene graph might allow prediction of potentially interesting gaze targets without requiring processing of fully rendered images, such an approach is limited to environments that are entirely virtual. Accounting for humans interacting with such environments becomes possible with heuristics that do not require knowledge of the scene graph and can instead operate at the pixel level.

Our main contributions have been two-fold. First, we studied *dynamic, interactive* video scenes whereas previous work on computational mechanisms for gaze direction has largely focused on still images or pre-recorded movies. Although the visual stimuli here were artificially generated, we believe they are similar enough to natural scenes to suggest that it is likely that the heuristics described here would also serve well in predicting gaze in truly natural scenes. Second, and perhaps most importantly, each of the heuristics that we tested has a tractable *computational implementation* that takes a time-varying 2-D pixel array as input and produces a time-varying prediction of where a human observer's gaze might be directed. Any apparent high-level behavior exhibited by the heuristic is simply attributable to low-level processing of the input; that is, the heuristics

do not require the visual input to be accompanied by explicit high-level labels for "objects" or "targets." Whereas behavioral studies have made great strides in understanding observers' gaze behaviors in just such object-based or agent-based terms [Land and Hayhoe 2001; Hayhoe et al. 2002; Hayhoe et al. 2003; Bailenson and Yee 2005], such findings become available to artificial visual systems only when there is an algorithm to extract the high-level information directly from the visual input. Accurate object recognition and semantic interpretation continues to be a "hard problem" in artificial vision, particularly for unconstrained visual input; while our approach has the drawback of lacking high-level scene understanding, it has the virtue of a straightforward computational implementation that can be applied to any visual input.

Acknowledgments

This work was supported by the National Geospatial-Intelligence Agency (NGA) and the Directorate of Central Intelligence (DCI). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

Bailenson, J., and Yee, N. 2005. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science in press*.

Finney, S. A. 2001. Real-time data collection in linux: A case study. *Behavior Research Methods, Instruments, and Computers 33*, 167–173.

Hayhoe, M., Ballard, D., Triesch, J., and Shinoda, H. 2002. Vision in natural and virtual environments. In *Proceedings of the symposium on Eye Tracking Research & Applications (ETRA)*, 7–13.

Hayhoe, M., Shrivastava, A., Mruczek, R., and Pelz, J. 2003. Visual memory and motor planning in a natural task. *Journal of Vision 3*, 1, 49–63.

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

Itti, L., Koch, C., and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11 (November), 1254–1259.

Itti, L., Dhavale, N., and Pighin, F. 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proceedings of SPIE 48th annual international symposium on optical science and technology*, 64–78.

Land, M., and Hayhoe, M. 2001. In what ways do eye movements contribute to everyday activities? *Vision Research 41*, 25-26, 3559–3565.

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

Peters, C., and O'Sullivan, C. 2003. Bottom-up visual attention for virtual human animation. In *Computer Animation and Social Agents 2003*, 111–117.

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

Privitera, C., and Stark, L. 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 9, 970–982.

Reinagel, P., and Zador, A. 1999. Natural scene statistics at the centre of gaze. *Network-Computation in Neural Systems 10*, 4 (November), 341–350.

Rensink, R. A. 2000. The dynamic representation of scenes. *Visual Cognition 7*, 17–42.

Sodhi, M., Reimer, B., Cohen, J., Vastenburg, E., Kaars, R., and Kirschenbaum, S. 2002. On-road driver eye movement tracking using head-mounted devices. In *Proceedings of the symposium on Eye Tracking Research & Applications (ETRA)*, 61–68.

Stampe, D. M. 1993. Heuristic filtering and reliable calibration methods for video based pupil tracking systems. *Behavior Research Methods, Instruments, and Computers 25*, 2, 137–142.

Terzopoulos, D., and Rabie, T. F. 1997. Animat vision: Active vision in artificial animals. *Videre: Journal of Computer Vision Research 1*, 1, 2–19.

Wolfe, J., and Horowitz, T. 2004. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience 5*, 6 (June), 495–501.

Yarbus, A. 1967. Eye movements during perception of complex objects. In *Eye Movements and Vision*, L. Riggs, Ed. Plenum Press, New York, NY.

Zetzsche, C., Schill, K., Deubel, H., Krieger, G., Umkehrer, E., and Beinlich, S. 1998. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In *From animals to animats, Proceedings of the fifth international conference on the simulation of adaptive behavior*, vol. 5, 120–126.