VISUAL COGNITION, 2006, 14 (4/5/6/7/8), 959–984

Psychology Press
Taylor & Francis Group

Quantitative modelling of perceptual salience at human eye position

Laurent Itti

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

We investigate the extent to which a simple model of bottom-up attention and salience may be embedded within a broader computational framework, and compared with human eye movement data. We focus on quantifying whether increased simulation realism significantly affects quantitative measures of how well the model may predict where in video clips humans direct their gaze. We hence compare three variants of the model, tested with 15 video clips of natural scenes shown to three observers. We measure model-predicted salience at the locations gazed to by the observers, compared to random locations. The first variant simply processes the raw video clips. The second adds a gaze-contingent foveation filter. The third further attempts to realistically simulate dynamic human vision by embedding the video frames within a larger background, and shifting them to eye position. Our main finding is that increasing simulation realism significantly improves the predictive ability of the model. Better emulating the details of how a visual stimulus is captured by a constantly rotating retina during active vision has a significant positive impact onto quantitative comparisons between model and human behaviour.

Over the past decades, visual psychophysics in humans and other primates have become a particularly productive technique to probe the mechanisms of visual processing, attentional selection, and visual search (Verghese, 2001; Wolfe, 1998). In typical visual psychophysics experiments, visual stimulus patterns are briefly presented to observers, for example on a computer monitor. Observers are instructed to describe their perception of the stimuli, for example by pressing one of several possible answer keys corresponding to alternate responses to a question asked by the experimenter (e.g., whether the stimulus was horizontal or slightly tilted off horizontal). Over the course

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

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

http://www.psypress.com/viscog DOI: 10.1080/13506280500195672

of many experimental trials, quantitative measures are collected and subjected to statistical analysis, to establish a relationship between the ability of observers to perceive some aspect of the presented stimuli and the hypothesis tested by the experimenter.

As visual psychophysics experiments become more sophisticated and stimuli more complex, the need for quantitative computational tools that can relate experimental outcomes to putative brain mechanisms is becoming more pressing than ever before. Indeed, in the presence of complex stimuli such as natural scenes, formulating hypotheses on how stimulus attributes may influence perception is hampered both by a difficulty in providing a quantitative formal description of the stimuli, and by the understanding that complex nonlinear interactions among the perceptual representations of diverse components of the stimuli at many processing levels will affect overall perception (Kofka, 1935; Sigman, Cecchi, Gilbert, & Magnasco, 2001).

A particularly successful experimental technique to evaluate perception of complex visual stimuli has been to track eye position of human subjects while they inspect visual displays. Using this technique, several studies have recently demonstrated how local image properties at the locations fixated by humans significantly differ from image properties at other locations in static images. These include studies by Zetzsche and colleagues (Barth, Zetzsche, & Rentschler, 1998; Zetzsche et al., 1998), who inferred from human eye tracking that the eyes preferentially fixate regions in greyscale images with multiple superimposed orientations, including corners. Similarly, Reinagel and Zador (1999) found that local spatial contrast of greyscale static images was significantly higher at the point of gaze than, on average, at random locations, whereas pairwise pixel correlations (image uniformity) were significantly lower. Privitera and Stark (2000) further 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, thus deriving bottom-up feature detectors that captured some of the local image properties which attracted eye movements. Extending on these purely local analyses, and considering colour scenes, Parkhurst, Law, & Niebur (2002) compared human scan paths to bottom-up saliency maps computed from static images (Itti, Koch, & Niebur, 1998). Thus, this study not only accounted for local image properties but also for long-range interactions among the cortical representations of distant image regions, which mediate visual salience and pop-out (Itti et al., 1998; Treisman & Gelade, 1980). This study revealed significantly elevated model-predicted salience at human fixations compared to random fixations, and more strongly so for the first few fixations on an image than for subsequent fixations.

Previous studies thus have abstracted most of the dynamics of human active vision, and have focused on intrinsic properties of local distributions of pixel intensities in an image. This contrasts with attempting a more realistic simulation of how the recent history of dynamic patterns of retinal stimulation, as the eye jumps from one image location to the next, may predict candidate target locations for a subsequent eye movement. As a first step in this direction, Parkhurst et al. (2002) noted a human bias for shorter saccades, and applied a spatial modulation filter to the static saliency map computed for each image, whereby the salience of locations increasingly further from current fixation was increasingly suppressed before salience distributions for the following fixation were analysed. Although this increased the model-predicted salience at human eye fixations compared to random fixations, the authors correctly noted that applying such foveation filter to the saliency map rather than the retinal inputs only is a coarse approximation of how active foveation may modulate salience.

Building on these previous studies, here we jointly use human eye tracking and a computational model of low-level visual perception to quantitatively evaluate dynamic attentional allocation onto visual stimuli. We develop a more realistic simulation framework, to transition from evaluating local *image properties* at human eye fixations to evaluating a measure of *perceptual saliency* at human eye fixations. For the purpose of this study, we thus operationally define perceptual saliency as depending not only on local image properties, but also on how these are captured by a foveated retina, and how their cortical representations interact over visual space and time. Thus, we attempt to more realistically account for a larger fraction of the details of how information present in an image may first be processed by the visual system of the observer, such as to eventually yield neural signals that may direct the observer's attention and eye movements. The main question addressed here is whether using such a realistic framework would significantly affect the outcomes of quantitative comparisons between model-predicted saliency maps and locations fixated by human observers. If it did, this would indicate that the details of active vision do play a non-negligible role in the selection of eye movements, and hence should be taken into account in future psychophysics studies.

Our approach relies upon and extends our previously proposed computational model of visual salience and bottom-up attentional allocation (Itti & Koch, 2000; Itti & Koch, 2001a; Itti et al., 1998). Previously, we have used this model in a "generative" mode, yielding predictions of the visual attractiveness of every location in a display (in the form of a graded topographic saliency map; Koch & Ullman, 1985) and predictions of attentional scan paths and eye movements (Itti, Dhavale, & Pighin, 2003). In contrast, here we use it in a "human-driven" or "servoed" mode: Using actual eye movement scan paths recorded from human subjects, we simulate

the retinal input received by the observers, process it through the low-level visual stages of the model, and continuously recompute the saliency map over the course of the human scan path. To effectively exercise and challenge the framework, we employ dynamic (video clip) natural stimuli rather than static imagery. We explore three variants of the model, with increasing realism. The first variant simply computes saliency maps from the raw video frames shown to human observers, similar to previous studies. The human eye movement recordings are then used to compare the model-predicted salience at human eye positions compared to random locations within the video frames. The second variant adds a foveation filter, by which each input frame is increasingly blurred with distance from current human eye position before it is processed by the model. Finally, the third and most realistic variant embeds the raw video frames into a background photograph of the experimental room and computer monitor, shifts the resulting image to centre it at human eye position, crops the shifted image to simulate a retinal field of view, and applies a foveation filter to the field of view before processing by our model.

The present study focuses on whether increasing the realism of model simulations significantly affects the measures of model-predicted salience at the image locations visited by the human eye compared to random locations. We here use our available model of bottom-up attention as an approximation to early visual processing in humans. However, it is important to keep in mind that the computations operated by this model represent only a very coarse approximation to a small subset of the many factors that influence attentional deployment onto a visual scene (Henderson & Hollingworth, 1999; Itti & Koch, 2001a; Rensink, 2000). Hence, we do not expect perfect agreement between model-predicted salience and human eye position. In particular, our bottom-up model as used here does not yet account, among others, for how the rapid identification of the gist (semantic category) of a scene may provide contextual priors to more efficiently guide attention towards target objects of interest (Biederman, Teitelbaum, & Mezzanotte, 1983; Friedman, 1979; Hollingworth & Henderson, 1998; Oliva & Schyns, 1997; Potter & Levy, 1969; Torralba, 2003); how search for a specific target might be guided top-down, for example by boosting visual neurons tuned to the attributes of the target (Ito & Gilbert, 1999; Moran & Desimone, 1985; Motter, 1994; Müller, Reimann, & Krummenacher, 2003; Reynolds, Pasternak, & Desimone, 2000; Treue & Maunsell, 1996; Treue & Trujillo, 1999; Wolfe, 1994, 1998; Wolfe, Cave, & Franzel, 1989; Yeshurun & Carrasco, 1998); or how task, expertise, and internal scene models may influence eye movements (Henderson & Hollingworth, 2003; Moreno, Reina, Luis, & Sabido, 2002; Nodine & Krupinski, 1998; Noton & Stark, 1971; Peebles & Cheng, 2003; Savelsbergh, Williams, van der Kamp, & Ward, 2002; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995;

Yarbus, 1967). Nevertheless, our hypothesis for this study is that a more realistic simulation framework might yield better agreement between human and model than a less realistic one. The present study quantifies the extent to which this hypothesis may be verified or rejected. The finding of significant differences between the three versions of our simulation framework would indicate that the details of foveation, eye movements, and background should not be ignored in the analysis of future psychophysics experiments.

METHODS

The proposed quantitative analysis framework allows us to derive an absolute measure of salience compounded over a human scan path, and to compare it to the same measure compounded over a random scan path. Although the absolute value of this measure is difficult to interpret, given all the factors influencing eye movements but not accounted for by the model, it allows us to rank the three variants of the model, and to suggest that the model variant with highest ranking may be the one that best approximates human bottom-up visual processing.

Human eye movement experiments

Eye movement recordings were collected from eight human observers watching a heterogeneous collection of 15 video clips, including outdoors scenes, video games, television newscasts, commercials, sports, and other content. Each clip comprised between 309 and 2083 frames (10.4–69.1 s), for a total of 13,770 distinct frames (totalling 7 min 37.0 s). The experimental protocol used to collect the data evaluated here has been previously described in detail (Itti, 2004; Itti, 2005). In short, stimuli were presented to four normal volunteer subjects on a 22-inch computer monitor (LaCie Corp; 640×480, 60.27 Hz double-scan, mean screen luminance 30 cd/m^2, room 4 cd/m^2) at a viewing distance of 80 cm ($28^\circ \times 21^\circ$ usable field-of-view). Subjects were instructed to attempt to "follow the main actors and actions" in the video clips, and thus were biased towards the scene elements that were the most important according to their current cognitive understanding of the scenes. The extent to which the selection of these cognitively important locations would be more or less predictable by variants of a simple bottom-up computational analysis of image pixels is a side issue addressed by the present study, with again our main focus being the influence of model realism onto agreement between model and humans. Eye position was tracked using a 240 Hz infrared video-based eyetracker (ISCAN, Inc model RK-464). Raw eye movement traces were remapped to screen coordinates using a nine-point calibration procedure (Stampe, 1993). Each calibration

point consisted of fixating first a central cross, then a blinking dot at a random point on a 3×3 matrix. After a nine-point calibration had been completed, five video clips were played. For every 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 to ensure accurate timing (Finney, 2001). Frame displays were hardware-locked to the vertical retrace of the monitor. Microsecond-accurate timestamps were stored in memory as each frame was presented, and later saved to disk to check for dropped frames. No frame drop occurred and all timestamps were spaced by $33{,}185 \pm 2\,\mu s$. During offline remapping of raw eye position to screen coordinates, data was discarded until the next calibration if residual errors greater than 20 pixels (0.90°) on any calibration point or 10 pixels (0.45°) overall remained.

The dataset used here is a subset of the entire dataset collected (which was for 50 clips and eight subjects each watching a subset of the 50 clips, so that four to six valid eye movement traces were available for each clip; Itti, 2004). The fifteen clips used here were randomly chosen so that we would use two clips from any category (e.g., outdoors, video games, TV news, etc.) that had at least two clips in the original dataset, except for the last one (the original dataset had 12 categories, of which 8 had two or more clips). In addition to the 14 clips thus obtained, we included one very simple synthetic clip of a coloured disk jumping to various locations on a static textured background, which is a useful control stimulus. Once the subset of 15 clips had been decided upon, the four observers who had watched the most clips were selected and alphabetically ordered, and the first three eye movement traces for each clip were selected. Hence, we used calibrated eye movement data for three subjects on each of the 15 clips, yielding a total of 45 calibrated eye movement traces. The reason of using a subset of the available data here was solely computational cost (a run of the three variants of the model for comparison to the 45 eye movement traces takes approximately two CPU-months, or 4 days on a 16-CPU Beowulf cluster of interconnected Linux PC computers).

Comparing model predictions to human eye movements

The computational model variants were used in the following manner. Each of the 45 human eye movement recordings was considered in turn. For a given recording, the corresponding video clip was presented to a model variant, at the same resolution (640×480) and frame rate (33.185 ms/frame) as it had been presented to human subjects. In the more sophisticated model variants, recorded human eye position was then used to drive the foveated

eye of the model and to generate a coarse simulation of the pattern of inputs actually received by the observer's retinas. For every human eye position recorded (240 Hz), several measurements of the internal state of the model were taken: Instantaneous model-predicted salience at current human eye position, maximum and average salience over the current video frame, and salience at a randomly chosen location in the image. Our analysis focuses on comparing how salience differed at human eye position compared to random locations, and on quantifying this effect so that the different variants of the model may be ranked according to their agreement with human eye movements.

Baseline model

To relate our findings to previous analyzes of local image properties at locations fixated by human observers, we begin with a simple model that analyses the raw video clips shown to observers, without including any realistic simulation of how human eye movements affect visual inputs that actually enter the eyes. Thus, the first model variant is mainly concerned with local *image* properties but disregards most of the byproducts of dynamic human *vision* (e.g., decrease of contrast sensitivity with eccentricity, motion transients due to eye movements, etc.).

At the core of the framework is our previously described model of bottom-up visual attention, which computes a topographic saliency map from the input images (Itti & Koch, 2000; Itti & Koch, 2001a; Itti et al., 1998; Koch & Ullman, 1985). Video input is processed in parallel by a number of multiscale low-level feature maps (Figure 1), which detect local spatial discontinuities in various visual feature channels using simulated centre-surround neurons (Hubel & Wiesel, 1962; Kuffler, 1953). Twelve neuronal features are implemented, sensitive to colour contrast (red/green and blue/yellow double-opponency, 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). The detailed implementation of these feature channels, thought indeed to guide human attention (Wolfe & Horowitz, 2004), has been described previously (Itti et al., 1998, 2003; Itti & Koch, 2001a). In the presentation of our results, these 12 features are combined into five categories: Colour, flicker, intensity, orientation, and motion. 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

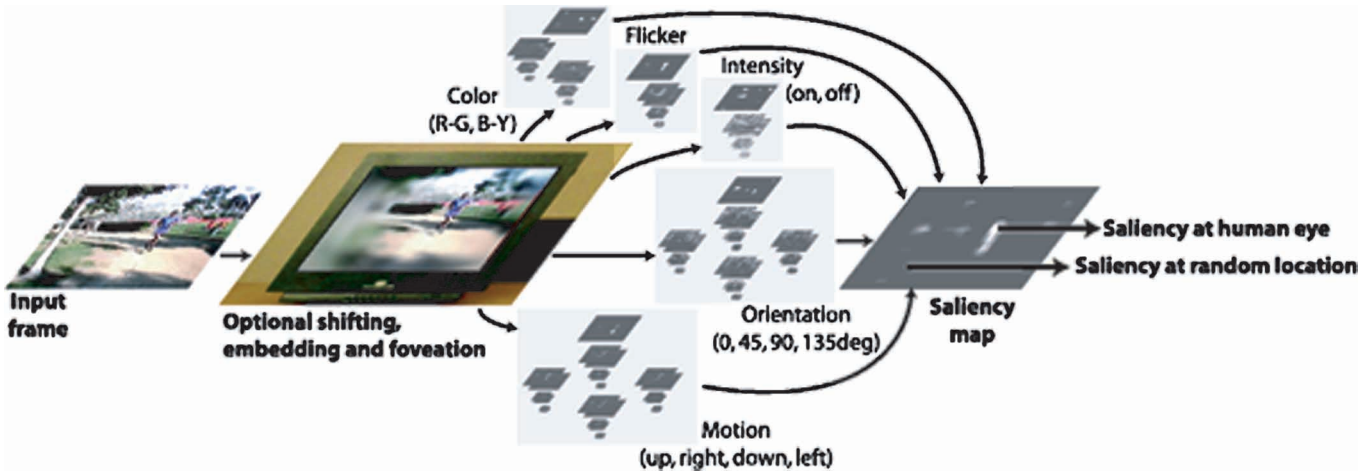

Figure 1. Overview of the computational framework. Depending on the model variant used, incoming video input may be used as it is, foveated around human eye position, or embedded within a background image, shifted so as to become centred at current human eye position, and foveated at the centre. Low-level visual features are then computed in a set of multiscale feature maps tuned to five classes of low-level visual properties (colour, intensity, orientation, flicker, motion). After nonlinear within-feature and across-feature competition for salience, all feature maps provide input to the saliency map. At each human eye position sample, measures are taken for the salience at eye position, the salience at a random location, and the maximum and average salience over the display area.

($\delta = \{3, 4\}$); thus, six feature maps are computed for each of the visual feature channels. 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; Itti et al., 1998). As a result, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings. All feature maps are then summed (Itti & Koch, 2001b) into the unique scalar input to the saliency map.

The saliency map is modelled as a two-dimensional layer of leaky integrator neurons at scale 4 (40×30 pixels given 640×480 video inputs) and with a time constant of 500 ms (Itti & Koch, 2000). Its main function is to provide temporal smoothing of the saliency values computed from each video frame. Consequently, in our 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 details the effects of stimulus–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, that is, 10,000 frames/s (Itti & Koch, 2000).

The spatial within-feature competition operated within each feature map is a crucial component of the model, mainly responsible for the model's ability to reproduce psychophysical "pop-out" effects observed behaviourally in visual search psychophysics (Treisman & Gelade, 1980; Wolfe, 1998). In a first approximation, the computational effect of the model's long-range, within-feature competitive 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). That is, the competitive interactions tend to enhance locations that are spatial outliers over the extent of the visual input. Hence, the salience of a patch of pixels at a given location in the visual field not only depends upon the local distribution of pixel values within that patch, but also on how that distribution compares to surrounding distributions. If the patch under consideration is similar to its neighbours (distant by up to approximately one-quarter of the width of the input image, with Gaussian decay; Itti & Koch, 2000), it will be strongly inhibited by its neighbours and will strongly inhibit them as well. This behaviour was included in the model to coarsely replicate nonclassical surround inhibition as observed in striate cortex and other early visual processing areas of mammals (Cannon & Fullenkamp, 1991; Levitt & Lund, 1997; Sillito, Grieve, Jones, Cudeiro, & Davis, 1995). Hence, while a vertical line segment would always excite a local vertical feature detector, irrespectively of the image contents far away from that stimulus, its salience will depend upon the possible presence of other vertical line segments throughout the entire visual field. If many other

vertical elements are present, the salience of the one under consideration will be low; but it will be much higher if no other vertical line element is present in the display (Itti & Koch, 2000). We have proposed several implementations for the long-range interactions, with varying degrees of biological plausibility and computational cost (Itti & Koch, 2000; Itti & Koch, 2001b; Itti et al., 1998). Here we use one iteration of the method introduced in Itti and Koch (2001b), which yields fairly sparse saliency maps as exemplified in Figure 4 (see later).

Foveation

To progress from local measures of image properties to measures of predicted perceptual salience at the point of gaze, a first improvement upon the simulation framework described above was to coarsely replicate the decrease of contrast sensitivity with eccentricity from the centre of fixation, as observed in humans (Spillmann & Werner, 1990; Wandell, 1995). To this end, we designed a simple foveation filter, which would blur the input video frames, increasingly so with distance from the current human eye position. Our results compare agreement between humans and model with and without this additional filter.

Recorded human eye position determined the location of the model's foveation centre. The foveation filter was achieved through interpolation across four levels of a colour Gaussian pyramid C_σ (Burt & Adelson, 1983) computed from each input frame, where the natural number σ represents spatial scale. Scale zero is the input colour image, and each subsequent scale is obtained by first blurring the image at the current scale by a 9×9 two-dimensional Gaussian kernel, and further decimating that low-pass-filtered image by a factor two horizontally and vertically:

$$\forall \sigma \geq 0, \quad C_{\sigma+1} = \downarrow^{x:2,y:2} (C_\sigma \otimes G_{9\times 9}) \tag{1}$$

where $\downarrow^{x:2,y:2}$ is the decimation (downsampling) operator, $\otimes$ is the convolution operator and $G_{9\times 9}$ is a (separable) 9×9 approximation to a Gaussian kernel. To determine which scale to use at any image location, we computed a 3/4-chamfer distance map $D(x,y)$ (Borgefors, 1991), encoding at every pixel with coordinates (x, y) for an approximation to the Euclidean distance between that pixel's location and a disc of 2° diameter (the model's fovea and perifovea), centred at the current eye position of the human observer:

$$\forall x, y, \quad \sigma(x, y) = K \times D(x, y) \tag{2}$$

where K is a scaling constant converting from distance into a fractional scale $\sigma(x,y)$. The fractional scale $\sigma(x,y)$ thus computed at every location in the image was used to determine the relative weights assigned to the immediately lower and higher integer scales represented in the pyramid C_σ, in a linear

interpolation scheme. Consequently, pixels close to the fovea were interpolated from high-resolution scales in the pyramid, while more eccentric pixels were interpolated from coarser-resolution scales. The process is illustrated in Figure 2. In a first step, we did not attempt to measure visual acuity of our subjects and to more accurately model the rate at which contrast sensitivity decreased with eccentricity for particular observers, although this could be included in a future implementation (Geisler & Perry, 2002). Although approximate, this model enhancement unveiled one additional contribution to perceptual saliency: Highly contrasted but small image regions far away from the observer's centre of gaze were unlikely to contribute much to the saliency map, since they were less or not detected by the low-level feature channels of the model once the foveation filter had been applied.

Input shifting and embedding

The realism of the simulation framework was further increased by dynamically shifting the input frames, so as to more faithfully simulate the changing patterns of retinal stimulation as the centre of gaze moved from one image location to another. That is, instead of considering that the model received inputs in the world-centred frame of reference attached to the stimulus frames, as implicitly assumed previously, we here transitioned to an eye-centred representation.

Hence, an explicit dissociation was made between retinotopic and world-centred (or head-centred since the head was fixed in our experiments) coordinate systems. The rationale for this enhancement was that the low-level visual processing of the model should operate not on the raw video frames, but on an approximation to the current retinal input received by the human observers. Thus, the first image processing step in the computational framework was to shift incoming raw video frames, such as to recentre them around the current human eye position. Hence, in the resulting images given as inputs to the model, human eye position was always at the centre of the field of view, and the raw video frames were shifted and pasted around that location. A background image consisting of a photograph of our experimental room and monitor (Figure 3) was used at retinal locations where there was no video clip input (e.g., at the top and left of the visual field when subjects fixated towards the top-left corner of the monitor). Finally, the image was cropped by a field of view larger than the raw video frames, centred at current eye position (Figure 3).

This transformation from stimulus to retinal coordinates unveiled two possible new contributions to perceptual salience that the simplified model did not account for: First, high contrast was often present at the boundary

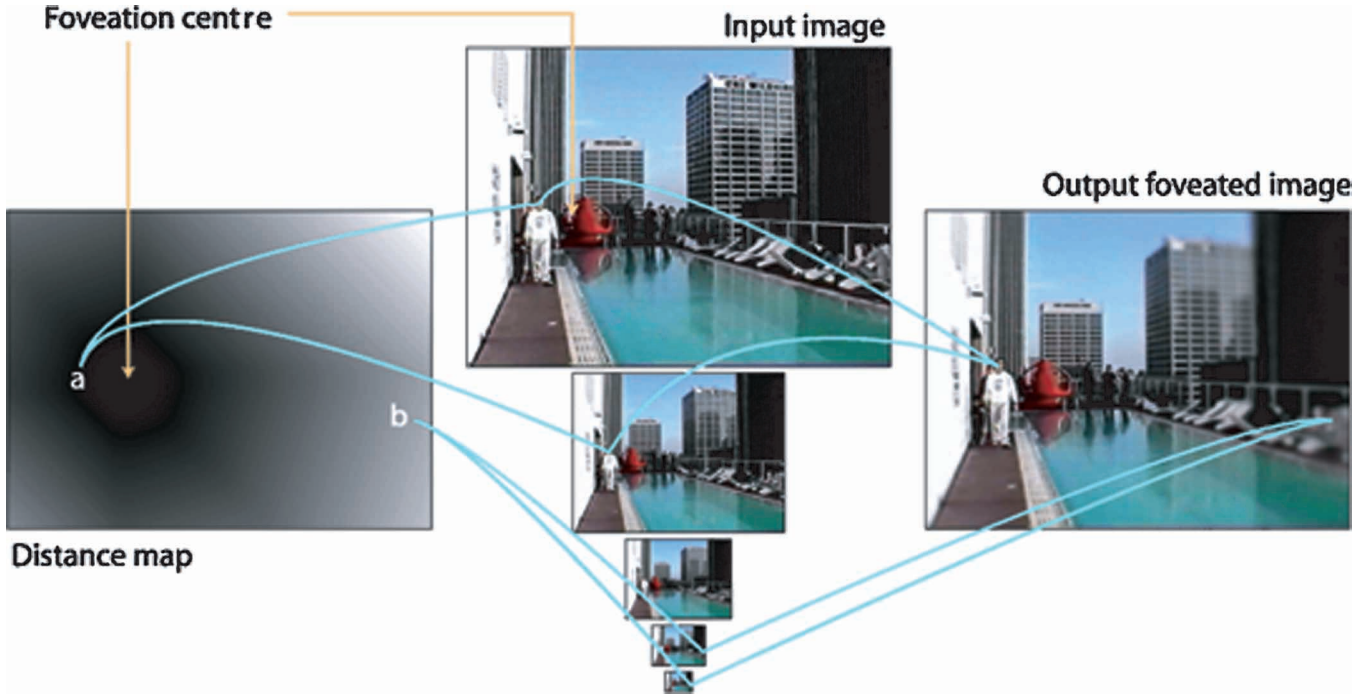

Figure 2. Foveation using a Gaussian image pyramid. Given a location for the foveation centre, the value at every location (x, y) in the distance map approximates the distance from the foveation centre to (x, y); brighter shades in the distance map correspond to larger distances. The appearance of (x, y) in the output foveated image is then read out from levels in the input image pyramid, which are increasingly coarse (and blurred) depending on the distance map value at (x, y). For example, the appearance of location a, close to the foveation centre, is interpolated from the two finest scales in the input pyramid, and hence is only lightly blurred. Conversely, the appearance of location b, more distant from the foveation centre, is interpolated from the two coarsest scales in the pyramid, and hence is highly blurred.

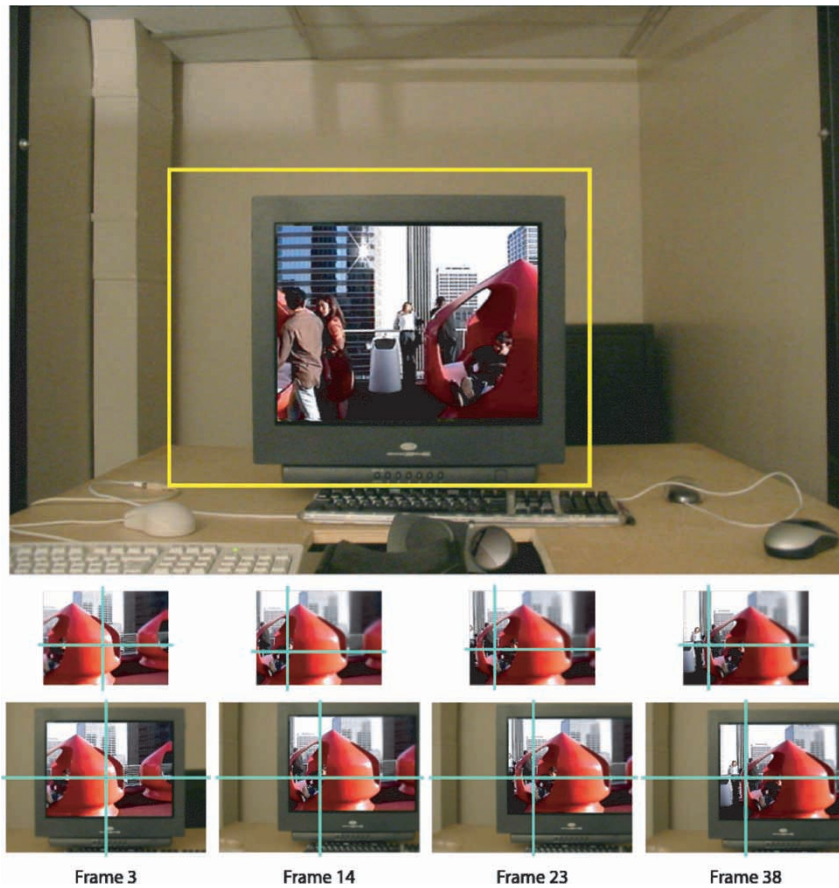

Figure 3. Embedding the video clips within a background. A 2048 × 1400 photograph of our eyetracker setup (top) was taken with a digital camera (here shown with slightly higher luminosity than actually used, for reproduction purposes). Incoming raw 640 × 480 frames from the video clips (middle; with current human eye position shown by the crosshairs) were pasted into this image, such as to appear within the monitor's screen area. Finally, the resulting image was cropped by a 1024 × 768 field of view centred at current human eye position (bottom, and also shown as a yellow rectangle in the top image), to simulate a human retinal field of view slightly larger than the monitor area. The resulting frames used as inputs to the shifted and foveated model thus had a fixed size of 1024 × 768 pixels, and depicted the room and monitor with the raw movie clip frames, shifting rapidly so that human eye position (crosshairs) would always remain at the centre of the field of view.

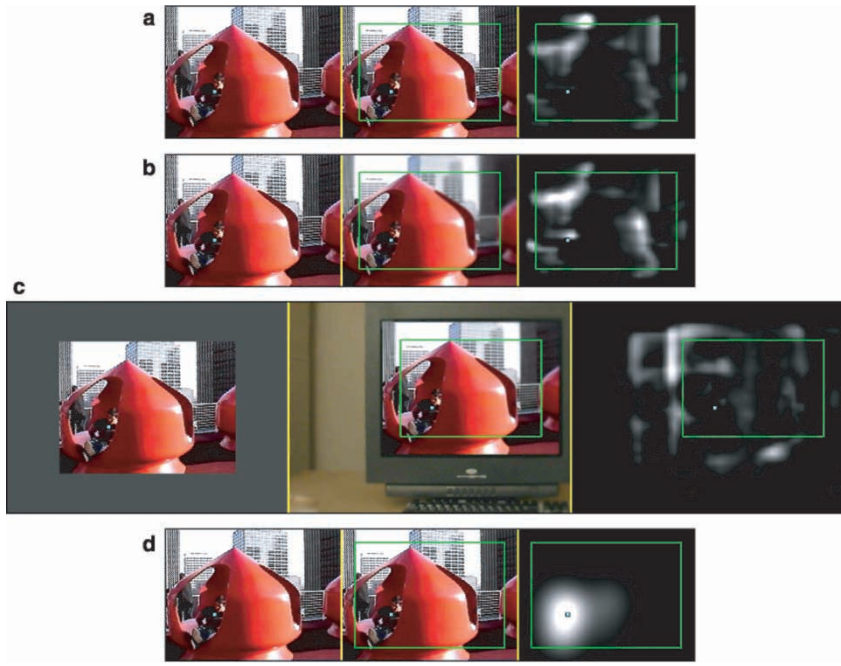

Figure 4. Example frame from a video clip, for the three model variants and the human-derived control model. The small cyan squares indicate current human eye position. (a) The simple model simply processes the raw video frames. (b) The second model variant applies a foveation filter centred at current human eye position before processing each frame. (c) The third model variant embeds the frame within a background, centres it to human eye position, crops it by a fixed field of view, and processes the contents of the field of view through the model. The green rectangles show the area over which maximum, average, and random salience were computed. (d) The human-derived control model uses a "saliency map" that contains three blobs, each corresponding to the instantaneous eye position of a human observer (with some temporal blurring provided by the internal dynamics of the saliency map). In the example frame shown, all three control observers are gazing at the same location as the fourth subject being tested (cyan square). This model provides a baseline for comparison with the bottom-up computational models.

between the shifted video frames and the background image, often strongly affecting model-predicted salience values within the video display area because of the long-range competitive interactions implemented in the model. Second, eye motion during human smooth pursuit and saccades significantly excited the model's low-level feature detectors sensitive to flicker and motion, thus adding an intrinsic motion component (due to the moving eye) in addition to extrinsic motion components (moving objects within the video clips).

Sample frames and saliency maps for the three model variants are shown in Figure 4.

RESULTS

At each human eye movement sample, several measurements were made on the model's current internal state, to quantify the extent to which salience at current eye position may compare to salience at other image locations.

The first of these measures was to read out salience and also low-level feature values (e.g., colour, orientation, etc.) at human eye position. To calibrate these measures, we also computed the maximum and average salience and feature values over the extent of the video display area, as well as salience and feature values at one random location (with uniform probability) within the display area. The goal of the random samples was to evaluate a baseline level of salience obtained by fixating a location by chance, for comparison to salience measured at human gaze location. Because we sampled a single point in the saliency and feature maps (which are at spatial scale 4, that is, downsized by a factor 16 horizontally and vertically compared to the original image), we expected that, on average over all frames, salience at random locations should be equal to the average salience over the entire frame. This was verified (see below), indicating that the random sampling worked properly.

There is one difficulty with defining what the random sampling area should consist of for the third variant of the model, which embeds the raw video frames within a larger background. With this model, the field of view over which feature and saliency maps are computed is larger than the actual video display (Figures 3 and 4). However, our human eye movement recordings contained no instance where the observers had looked either outside or right at the edges of the monitor's video display area. This was not a limitation of our eyetracking apparatus, which is capable of tracking over wider fields of view than the $28^\circ \times 21^\circ$ of our display monitor. We hence assumed that observers used top-down knowledge of the extent of the visible screen area to restrict their eye movements within that area. Under these conditions, it would seem an unfair comparison to allow random saccades to be distributed anywhere within the larger field of view. On the one hand, the background around the monitor often contained no or very few highly salient objects; including this area within the random sampling area hence would tend to artificially lower the average salience at random locations. On the other hand, the edges of the visible screen area often were salient, due to high contrast between the contents of the video frames and the dark grey plastic enclosure of the monitor, which would tend to artificially increase the average salience at random locations. Consequently, we here decided to only consider for random sampling (and for the computation of maximum and average salience) an area corresponding to the visible video display area, minus a border of three pixels at the scale of the saliency map (48 pixels at the scale of the original video frames; green rectangles in Figure 4).

Given this, the comparison between human and random samples concerns the following question: Given that observers knew to restrict their gaze to within the video display area, would the addition, in the simulations, of a surrounding background that observers never looked at influence the distribution of salience within the video display in a significant manner? If the model (and, presumably, the low-level visual processing stages of our observers) operated purely local processing, no influence would be expected. However, since model-predicted salience of a location within the video display area may be modulated by the presence of objects outside that area (due to the nonlinear spatial competition for salience at the core of our model), some significant effect was possible. It is important to remember that our goal in this study is not to make a point about the particular bottom-up visual attention used here, but rather to use this model to evaluate whether details like an ignored background scenery may significantly influence quantitative comparisons between model and human data.

The results of our measures are shown in Table 1. To allow comparison across model variants, the results are expressed in terms of a metric defined by the ratios:

$$F(x,y,t) = \frac{F_{\max}(t) - F(x,y,t)}{F_{\max}(t) - F_{avg}(t)} \tag{3}$$

where $F(x, y)$ is the feature (or salience) value at location (x, y) and time t, $F_{\max}(t)$ is the maximum such value over the sampling area and at instant t, and $F_{\text{avg}}(t)$ is the average such value over the sampling area and same instant t. The motivation for using this metric is that it is more robust to varying dynamic range and varying baselines than other, simpler metrics like $F(x,y,t)/F_{\text{avg}}(t)$ or $F(x,y,t)/F_{\max}(t)$.

The values shown in Table 1 are the values of the above metric, averaged over all video clips and observers (45 eye movement traces and a total of 324,036 valid eye position samples), for both human and random samples. A compound measure of 0% would indicate perfect agreement between humans and model, whereby humans would always gaze exactly at the single most salient pixel within the entire sampling area. In contrast, a compound measure of 100% would indicate that humans did not gaze at salient locations more than expected by chance, and measures above 100% would indicate that humans preferentially gazed at locations predicted to be of salience lower than average.

Clearly, a score of 0% is not possible as different observers often gazed to different locations while watching the same video clip. This intersubject variability makes it impossible for a model to highlight a single most salient pixel in the display that always corresponds to instantaneous human eye position. However, scores closer to 0% indicate models that more often

TABLE 1

Measures of agreement between model and humans (second column) and between model and random (third column), for saliency as well as individual visual features implemented in the model. A value of 0% would indicate that salience (or feature value) at the human (or random) sampling point was always the maximum over the entire display, while a value of 100% would indicate that it was average. In practice, 0% is unattainable because interobserver variability makes it impossible for a model to always exactly pinpoint a single best location for gaze. The score of 52.21% obtained when three humans predict one represents a more practical lower bound. The comparisons between human and random distributions of salience or feature values relied on a nonparametric sign test (fourth column). Finally, the fifth and sixth columns compare the values obtained for humans across two variants of the model, also using a sign test.

Visual feature	*Average human* $\frac{F_{\max}(t) - F(x,y,t)}{F_{\max}(t) - F_{avg}(t)}$	*Average random* $\frac{F_{\max}(t) - F(x,y,t)}{F_{\max}(t) - F_{avg}(t)}$	*Compare human to random*	*Compare human to human NS.NF*	*Compare human to human NS.F*
Three humans predict one:					
Humans	**52.21%**	**100.02%**	$p < 10^{-10}$		
No shifting/embedding, no foveation (NS.NF):					
Salience	**87.51%**	**99.97%**	$p < 10^{-10}$		
Colour	93.38%	100.00%	$p < 10^{-10}$		
Flicker	93.71%	99.98%	$p < 10^{-10}$		
Intensity	97.54%	99.98%	$p < 10^{-10}$		
Orientation	93.80%	100.02%	$p < 10^{-10}$		
Motion	93.73%	100.01%	$p < 10^{-10}$		
No shifting/embedding, foveation (NS.F):					
Salience	**81.85%**	**100.05%**	$p < 10^{-10}$	$p < 10^{-10}$	
Colour	93.44%	100.03%	$p < 10^{-10}$	$p < 10^{-10}$	
Flicker	87.13%	99.98%	$p < 10^{-10}$	$p < 10^{-10}$	
Intensity	97.18%	99.97%	$p < 10^{-10}$	$p \geq 0.1$	
Orientation	87.02%	99.97%	$p < 10^{-10}$	$p < 10^{-10}$	
Motion	92.01%	99.98%	$p < 10^{-10}$	$p < 10^{-10}$	
Shifting/embedding, foveation (S.F):					
Salience	**77.24%**	**99.94%**	$p < 10^{-10}$	$p < 10^{-10}$	$p < 10^{-10}$
Colour	95.85%	99.96%	$p < 10^{-10}$	$p < 10^{-10}$	$p < 10^{-10}$
Flicker	84.41%	100.03%	$p < 10^{-10}$	$p < 10^{-10}$	$p < 10^{-10}$
Intensity	97.34%	99.96%	$p < 10^{-10}$	$p < 10^{-10}$	$p < 10^{-10}$
Orientation	89.82%	99.99%	$p < 10^{-10}$	$p < 10^{-10}$	$p < 10^{-10}$
Motion	90.11%	99.98%	$p < 10^{-10}$	$p < 10^{-10}$	$p < 10^{-10}$

correlate with human behaviour. To provide a practical lower bound for scores that may be attainable by a computational model, we applied our scoring scheme to an alternate model, whose saliency map was derived from eye movement traces of three humans and tested against a fourth one. In this human-derived model, the "saliency map" contained three continuously moving Gaussian blobs (standard deviation $\sigma = 0.7^\circ$), which followed in real-time the current eye positions of three human subjects who had watched

the video clip of interest. By sampling this human-derived saliency map along the gaze locations of a fourth observer, we obtained a score of 52.21% for how well that fourth observer could have been predicted by the other three (Table 1). The human-derived model encompasses both bottom-up and top-down factors, some of which are well beyond the capabilities of our computational bottom-up saliency models (e.g., building a cognitive understanding of the story depicted in a video clip). Our computational models are hence expected to score somewhere between 100% (chance level) and approximately 50% (a remarkably good model that could predict human eye movements as well as other humans could).

Results for the first bottom-up saliency model variant (no shifting/embedding of the video frames, no foveation filter) confirm with dynamic video stimuli previous results indicating that human observers preferentially gaze at locations with model-predicted salience higher than average (Parkhurst et al., 2002), with a compound metric value of 87.51%. In contrast, the compounded metric for random samples was very close to 100%, as expected above. A nonparametric sign test of whether human and random measures for all 324,036 valid eye position samples could have been drawn from distributions with same median suggested highly significant differences between human and random distributions ($p < 10^{-10}$). It is interesting to note that the average metric was better for salience than for any of the individual visual features contributing to salience. This suggests that some of the locations visited by our observers were salient according to the model not only for a single feature (e.g., high colour contrast) but for combined features (e.g., high colour contrast and high motion energy). The figure of 87.51% for salience suggests that our simple, purely bottom-up model does predict locally higher salience at image regions likely to be visited by human observers, in a highly significant manner. But, obviously, there is more to human vision than bottom-up salience: 87.51% is rather far from the score near 50% achieved by our human-derived model. This point is further explored in our discussion; here we accept this figure as a baseline corresponding to a low-realism version of the simulation framework, and compare it to the other two model variants with increased realism.

Results for the second model variant (adding a foveation filter around current human eye position) overall improved the metric for humans (down to 81.85%), whereas random figures remained close to 100%. Comparing human metric values for the various features between the first and second model variants suggested a mixed pattern of changes, though all were highly significant except for the intensity feature. Colour, intensity, and motion energy metrics were little affected by the addition of a foveation filter, whereas flicker and orientation improved more substantially. Hence, the model benefited from foveation predominantly in that it reduced the competitive influence of previously salient fine oriented textures and flicker

far from fixation (often present due to the fairly low quality of our video clips, digitized from analogue, interlaced NTSC video sources), which would tend to lower salience at fixation due to the long-range interactions. Irrespective of the details of the different feature responses, and keeping in mind that the model used here is only a very coarse approximation to low-level human vision and that its internals should not be overinterpreted, comparing first and second model variants showed that adding a foveation filter had a highly significant outcome onto the quantitative measure of agreement between model and humans.

Results with the third variant of the model (including embedding the video frames within a larger background, shifting to eye position, cropping around that position by a field of view larger than the video display area, and foveating the resulting image) fairly surprisingly suggested even better agreement between model and humans (metric down to 77.24% for the salience values). We believe this result was surprising because shifting the input, in a coarse attempt to simulate the dynamic motions of a retina, makes the task of computing perceptual salience with a computational model much more difficult, as large motion transients and smearing of the saliency map during saccades may occur (see Discussion). Interestingly, the better agreement for the salience measure was obtained despite the fact that slightly worse agreement was obtained, for this third variant than for the second, for the colour, intensity, and orientation features. Remarkably, the two features we had originally thought would suffer from rapidly shifting the input (and the associated high transients), flicker and motion, actually improved in their agreement with humans between the second and third variants of the model. This suggests that the simulation operated with the third variant may actually more closely approximate the patterns of motion and flicker signals received by a moving human retina. Again, one should be careful, however, not to overinterpret these detailed results and to keep in mind that the model only simulates a very approximate and small fraction of human visual processing. As for the second variant, however, the high-level conclusion for the third variant of the model was that the increased realism of the simulation highly significantly affected the outcome of the comparison between model and humans.

To summarize, Table 2 shows the results of a one-way ANOVA test for whether model variant was a factor that significantly influenced our metric, for both humans and random. The test suggested that the metric for salience as well as for all features was highly significantly affected for humans, but not nearly as much for random. Note, however, that the random salience metric was slightly affected by model variant, $F(2,\ 972{,}105) = 3.82$ (three groups for three model variants, and $3 \times 324{,}036 = 972{,}108$ data samples where valid measurements could be sampled), which would be considered a significant result in typical statistical analysis (with the conclusion that

TABLE 2
Analysis of variance (ANOVA) to test whether model variant was a factor in our measures of agreement between models and humans (ANOVA, human) and between models and random (ANOVA, random), for saliency as well as the individual visual features implemented in the models. All *F*-values reported here are for three groups (corresponding to the three model variants) and $3 \times 324{,}036 = 972{,}108$ data points where valid measurements could be sampled, hence should be read as $F(2, 972105)$.

Feature	*ANOVA, Human*		*ANOVA, Random*	
salience	$F = 9{,}803.86$	$p < 10^{-10}$	$F = 3.82$	$p \geq .02$
color	$F = 1{,}015.15$	$p < 10^{-10}$	$F = 2.04$	$p \geq .13$
flicker	$F = 11{,}275.84$	$p < 10^{-10}$	$F = 2.11$	$p \geq .12$
intensity	$F = 32.17$	$p < 10^{-10}$	$F = 0.04$	$p \geq .96$
orientation	$F = 5{,}761.22$	$p < 10^{-10}$	$F = 0.66$	$p \geq .52$
motion	$F = 2{,}194.56$	$p < 10^{-10}$	$F = 0.61$	$p \geq .54$

model variant was a factor, with significance $p < .02$, rather than the conclusion that it was not a factor, with $p \geq .02$, as reported in Table 2). Indeed, one difference which affected random sampling in the third variant of the model was that rarely only a portion of the video display area fell within the field of view of the model (e.g., frames 23 and 38 in Figure 3 are slightly cut off on the right side as the human eye position is close to the left side), resulting in a slightly truncated sampling window over which maximum, average, and random salience were computed. Nevertheless, the random metric was clearly affected to an extent that simply does not compare with the extremely high *F*-values found for the human metric. Of all features, the metric for intensity was the least affected by model variants, with colour a far second. The high-level conclusion here again is that the level of detail used for the simulations very highly significantly affected the agreement between humans and model, but very little affected that between random sampling and the model.

Finally, a breakdown histogrammed by metric values for the salience measures is shown in Figure 5. This shows how, for all three model variants, the basic difference between humans and random was that the distributions of metric values for humans had heavier tails towards 0% (salience at point of gaze higher than on average over the video display area), whereas the distributions for random were more concentrated around 100% (salience at sampled location close to the average over the video display area). This difference is increasingly emphasized from the first to second and to third model variants, hence showing in a graphical manner why the metric values reported in Table 1 also became better with increased simulation realism.

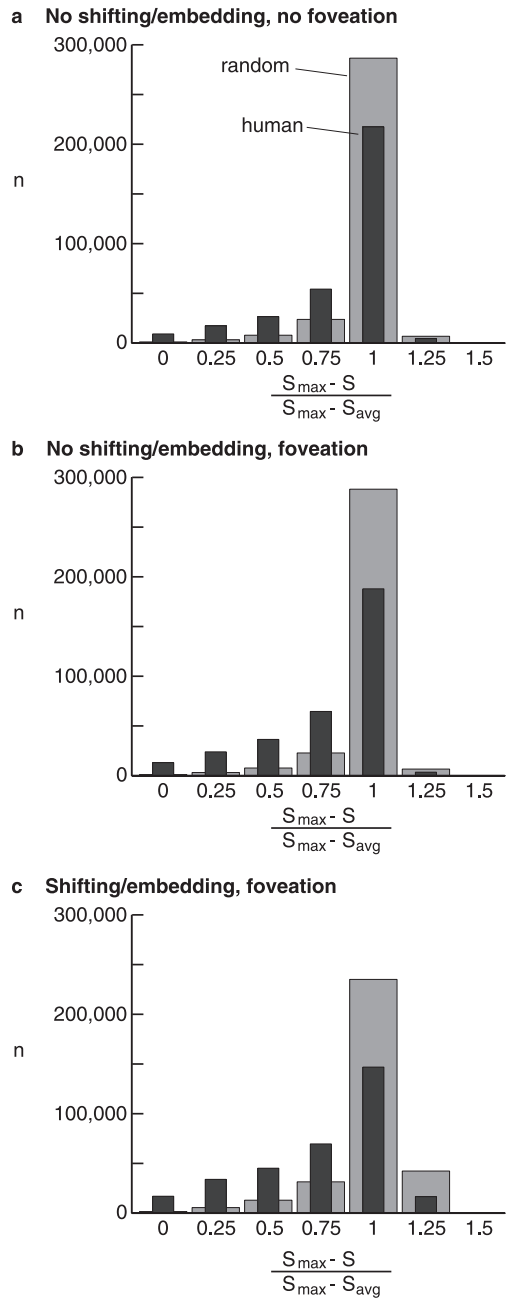

Figure 5. Histograms of the distributions of our metric values for humans (thin dark grey bars) and random (wide light grey bars), for the three model variants.

DISCUSSION

This study represents a first attempt at investigating whether increased simulation realism could significantly affect quantitative comparisons between a computational model and human behavioural data.

It is interesting that, of all features implemented in the model, intensity contrast seemed to be overall the least affected by increased simulation realism. Indeed, the computation of that feature is instantaneous and memoryless in the model, performed independently on every frame. This could explain why intensity contrast was less affected by input shifting. However, the colour and orientation features also are instantaneous and memoryless in our model (as opposed to the flicker and motion features, which are computed from differential comparisons between successive video frames) but were affected more strongly. One positive outcome of this observation is that the conclusions of previous eye movement studies (Barth et al., 1998; Reinagel & Zador, 1999; Zetzsche et al., 1998), which largely focused on contrast measures and greyscale static images, might not have been affected too strongly by lower simulation realism.

There are many shortcomings and limitations to the implementation of our model, which should be considered a coarse first pass. First, we did not attempt to calibrate our foveation filter to the visual acuity profiles of our observers (Geisler & Perry, 2002). Second, the 1024×768 field of view used in the third variant of our model, corresponding to approximately $45^\circ \times 33^\circ$ of visual angle, was much smaller than a typical $160 \times 175^\circ$ subtended by each human retina (Wandell, 1995). There were two main reasons which restricted our model's field of view: One was computer memory available to run the simulations (remember that 72 feature maps are computed from the input image); the other was that obtaining a background image of this width would require special apparatus, and a more sophisticated method to simulate the rotation of the eye than the planar shifting used here in a first approximation. Thus, our study is limited in that the possible influence of objects or otherwise salient stimuli in the very far visual periphery is not accounted for.

Overall, the metric values for all three variants of the model, ranging from 77.24% to 87.51%, were honourable compared to the score near 52% of the human-derived model—in particular, the fully detailed model scores approximately half-way between the human-derived model and chance. As briefly reviewed in our introduction, we did not expect perfect agreement given all the additional factors (biasing by the gist of a scene, top-down guidance of search, etc.) that contribute to eye movements and are not simulated by the model (Itti, 2003). It is interesting to note that the additional machinery implemented in our model compared to, for example, a simple local contrast detector, significantly improved the correlation between

humans and model (the metric values for salience were always better than for intensity contrast alone). Although it is not our main aim here to advocate that the specific bottom-up saliency model tested in this study can account for a significant fraction of human gaze allocation, the fact that the models perform significantly below 100% is interesting as it suggests that bottom-up salience plays a sustained role in attracting attention in dynamic video scenes. Another study explores this question in further detail (Itti, in press).

We found it surprising that the third variant of the model performed best. Indeed, as the input image rapidly shifts during saccades in this model variant, possibly large motion transients may be elicited, which could easily overwhelm the saliency map. Remarkably, however, these were seldom observed in our dynamic saliency maps, suggesting that the long-range competition for salience implemented in our model was very efficient at suppressing full-field motion transients. Indeed, shifting the entire input at a given velocity and in a given direction would excite a single motion feature map (tuned to that velocity and direction) throughout the entire visual field. Strong local motion responses would thus be elicited in that feature map, but at many locations in the visual field, so that the entire feature map would become inhibited by the long-range competition mechanism because not containing any single location that significantly differs from most other locations. A consequence of this for human vision is that, at least in our model, an explicit mechanism for top-down saccadic suppression may be unnecessary (Thiele, Henning, Kubischik, & Hoffmann, 2002), although this remains to be tested (i.e., maybe even better agreement between model and humans could be obtained after addition of explicit saccadic suppression to the model). Additional difficulties which would tend to make the simulation with the third model variant more difficult include smearing of the low-pass filtered saliency map as the input shifted (which we often observed when inspecting the dynamic saliency maps predicted for various video clips), competition between possibly salient objects outside the video screen area and objects within that area, and lack of sufficient persistence to allow for salience to build up over time (remember that the saliency map is modelled as leaky integrator neurons, which respond better when inputs are somewhat stable). By computer vision standards, these may be regarded as artifacts, limitations of the feature detectors, or fundamental problems that inevitably deteriorate the output of the model. However, our results suggest that they actually might be a feature rather than a shortcoming of the model, in that biological early vision is subjected to similar situations as the eye moves, in a manner that eventually will affect biological computation of salience and eye movements.

It is important to again stress that one should not overinterpret the absolute quantitative measures of agreement between model variants and human scan paths. Indeed, these absolute numbers are highly dependent upon software implementation details of our model, various model

parameters that we have here not attempted to tune, such as the strength of the long-range interactions, and even architectural choices, such as which preattentive visual features may guide attention (Wolfe & DiMase, 2003). Thus, the important conclusion of our work is not in the absolute performance value of each model variant, but in the comparison between the three variants. The fact that increased realism affected the comparison outcomes in such a significant manner stresses how attempting simulations that are as realistic as possible is important when using computational models to interpret empirical data. Translating these findings to human vision, in particular for psychophysical studies like visual search, our results strongly caution against an often fairly intuitive interpretation of the data in terms of the intrinsic local features of the target and distractor items. Instead, the outcome of a search may much more strongly be influenced than one may think at first by the spatiotemporal dynamics of eye movements operated over the course of the search.

REFERENCES

Barth, E., Zetzsche, C., & Rentschler, I. (1998). Intrinsic two-dimensional features as textons. *Journal of the Optical Society of America. Part A, Optics, and 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, Memory, and Cognition*, *9*(3), 411–429.

Borgefors, G. (1991). Distance transformations in digital images. *CVGIP: Image Understanding*, *54*, 301.

Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, *31*, 532–540.

Cannon, M. W., & Fullenkamp, S. C. (1991). Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Research*, *31*(11), 1985–1998.

Finney, S. A. (2001). Real-time data collection in Linux: A case study. *Behavior 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.

Geisler, W. S., & Perry, J. S. (2002). *Real-time simulation of arbitrary visual fields*. Paper presented at the ACM symposium on Eye Tracking Research and Applications.

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 Sciences*, *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.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, *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. (2003). Modeling primate visual attention. In J. Feng (Ed.), *Computational neuro-science: A comprehensive approach* (pp. 635–655). Boca Raton, FL: CRC Press.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*(10), 1304–1318.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*, 1093–1123.

Itti, L., Dhavale, N., & Pighin, F. (2003, August). Realistic avatar eye and head animation using a neurobiological model of visual attention. In B. Bosacchi, D. B. Fogel, & J. C. Bezdek (Eds.), *Proceedings of the SPIE 48th Annual International Symposium on Optical Science and Technology* (Vol. 5200, pp. 68–78). Bellingham, WA: SPIE Press..

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 Reviews 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.

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

Kofka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt & Brace.

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

Levitt, J. B., & Lund, J. S. (1997). Contrast dependence of contextual effects in primate visual cortex. *Nature*, *387*(6628), 73–76.

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 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.

Müller, H. J., Reimann, B., & Krummenacher, J. (2003). Visual search for singleton feature targets across dimensions: Stimulus- and expectancy-driven effects in dimensional weighting. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(5), 1021–1035.

Nodine, C. F., & Krupinski, E. A. (1998). Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. *Academic Radiology*, *5*(9), 603–612.

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

Oliva, 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.

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.

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

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

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

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.

Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and Gestalt rules. *Proceedings of the National Academy of Sciences, USA*, *98*(4), 1935–1940.

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

Spillmann, L., & Werner, J. S. (1990). *Visual perception: The neurophysiological foundations*. San Diego, CA: Academic Press.

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.

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.

Thiele, A., Henning, P., Kubischik, M., & Hoffmann, K. P. (2002). Neural mechanisms of saccadic suppression. *Science*, *295*(5564), 2460–2462.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America. Part A, Optics, and 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., & Maunsell, J. H. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, *382*(6591), 539–541.

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

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron*, *31*(4), 523–535.

Wandell, B. (1995). *Foundations of vision*. Sunderland, MA: Sinauer Associates.

Wolfe, J. (1998). Visual Search. In H. Pashler (Ed.), *Attention*. London: University College London Press.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, *1*, 202–238.

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 Psychology: Human Perception and Performance*, *15*(3), 419–433.

Wolfe, J. M., & DiMase, J. S. (2003). Do intersections serve as basic features in visual search? *Perception*, *32*(6), 645–656.

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

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 *From animals to animals: Proceedings of the fifth international conference on Simulation of Adaptive Behavior* (Vol. 5, pp. 120–126). Cambridge, MA: MIT Press.