VISUAL COGNITION, 2005, 12 (6), 1093–1123

Psychology Press
Taylor & Francis Group

Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes

Laurent Itti

Departments of Computer Science, Pyschology, and Neuroscience Graduate Program, University of Southern California, Los Angeles, USA

We investigated the contribution of low-level saliency to human eye movements in complex dynamic scenes. Eye movements were recorded while naive observers viewed a heterogeneous collection of 50 video clips (46,489 frames; 4–6 subjects per clip), yielding 11,916 saccades of amplitude $\geq 2^{\circ}$. A model of bottom-up visual attention computed instantaneous saliency at the instant each saccade started and at its future endpoint location. Median model-predicted saliency was 45% the maximum saliency, a significant factor 2.03 greater than expected by chance. Motion and temporal change were stronger predictors of human saccades than colour, intensity, or orientation features, with the best predictor being the sum of all features. There was no significant correlation between model-predicted saliency and duration of fixation. A majority of saccades were directed to a minority of locations reliably marked as salient by the model, suggesting that bottom-up saliency may provide a set of candidate saccade target locations, with the final choice of which location of fixate more strongly determined top-down.

Humans and other mammals make extensive use of rapid eye movements to direct the highest-resolution region of their foveated eyes towards locations and objects of current behavioural interest. A productive approach for gaining a better understanding of foveated vision at the system level has been to use eyetracking devices, to evaluate image statistics at the locations visited by the eyes (Motter & Belky, 1998; Shen, Reingold, & Pomplun, 2000). In the context of natural scenes, most such research has thus far largely focused on characterizing local image properties at fixated locations. For instance, Zetzsche et al. (Barth, Zetzsche, & Rentschler, 1998; Zetzsche, Schill, Deubel, Krieger, Umkehrer, & Beinlich, 1998) inferred from human eyetracking recordings that the eyes preferentially fixated regions with multiple superimposed orientations,

Please address all correspondence to: L. Itti, Departments of Computer Science, Psychology, and Neuroscience Graduate Program, University of Southern California, Los Angeles, CA 90089-2520, USA. Email: itti@usc.edu

Supported by NSF, NEI, NIMA, the Zumberge Fund, and the Charles Lee Powell Foundation.

http://www.tandf.co.uk/journals/pp/13506285.html DOI:10.1080/13506280444000661

like corners. These authors then proceeded to derive nonlinear local operators to detect these regions. In a related study, Reinagel and Zador (1999) analysed the local grey-level distributions along eye scanpaths generated by humans while free-viewing greyscale images. They found the spatial contrast at fixated locations to be significantly higher than, on average, at random locations, and the pairwise pixel correlations (image uniformity) to be significantly lower. A complementary approach was proposed by Privitera and Stark (2000), who computed the linear combination of a collection of image processing operators (e.g., local cross detector, Laplacian of Gaussian, local entropy measure, etc.) that maximized overlap between regions of high algorithmic responses and regions fixated by human observers. Their study provided a new tool for creating feature detectors that would capture some of the local image properties that attracted the observer's attention and eye movements.

One extension of these approaches, further investigated here, consists of explicitly taking into account the long-range interactions across possibly distant locations in the visual field, as extensively documented in monkey electrophysiology and human psychophysics (Cannon & Fullenkamp, 1991; Gallant, Connor, & van Essen, 1998; Gilbert & Wiesel, 1989; Li & Gilbert, 2002; Sillito, Grieve, Jones, Cudeiro, & Davis, 1995; Polat & Sagi, 1993). That is, instead of attempting to characterize fixated locations solely in terms of their local image properties, computational models have been applied to correlate human fixations to measures of context-dependent saliency (conspicuity) of objects embedded within background clutter. A recent example is the study of Parkhurst, Law, and Niebur (2002), which compared eye movement scanpaths recorded from human observers presented with static colour scenes to a saliency map derived from a computational model of bottom-up attention (Itti, Koch, & Niebur, 1998). An important aspect of the model used in this study is that it explicitly incorporates a spatial competition for saliency, by which foci of high neural activity at possibly distant locations in the visual field compete for representation in the saliency map (Itti & Koch, 2001b). Parkhurst et al. found a significantly elevated model-predicted saliency at human fixations compared to random fixations (also see Peters, Itti, & Koch, 2002, for related results), an effect which was stronger for the first few fixations on a given image than for subsequent fixations.

Bottom-up saliency is only one of the many factors that contribute to the spatiotemporal deployment of attention and eye movements over complex dynamic visual scenes (Henderson & Hollingworth, 1999; Itti & Koch, 2001a; Rensink, 2000). The rapid understanding of the scene's gist and rough layout is also thought to provide priors on where objects of current behavioural interest may be located and to facilitate their recognition; for example, a vase is likely to be located on a tabletop (Friedman, 1979; Hollingworth & Henderson, 1998; Oliva & Schyns, 1997; Potter & Levy, 1969; Torralba, 2003), but see (Biederman, Teitelbaum, & Mezzanotte, 1983). Further, as specific objects are

searched for, low-level visual processing may be biased for the features of these objects (Ito & Gilbert, 1999; Moran & Desimone, 1985; Motter, 1994; Reynolds, Pasternak, & Desimone, 2000; Treue & Martinez Trujillo, 1999; Treue & Maunsell, 1996; Yeshurun & Carrasco, 1998). This top-down modulation of bottom-up processing results in an ability to guide search towards targets of interest (Wolfe, 1994, 1998; Wolfe, Cave, & Franzel, 1989). In more general terms, task has been widely shown to affect eye movements (Peebles & Cheng, 2003; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Yarbus, 1967), and training and general expertise as well (Moreno, Reina, Luis, & Sabido, 2002; Nodine & Krupinski, 1998; Savelsbergh, Williams, van der Kamp, & Ward, 2002). In addition, eye movements may often be executed in a purely volitional manner; for example, recognizing one object and knowing its typical spatial relationship to another target object may allow an observer to orient towards the target. Reading is an important example where eye movements are strongly influenced by the known spatial arrangement of words on a page and by sentence parsing and comprehension processes (Legge, Hooren, Klitz, Mansfield, & Tjan, 2002; Reichle, Rayner, & Pollatsek, 2003). Similarly, memorizing the locations of objects found along an initial exploratory scanpath on a new scene may allow observers to later orient back to some of these objects in an efficient manner (Henderson & Hollingworth, 2003; Hollingworth, Williams, & Henderson, 2001). Finally, eye movements from different observers exhibit different idiosyncrasies, which may result from possibly different internal world representations (Noton & Stark, 1971), different search strategies, and other factors (Andrews & Coppola, 1999).

Given all the factors that contribute to eye movements, it is hence expected and generally widely accepted that top-down influences may much more strongly determine gaze allocation than bottom-up influences. Our focus in this study is an attempt to quantify the contribution of low-level saliency to human scanpaths over complex dynamic scenes, and to analyse which low-level visual features contributing to saliency (such as colour, intensity, motion) are more strongly correlated with human eye movements. Our main contributions are two-fold: First, we study dynamic scenes (video clips); previous work has focused on static scenes (still images). Adding a temporal dimension brings a number of complications in trying to realistically predict saliency with a computational model. Also, when viewing natural video scenes in a realistic setting, top-down influences certainly are expected to be the largely predominant factor in determining the next fixation location, such that it is unclear whether bottom-up saliency would play a significant role at all. Second, we push the analysis beyond the observation that, overall, model-predicted saliency is elevated at the locations of human eye fixations, in particular by analysing the distribution of saliency at eye fixations and by attempting to tease out which low-level visual features were strongest at locations fixated by humans.

METHODS

We recorded eye movements from human observers watching a heterogeneous collection of 50 video clips, and segmented the recordings into saccade, blink and fixation/smooth-pursuit periods. For each human recording, we then employed a neurobiological model of bottom-up attention to explicitly predict bottom-up saliency at every scene location. At the beginning of each human saccade, we sampled the model-predicted saliency around the location of the future endpoint of that saccade. Our analysis focuses on comparing these saliency samples at saccade targets to maximum and randomly sampled saliency measures. These measures were analysed to determine the extent to which humans tended to saccade towards locations in their visual field that the model predicted were salient.

Our purpose in this study is not to attempt to maximize agreement between humans and model, but to quantitatively estimate the contribution of bottom-up processing to everyday eye movements. We hypothesized that if saliency was higher at saccade targets compared to other locations in the visual field, it would suggest that saliency may have contributed to choosing the saccade targets. Because of the many other factors mentioned above that influence eye movements, most of which are generally considered to be top-down, our starting hypothesis was that the fraction of eye movements directed to the most salient location in the scene would probably be very small—that is, top-down factors would dominate in determining the spatiotemporal allocation of gaze. One of our goals was to estimate that fraction in a quantitative manner. Here we assumed that our available computational model was a reasonable approximation of bottom-up processing, but we did not attempt to tune it to the specific data used in this study.

Human eye movement recording

Human subjects were naïve to the purpose of the experiment and were USC students and staff (three females, five males, mixed ethnicities, ages 23–32, normal corrected or uncorrected vision). They were instructed to watch the video clips, and to attempt to follow the main actors and actions presented in the clips, as they would later be asked some general questions about what they had watched. It was emphasized that the question would not pertain to small details (e.g., specific small objects, colours of clothing, or text messages), but would instead help us evaluate their general understanding of the contents of the clips. These instructions had two main purposes. First, we avoided instructionless free viewing, since in preliminary testing this had often yielded largely idiosyncratic patterns of eye movements, as subjects would lose interest and disengage from the experiment over time. Second, as mentioned in our introduction, our purpose

in this study was not to attempt to maximize agreement between human eye movements and model predictions, but rather to quantify the extent to which low-level bottom-up saliency computed from responses of very simple feature detectors may contribute to the guidance of eye movements in normal viewing conditions. This is why we asked subjects to ''follow the main actors and actions'' rather than to ''attend to the most salient image locations'', and why we did not explain to subjects our computational notion of bottom-up saliency. While the notion of main actor is essentially a top-down one, requiring an evaluation of all actors and a decision of which one is the most interesting, our data hence quantifies the extent to which low-level saliency may bias this decision. While the bulk of our results should be interpreted carefully given our task definition, we also tested one control subject to whom no other instructions than to watch the video clips and enjoy were given (subject PA, not included in any of the group analyses). The procedure was approved by USC's Internal Review Board, and informed consent was obtained from all subjects. A set of calibration points and clips not part of the experiment were shown to familiarize the subjects with the displays.

Stimuli were presented on a 22-inch computer monitor (LaCie Corp; 640 $\times$ 480, 60.27 Hz double-scan, mean screen luminance 30 cd/m^2, room 4 cd/m^2). Subjects were seated on an adjustable chair at a viewing distance of 80 cm (28° $\times$ 21° usable field-of-view) and rested on a chinrest. A nine-point eye-tracker calibration was performed every five clips. Each calibration point consisted of fixating first a central cross, then a blinking dot at a random point on a 3 $\times$ 3 matrix. The experiment was self-paced and subjects could stretch before any nine-point calibration. For every video clip, subjects fixated a central cross, pressed a key to start, at which point the eyetracker was triggered, the cross blinked for 1206 ms, and the clip started. Stimuli were presented on a Linux computer, under SCHED_FIFO scheduling (process would keep 100% of the CPU for as long as needed; Finney, 2001). Each clip (MPEG-1 encoded) was entirely preloaded into memory. Frame displays were hardware-locked to the vertical retrace of the monitor (one movie frame was shown for two screen retraces, yielding a playback rate of 30.13 fps). Microsecond-accurate (Finney, 2001) timestamps were stored in memory as each frame was presented, and later saved to disk to check for dropped frames. No frame drop ever occurred and all timestamps were spaced by 33,185 $\pm$ 2 μs.

Eye position was tracked using a 240 Hz infrared-video-based eyetracker (ISCAN, Inc. model RK-464). Point of regard (POR) was estimated from comparative tracking of both the centre of the pupil and the specular reflection of the infrared light source on the cornea. This technique renders POR measurements immune to small head translations (tested up to $\pm$10 mm in our laboratory). Thus, no stricter restraint than a chinrest was necessary, which is important as head restraint has been shown to alter eye movements (Collewijn,

Steinman, Erkelens, Pizlo, & van der Steen, 1992). All analysis was performed offline. Linearity of the POR-to-stimulus coordinate mapping was excellent, as previously tested using a 7 × 5 calibration matrix, justifying the use of a 4 × 3 matrix here. The eyetracker calibration traces were filtered for blinks, then automatically segmented into two fixation periods (the central cross, then the flashing point), or discarded if that segmentation failed a number of quality control criteria. From the nine (or fewer) resulting calibration points, an affine POR-to-stimulus transform was computed in the least-square sense, outlier calibration points were eliminated, and the affine transform was recomputed. If fewer than six points remained after outlier elimination, recordings were discarded until the next calibration. A thin-plate-spline nonlinear warping algorithm was then computed to account for any small residual nonlinearity (Bookstein, 1989). Only 25% of the thin-plate-spline deformation was applied to the data (0% would induce no warping, 100% would ensure a zero residual error on all calibration points), as observers often slightly jittered around a given calibration point. Thus, we assumed almost perfect linearity of the tracker, but possible noise in individual calibration traces. Data was discarded until the next calibration if residual errors greater than 20 pixels on any calibration point or 10 pixels overall remained. Eye traces for the five video clips following a calibration were remapped to screen coordinates, or discarded if they failed some 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). Calibrated eye traces were visually inspected when superimposed with the corresponding video clips, but none was discarded based upon that subjective inspection.

Fifty video clips were selected from a database of eighty-five, with the only selection criterion of maximizing diversity. All clips had been digitized from analogue interlaced NTSC video sources using a consumer-grade framegrabber (WinTV Go, Hauppage, Inc.) and no attempt was made at deinterlacing or colour-correcting them. The clips included:

beverly: Daytime outdoors scenes filmed at a park in Beverly Hills
gamecube: Various video games (first-person, racing, etc.)
monica: Outdoors day/night scenes at the Santa Monica Promenade
saccadetest: A synthetic disk drifting on a textured background
standard: Daylight scenes filmed at a crowded openair rooftop bar
tv-action: An action scene from a television movie
tv-ads: Television advertisements
tv-announce: A television programmes announcement
tv-music: A music video interleaved with some football scenes
tv-news: Various television newscasts
tv-sports: Televised basketball and football games
tv-talk: Television talk-shows and interviews.

Each clip comprised between 164 and 2814 frames (5.5–93.9 s), for a total of 46,489 distinct frames (25 min and 42.7 s). Selection of clips was made prior to the experiments and was not altered after recording and data analysis. Subjects viewed each clip at most once, to ensure that they were naïve to its contents. Five subjects viewed all clips and three only viewed a few; after our quality-control criteria were applied, calibrated eye movement data was available for four to six subjects on each clip, yielding a total of 235 calibrated eye movement traces.

The calibrated eye movement data was further segmented into saccades, eye blinks, and fixation/smooth pursuit periods (Sparks, 2002). In a preprocessing step, eye blinks were eliminated from the recordings. Blinks were characterized in our recordings by a measured pupil diameter falling to zero. These periods were marked in the datasets, so that visual input to the model would be turned off during blinks. Our dataset contained occurrences of blinked saccades, that is, blinks (as defined by the measured pupil diameter of zero) with distant eye position at the end of the blink period compared to its beginning. This slightly complicated the extraction of saccades from the dataset, as, typically, high-velocity saccadic eye motion preceded and followed these particular blinks (that is, these blinks occurred within a fraction of the temporal extent of a saccade). We note that, behaviourally, blinking during saccades seems an efficient strategy in terms of minimizing the total amount of time during which visual information will not be reliably exploitable by the central nervous system. For this reason, in our preprocessing, we linearly interpolated eye position during these blinked saccades, so that they could be detected as legitimate saccades in the next processing step. Similarly, we also linearly interpolated eye position during normal blinks, so that fixation periods could be reliably detected. Thus, the output of this preprocessing step consisted of fully continuous eye position traces, with blink periods marked so that visual input would be turned off in the model.

The segmentation of saccades was based on instantaneous eye velocity computed from a low-pass-filtered version of the blink-interpolated eye movement traces. Although more sophisticated methods have been proposed (Salvucci & Goldberg, 2000), this simple approach worked reliably with our high sampling rate of 240 Hz. Periods of the recordings where smoothed eye velocity exceeded $20^\circ/s$ were marked as candidate saccade periods. The eye position at the end of each saccade was considered the saccade target location in our analysis. Candidate saccades of amplitude smaller than 2° (a threshold chosen rather arbitrarily) were not considered as saccades in our analysis, but were instead reverted to the pool of fixation/smooth pursuit periods. The rationale for this decision to not consider very small saccades was that they were unlikely to indicate a shift of foveation from one object to another (possibly salient) object, but rather, we hypothesized, would mainly reflect small adjustments of eye position onto various parts of a same object (no matter its saliency). Thus, very

small saccades would be of little interest in determining whether humans tended to orient towards salient objects or not.

In summary, our data recording and segmentation yielded a total of 11,916 saccades, for each of which the starting eye position, starting time, and target location were known.

Bottom-up attention model

The model computes a topographic saliency map (Figure 1), which quantitatively predicts how conspicuous every location in the input scenery is. Its implementation has been previously described in details (Koch & Ullman, 1985; Itti & Koch, 2000, 2001a; Itti et al., 1998). Retinal input is processed in parallel by 12 multiscale low-level feature maps, which detect local spatial discontinuities using simulated centre-surround neurons (Hubel & Wiesel, 1962; Kuffler). The 12 neuronal features implemented are sensitive to colour 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) (Itti & Koch, 2001a; Itti et al., 1998). Centre and surround scales are obtained using dyadic pyramids with nine levels (from level 0, the original image, to level 8, reduced by a factor 256 horizontally and vertically). Centre-surround differences are then computed as pointwise differences across pyramid levels, for combinations of three centre scales ($c = \{2, 3, 4\}$) and two centre-surround scale differences ($\delta = \{3, 4\}$); thus, six feature maps are computed for each of the 12 features, yielding a total of 72 feature maps. After long-range competition for salience, all feature maps are summed (Itti & Koch, 2001b) into the unique scalar saliency map that guides attention. The saliency map is at scale 4 (40×30 pixels given our 640×480 video frames).

Each feature map is endowed with internal dynamics that operate a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition (Itti & Koch, 2001b). These long-range within-feature competitive interactions are a crucial component of our computational framework, as they are mainly responsible in the model for ''pop-out'' effects observed behaviourally in visual search psychophysics (Treisman & Gelade, 1980; Wolfe, 1998). In a first approximation, the computational effect of these interactions is similar to applying a soft winner-take-all to each feature map before it is combined into the saliency map (Itti & Koch, 2001b). Consider for instance a simplistic scene with a single red disc and a single blue disc on a black background. Our model would predict that both discs are equally salient, if the red and blue colours are matched for luminance and their chrominances are optimally tuned to the model's red/green and blue/yellow feature detectors, respectively. The red disc would be mainly represented in the red/green feature maps with approximately matching centre-surround receptive field sizes, while

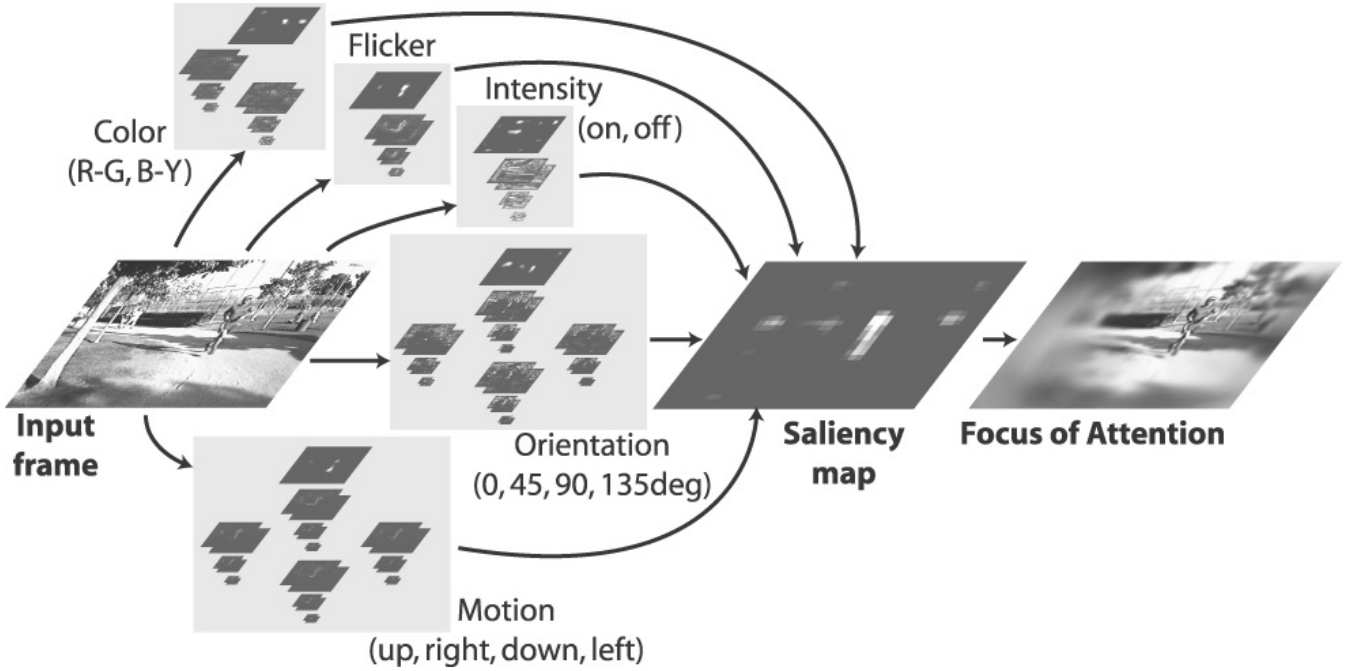

Figure 1. Overview of the model. Low-level visual features are computed from incoming input frames in a set of multiscale feature maps tuned to colour opponencies, intensity contrast, orientation contrast, temporal change (flicker) and oriented motion energy. After strong nonlinear within-feature competition for saliency, all feature maps provide input to the unique saliency map.

the blue disc would mainly excite blue/yellow feature maps with matching receptive field sizes. Starting from this baseline image, now consider the addition of three additional red discs identical to the first one and placed at distant locations from the initial two discs. The model's long-range interactions in the red/green feature maps would yield competition among the four red discs present in this new display, resulting in reduced saliency for each of the red discs. In contrast, no such competition would be triggered in the model's blue/yellow feature maps, and the blue disc would remain as salient as it initially was. Resultingly, after the various feature maps are summed into the saliency map, the blue disc would be more salient than any of the red discs, and would be the most salient location in the display (pop-out effect). Thus, although our low-level feature detectors do respond to local image properties, the addition of nonlinear competition for saliency in our model renders the predicted saliency of a given object dependent on every other object and on background clutter present in the input images (Itti & Koch, 2001b).

We have previously proposed several implementations for these long-range interactions (Itti & Koch, 2000, 2001b; Itti et al., 1998), yielding saliency maps with different overall sparseness (Itti, 2004). While Parkhurst et al. (2002) used fairly weak interactions, yielding relatively dense saliency maps where many salient locations are highlighted (Itti et al., 1998), here we use one iteration of the iterative scheme proposed by Itti and Koch (2001b), which yields much sparser saliency maps that typically highlight only 3–5 locations in the entire image.

The saliency map itself is modelled as a two-dimensional layer of leaky integrator neurons (Itti & Koch, 2000). It receives as input the sum of all feature maps, after that sum is subjected one more time to nonlinear competition for saliency. Its main function is to provide temporal smoothing of the saliency values, and no further interactions across distant locations are implemented at this stage. Consequently, in this study we directly evaluate the saliency values at current human eye position, even though some reaction delay exists between appearance of a stimulus element and the possible decision to orient the eyes towards that location. Thus, we here do not investigate in detail the effects of stimulus-to-response latency, but instead temporally low-pass filter the saliency map. The integrators are simulated using difference equations at a temporal resolution of 0.1 ms of simulated time (Itti & Koch, 2000).

Comparing model predictions to human eye movements

The computational model was used in the following manner to evaluate bottom-up saliency at the target locations of human saccades. Each of the 235 human eye movement recordings was considered in turn. For a given recording, the

corresponding video clip was presented to the model, at the same resolution and framerate as it had been presented to human subjects.

It is important to note that in this study we do not attempt to replicate many of the details of human foveated active vision. In particular, the video images are processed in their native frame of reference (as opposed to a retinal frame of reference that would be centred at current human eye position), and previously explored model components (Itti, Dhavale, & Pighin, 2003) for foveation, overt inhibition-of-return (IOR), and saccadic suppression were not used here. The rationale for this decision was that some of these factors may have top-down components (e.g., IOR) which would make the interpretation of our results in terms of bottom-up processing more difficult. Rather than employing a hybrid framework (e.g., including foveation but not IOR nor coordinate transforms) like, for example, Parkhurst et al. (2002) did, in the present study we attempt to compute a pure form of stimulus saliency, which is an intrinsic property of the video stimuli and is independent of the observer. It is understood that the actual bottom-up computations performed by humans observers are modulated by foveation, coordinate transforms, short-term visual memory, and other factors, so that the internal representation of saliency may substantially differ from the intrinsic image saliency computed here. Hence the saliency computed here does not so much attempt to replicate the internal representation of saliency in the human brain than the external intrinsic stimulus saliency.

RESULTS

At the beginning of each human saccade, several measurements were made on the model's current internal state, to quantify the extent to which that human saccade was aimed towards a location of high model-predicted saliency. The following samples were taken:

- S_h: Saliency at human eye position, computed as the maximum over a circular aperture of diameter 5.6° (nine pixels in the saliency map) of the model's dynamical saliency map sampled at the moment a saccade starts and around the location of the future endpoint of that saccade. Thus, a high value of S_h indicates that, when it starts, a saccade is targeted towards a location that is highly salient at that moment.
- S_r: Saliency at a random location, computed in exactly the same manner as S_h except that a random endpoint within the image (with uniform probability) is considered rather than a given human saccade endpoint. Thus, a high value of S_r indicates that, at the moment a saccade starts, saliency is high around a randomly chosen location within the image. This provides a control for the evaluation of human saccades, by estimating the overall spatial density of active locations in a saliency map, i.e., the degree of sparseness or selectivity of the map.

- S_{max}: Maximum saliency over entire frame, computed as the maximum over the spatial extent of the entire dynamical saliency map, at the same instant as the other measurements were taken. This provides a very strict test on the saliency of a saccade target, in that $S_h = S_{max}$ if and only if the saccade is directed towards the currently most salient location in the entire visual field. It also allows us to compare saliency values across frames and video clips, by normalizing S_h, and S_r relative to S_{max}.

Analysis by saccade amplitude

A first observation concerns the distribution of saccade amplitudes, which shows a strong bias for shorter saccades (Figure 2). Given this bias, we next asked whether saccade length would be a factor in our measures of saliency at the endpoint of saccades. Figure 3a shows a histogram of the ratios of saliency sampled, at the time each human saccade began, around the location of the future saccade endpoint (S_h), compared to the maximum saliency over the entire saliency map at the same moment (S_{max}). For comparison, the histogram of the same ratios but using the saliency at random saccade endpoint locations (S_r) is

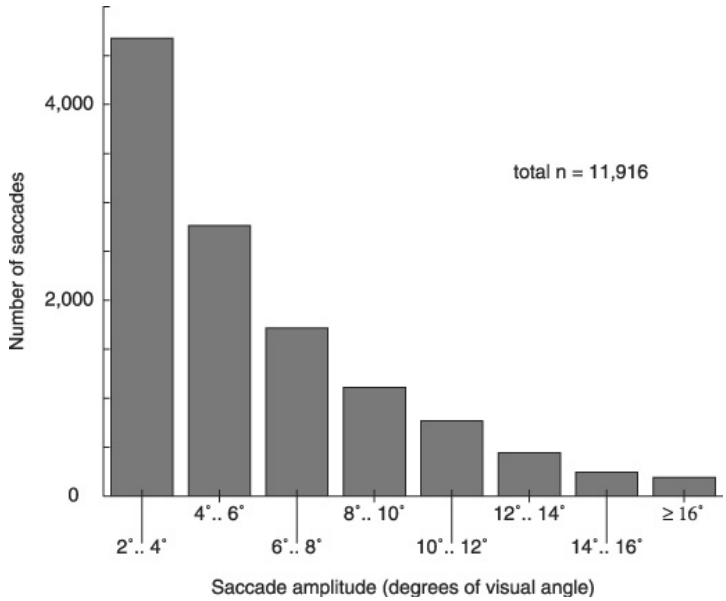

Figure 2. The distribution of observed saccade amplitudes during our experiments shows a strong bias for shorter saccades. Saccades shorter than 2° were treated as fixation/smooth-pursuit in our analysis. The count for each bar is inclusive of the lower bound but exclusive of the higher bound (e.g., the bar labelled 8°.. 10° shows the number of saccades where 8° ≤ amplitude < 10°).

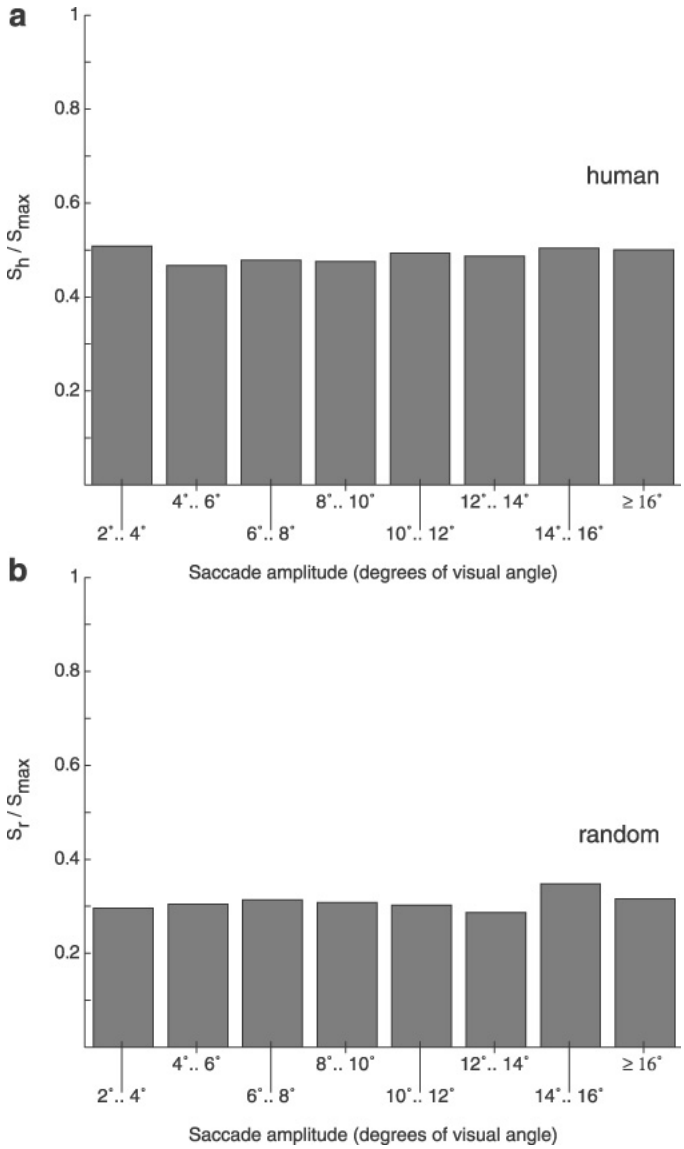

Figure 3. Distribution, by saccade amplitude, of the ratios of saliency at human saccade target locations to the maximum saliency (S_h/S_{max}; panel a), and of the ratios of saliency at random locations to the maximum saliency (S_r/S_{max}; panel b). For each saccade, all measurements were taken at the moment the saccade began. Overall, S_h/S_{max} was higher than S_r/S_{max}, in a highly significant manner (see text). Shorter and longer human saccades show slightly higher S_h/S_{max} (significant effect in a one-way ANOVA; see text), but random saccades do not. Note that the histogram in panel b is indexed by human saccade amplitude like that of panel a (so that a given amplitude range includes the same number of saccades in both panels), not by random saccade amplitude.

shown in Figure 3b. Looking at the first order statistics of the plotted distributions, saliency around saccade endpoints was higher than around random locations, with a median S_h/S_{max} of 0.45 over all 11,916 saccades, compared to a median S_r/S_{max} of 0.22 (Table 1). A paired-sample nonparametric sign test rejected the null hypothesis that the samples S_h/S_{max} and the samples S_r/S_{max} could have been drawn from distributions with same medians, in a highly significant manner ($p < 10^{-232}$). This difference was also highly significant for all saccade amplitude ranges shown in Figure 3. That is, the sign test suggested a significant difference with $p < 10^{-10}$ or better also when applied only to saccades with amplitudes between 2° and 4°, as well as when applied to saccades with amplitudes in any of the ranges shown in Figure 3. A one-way ANOVA suggested that human saccade amplitude (binned into the eight ranges of Figure 3) had a significant effect onto S_h/S_{max}, $F(7, 11908) = 5.22$, $p < 10^{-5}$, eight groups and 11,916 saccades, but not onto S_r/S_{max}, $F(7, 11908) = 1.86$, n.s. Indeed, very short and very long saccades tended to exhibit slightly higher S_h/S_{max}, but that trend was not present in the S_r/S_{max} data.

A breakdown of the data into various subsamples is shown in Table 1. Overall, for all subsamples except one (the clip tv-announce), median S_h/S_{max} was still significantly elevated compared to median S_r/S_{max}. This breakdown as well as the previous breakdown by saccade amplitude suggest that our main finding of an overall elevated S_h/S_{max} compared to S_r/S_{max} was not primarily driven by a subset of the recorded saccades; instead, significant differences between humans and random were found for all saccade amplitudes, all observers, and all but one classes of video clips. This suggests a strong contribution of intrinsic image saliency, as defined by our computational model, to human eye movements in a wide range of situations. In particular, the first, first five, and first ten saccades onto a given video clip corresponded to higher ratios of medians, indicating a stronger correlation between bottom-up saliency and human saccade targets during the first few seconds of viewing a new video clip. Further comparisons between the various subgroups of saccades presented in Table 1 are presented in the following sections.

Our control subject (PA in Table 1), studied independently and not part of any of the group analyses, was given no instruction to follow the main actors and actions in the video clips, yet fell within the range of variability of other subjects. However, this subject showed slightly higher overall saccade frequency (approximately 2.4 saccades/s on average while the average of all other subjects was approximately 1.7 saccades/s). This seems reasonable as the instruction of ''following'' the main actors and actions may have incited the other subjects to lock onto these scene elements for extended periods of time.

This first set of results confirms with dynamic scenes previous studies which suggested that, overall, humans look are image regions of higher saliency than would be expected by chance (Parkhurst et al., 2002). To summarize, median S_h/S_{max} was about twice the median S_r/S_{max}, a factor significantly greater than unity.

TABLE 1
Average and median S_h/S_{max} and S_r/S_{max} for various subsamples of our overall pool of saccades

Samples	*n*	*Median* S_h/S_{max}	*Median* S_r/S_{max}	*Sign test*	*Ratio of medians*	*Kullback-Leibler distance*
All	11,916	0.45	0.22	$p < 10^{-232}$	2.03	0.195 ± 0.006
First	235	0.57	0.24	$p < 4 \times 10^{-16}$	2.32	0.414 ± 0.062
First 5	1,173	0.48	0.21	$p < 3 \times 10^{-52}$	2.31	0.279 ± 0.021
First 10	2,322	0.49	0.22	$p < 2 \times 10^{-90}$	2.25	0.278 ± 0.015
CZ	2,369	0.46	0.22	$p < 5 \times 10^{-66}$	2.12	0.219 ± 0.016
JC	1,527	0.41	0.23	$p < 8 \times 10^{-34}$	1.76	0.147 ± 0.011
JZ	1,965	0.44	0.21	$p < 9 \times 10^{-61}$	2.10	0.186 ± 0.015
NM	578	0.47	0.22	$p < 3 \times 10^{-24}$	2.19	0.267 ± 0.031
ND	1,602	0.45	0.23	$p < 2 \times 10^{-42}$	2.00	0.197 ± 0.014
RC	3,006	0.47	0.22	$p < 7 \times 10^{-95}$	2.12	0.218 ± 0.013
VN	731	0.48	0.23	$p < 2 \times 10^{-29}$	2.07	0.237 ± 0.027
VC	138	0.28	0.20	$p < 7 \times 10^{-4}$	1.42	0.101 ± 0.041
beverly	593	0.55	0.15	$p < 6 \times 10^{-39}$	3.71	0.522 ± 0.045
gamecube	3,442	0.50	0.17	$p < 2 \times 10^{-130}$	2.87	0.281 ± 0.014
monica	770	0.52	0.23	$p < 2 \times 10^{39}$	2.28	0.303 ± 0.029
saccadetest	72	1.00	0.00	$p < 2 \times 10^{-8}$	∞	1.284 ± 0.338
standard	876	0.43	0.27	$p < 3 \times 10^{-15}$	1.58	0.116 ± 0.014
tv-action	105	0.53	0.22	$p < 6 \times 10^{-6}$	2.43	0.408 ± 0.101
tv-ads	844	0.45	0.26	$p < 4 \times 10^{-18}$	1.73	0.161 ± 0.017
tv-announce	153	0.38	0.35	n.s.	1.08	0.083 ± 0.028
tv-music	244	0.41	0.34	$p < .02$	1.23	0.146 ± 0.030
tv-news	2,556	0.37	0.23	$p < 8 \times 10^{-48}$	1.60	0.114 ± 0.009
tv-sports	1,292	0.46	0.26	$p < 2 \times 10^{-39}$	1.76	0.145 ± 0.015
tv-talk	989	0.46	0.23	$p < 10^{-34}$	1.98	0.233 ± 0.020
PA	3,395	0.43	0.22	$p < 6 \times 10^{-89}$	1.99	0.172 ± 0.011

Sign test: Significance level with which the null hypothesis was rejected that the samples S_h/S_{max} and S_r/S_{max} could have been drawn from distributions with equal median, as tested with a nonparametric paired-sample sign test. Ratio of medians: Ratio of the median S_h/S_{max} to the median S_r/S_{max}. Kullback-Leibler distance: KL distance between the histogrammed distributions of S_h/S_{max} and of S_r/S_{max} (higher distances indicate more different distributions). All: All 11,916 saccades recorded in our experiments. First: all the first saccades of our 235 individual recordings (50 clips and 4–6 subjects per clip). First 5, First 10: All the first five and first ten saccades of our 235 recordings (some recordings had fewer than five or ten saccades). CZ to VC: Breakdown by human subject. beverly to tv-talk: Breakdown by category of video clips. Median S_h/S_{max} is significantly higher than median S_r/S_{max} for all but one clip (tv-announce). PA: Data from a control subject (not included in any other analysis) who was given no instructions before watching the video clips.

Analysis by ratios to maximum saliency

The observation of elevated median saliency at human saccade targets compared to random targets is a fairly crude measure that does not take into account possible differences in the shapes of the human and random distributions that would not affect the median.

Hence, we further pushed our analysis to determine the fraction of saccades that had been made to locations of various saliency relative to the maximum saliency. In Figure 4, saccades were binned by their saliency ratio to the maximum saliency, both for human and corresponding random saccades. One striking feature in this figure is that the number of saccades made to absolutely not salient locations was much lower for human than random saccades; conversely, the number of saccades made to the single most salient location in the display was much higher for human than random saccades.

To quantify these differences in shape of distributions, we computed the Kullback-Liebler distance between the histogrammed distribution of $S_h/S_{\max}$ values and that of $S_r/S_{\max}$ values. The Kullback-Leibler (KL) distance is a general metric by which shape similarity between two arbitrary distributions may be measured (Kullback, 1959). For two discrete distributions X and Y with probability density functions x_k and y_k, the KL distance between X and Y is nothing else than the relative entropy of X with respect to Y:

$$KL(X, Y) = \sum_k x_k \log\left(\frac{x_k}{y_k}\right) \tag{1}$$

Note that this actually is not a distance, as it is not symmetric. Hence, in what follows, the shorthand denomination of ''KL distance'' always refers to the KL distance from the human $S_h/S_{\max}$ distribution (X) to the random $S_r/S_{\max}$ distribution (Y).

Table 1 (rightmost column) reports the KL distance for all saccades as well as the various subgroups of saccades studied. A larger KL distance indicates that the distribution of relative saliency at human saccade targets (as shown, for example, in Figure 4a for all saccades) was more different from the distribution of relative saliency at random saccade targets (as shown in Figure 4b for all saccades). A KL distance of zero indicates no difference between the two distributions under consideration. To obtain an estimate of the relevant accuracy at which the KL distances should be interpreted, we repeated the random saccade generation process 100 times, and each time measured the corresponding KL distances; the standard deviation computed over the 100 repeated distance measurements for each subgroup of saccades then quantifies the dependence of the KL metric onto the random saccade generation process.

The KL distance basically followed the same trend as the ratio of medians for the saccade subgroups shown in Table 1 (although it will not always be the

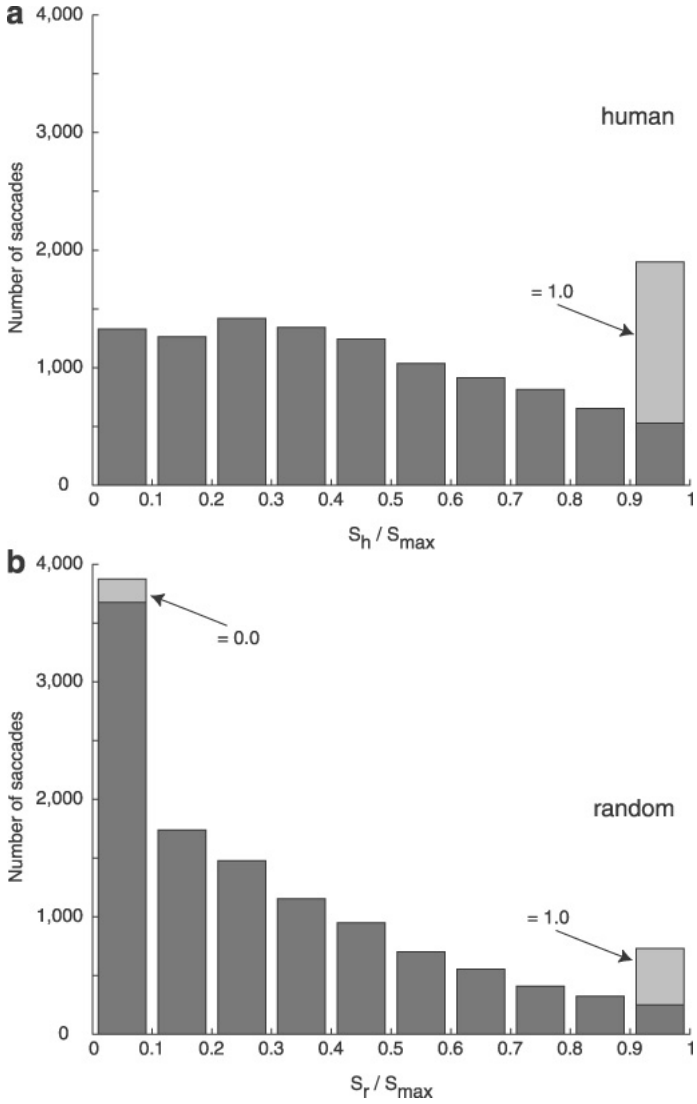

Figure 4. Distributions, by ratios S_r/S_{max} (panel a) and S_r/S_{max} (panel b) of human and random saccades. Each bar includes the lower bound and excludes the higher bound (e.g., the bar between 0.4 and 0.5 is for $0.4 \leq$ ratio < 0.5). In lighter shades are those saccades with ratio < 0.1 where the absolute saliency was zero (noted as ''=0.0''; saccades to an absolutely not salient location), and those saccades with ratio ≥ 0.9 where the absolute saliency was S_{max} (noted as ''=1.0''; saccades to the single most salient location). The histogram for human saccades show a largely increased number of saccades to the most salient location and a largely decreased number of saccades to not salient locations, compared to the random saccades, as further analysed in Table 2.

case in following sections). All KL distances in the table were significantly higher than zero (*t*-test, $p < .01$ or better). The first few saccades showed higher KL distance in addition to the higher ratios of medians noted above. Distributions of saliency at saccade targets for every observer taken individually were also significantly different from the random distributions. The same result was observed with subgroups of video clips. The *saccadetest* clip consisted of a single coloured disc jumping to various places on a uniformly textured blue background, and provided a simple control as there always was only salient location to look at. For this clip, median $S_h/S_{\max}$ was 1.0, indicating that more than half of the human saccades had been directed to the single most salient location in the display (the jumping disc); in contrast, median $S_r/S_{\max}$ was 0.0, indicating that more than half of the random saccades had been made to locations of zero model-predicted saliency. The corresponding KL distance was 1.284 ± 0.338, which gives an estimate of the values that may be obtained in ideal conditions. Values obtained for other subgroups were generally lower, indicating again that, although strong, bottom-up influences were clearly not the only driver of eye movements, and top-down influences also played a strong role as most other clips contained many potentially interesting objects and actors.

A breakdown by ratios to maximum saliency is presented in Table 2. For a given threshold value (0.25, 0.75, or 1.00), this table shows the number of human and of random saccades targeted to regions with relative saliency above threshold. A first interesting point from this analysis is that the frequency of saccades directed to the single most salient location in the saliency map was small, only 11.5% of all saccades. This suggests that a simple strategy consisting of directing the eyes to the most active location in our saliency map would only account for a small fraction of the observed eye movement traces, though in a very reliable manner (only 4.0% of the random saccades were targeted to the most salient location, a factor of almost three times smaller than the human saccades). Given the sparseness of our saliency maps, as exemplified in Figures 5 and 6, it is reasonable to consider that a relative saliency value of 0.25 or more indicates a location that was reliably detected by the model as salient (regions with relative saliency below 0.25 essentially appear as black in the figures). Considering all saccades, 72.3% of the human saccades were targeted towards regions above that threshold, compared to 46.6% of random saccades. While it probably is a wrong conclusion that all of these human saccades were primarily driven bottom-up, certainly it seems that by and large human subjects were more likely to target regions marked by the model as salient than would be expected by chance. Further considerations of bottom-up compared to top-down influences are presented in the Discussion.

A second message of our study, thus, is that the proportion of saccades to the most salient location in the display was small (11.5%) but overall much greater than what would have been expected by chance alone. The proportion of human

TABLE 2
Breakdown by ratios S_h/S_{max} and S_r/S_{max}

		S_h/S_{max}			S_r/S_{max}		
Samples	*n*	*≥ 0.25*	*≥ 0.75*	*= 1.00*	*≥ 0.25*	*≥ 0.75*	*= 1.00*
All	11,916	72.3%	24.5%	11.5%	46.6%	10.6%	4.0%
First	235	85.1%	33.2%	14.9%	49.4%	12.3%	5.1%
First 5	1,173	77.7%	26.6%	13.2%	44.8%	10.6%	4.3%
First 10	2,322	77.6%	26.9%	13.2%	46.3%	10.7%	4.3%
CZ	2,369	72.8%	26.9%	12.7%	46.2%	11.7%	4.7%
JC	1,527	69.9%	21.0%	8.8%	47.9%	10.3%	3.5%
JZ	1,965	71.5%	24.2%	11.9%	45.5%	10.0%	3.0%
NM	578	74.7%	28.2%	13.0%	45.8%	10.0%	4.2%
ND	1,602	72.5%	23.5%	12.1%	47.3%	11.1%	4.2%
RC	3,006	73.1%	25.2%	11.6%	46.7%	10.0%	4.1%
VN	731	76.3%	23.9%	11.4%	47.2%	10.7%	4.5%
VC	138	54.3%	7.2%	2.2%	42.0%	10.1%	5.1%
beverly	593	79.1%	35.4%	21.1%	39.3%	9.9%	4.4%
gamecube	3,422	70.5%	29.5%	16.0%	41.2%	10.1%	4.0%
monica	770	81.0%	28.3%	10.1%	47.0%	11.2%	4.2%
saccadetest	72	68.1%	65.3%	54.2%	12.5%	8.3%	6.9%
standard	876	72.8%	17.8%	5.0%	53.0%	11.6%	4.0%
tv-action	105	86.7%	32.4%	8.6%	46.7%	11.4%	3.8%
tv-ads	844	74.2%	22.7%	8.8%	51.4%	10.7%	3.2%
tv-announce	153	69.3%	9.2%	2.6%	61.4%	14.4%	5.2%
tv-music	244	77.0%	17.2%	5.3%	61.1%	11.9%	3.3%
tv-news	2,556	67.9%	18.0%	8.1%	47.2%	10.7%	4.4%
tv-sports	1,292	74.1%	21.8%	8.6%	50.9%	11.1%	3.6%
tv-talk	989	72.9%	25.2%	12.0%	48.4%	9.5%	3.8%
PA	3,395	69.6%	23.8%	10.5%	45.5%	10.0%	3.7%

The row labels are the same as in Table 1.

saccades directed to overall salient locations was large (72.3%) and also much larger than expected by chance (a factor of 1.55).

Contributions of individual low-level features

To study whether particular features in the model would be better predictors of human saccadic targeting than others, we repeated the analysis, but using variants of the model which only included one feature channel. The feature channels thus considered separately were: color contrast (red/green and blue/yellow double opponencies, combined), intensity contrast (light on dark and dark on light, combined), orientation 0°, 45°, 90°, and 135°, combined), flicker

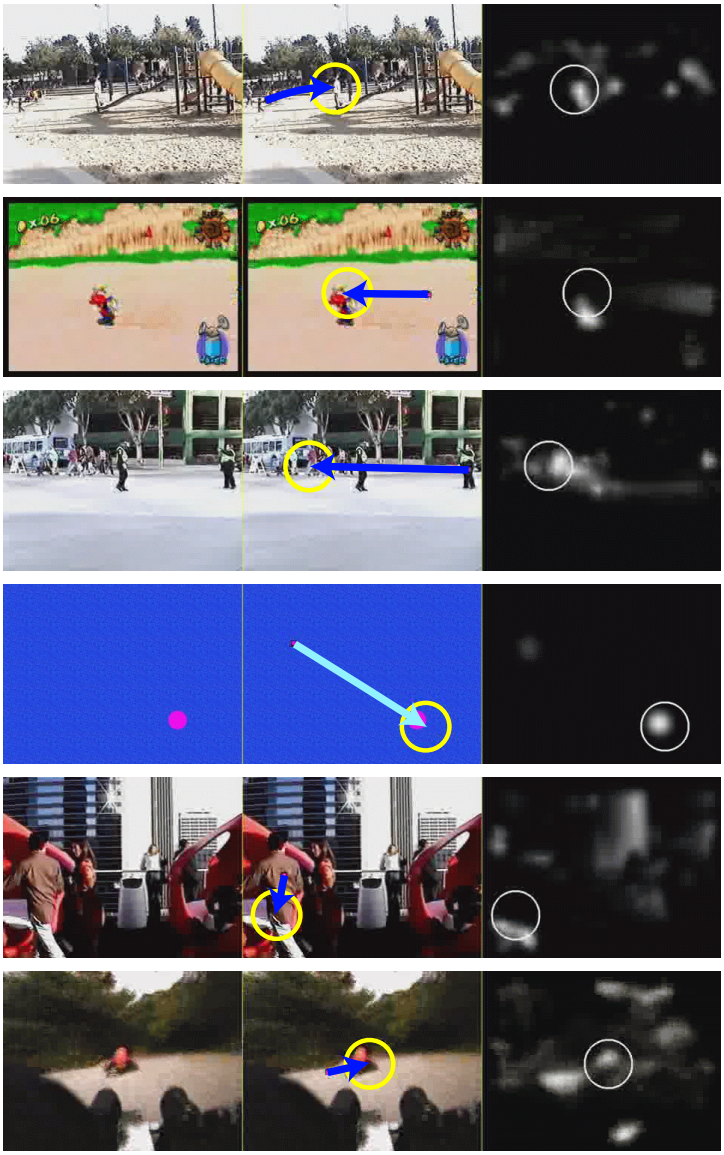

Figure 5. Example frames from our video clips. From left to right: Input to the model, input plus explanatory markers generated by the model, and instantaneous saliency map. From top to bottom: beverly01 (frame 31/490), gamecube04 (frame 914/2083), monica03 (frame 483/1526), saccadetest (frame 63/516), standard03 (frame 112/515), and tv-action01 (frame 142/567). During each frame shown here, a human saccade was initiated, which would in subsequent frames follow the arrow shown on each frame. Yellow circle at arrowhead (and also white circle in saliency map) shows the circular aperture within which saliency at the future location of saccade target was sampled.

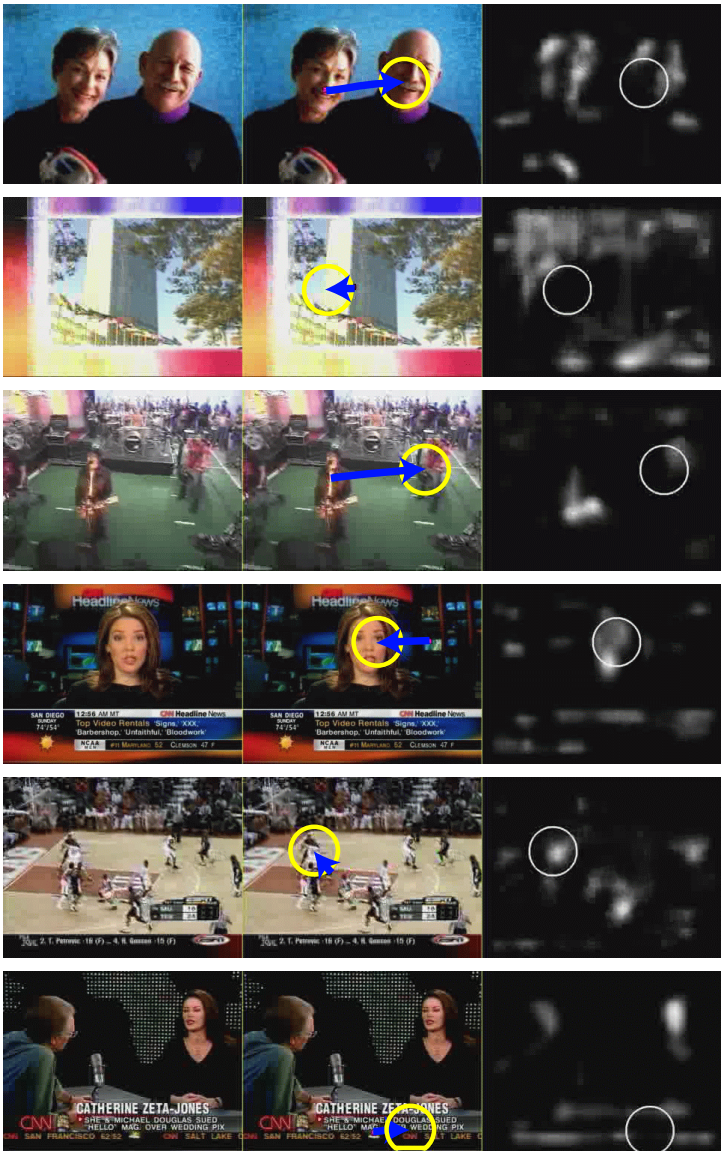

Figure 6. More example frames from our video clips. Format is identical to Figure 5. From top to bottom: tv-ads02 (frame 209/387), tv-announced01 (frame 18/434), tv-music01 (frame 909/1022), tv-news03 (frame 70/1444), tv-sports01 (frame 379/579), and tv-talk01 (frame 121/1651).

TABLE 3
Comparison among models

Model	*Median* $S_h/S_{\max}$	*Median* $S_r/S_{\max}$	*Sign test*	*Ratio of medians*	*Kullback-Leibler distance*
CIOFM	0.45	0.22	$p < 10^{-232}$	2.03	0.195 ± 0.006
C	0.29	0.13	$p < 5 \times 10^{-146}$	2.20	0.075 ± 0.003
I	0.33	0.15	$p < 3 \times 10^{-167}$	2.14	0.089 ± 0.003
O	0.32	0.16	$p < 2 \times 10^{-162}$	1.98	0.084 ± 0.004
F	0.33	0.09	$p < 10^{-232}$	3.71	0.178 ± 0.005
M	0.34	0.09	$p < 10^{-232}$	3.70	0.176 ± 0.005

CIOFM: Model with all feature channels. C: Colour only; I: Intensity only; O: Orientation only; F: Flicker only; M: Motion only. Columns are a subset of those in Table 1.

(absolute difference between successive frames), and motion energy (up, down, left, and right, combined).

Results are presented in Table 3. Flicker and motion each taken separately yielded higher ratios of medians and KL distances than intensity, colour, and orientation taken separately. In fact, the ratios of medians for flicker and motion were even higher than for the full model which includes all features. To a large extent, this was due to the increased sparseness of the motion and flicker saliency maps compared to the full-model saliency maps. Indeed, the number of human saccades targeted to the most salient location was similar (12.1% for the motion-only model and 12.1% for the flicker-only, compared to 11.5% for the full model reported in Table 2, with random saccades all around 4.0%). In contrast, there were many more random saccades targeted to nonsalient locations, hence driving the median saliency at random saccade endpoints down, and the ratio of medians up. The KL distance is useful here to provide a finer analysis of which version yielded the largest differences in saliency distribution between human and random scanpaths. Motion and flicker, each individually, yielded KL distances quite close to that obtained with the full model, although significantly lower, as shown in Table 4. These results indicate that motion and image transients are much more reliable predictors of human saccadic targeting than colour, intensity, or orientation contrasts, with the best predictor being the sum of all these features.

Duration of subsequent fixations

A last analysis was to determine whether the duration of a fixation following a saccade would correlate with the saliency at that saccade's endpoint. With our video clips, average duration of fixation was short, approximately 135 ms (although with a long tail towards longer durations). Figure 7 shows the average duration of fixation following saccades to locations with given model-predicted

TABLE 4
Contributions of features

	CIOFM	*C*	*I*	*O*	*F*	*M*
CIOFM	—					
C	$p < 4 \times 10^{-23}$	—				
I	$p < 4 \times 10^{-23}$	$p < 3 \times 10^{-22}$	—			
O	$p < 4 \times 10^{-23}$	$p < 3 \times 10^{-18}$	$p < 4 \times 10^{-8}$	—		
F	$p < 1 \times 10^{-20}$	$p < 4 \times 10^{-23}$	$p < 4 \times 10^{-23}$	$p < 4 \times 10^{-23}$	—	
M	$p < 3 \times 10^{-22}$	$p < 4 \times 10^{-23}$	$p < 4 \times 10^{-23}$	$p < 4 \times 10^{-23}$	n.s.	

Reported are the p values with which the null hypothesis was rejected for a t-test that any two model variants had yielded the same KL distance. All model variants were significantly different from each other, except for M (motion) compared to F (flicker). Notations are as in Table 3.

saliency, grouped in 10 bins by saliency. A one-way ANOVA suggested no significant effect of saliency onto duration, both for human saccades, $F(9, 11906) = 0.75$, n.s., 11,916 saccades and 10 bins, and for random saccades, $F(9, 11906) = 1.64$, n.s. Only considering the 2306 saccades with duration 200 ms or more did not alter this conclusion: Human, $F(9, 2296) = 0.76$, n.s.; random: $F(9, 2296) = 1.68$, n.s. Hence, for our video clips, subjects, and viewing instructions, we found no correlation between the model-predicted saliency of a saccade target location and the duration of the subsequent fixation onto that location.

DISCUSSION

Our study confirms previous findings that humans overall tend to look at salient objects in their visual environment. In particular, our study confirms and reinforces, under conditions of sustained viewing and dynamic colour scenes, the findings of Parkhurst et al. (2002) using an earlier version of our model and static scenes. In addition, we here have pushed the analysis further, by specifically counting those saccades that were made to the single most salient location in the scene, and using the KL distance to quantify how the distributions of saliency at human and random saccade endpoints differed. With the simulation framework used here, we found that occurrences of saccades to the most salient location in the scene were rare, although much more frequent than expected by chance. Motion and flicker were found to be better correlated with human saccades than colour, intensity, and orientation, but not as good as all features combined in terms of KL distance. No correlation was found between the model-predicted saliency of a location and the amount of time during which that location was subsequently fixated. Subjective visual analysis of the eye movement traces yielded a number of additional observations, illustrated in Figures 5

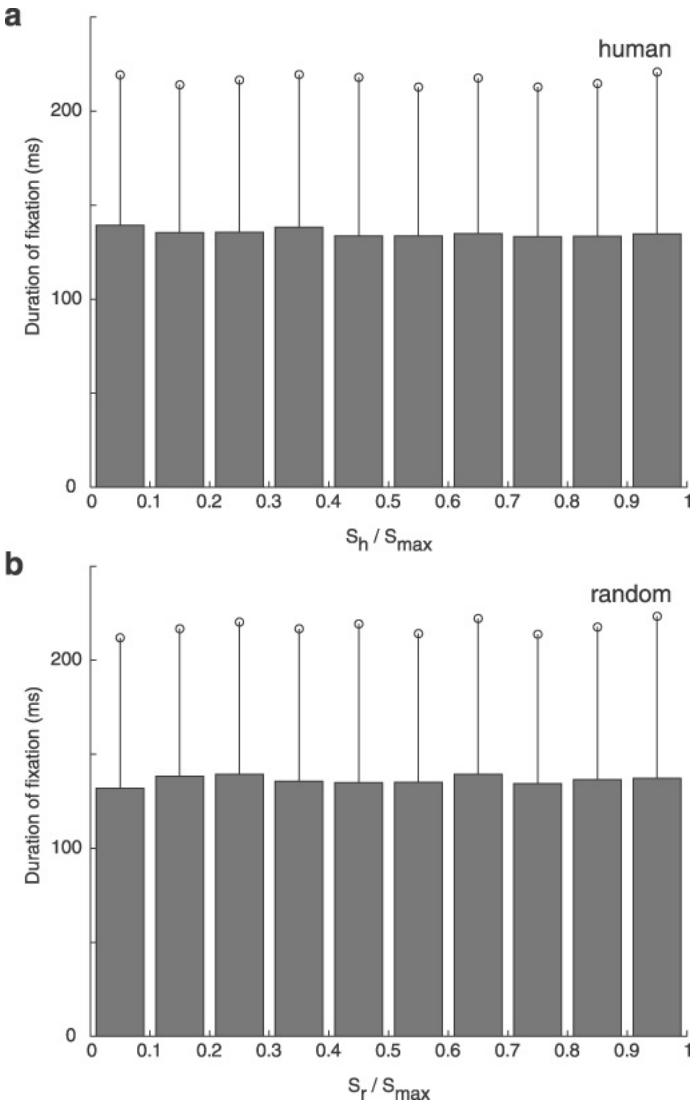

Figure 7. Distribution of the durations of fixations (mean + standard deviation) following saccades targeted towards locations of a given saliency. We found no significant effect of target saliency onto subsequent fixation duration (one-way ANOVA; see text).

and 6 and discussed here, which may serve as guidelines for extensions of the present study.

We used a fairly small saliency sampling window (5.6° diameter), and in some instances it was clear that the observer had saccaded towards a salient object, but had targeted such part of the object that the saliency sample at the saccade endpoint had partially missed the representation of the object in our low-resolution saliency map (e.g., second and fifth rows of Figure 5). Using a larger sampling window would certaintly address this problem, but at the cost of possibly artificially increasing the saliency reading when observers had saccaded to a nonsalient object (e.g., a larger radius of the circle in the first or second rows of Figure 6 would increase the sampled saliency value, while in these examples clearly the observer had saccaded to a location that the model considered not salient). An additional complication in this respect was that humans sometimes were fairly inaccurate in targeting their saccades, especially for long ones (e.g., fourth row of Figure 5; this was not an inaccuracy of the eyetracking, as the saccade shown here was followed by another short saccade that better centred the eye onto the object). One limitation of the current study which will need to be further explored is the lack of a systematic investigation of how sampling window size would affect the results reported here. Our comparison between human and random saccades, however, aims at minimizing the effects of window size onto our conclusions (as different window sizes would affect both the human and random histograms).

We measured a minority but not negligible number of human saccades targeted to locations deemed not salient by the model (27.7% of all saccades human saccades were to locations with $S_h/S_{\max} < 0.25$). In some cases, these saccades were predictive of moving object trajectories; that is, instead of focusing onto a moving object, observers looked at an empty location slightly ahead of that object. This top-down predictive capability is not captured by our computational model. Building neural models that can embody it remains an open issue (although many mathematical models, such as Kalman filters, may be used to solve this problem).

Our measure of saliency was very strict, in that a unique sample was taken at the precise moment when each saccade began. An interesting extension of our analysis will be to investigate a more relaxed question, such as whether a given saccade target location has been among the five most salient locations at any given time during the second preceding a given saccade. Intuitively, we often are confronted with situations where several objects are of interest at a given moment, but we employ a serial strategy to focus onto each in turn. Adding some account for such memory process in our measure of saliency may reveal a larger fraction of saccades made to objects that were among the few most salient at some point in the recent past. One difficulty in developing such more relaxed measure lies in the necessity to be able to reliably segment the few most salient objects in the saliency map, on a continual basis.

In the present study we have decided to remain agnostic to putative saccade generation models, which has limited how far the analysis could be pushed. Our results suggest, however, that a winner-take-all strategy often used in covert attention models (Itti & Koch, 2001a), directing gaze towards the most salient location in the display, would only account for a small fraction of human saccades (around 12%). To estimate the fraction of human and random saccades directed to salient locations, we have explored three arbitrary $S_h/S_{\max}$ thresholds of 0.25, 0.75 and 1.00 in Table 2. It is important to realize that these indeed are arbitrary and are here only intended to summarize, for many subgroups of saccades, the more detailed picture provided by the distributions in Figure 4 for all saccades. Furthermore, applying a nonlinearity to the saliency values would in many cases change the figures in Table 2. Hence, while it is certainly tempting to attribute some of the saccades to bottom-up influences, and the rest to top-down influences, such attribution may not be very meaningful. Indeed, changing the cut-off threshold, or applying a nonlinearity to the saliency values, could yield a wide range of bottom-up/top-down ratios. For example, if one considers that only saccades to the maximum saliency can be attributed to bottom-up, then the ratio is 11.5% bottom-up for 88.5% top-down, or approximately 1:8. In contrast, if all saccades to regions reliably marked by the model as salient ($S_h/S_{\max} \geq 0.25$) are attributed to bottom-up, then the ratio becomes 72.3% to 27.7% or approximately 2.5:1. The value of 0.25 seems to be an appropriate threshold, since, given the sparseness of our saliency maps exemplified in Figures 5 and 6, any location with saliency above that threshold is likely to be a nonaccidental detection. More quantitatively, it is interesting to note from Table 2 that a majority of random saccades were to locations with $S_r/S_{\max} < 0.25$ (for all saccades, all first, first 5, and first 10 saccades, and all observers taken individually, although not for all classes of video clips taken individually), confirming that overall the number of pixels (or surface area) in the saliency map above that threshold were in minority (especially given our aperture-based sampling scheme).

In reality, it is likely that bottom-up and top-down influences simultaneously contribute to directing every saccade. Intuitively, it is clear that top-down influences can be very strong—for instance, we can command our gaze to any location we like, we can make eye movements based purely on cognitive reasoning (e.g., from knowing the typical spatial relationship between the object of current fixation and a desired object), and there are many other top-down factors that have been clearly shown to influence eye movements, as reviewed in introduction. Yet, our results suggest that, if one is to accept the threshold of 0.25 just discussed, a comfortable majority of human saccades (72.3% overall) was directed to a minority of locations that had been highlighted in our bottom-up saliency map (and to which a minority of 46.6% of random saccades went). This relative rarity of occurrence of saccades to locations that were not bottom-up salient suggests that the bottom-up saliency map may be used as a mask,

highlighting a set of potentially interesting locations in the scene, with top-down influences mainly responsible for deciding upon one specific location among these candidates. This hypothesis seems well in line with our finding that the most salient location in the display was not attracting gaze very often; hence it may only have been one among several bottom-up candidate locations, with the choice of which candidate would eventually be fixated primarily mediated top-down rather than by bottom-up saliency.

Our discussion thus far is compatible with a number of saccade generation models. For example, the area activation model of Pomplun, Shen, and Reingold (2003), proposed in the context of visual search, suggests that scanpaths visit peaks of an activation map that are above a given threshold, with the selection of the next peak based on proximity to current fixation rather than absolute peak activity. The activation map in that model is not purely bottom-up, but substantially modulated by the nature of the search task and of target and distractor items; thus, it may be considered a combined bottom-up and top-down map. The greedy choice of the next saccade target based on proximity brings the idea that there are also other factors to consider, as this mechanism may not be considered purely top-down nor bottom-up—rather, it is an additional bias that may influence saccade target selection in cases where one is faced with a number of equally bottom-up salient and top-down interesting candidate targets (such as the distractor items in cases of difficult visual search). Similarly, Rao, Zelinsky, Hayhoe, and Ballard (2002) select saccade targets by subjecting an activation map to a softmax operator (which further emphasizes strong activation peaks and weakens regions of lower activation) and then aiming for the centre of gravity of the transformed map. This model, in particular by reproducing centre-of-gravity effects observed experimentally when several items simultaneously attract gaze, suggests that in general the entire spatial distribution of values in the saliency (or activation) map is likely to contribute to saccadic orienting, rather than only a selected subset of locations or the single location of maximum saliency. In previous work, we have proposed that attention and eye movements may be guided by an attention guidance map (AGM) defined as the pointwise product between a bottom-up saliency map (SM) and a top-down task-relevance map (TRM; Navalpakkam & Itti, 2002). Hence, a location would have to be both somewhat salient and somewhat relevant to become a candidate saccade target. The SM might be modulated top-down using feature weighing similar to Guided Search (Wolfe et al., 1989), and the TRM could be populated based on information from the gist and rough layout of the scene (Hollingworth & Henderson, 1998) as well as top-down priors on the likely locations of desired objects (Torralba, 2003). Our present observations are compatible with this hypothesis (also see Rensink, 2000), although many of the details still need to be elucidated.

These considerations reinforce the idea that a metric like the KL distance, which considers dissimilarity of shape of the full distributions of saliency

at human compared to random saccade targets, is a more appropriate metric for comparing models than, for example, the simpler ratio of median saliencies. Indeed, even if the distribution of model-predicted activation at human saccade targets were to differ from random by exhibiting a larger number of human saccades towards low model-predicted activation values, that information could still be exploited by a saccade generation mechanism. The mechanism would just have to reflect the fact that low-activation regions actually should be strong attractors of gaze. Thus, in principle at least, any shape difference between the human and random distributions obtained for a given model in a manner similar to that of Figure 4 could be exploited to develop a gaze generation model, not just a shift towards higher values. Under these conditions, we found that models based purely on motion energy or on temporal change (flicker) performed much better than models relying solely on colour, intensity, or orientation information. While it has often been posited that motion was deemed to be one of the strongest attractors of attention, our study allows us to quantify this in terms of KL distance: The distance between human and random obtained for the motion and flicker models was about twice that obtained for the colour, intensity, or orientation models. Yet, in terms of KL distance, the best model remained that which incorporated all features. This is not surprising, as strong motion cues were not always present in our stimuli. During periods of rather still video content, colour, intensity, and orientation certainly were better predictors of human saccade targets than flicker and motion which essentially yielded no output during these periods.

As reviewed in our introduction, bottom-up saliency is only one of the many factors that may influence eye movements. The main contribution of our study is to attempt to quantify the extent to which it may continually contribute to human eye movements. Our general conclusion in this respect is that a comfortable majority of human saccades were made to a minority of fairly salient locations in the saliency map, suggesting that bottom-up saliency plays a non-negligible role in determining saccade targets. In addition, we found that few saccades were directed to the most salient location in the visual input, hence suggesting that absolute saliency may play only a minor role in picking one target among a set of fairly salient candidates, a process likely to be primarily driven top-down. We found that motion energy and temporal change were strong attractors of attention, though the model that also included colour, intensity, and orientation features performed overall best. Finally, we did not find any evidence of correlation between saliency of a saccade target and duration of subsequent fixation of that target, suggesting that bottom-up saliency does not contribute to duration of fixation. All these observations are compatible with the idea that a combination of bottom-up and top-down influences determines every saccade target location, rather than some saccades being primarily directed bottom-up while others are primarily targeted top-down.

REFERENCES

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, *39*(17), 2947–2953.

Barth, E., Zetzsche, C., & Rentschler, I. (1998). Intrinsic two-dimensional features as textons. *Journal of the Optical Society of America: Part A. Optics Image Science and Vision*, *15*(7), 1723–1732.

Biederman, I., Teitelbaum, R. C., & Mezzanotte, R. J. (1983). Scene perception: A failure to find a benefit from prior expectancy or familiarity. *Journal of Experimental Psychology: Learning of Memory Cognition*, *9*(3), 411–429.

Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Translations on Pattern Analysis and Machine Intelligence*, *11*, 567–585.

Cannon, M. W., & Fullenkamp, S. C. (1991). A transducer model for contrast perception. *Vision Research*, *31*(6), 983–998.

Collewijn, H., Steinman, R. M., Erkelens, C. J., Pizlo, Z., & van der Steen, J. (1992). Effect of freeing the head on eye movement characteristics during three-dimensional shifts of gaze and tracking. In A. Berthoz, W. Graf, & P. P. Vidal (Eds.), *The head-neck sensory motor system*. Oxford, UK: Oxford University Press.

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

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*(3), 316–355.

Gallant, J. L., Connor, C. E., & van Essen, D. C. (1998). Neural activity areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport*, *9*(9), 2153–2158. (Reprinted with corrections from *Neuroreport*, 1998, *9*(1), 85–90)

Gilbert, C. D., & Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, *9*(7), 2432–2442.

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

Henderson, J. M., & Hollingworth, A. (2003). Global transsaccadic change blindness during scene perception. *Psychological Science*, *14*(5), 493–497.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, *127*(4), 398–415.

Hollingworth, A., Williams, C. C., & Henderson, J. M. (2001). To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin and Review*, *8*(4), 761–768.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Ito, M., & Gilbert, C. D. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, *22*(3), 593–604.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*, 1304–1318.

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

Itti, L., & Koch, C. (2001a). Computational modeling of visual attention. *Nature Review Neuroscience*, *2*(3), 194–203.

Itti, L., & Koch, C. (2001b). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*(1), 161–169.

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

Itti, L., Dhavale, N., & Pighin, F. (2003, Aug.). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proceedings of the SPIE 48th annual international symposium on Optical Science and Technology* (pp. 64–78). San Diego, CA: SPIE.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human neurobiology*, *4*(4), 219–227.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, *16*, 37–68.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Research*, *42*(18), 2219–2234.

Li, W., & Gilbert, C. D. (2002). Global contour saliency and local colinear interactions. *Journal of Neurophysiology*, *88*(5), 2846–2856.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784.

Moreno, F. J., Reina, R., Luis, V., & Sabido, R. (2002). Visual search strategies in experienced and inexperienced gymnastic coaches. *Perceptual and Motor skills*, *95*(3, Pt. 1), 901–902.

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, *14*(4), 2178–2189.

Motter, B. C., & Belky, E. J. (1998). The guidance of eye movements during active visual search. *Vision Research*, *38*(12), 1805–1815.

Navalpakkam, V., & Itti, L. (2002). A goal oriented attention guidance model. *Lecture Notes in Computer Science*, *2525*(Nov.), 453–461.

Nodine, C. F., & Krupinski, E. A. (1998). Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. *Academic Radiology*, *5*, 603–612.

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, *171*(968), 308–311.

Olivia, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*(1), 72–107.

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

Peebles, D., & Cheng, P. C. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, *45*(1), 28–46.

Peters, R. J., Itti, L., & Koch, C. (2002, Nov.). Eye movements are influenced by short-range interactions among orientation channels. In *Proceedings of the Society for Neuroscience annual meeting (SFN'02)* (p. 715.12). Orlando, FL: Society for Neuroscience.

Polat, U., & Sagi, D. (1993). Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. *Vision Research*, *33*(7), 993–999.

Pomplun, M., Shen, J., & Reingold, E. M. (2003). Area activation: A computational model of saccadic selectivity in visual search. *Cognitive Science*, *27*, 299–312.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15.

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

Rao, R. P., Zelinsky, G., Hayhoe, M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*(11), 1447–1463.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioural Brain Sciences*, *26*(4), 445–476; discussion 477–526.

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

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*(1/2/3), 17–42.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703–714. Comment in: *Neuron*, 2000, *26*(3), 548–550.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications symposium* (pp. 71–78). New York: ACM Press.

Savelsbergh, G. J., Williams, A. M., van der Kamp, J., & Ward, P. (2002). Visual search, anticipation and expertise in soccer goalkeepers. *Journal of Sports Sciences*, *20*(3), 279–287.

Shen, J., Reingold, E. M., & Pomplun, M. (2000). Distractor ratio influences patterns of eye movements during visual search. *Perception*, *29*(2), 241–250.

Sillito, A. M., Grieve, K. L., Jones, H. E., Gudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, *378*(6556), 492–496.

Sparks, D. L. (2002). The brainstem control of saccadic eye movements. *Nature Reviews Neuroscience*, *3*(12), 952–964.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

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

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treue, S., & Martinez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579.

Treue, S., & Maunsell, J. H. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, *382*(6591), 539–541.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision Research*, *34*(9), 1187–1195.

Wolfe, J. M. (1998). Visual memory: What do you know about what you saw? *Current Biol*, *8*(9), R303–R304.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychological Human Perception Performance*, *15*(3), 419–433.

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

Yeshurun, Y., & Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, *396*(6706), 72–75.

Zetzsche, C., Schill, K., Deubel, H., Krieger, G., Umkehrer, E., & Beinlich, S. (1998). investigation of a sensorimotor system for saccadic scene analysis: An integrated approach. In *Proceedings of the fifth international conference on Simulation of Adaptive Behavior: Vol. 5 From animals to animats* (pp. 120–126). Cambridge, MA: MIT Press.