Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention

Laurent Itti

***Abstract*—We evaluate the applicability of a biologically-motivated algorithm to select visually-salient regions of interest in video streams for multiply-foveated video compression. Regions are selected based on a nonlinear integration of low-level visual cues, mimicking processing in primate occipital, and posterior parietal cortex. A dynamic foveation filter then blurs every frame, increasingly with distance from salient locations. Sixty-three variants of the algorithm (varying number and shape of virtual foveas, maximum blur, and saliency competition) are evaluated against an outdoor video scene, using MPEG-1 and constant-quality MPEG-4 (DivX) encoding. Additional compression radios of 1.1 to 8.5 are achieved by foveation. Two variants of the algorithm are validated against eye fixations recorded from four to six human observers on a heterogeneous collection of 50 video clips (over 45 000 frames in total). Significantly higher overlap than expected by chance is found between human and algorithmic foveations. With both variants, foveated clips are, on average, approximately half the size of unfoveated clips, for both MPEG-1 and MPEG-4. These results suggest a general-purpose usefulness of the algorithm in improving compression ratios of unconstrained video.**

***Index Terms*—Bottom up, eye movements, foveated, saliency, video compression, visual attention.**

I. Introduction and Background

AN INCREASINGLY popular approach to reduce the size of compressed video streams is to select a small number of interesting regions in each frame and to encode them in priority. This spatial prioritization scheme relies on the highly nonuniform distribution of photoreceptors on the human retina, by which only a small region of $2 - 5^\circ$ of visual angle (the fovea) around the center of gaze is captured at high resolution, with logarithmic resolution falloff with eccentricity [1]. Thus, the rationale is that it may not be necessary or useful to encode each video frame with uniform quality, since human observers will crisply perceive only a very small fraction of each frame, dependent upon their current point of fixation. In a simple approach (used here), priority encoding of a small number of image regions may decrease overall compressed file size by tolerating additional degradation in exchange for increased compression outside the priority regions. In more sophisticated approaches, priority encoding may serve to temporally sequence content delivery (deliver priority regions first), or to continuously scale video quality depending on available transmission bandwidth (encode priority regions as the core of a compressed stream, with additional details transmitted only as additional bandwidth is available [2]–[4]).

The selection of priority regions remains an open problem. Recently, key advances have been achieved in at least two contexts. First, real-time interactive gaze-contingent foveation for video transmission over a bandwidth-limited communication channel, and, second, priority encoding for general-purpose noninteractive video compression. Gaze-contingent video transmission typically uses an eye-tracking device to record eye position from a human observer on the receiving end, and applies in real time a foveation filter to the video contents at the source [5]–[11]. Thus, most of the communication bandwidth is allocated to high-fidelity transmission of a small region around the viewer's current point of regard, while peripheral image regions are highly degraded and transmitted over little remaining bandwidth. This approach is particularly effective, with observers often not noticing any degradation of the signal, if well matched to their visual system and viewing conditions. Even in the absence of eye tracking, this interactive approach has demonstrated usefulness, for example when there exists a set of fixed priority regions, or when observers explicitly point to priority regions [12]. Further, online analysis of the observer's patterns of eye movements may allow more sophisticated interactions than simple foveation (e.g., zooming-in and other computer interface controls [13]). However, extending this approach to general-purpose noninteractive video compression presents severe limitations.

In the context of general-purpose video compression, indeed, it is assumed that a single compressed video stream will be viewed by many observers, at variable viewing distances, and in the absence of any eye tracking or user interaction. Very high inter-observer variability then precludes recording a single eye movement scanpath from a reference observer and using it to determine priority regions in the video clip of interest. Recording from several observers and using the union of their scanpaths partially overcomes this limitation [14], but at a prohibitive cost: An eye-tracking setup, population of observers, and time-consuming experiments are required for every new clip to be compressed.

Algorithmic methods, requiring no human testing have the potential of making the process practical and cost-effective [15], [16]. Computer vision algorithms have, thus, been proposed to automatically select regions of high encoding priority. Of particular interest here, several techniques rely on known properties of the human visual system to computationally define perceptually important image regions (e.g., based on object size, contrast, shape, color, motion, or novelty [17]–[19]). This type

Manuscript received June 24, 2003; revised June 24, 2004. This work was supported by NSF, NEI, NIMA, the Zumberge Fund, and the Charles Lee Powell Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fernando M. B. Pereira.

The author is with the Departments of Computer Science, Psychology and Neuroscience Graduate Program, University of Southern California, Los Angeles, CA 90089-2520 USA (e-mail: itti@usc.edu).

Digital Object Identifier 10.1109/TIP.2004.834657

of approach has been particularly successful, and those properties which are well defined (e.g., contrast sensitivity function, importance of motion, and temporal masking effects) are already implemented in modern video and still-image codecs [20], [21]. A limitation of these approaches, however, is that the remaining properties of human vision are difficult to algorithmically implement (e.g., evaluating object size and shape requires that first object segmentation be solved in a general manner). In contrast, an important contribution of the present study is to evaluate the applicability of a computational model that mimics the well-known response characteristics of low-level visual processing neurons in the primate brain, rather than attempting to implement less well-defined, higher-level visual properties of objects and scenes. In addition, many of the existing computational algorithms have typically been developed for specific video content (e.g., giving preference to skin color or facial features, under the assumption that human faces should always be present and given high priority [4], [22], or learning image processing algorithms that maximize overlap with human scanpaths on a specific set of images [23]) and, thus, are often not universally applicable. Instead, our model makes no assumption on video contents, but is strongly committed to the type of neuronal response properties found in early visual areas of monkeys and humans. Finally, computational algorithms have thus far typically been demonstrated on a small set of video clips and often lack validation against human eye movement data. Another important contribution of our study is, hence, to widely validate our algorithm against eye movements of human observers [24], [25].

Validating an algorithm against human behavior is a challenging test, as human eye movements are influenced by many factors, often tied to higher cognitive understanding of semantic and affective scene contents rather than low-level image properties [26]–[30]. These factors include recognizing the scene's gist (broad semantic category, such as indoors or outdoors) and layout, which may provide priors on the probable locations of objects of interest and facilitate their recognition [31]–[36]. In addition, when specific objects are actively searched for or ignored, low-level visual processing may be biased for or against the visual features of these objects [37]–[42], resulting in a top-down guidance of visual attention toward objects of interest [43], [44]. Task, training, and general expertise also affect eye movements, in part as the recognition of one object may provide clues as to the location of another [26], [45]–[50]. Similarly, memorizing objects found along an initial exploratory scanpath may allow observers to later efficiently orient back to some of these objects [51], [52]. Finally, different observers exhibit different eye movement idiosyncrasies, possibly resulting from different internal world representations [53], search strategies, personal preferences, culture, and other factors [54]. Thus, a last important contribution of the present study is to investigate whether a simple visual processing algorithm, which does not attempt to understand the semantic contents of a given video clip, may reasonably well correlate with human eye movements.

In the following, we start by describing our neurobiological model of visual attention, which automatically selects regions of high saliency (conspicuity) in unconstrained video inputs, without requiring any per-clip tuning. As the low-level visual processing at the basis of this model has been previously described in details [29], [55]–[57], we focus here on specific new model additions for the prediction of priority regions in video streams. We then proceed with a systematic investigation of 63 variants of the algorithm, demonstrating a range of achievable tradeoffs between compressed file size and visual quality. We then validate the algorithm, for two of the 63 settings, on a heterogeneous collection of 50 video clips, including synthetic stimuli, outdoors daytime and nighttime scenes, video games, television commercials, newscast, sports, music video, and other content. Using eye movement recordings from eight human subjects watching the unfoveated clips (four to six subjects for each clip), we show that subjects preferentially fixate locations which the model also determines to be of high priority, in a highly significant manner. We finally compute the additional compression ratios obtained on the 50 clips using the foveation centers determined by our model, demonstrating the usefulness of our approach to the fully automatic determination of priority regions in unconstrained video clips.

II. Attention Model and Foveation

The model computes a topographic saliency map (Fig. 1), which indicates how conspicuous every location in the input image is [29], [58]. Retinal input is processed in parallel by multiscale low-level feature maps, which detect local spatial discontinuities using simulated center-surround neurons [59], [60]. Twelve neuronal features are implemented, sensitive to color contrast (red/green and blue/yellow, separately), temporal flicker (onset and offset of light intensity, combined), intensity contrast (light-on-dark and dark-on-light, combined), four orientations (0°, 45°, 90°, 135°), and four oriented motion energies (up, down, left, and right), as previously described [29], [55], [61]. The extent to which the low-level features used here attract attention in humans and monkeys has been previously investigated in details [29], [44], [62]. Center and surround scales are obtained using dyadic pyramids with nine scales (from scale 0, the original image, to scale 8, the image reduced by a factor 256). Center-surround differences are then computed as pointwise differences across pyramid scales, for combinations of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($\delta = \{3, 4\}$); thus, six feature maps are computed for each of the 12 features, yielding a total of 72 feature maps. 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 [57]. Resultingly, 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 [57] into the unique scalar saliency map that guides attention (Fig. 1).

The basic operation of the algorithm is as follows in the context of video compression: A dynamic saliency map is computed as described above, over the entire duration of each video clip. In one variant of the algorithm, a snapshot of the saliency map at each frame directly determines the priority to be given to every spatial location in the frame, after normalization by a squashing function and temporal smoothing [Fig. 2(a)]. For every pixel with image coordinates $X = (x, y)$ and instanta-

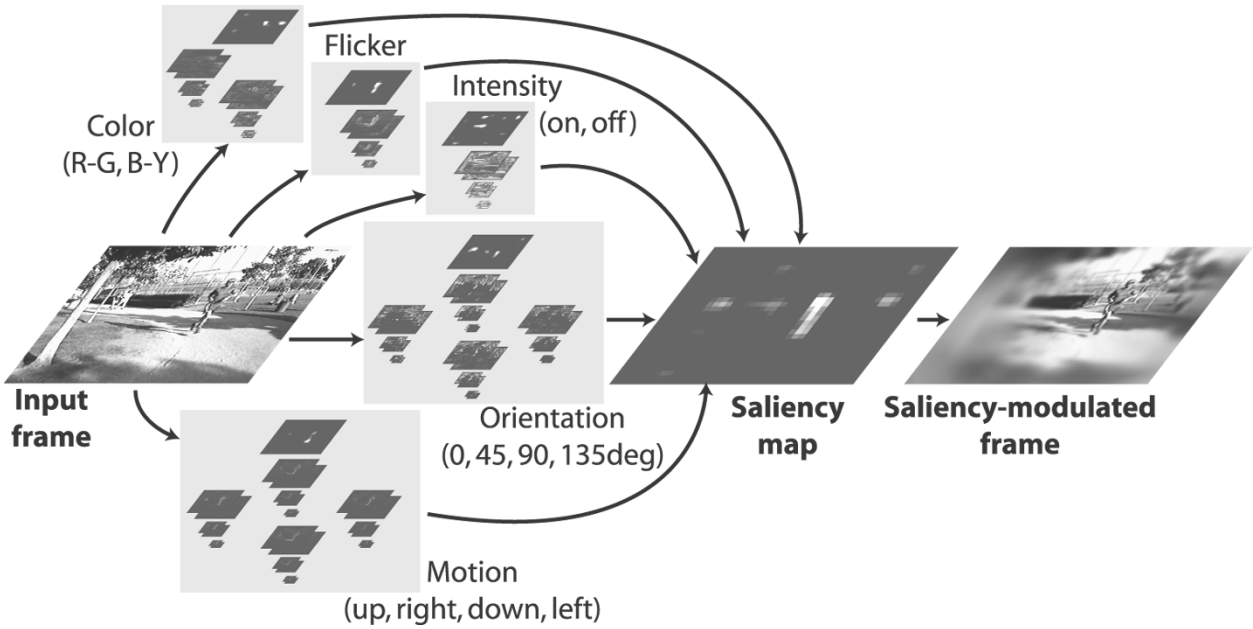

Fig. 1. Overview of the model. Inputs are decomposed into multiscale analysis channels sensitive to low-level visual features (two color contrasts, temporal flicker, intensity contrast, four orientations, and four directional motion energies). After strong nonlinear competition for saliency, all channels are combined into a unique saliency map. This map either directly modulates encoding priority (higher priority for more salient pixels), or guides several virtual foveas toward the most salient locations (highest priority given to fovea centers).

neous saliency value $s_t(X)$ at frame t, the saliency value is first squashed by a remapping function r_t, to downplay values below average and give more dynamic range to values above average. r_t is a fourth-order sigmoidal interpolation polynomial satisfying $r_t(a_t) = d_t$, $r_t(b_t) = e_t$, $r_t(c_t) = f_t$, $r'_t(a_t) = 0$ and $r'_t(c_t) = 0$ with $a_t < b_t < c_t$, and $d_t < e_t < f_t$. Thus, r_t has horizontal slope at both ends of an input range of values $[a_t \dots c_t]$ and remaps that range to a new range $[d_t \dots f_t]$ while also remapping a midpoint b_t to value e_t. Adjustment of this midpoint allows to give more emphasis to higher or lower values in the remapping. The functional expression for r_t is derived from the above constraints, yielding

$$\forall t \in \mathcal{N}, x \in \mathcal{R}\ r_t(x) = \frac{1}{Z_t}(P_t + Q_t x + R_t x^2 + S_t x^3 + T_t x^4) \quad \text{with} \tag{1}$$

$$\begin{aligned} P_t = {} & c_t^2(b_t(b_t - c_t)^2 d_t(-4a_t^2 + 3a_t b_t + 2a_t c_t - b_t c_t) \\ & + a_t^2(a_t - c_t)^3 e_t) + a_t^2(a_t - b_t)^2 b_t(a_t(b_t - 2c_t) \\ & + c_t f_t(-3b_t + 4c_t)) \end{aligned} \tag{2}$$

$$\begin{aligned} Q_t = {} & 2a_t c_t(c_t^4(e_t - d_t) - (3b_t^4 - 4b_t^3 c_t)(d_t - f_t) + (2a_t^3 c_t - a_t^4) \\ & (e_t - f_t) + 2a_t(c_t^3(d_t - e_t) + (2b_t^3 - 3b_t^2 c_t)(d_t - f_t))) \end{aligned} \tag{3}$$

$$\begin{aligned} R_t = {} & (a_t^5 + a_t^4 c_t - 8a_t^3 c_t^2)(e_t - f_t) + (a_t + c_t)(c_t^4(d_t - e_t) \\ & + (3b_t^4 - 4b_t^3 c_t)(d_t - f_t)) - 4a_t^2(2c_t^3(d_t - e_t) \\ & + (b_t^3 - 3b_t c_t^2)(d_t - f_t)) \end{aligned} \tag{4}$$

$$\begin{aligned} S_t = {} & 2(c_t^4(e_t - d_t) + 2a_t^2(b_t - c_t)^2(d_t - f_t) + 2b_t^2 c_t^2(d_t - f_t) \\ & + 2a_t^3 c_t(e_t - f_t) + b_t^4(f_t - d_t) + a_t^4(f_t - e_t) \\ & + 2a_t c_t(c_t^2(d_t - e_t) + b_t^2(d_t - f_t) + 2b_t c_t(f_t - d_t))) \end{aligned} \tag{5}$$

$$\begin{aligned} T_t = {} & -3a_t b_t^2 d_t + 2b_t^3 d_t + 6a_t b_t c_t d_t - 3b_t^2 c_t d_t - 3a_t c_t^2 d_t \\ & + c_t^3 d_t + a_t^3 e_t - 3a_t^2 c_t e_t + 3a_t c_t^2 e_t - c_t^3 e_t \\ & - f_t(a_t - b_t)^2(a_t + 2b_t - 3c_t) \end{aligned} \tag{6}$$

$$Z_t = (a_t - b_t)^2(a_t - c_t)^3(b_t - c_t)^2. \tag{7}$$

In our implementation, for a given t, denoting by m_t the minimum saliency value over the current frame, by M_t the maximum and by A_t the average

$$\begin{aligned} & a_t = d_t = m_t,\ c_t = f_t = M_t, \\ & b_t = 0.5(A_t - m_t),\ e_t = 0.55(A_t - m_t). \end{aligned} \tag{8}$$

Encoding priority $p_t(X)$ of an image location X at frame t is finally computed as the averaged squashed saliency value over $k = 8$ successive frames

$$p_t(X) = \frac{1}{k} \sum_{u=t}^{t+k-1} r_u\left(s_u(X)\right). \tag{9}$$

Alternatively, a small number of discrete virtual foveas endowed with mass, spring (stiffness k), and fluid friction (coefficient μ) dynamics attempt to track a collection of most salient objects, with each fovea modeled as a unit mass at one end of a spring with zero rest length, and a salient location serving as an anchor point at the other end. Given the current location $X_t(j) = (x_t(j), y_t(j))$ of the j-th fovea and an anchor location $X_t^A(j) = (x_t^A(j), y_t^A(j))$, at every time step dt

$$\begin{aligned} X_{t+dt}(j) = {} & 2X_t(j) - X_{t-dt}(j) \\ & + dt^2 k(X_t^A(j) - X_t(j)) - \mu(X_t(j) - X_{t-dt}(j)) \end{aligned} \tag{10}$$

with, in our implementation, $dt = 0.1$ ms, $k = 10{,}000$ N/pixel, $\mu = 2{,}000$ N$\cdot$ s/pixel, and an implicit mass of 1 kg at $X_t(j)$ (hence the large value for k, so that significant motion may be achieved in a few milliseconds). At every frame, a new spring anchor point $X_t^A(j)$ is set for every fovea j at a salient image location that maximizes a correspondence score for that fovea. Correspondence between the n most salient locations on a given frame and the p foveas from the previous frame is established

Fig. 2. Examples of predicted priority maps for two settings of our model. (a) Maxnorm feature normalization, continuous priority map (0 virtual foveas), foveation pyramid depth 4. (b) Fancynorm, three foveas, depth 6. Top left: original frame. Top right: foveated frame (with fovea centers marked by yellow squares when using discrete foveas). Bottom left: saliency map (brighter for higher saliency). Bottom right: priority map (darker for higher priority).

through an exhaustive scoring of all $n \times p$ possible pairings between a new salient location $i \in \{1 \ldots n\}$ and an old foveation center $j \in \{1 \ldots p\}$ (typically, p is fixed and $n = p + 4$ to ensure robustness against varying saliency ordering from frame to frame). The correspondence score $c_t(i, j)$ to be maximized combines four criteria

$$c_t(i,j) = -\alpha||X_t(i) - X_t(j)|| - \beta||V_t(i) - V_t(j)|| - \gamma|i - j| + \delta_t s_t(X_t(i)) \quad (11)$$

where $||X_t(i) - X_t(j)||$ is the Euclidean spatial distance between locations $X_t(i)$ and $X_t(j)$ of i and j, $||V_t(i) - V_t(j)||$ is the Euclidean distance between 72-component feature vectors constructed by reading out (with bilinear interpolation) values from the 72 feature maps at the locations of i and j, $|i - j|$ is a penalty term that discourages permuting previous pairings by encouraging a fixed ordered pairing, and $s_t(X_t(i))$ is the saliency of i. In our implementation, $\alpha = 100/\sqrt{w^2 + h^2}$ where w and h are the image width and height, $\beta = 0.5$ (center-surround features vary between -255 and 255), $\gamma = 10$ and $\delta_t = 100/s_t(1)$. Combined, these criteria tend to assign foveas to salient locations in decreasing order of saliency, unless feature similarity is strong enough to warrant a different ordering (e.g., when a tracked object switches from being the second most salient in one frame to the fourth most salient in the following frame). Video compression priority $p_t(X)$ at every location X is then derived from the squashed distance to the closest fovea center [Fig. 2(b)], as computed using a 3/4 chamfer distance transform [63]

$$p_t(X) = r_t\left(d_{\max} - d_{3/4}(X)\right) \quad (12)$$

where $d_{\max} = 2\max(w, h)$ and $d_{3/4}$ is computed as in [63]. For further implementation details, we refer the reader to the source code of our algorithm, freely available online upon request [64].

The dynamics of the virtual foveas do not attempt to emulate human saccadic (rapid ballistic) eye movements [65], as those rapid and often jerky motions would create highly visible artifacts in the compressed streams. Rather, the virtual foveas attempt to track salient objects in a smooth and damped manner so that the foveation filter does not change too abruptly from frame to frame. Also, perfect tracking of a given set of objects was not desirable in our context where it is very important to rapidly focus onto new salient objects, even though that often means losing track of some current object. This is why we have not used stronger trackers like particle filters [66]. Thus, our correspondence and tracking algorithm compromises between reliably tracking the few most salient objects, and time-sharing remaining foveas among a larger set of less salient objects. A key aspect of our approach is that is makes no assumption on the type of video streams to be processed or their contents (e.g., presence of human faces). Instead, our low-level features maps are computed according to electrophysiological evidence for certain types of neural feature detectors and spatial interactions in the primate retina, lateral geniculate nucleus of the thalamus, primary visual cortex, and posterior parietal cortex [29]. Similarly, our foveation technique relies on spatial and feature-space distances rather than, for example, predefined object templates.

To evaluate the algorithm, we here simply use it as a front end, applied before standard video compression [both MPEG-1 and MPEG-4 (DivX) were tested]: a spatially-variable blur is applied to the input frames such that lower-priority regions are more strongly blurred. Although this is not optimal in terms of expected file size gains, it has the advantage of producing compressed streams that are compatible with existing decoders and to render the spatial prioritization computed by the algorithm obvious upon visual inspection. This method should be regarded as a worst-case scenario for two reasons. First, as virtual foveas move, the appearance of static objects far from the foveas changes, requiring continuous re-encoding of those changes. Second, even when the foveas remain static for some time, peripheral moving objects receive variable amounts of blur, depending on distance to the closest fovea. This defeats motion compensation in the encoder, yielding continuous re-encoding of these moving objects. Video codecs have been proposed to address these problems inherent to any foveated video compression technique (e.g., encode high-priority regions first, then lower-priority regions, in a continuously-variable-bitrate encoding scheme [4]). To simplify the visual evaluation of our algorithm and to evaluate whether our technique might prove useful even with standard codecs, however, here we use standard MPEG-1 and MPEG-4 encoding and simple spatially-variable blur of the video stream prior to compression. Any file size gain obtained despite these limitations would, hence, represent the promise that even better size gains should be obtained with a video codec that would truly use the model's priority maps to *prioritize* encoding.

On one example clip, we explore variations of the algorithm (Table I); namely, we vary the following.

1) **Number of foveas:** When using virtual foveas.
2) **Saliency interactions:** We explore three methods by which conspicuous regions spatially compete with each other for saliency, previously described in details [55], [57]: **Maxnorm** (normalization in each feature channel by the squared difference between global maximum and average of all other local maxima), **FancyOne** (one iteration of a detailed model of nonclassical surround interactions in primary visual cortex), and **Fancy** (ten iterations of the former). Maxnorm yields smoother, more continuous saliency maps [Fig. 2(a)], while the other two yield increasingly sparser saliency maps, with only a few sharp peaks of activity [Fig. 2(b)].
3) **Object segmentation:** When using virtual foveas, their centers are represented either by a disk of fixed radius (30 pixels) or are dynamically shaped to coarsely match the shape of the attended object, thus resulting in so-called "object-shaped foveas." The object segmentation technique used for this purpose relies on a simple region-growing algorithm that is applied in the feature map that is the most active at the attended location and, thus, is predominantly responsible for making that location salient. Additional details on this technique have been previously described [67].

TABLE I
COMPRESSED FILE SIZE OF FOVEATED RELATIVE TO ORIGINAL CLIP FOR BEVERLY03 (FIG. 2), VARYING FOVEATION TECHNIQUE (CONTINUOUS BLUR BASED ON ENTIRE SALIENCY MAP, OR 1–5 DISCRETE FOVEAS) AND SALIENCY COMPETITION TECHNIQUE (MAXNORM, FANCYONE OF FANCY)

	Blur Pyramid Depth 2				Blur Pyramid Depth 4				Blur Pyramid Depth 6			
	cont.	1 fov	3 fov	5 fov	cont.	1 fov	3 fov	5 fov	cont.	1 fov	3 fov	5 fov
	— MPEG-1 compression, circular foveas —											
Maxnorm	64.2%	53.2%	60.3%	64.0%	37.6%	31.5%	38.7%	42.8%	30.3%	25.7%	32.5%	36.4%
FancyOne	42.3%	53.2%	61.7%	65.8%	27.2%	31.3%	39.6%	44.1%	23.7%	25.5%	33.4%	37.6%
Fancy	35.9%	53.7%	62.9%	65.6%	20.7%	31.0%	40.3%	43.7%	17.0%	25.3%	33.7%	37.2%
	— MPEG-1 compression, object-shaped foveas —											
Maxnorm		71.7%	87.5%	94.0%		51.0%	74.7%	86.6%		44.5%	69.6%	83.1%
FancyOne		60.1%	73.4%	78.7%		38.6%	53.0%	60.5%		32.5%	46.3%	54.0%
Fancy		57.1%	68.7%	71.1%		34.6%	46.9%	49.7%		28.7%	40.1%	42.9%
	— Constant-quality MPEG-4 (DivX) compression, circular foveas —											
Maxnorm	55.6%	44.0%	52.7%	57.1%	30.7%	25.3%	33.5%	37.9%	25.0%	20.2%	27.9%	32.2%
FancyOne	32.0%	43.7%	54.0%	59.2%	22.8%	24.9%	34.1%	39.1%	19.7%	19.9%	28.7%	33.5%
Fancy	24.0%	44.4%	55.4%	58.7%	14.7%	24.6%	34.7%	38.6%	11.8%	19.6%	29.0%	32.9%
	— Constant-quality MPEG-4 (DivX) compression, object-shaped foveas —											
Maxnorm		65.4%	84.7%	92.7%		45.9%	71.8%	85.2%		40.4%	67.5%	82.4%
FancyOne		51.9%	67.9%	74.2%		32.9%	48.6%	56.8%		27.7%	43.0%	51.3%
Fancy		48.7%	62.4%	65.1%		28.8%	42.1%	45.1%		23.6%	36.3%	39.3%

4) **Blur pyramid depth:** Blurring is achieved through computation of a Gaussian pyramid from each input frame and subsequent readout of every pixel from a pyramid level that depends on the priority assigned to that pixel (with trilinear interpolation). The deeper the pyramid, the higher the amount of blur applied at locations of low priority.

The encoder settings are shown in Fig. 3. For MPEG-1, we simply used the default settings of the mpeg_encode program [68]. For MPEG-4, the default settings of the mencoder program [69] are for constant-bitrate, variable-quality encoding, which is not appropriate for size comparisons since the encoder would vary quality to always produce streams with approximately same bitrate. Thus, we used constant-quality, variable-bitrate and otherwise default settings for this encoder. We used a medium MPEG-4 quality setting, to approximately match the MPEG-1 filesizes on the unfoveated clips.

Overall, the results in Table I indicate a high degree of flexibility of the approach, with additional compression ratios for the foveated clips ranging from 1.1 to 8.5 compared to the unfoveated clips. With pyramid depths 2 and 4, the compressed clips are pleasant to watch, especially in the continuous mode (using the whole saliency map to derive blur rather than discrete foveas). With depth 2, it is often difficult to notice any obvious blurring, though additional compression ratios of up to 2.8 with MPEG-1 and 4.2 with DivX are achieved. With depth 6, size gains are maximized but blurring artifacts are obvious. This suggests an application to long-term archival, where apparent artifacts may be tolerated as long as the main contents of the clip can be understood. Object-shaped foveas resulted in better clarity, at the cost of lower size gains compared to the circular foveas (since object shapes often extended to much larger regions than the fixed-size circular apertures). The dependence of size gain on model parameter settings followed the same trends for both encoders. The resulting clips can be visually inspected online [70].

III. HUMAN EYE TRACKING

To validate our approach, we compared the computed locations of high priority from our algorithm to the gaze locations from eight human observers watching the *unfoveated* clips.

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 or corrected-to-normal vision). They were instructed to watch the video clips, and to attempt to follow the main actors and actions, as they would be later asked some general questions about what they had watched. It was emphasized that the questions would not pertain to small details (e.g., specific small objects, colors of clothing, identities of persons, or contents of text messages) but would instead help us evaluate their general understanding of the contents of the clips. The choice of task instructions given to our subjects was motivated by two factors: first, we avoided instructionless free viewing, since we believe that it often yields largely idiosyncratic patterns of eye movements, as subjects may attempt to

Video grabbing:
From interlaced NTSC video composite sources, 640×480, 29.97 fps, YV12 uncompressed. Resulting video streams converted to series of PPM frames.

MPEG-1 encoding (mpeg_encode):

```
PATTERN                  IBBPBBPBBPBBPBB
BASE_FILE_FORMAT         PPM
GOP_SIZE                 30
SLICES_PER_FRAME         1
PIXEL                    HALF
RANGE                    10
PSEARCH_ALG              LOGARITHMIC
BSEARCH_ALG              CROSS2
IQSCALE                  8
PQSCALE                  10
BQSCALE                  25
FORCE_ENCODE_LAST_FRAME  1
REFERENCE_FRAME          ORIGINAL
```

MPEG-4 (DivX) encoding (mencoder):

```
a) convert frames to lossless JPEG
b) mencoder  -noskip                              # do not drop frames
             -ovc lavc                            # use libavcodec codecs
             -lavcopts vcodec=mpeg4:vqscale=10    # DivX, constant quality
             -mf on:type=jpeg:fps=30              # series-of-frames input
             "frame*.jpg"                         # input frames
             -o movie.avi                         # output video
```

Fig. 3. Video grabbing and encoding.

find what the true undisclosed purpose of the experiment may be, or simply lose interest and disengage from the experiment over time. Second, our purpose was to evaluate the applicability of our model to predicting regions that should be given high priority during compression, that is, regions of high *cognitive importance* for scene understanding, such as those which subjects would consider were the main actors and actions in a scene. We believe that our instructions did not explicitly bias subjects toward low-level salient image locations. During the discussions that led subjects to sign an informed consent to participate to the experiments (as approved by USC's Internal Review Board), we were careful to avoid words including "saliency," "attention," "center-surround features," or "image processing," and to never mention our model. Nevertheless, one should be aware that our decision to give explicit task instructions to our subjects may limit the general applicability of our results, to the extent that biasing subjects toward the main actors and actions may have implicitly biased them toward low-level salient image locations. Given all the factors that contribute to guiding eye movements in humans (see Introduction), however, we believe that our results in the following section evaluate the extent to which a task-independent, pixel-based image processing algorithm, devoid of explicit notions of actors or actions, may yield a selection of scene locations that is comparable to the selection operated by human subjects trying to build a cognitive interpretation of the main contents of a video clip. 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" computer monitor (LaCie Corp; 640×480, 60.27 Hz double-scan, mean screen luminance 30 cd/m^2, room 4 cd/m). Subjects were seated on an adjustable chair at a viewing distance of 80 cm ($28^\circ \times 21^\circ$ usable field-of-view) and rested on a chin-rest. A nine-point eye-tracker calibration was performed every five clips [71]. Each calibration point consisted of fixating first a central cross, then a blinking dot at a random point on a 3×3 matrix. The experiment was self-paced and subjects could stretch before any nine-point calibration. Subjects fixated a central cross, pressed a key to start, at which point the eye-tracker was triggered, the cross blinked for 1 206 ms, and the clip started. Stimuli were presented on a Linux computer, under SCHED_FIFO scheduling (process would keep 100% of the CPU as long as needed [72]). Each unfoveated 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 [72] 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 $33\ 185 \pm 2\ \mu$s.

Eye position was tracked using a 240-Hz infrared-video-based eye tracker (ISCAN, Inc., model RK-464). This machine estimates point of regard (POR) in real time from comparative tracking of both the center 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). All analysis was performed off line. Linearity of the machine's POR-to-stimulus coordinate mapping was excellent, as previously tested using a 7×5 calibration matrix in our laboratory, justifying a 3×3

matrix here. The eye-tracker calibration traces were filtered for blinks and segmented into two fixation periods (the central cross, then the flashing point), or discarded if that segmentation failed a number of quality control criteria. 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 was then applied to account for any small residual nonlinearity [73]. 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 clips following a calibration were remapped to screen coordinates, or discarded if they failed some quality control criteria (excessive eye-blinks, motion, eye wetting, or squinting). Calibrated eye traces were visually inspected when superimposed with the clips, but none was discarded based upon that subjective inspection. Although we had no external reference to quantify the accuracy of the calibrated traces, overall the quality seemed remarkable with this eye-tracker (e.g., subjects tracking the ten-pixel-wide head of a person running at a distance, in clip beverly08).

Fifty video clips were selected from a database of 85, with as only selection criterion to maximize diversity. All clips had been digitized from their analog interlaced NTSC video source (Fig. 3) using a consumer-grade framegrabber (WinTV Go, Hauppage, Inc.) and no attempt was made at de-interlacing or color-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 open-air rooftop bar;
tv action	an action scene from a television movie;
tv ads	television advertisements;
tv announce	a television program's announce;
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.

All clips and the corresponding algorithmic multifoveation results and human eye movements may be examined online [70]. Clips had between 164 and 2 814 frames (5.5 s to 93.9 s). 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 (Table II). Note that the figures for beverly03 slightly differ between Tables I and II due to different playback frame rates (29.97 fps versus 30.13 fps).

To interpret our results, it is useful to note that (not unexpectedly) the average recommended blur (over the extent of each frame, then over all frames) closely matched a compound measure of local blur over a random scanpath and also closely matched a compound measure of blur over a human scanpath if the priority map was randomly scrambled (not shown). Thus, in the **%avg at eye** columns of Table II, a value of 100% or more would indicate that humans did not look at regions of high model priority (low suggested blur) more than a random scanpath (or, equivalently, that a randomly scrambled priority map would predict human fixations as well as the model's priority map). Conversely, a value of 0% would indicate that humans always gazed exactly at regions of maximum priority (no suggested blur). Remarkably, for all but two clips, we found a highly significant agreement between model priority and human fixations. Agreement was independently evaluated for each clip using a one-tailed t-test for the null hypothesis that the **%avg at eye** figures could have been drawn from a distribution with mean 100% (i.e., human scanpaths correlated with model priority no better than random scanpaths). When using *Maxnorm* feature normalization and continuously-variable blur, the hypothesis was rejected for every clip, with $p < 0.01$ or better. When using *Fancy* normalization and three circular foveas, the hypothesis was also rejected with $p < 0.01$ or better for 48 of the 50 clips. Figs. 4 and 5 show sample frames.

A breakdown by visual features in Table III indicated that flicker and motion energy were the most elevated at human eye fixations compared to on average over the entire frame, in agreement with the widely recognized importance of motion in capturing attention [29]. However, all features were significantly elevated at human eye locations and, thus, none was useless ($p < 0.001$ for each feature on a one-tailed t-test that the mean ratio of feature values at eye to the average was greater than unity).

It is interesting to compare our model to a much simpler, center-weighted model, since previous studies have indicated that the distributions of fixations typically observed as humans inspect visual scenes show a strong bias for the center of the display [14]. Fig. 6(a) indicates that indeed, overall, the distribution of fixations over all observers and clips was strongly biased toward the display center. On individual clips, however, this was not always true, as exemplified in Fig. 6(b)–(d). Thus, while a center-weighted model may perform well overall, it will fail on some clips, while our model performed well on all clips, except for two with one model variant. This, however, suggests an interesting extension of our model, by which the distributions of human eye fixations could be used as prior biasing maps, if the model was to be applied to specific video content of fixed layout (e.g., the highly structured tv news clips) [12].

IV. Discussion

Our main focus in this study is not a specific video compression technique, but the evaluation of our biological model of attention to compute priority maps. Simply blurring the frames before compression is a worst-case scenario in terms of expected size gains, but is useful for inspecting our results.

In Table I, we have shown a high degree of flexibility of our paradigm, yielding a wide range of additional compression

TABLE II

AGREEMENT BETWEEN HUMAN EYE MOVEMENTS AND MODEL PRIORITY MAPS, FOR TWO ALGORITHM VARIANTS (FIG. 2). **NSUBJ**: FOUR TO SIX HUMAN SUBJECTS VIEWED EACH CLIP. **%AVG AT EYE**: COMPOUNDED RATIOS BETWEEN MODEL-SUGGESTED BLUR AT HUMAN EYE FIXATIONS AND AVERAGE BLUR OVER ENTIRE FRAME. **MPG1**, **DIVX**: RATIO (%) OF COMPRESSED SIZES FOR FOVEATED VERSUS UNFOVEATED CLIPS, USING MPEG-1 AND DIVX ENCODING

			Maxnorm, 0 fov, depth 4			Fancynorm, 3 fov, depth 6		
Clip	frames	subj	%avg at eye	mpg1	divx	%avg at eye	mpg1	divx
beverly01	490	5	52.3±18.5 (p <0.005)	36.4	28.0	41.4±13.2 (p <0.0005)	38.1	30.9
beverly03	481	5	42.9± 9.5 (p <0.00025)	37.6	30.7	21.5± 4.1 (p <0.00005)	33.9	29.1
beverly05	546	4	40.3± 8.7 (p <0.0005)	51.3	42.1	43.5± 9.3 (p <0.001)	48.6	38.5
beverly06	521	4	39.0± 9.9 (p <0.001)	40.1	29.2	23.9±11.0 (p <0.0005)	44.3	33.4
beverly07	357	4	48.5±11.8 (p <0.005)	25.7	17.0	31.2± 3.9 (p <0.00005)	32.3	25.9
beverly08	237	5	67.7± 5.6 (p <0.00025)	25.5	15.6	36.6±15.5 (p <0.0005)	34.2	26.5
gamecube02	1819	6	55.1± 5.0 (p <0.00005)	73.0	66.3	42.5± 4.7 (p <0.00005)	65.4	65.0
gamecube04	2083	4	50.5± 1.4 (p <0.00005)	77.7	68.1	27.9± 9.6 (p <0.0005)	73.0	66.3
gamecube05	213	6	37.3± 7.3 (p <0.00005)	82.2	70.9	15.1± 5.8 (p <0.00005)	88.7	87.1
gamecube06	2440	6	35.7± 7.8 (p <0.00005)	58.9	54.2	36.4± 6.6 (p <0.00005)	52.8	43.5
gamecube13	898	5	72.9± 3.0 (p <0.00005)	54.6	45.5	52.3± 8.8 (p <0.00025)	50.8	42.5
gamecube16	2814	4	81.0± 7.5 (p <0.01)	57.2	47.1	38.8± 7.9 (p <0.0005)	53.6	43.2
gamecube17	2114	5	44.6± 8.3 (p <0.00025)	82.5	70.2	31.2±11.0 (p <0.00025)	83.2	70.5
gamecube18	1999	5	52.8± 4.0 (p <0.00005)	74.4	65.2	29.1± 8.5 (p <0.00005)	74.1	67.4
gamecube23	1429	4	28.6± 9.9 (p <0.0005)	59.6	54.0	30.1±14.5 (p <0.005)	58.4	52.8
monica03	1526	5	53.3±13.8 (p <0.001)	51.2	41.2	40.3± 3.3 (p <0.00005)	46.1	37.4
monica04	640	5	47.0± 6.8 (p <0.00005)	47.6	38.4	23.2± 6.3 (p <0.00005)	43.8	36.7
monica05	611	4	60.8± 6.8 (p <0.001)	43.4	33.9	58.5± 6.3 (p <0.0005)	40.7	32.1
monica06	164	4	23.8± 6.5 (p <0.00025)	47.6	39.6	41.7±12.3 (p <0.005)	43.0	34.6
saccadetest	516	5	29.2±10.3 (p <0.00025)	24.8	64.0	14.4±10.8 (p <0.00005)	34.4	95.5
standard01	254	4	58.0± 6.5 (p <0.0005)	49.9	45.4	70.1±19.3 (p <0.05)	38.5	34.7
standard02	515	5	51.0± 4.4 (p <0.00005)	49.5	40.3	52.4±11.0 (p <0.0005)	42.0	34.2
standard03	309	4	72.8±10.5 (p <0.01)	52.3	41.6	78.8± 9.5 (p <0.025)	46.5	38.6
standard04	612	5	81.3± 6.1 (p <0.005)	47.7	37.0	77.0± 3.4 (p <0.00025)	40.6	30.3
standard05	483	5	52.2± 1.4 (p <0.00005)	52.7	44.5	53.6± 8.1 (p <0.00025)	43.3	38.0
standard06	434	5	63.3± 2.9 (p <0.00005)	52.0	44.0	67.1±14.3 (p <0.005)	40.9	34.4
standard07	177	4	43.2± 4.9 (p <0.00025)	43.0	32.9	42.8± 9.8 (p <0.001)	39.3	31.5
tv-action01	567	4	39.0± 1.5 (p <0.00005)	47.0	32.1	20.0± 4.3 (p <0.00005)	42.4	27.9
tv-ads01	1077	4	79.4± 5.4 (p <0.005)	59.2	52.8	52.4±10.5 (p <0.005)	65.2	62.8
tv-ads02	387	4	60.3± 9.3 (p <0.005)	52.4	41.7	48.4± 5.4 (p <0.00025)	57.0	47.6
tv-ads03	841	5	64.9±14.3 (p <0.005)	48.8	39.7	44.2±15.6 (p <0.001)	46.3	37.1
tv-ads04	313	5	43.9± 3.3 (p <0.00005)	56.1	50.8	44.0± 9.6 (p <0.00025)	53.8	46.0
tv-announce01	434	4	78.0± 2.2 (p <0.00025)	60.6	52.4	51.2±10.6 (p <0.005)	59.0	51.8
tv-music01	1022	5	59.9± 3.9 (p <0.00005)	51.6	42.2	51.0± 5.6 (p <0.00005)	45.8	36.4
tv-news01	918	5	59.5± 3.9 (p <0.00005)	46.9	32.7	70.3± 5.2 (p <0.00025)	51.3	39.0
tv-news02	1058	6	60.0± 2.5 (p <0.00005)	57.5	58.6	67.5±14.2 (p <0.005)	62.2	69.8
tv-news03	1444	5	71.8± 1.9 (p <0.00005)	53.7	50.7	67.0± 9.1 (p <0.001)	62.8	74.8
tv-news04	491	5	33.5± 9.5 (p <0.00005)	52.1	53.0	31.4± 8.0 (p <0.00005)	55.7	56.0
tv-news05	1341	5	60.7± 4.3 (p <0.00005)	65.3	64.0	46.1± 5.6 (p <0.00005)	70.1	73.1
tv-news06	1643	5	72.9± 7.0 (p <0.0005)	56.2	51.2	78.9±11.1 (p <0.01)	62.4	68.2
tv-news09	1176	4	69.3± 6.4 (p <0.005)	53.4	48.7	87.8±16.8 (p <0.15)	59.2	60.1
tv-sports01	579	5	67.1± 6.1 (p <0.00025)	46.0	38.3	51.4±10.8 (p <0.0005)	43.8	38.9
tv-sports02	444	4	72.0± 9.7 (p <0.01)	53.4	49.7	56.5±11.5 (p <0.005)	54.0	51.9
tv-sports03	1460	5	52.7± 5.6 (p <0.00005)	46.7	36.4	50.3± 7.7 (p <0.00025)	43.5	34.2
tv-sports04	982	4	79.4± 5.2 (p <0.005)	43.3	34.2	56.3± 4.8 (p <0.00025)	43.9	36.2
tv-sports05	1386	6	55.0± 6.1 (p <0.00005)	50.5	38.4	41.4± 3.0 (p <0.00005)	45.6	34.3
tv-talk01	1651	4	37.7± 3.5 (p <0.00005)	65.7	62.1	42.8±13.2 (p <0.005)	61.3	64.0
tv-talk03	783	5	56.4± 4.2 (p <0.00005)	44.8	36.1	43.3± 5.7 (p <0.00005)	43.7	36.6
tv-talk04	1258	5	51.8± 4.2 (p <0.00005)	36.8	25.6	55.5±10.9 (p <0.0005)	42.7	34.8
tv-talk05	552	4	64.9± 1.3 (p <0.00005)	38.1	28.0	70.2± 9.7 (p <0.005)	41.9	34.4
SUMMARY	46489	4.7	55.1± 7.3	51.7	44.5	46.5± 9.7	50.9	46.3

ratios (from 1.1 to 8.5) and associated visual degradation, depending on settings. In Table II, foveated clip size was approximately half of the unfoveated size on average. Smallest sizes were obtained for the simplest clips, where only one small object moved on a static background (e.g., saccadetest, gamecube05). There was no evidence for any systematic covariation of size gains between the two variants of the algorithm. Typically, using a continuous priority map yielded higher compression when a small number of objects moved on a static background (e.g., saccadetest, beverly07). Indeed, the amount of blur applied to the background would remain fairly unchanged in this case, but would greatly vary when discrete foveas focused on the moving

Fig. 4. Additional examples of model-predicted priority maps for the first variant of the model [continuously variable blur, Maxnorm, blur pyramid depth 4, as in Fig. 2(a)]. Current eye position of one human observer watching the original unfoveated clip is shown (arrow) for comparison.

objects (but see below, for the more complex scenario where additional objects enter and leave the field of view). When significant full-field motion was present, however, three discrete foveas often performed better (e.g., panning camera motion in some "standard" clips), because they would select at most three regions while the continuous map would select many more. Note that our low-level motion detectors, however, are naturally insensitive to full-field translational motion, due to the global spatial competition for saliency operated in each feature map [57], [61]. A possible future extension of our algorithm consists of allowing the number of foveas to vary from frame to frame, although one difficulty will be to ensure that the appearance and disappearance of additional foveas does not create highly visible artifacts in the compressed stream.

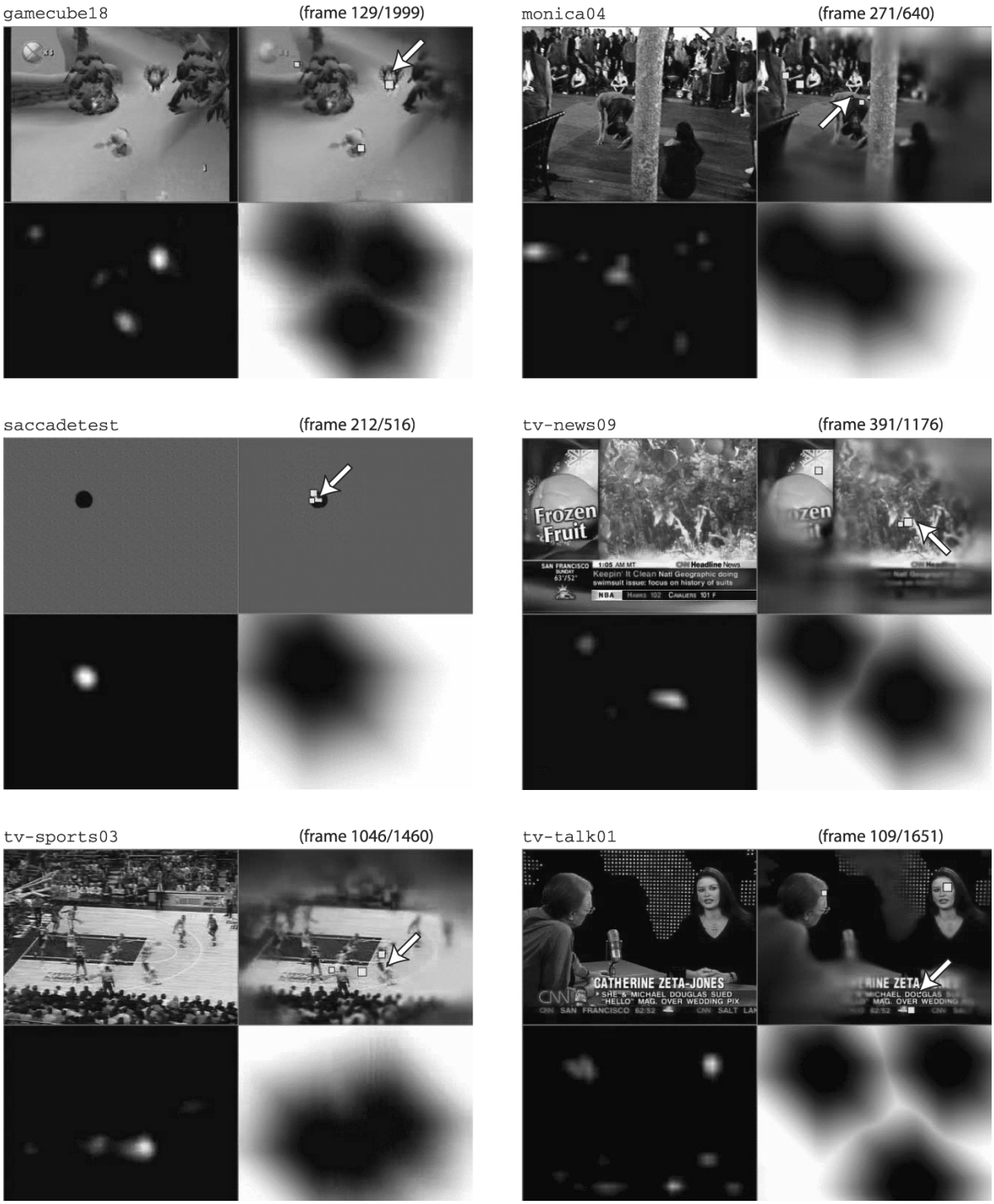

Fig. 5. Additional examples of model-predicted priority maps for the second variant of the model [three discrete foveas, Fancynorm, blur pyramid depth 6, as in Fig. 2(b)]. Current eye position of one human observer watching the original unfoveated clip is shown (arrow) for comparison.

An interesting feature of the results in Table II was that both variants of the algorithm performed equally well in terms of average additional compression ratios, yet one used a blur pyramid of depth 4 (and continuous blur) while the other used stronger depth 6 (and three foveas). For general application, when it is unknown whether only small objects or the full field of view will substantially change over time, the variant with depth 4 yields lower maximum possible blur and overall more visually pleasant results. Hence, our results argue somewhat against the traditional foveation paradigm and the binary selection of a small number of regions of interest [4], [12]. Instead, we suggest that a continuous measure of interest for every pixel in the frame is a more efficient strategy, since depth 4 in that case yielded on average same compression gains but substantially less-visible artifacts than depth 6 with the traditional approach.

TABLE III

BREAKDOWN BY FEATURE CATEGORIES. RATIOS (MEAN ± STANDARD ERROR OF THE MEAN) OF THE FEATURE VALUES AT HUMAN EYE POSITION TO THE AVERAGE VALUE OVER THE ENTIRE FRAME, COMPOUNDED OVER ALL 1 658 161 VALID EYE MOVEMENT SAMPLES, FOR EACH FEATURE CATEGORY (COLOR, RED/GREEN AND BLUE/YELLOW MAPS COMBINED; FLICKER; INTENSITY, ORIENTATION, 0°, 45°, 90°, 135° COMBINED; AND MOTION ENERGIES, UP, DOWN, LEFT AND RIGHT COMBINED). ALL FEATURES WERE SIGNIFICANTLY HIGHER AT HUMAN EYE POSITIONS THAN ON AVERAGE OVER THE DISPLAY (RATIOS LARGER THAN UNITY, $p < 0.001$ ON A ONE-TAILED T-TEST FOR EVERY FEATURE), WITH FLICKER AND MOTION ENERGY BEING THE MOST ELEVATED. RATIOS ARE OVERALL HIGHER AND MORE VARIABLE FOR THE FANCY MODEL VARIANT, WHICH HAS SPARSER FEATURE MAPS THAN THE MAXNORM VARIANT (FIG. 2)

Model Variant	Color	Flicker	Intensity	Orientation	Motion
Maxnorm	1.775 ± 0.002	2.023 ± 0.002	1.671 ± 0.003	1.760 ± 0.001	2.445 ± 0.003
Fancy	4.555 ± 0.012	6.141 ± 0.015	3.444 ± 0.011	3.801 ± 0.010	6.196 ± 0.016

When using three discrete foveas, sometimes the motion of the foveas induced so much change in textured backgrounds that any size gain due to foveation was offset by the need to continuously re-encode the background (e.g., saccadetest and some gamecube clips). This is a clear limitation of the fairly simplistic blurring technique used here to visualize model predictions. When using continuous blur, sometimes a similar problem existed, as, inherently, the saliency of an object depends on every other object in the scene (e.g., a bright red object will not be salient if many other bright red objects are present in the image). Consequently, the saliency of a static object typically varied as other salient objects entered or left the field of view, requiring re-encoding of the static object to reflect changes in its appearance with varying saliency and blur. A simple solution to this problem could be to take the maximum, at each frame, between previous and current priority maps, in regions where no significant change is detected. Thus, once a static object has been encoded crisply, it would remain crisp as long as it remains static (at very little cost in the compressed stream, since current appearance would be accurately predicted by previous frames). However, it may be more useful to instead replace our blurring scheme used for testing by a more sophisticated prioritization scheme and dedicated video codec [4]. These limitations in the applicability of blur-based foveation to video compression may be somewhat reduced by the use of a biresolutional foveation blur instead of the continuously-variable blur used here [11].

Overall, we found surprisingly strong agreement between our model's predictions and the scanpaths recorded from human observers, with motion and flicker cues providing the best agreement, but other cues in the model showing strong responses at human eye fixations as well. It is important to note that our measure of agreement required that humans fixated a given location *at the same time* as the model did, in order for this fixation to increase the measure of agreement between humans and model. Our use of multiple foveas certainly helped reduce the number of cases where humans and model were essentially interested in the same objects but in a different temporal order, but did not entirely eliminate them.

Another difference which lowered the agreement between humans and model was that humans often fixated small details (e.g., fineprint text tickers in the tv-news clips), while other actors or actions were more salient (e.g., the anchor). This is perfectly reasonable and indicates that sometimes there was a mismatch between regions that were low-level salient as computed by our model, but largely irrelevant to cognitive scene understanding as attempted by our observers (e.g., subjects guessed that the salient anchor would just speak, with a low probability of making interesting faces, and decided that the text might be more informative although far less salient, in particular since our clips had no soundtrack). Given this, it is remarkable that such good agreement between humans and model was obtained on our wide variety of clips. Our results indicate that throughout each clip, bottom-up saliency based on low-level analysis of pixels strongly correlated with the cognitive decision of what constituted the main actors and actions. This contrasts with a more commonly agreed view in which bottom-up attentional guidance is believed to be active only for a short period after the presentation of a new (static) scene, with more cognitive top-down guidance taking over afterwards [25]. With rapidly changing video stimuli, our results reinforce the idea of a continual contribution of bottom-up cues to the guidance of attention. Thus, although our model is bottom-up and has no explicit notion of actor or action, its applicability to the selection of regions that are important for cognitive scene understanding is well supported by our data.

One last difference between humans and model was the often predictive nature of human gaze shifts, while our model only follows salient objects. In several gamecube clips, indeed, humans focused on the empty path before a fast-running hero, probably to ensure that it remained clear of obstacles. This is another limitation of our model, which may be solved using more sophisticated trackers for our virtual foveas, with predictive capabilities. Note, however, that since usually the path was empty and often featureless, its representation in the foveated clips was not much degraded by blurring.

Our model performed generally well on every clip, with **%avg at eye** figures in Table II significantly below 100% on every clip for the first model variant and on 48/50 clips for the second. In contrast, we saw in Fig. 6 that a simpler center-weighted model would be expected to perform well overall, but not necessarily on every clip. Indeed, our model does not make any *a priori* assumption as to the spatial distribution of priority regions. As mentioned in introduction, however, it is known that the rapid recognition of the gist of a scene may impose such prior distribution as to the likely locations of objects of interest [36]. Such distribution may also be learned over time in the form of a "task-relevance map" [74] that may filter out salient locations known to be of little task relevance (e.g., in the gamecube clips, salient health indicators strongly attracted the model but were largely ignored by subjects). A promising direction for future research is to attempt to learn the association between scene gists and eye movement density maps like those in Fig. 6, and to use those as modulators to the model's saliency map.

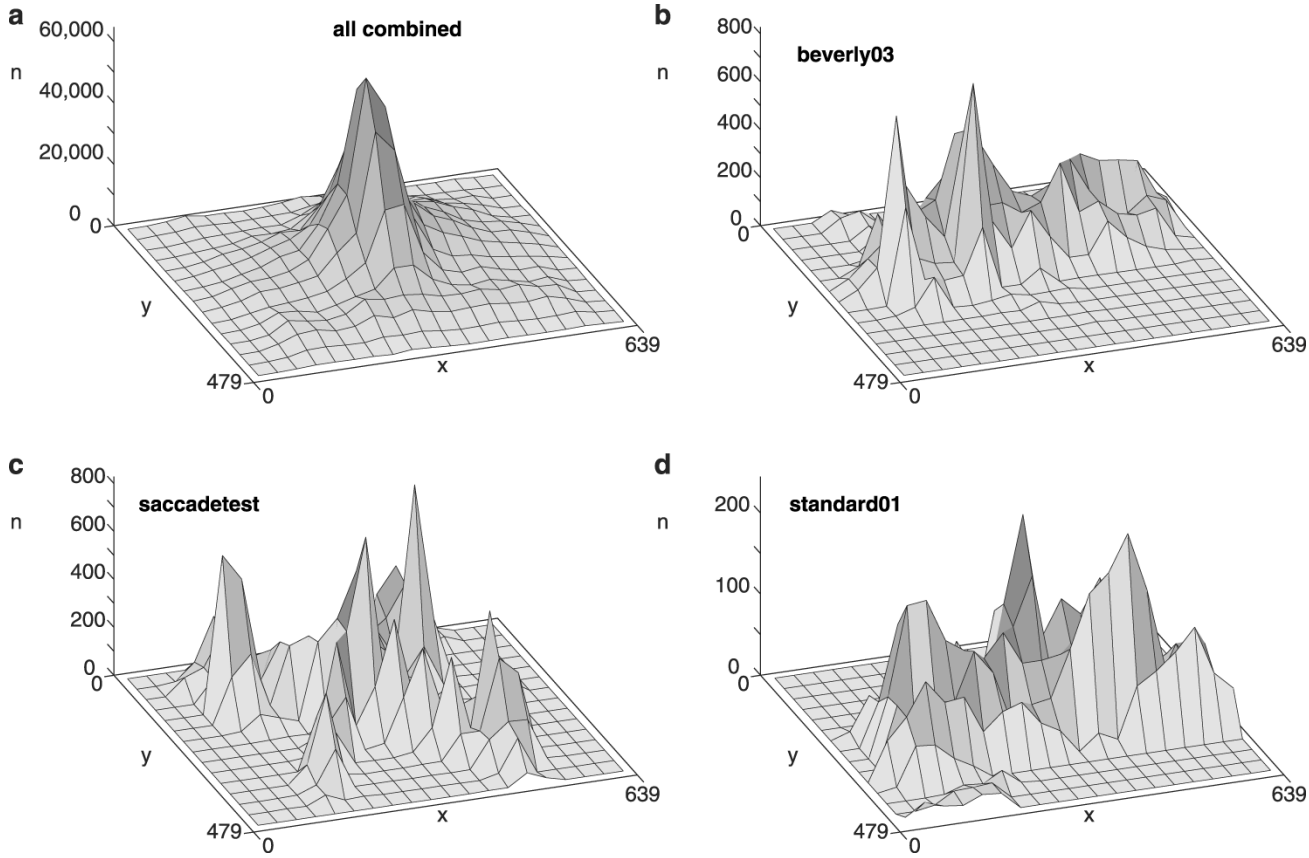

Fig. 6. Distribution of the number n of our 1 658 161 valid human eye movement samples (recorded at 240 Hz) over the (x, y) extent of the video frames, histogrammed over 32×32 image tiles. (a) The overall distribution for all subjects and clips shows a strong bias toward the center of the display. However, individual clips may exhibit very different distributions, in a manner not predictive of how well our algorithm will perform. (b) Distribution for clip beverly03 (see sample frames in Fig. 2) where very good agreement was found between the three-fovea model and humans. (c) Distribution for clip saccadetest (see sample frames in Fig. 5), also in excellent agreement. (d) Distribution for clip standard01 (see sample frames in Fig. 4), where worse agreement was found between model and humans.

Overall, our study demonstrates that our biological model of bottom-up attention highly significantly correlates with human eye movements on unconstrained video inputs. Importantly, this was observed although the model contains nothing in its design or implementation that specializes it to any specific video content or pre-defined objects of interest. Both with MPEG-1 and MPEG-4 encoding, substantial compressed file size reductions were achieved by foveation. Our study thus argues that our model is applicable to automatic spatiotemporal prioritization of arbitrary video contents for compression.

REFERENCES

[1] B. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer, 1995.

[2] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[3] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 3, pp. 572–588, Sept. 1994.

[4] Z. Wang, L. G. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Processing*, vol. 12, pp. 243–254, Feb. 2003.

[5] P. M. J. van Diepen, P. De Graef, and J. van Rensbergen, "On-line control of moving masks and windows on a complex background using the atvista videographics adapter," *Behav. Res. Meth.*, vol. 26, pp. 454–460, 1994.

[6] P. T. Kortum and W. S. Geisler, "Implementation of a foveated image-coding system for bandwidth reduction of video images," *Proc. SPIE*, vol. 2657, pp. 350–360, 1996.

[7] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Improving the performance of MPEG coders using adaptive regions of interest," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 928–934, Dec. 1998.

[8] S. Lee, A. C. Bovik, and Y. Y. Kim, "Low delay foveated visual communications over wireless channels," in *Proc. IEEE Int. Conf. Image Processing*, 1999, pp. 90–94.

[9] U. Rauschenbach and H. Schumann, "Demand-driven image transmission with levels of detail and regions of interest," *Comput. Grpah.*, vol. 23, no. 6, pp. 857–866, 1999.

[10] E. M. Reingold and L. C. Loschky, "Saliency of peripheral targets in gaze-contingent multiresolutional displays," *Behavior Res. Meth., Instrum. Comput.*, vol. 34, no. 4, pp. 491–499, 2002.

[11] D. J. Parkhurst and E. Niebur, "Variable-resolution displays: A theoretical, practical, and behavioral evaluation," *Human Factors*, vol. 44, no. 4, pp. 611–629, 2002.

[12] W. S. Geisler and J. S. Perry, "A real-time foveated multi-resolution system for low-bandwidth video communication," *Proc. SPIE*, vol. 3299, pp. 294–305, 1998.

[13] J. H. Goldberg and J. C. Schryver, "Eye-gaze-contingent control of the computer interface: Methodology and example for zoom detection," *Behav. Res. Meth., Instrum. Comput.*, vol. 27, no. 3, pp. 338–350, Aug. 1995.

[14] L. B. Stelmach, W. J. Tam, and P. J. Hearty, "Static and dynamic spatial resolution in image coding: An investigation of eye movements," *Proc. SPIE*, vol. 1453, pp. 147–152, 1991.

[15] A. T. Duchowski and B. H. McCormick, "Pre-attentive considerations for gaze-contingent image processing," *Proc. SPIE*, vol. 2411, pp. 128–139, 1995.

[16] ——, "Modeling visual attention for gaze-contingent video processing," in *Proc. 9th Image and Multidim.l Signal Proc. Workshop*, 1996, pp. 130–131.

[17] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image using a model of the human visual system," in *Proc. Int. Conf. Pattern Recognition*, Aug. 1998, pp. 701–704.

[18] X. Yang and K. Ramchandran, "A low-complexity region-based video coder using backward morphological motion field segmentation," *IEEE Trans. Image Processing*, vol. 8, pp. 332–345, Mar. 1999.
[19] F. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," in *Proc. Picture Coding Symp.*, Apr. 2001, pp. 101–104.
[20] W. Osberger, A. J. Maeder, and N. Bergmann, "Perceptually based quantization technique for MPEG encoding," *Proc. SPIE*, vol. 3299, pp. 148–159, 1998.
[21] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of JPEG-2000," in *Proc. Data Compression Conf.*, 2000, pp. 523–544.
[22] E. Mandel and P. Penev, "Facial feature tracking and pose estimation in video sequences by factorial coding of the low-dimensional entropy manifolds due to the partial symmetries of faces," in *Proc. 25th IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. IV, June 2000, pp. 2345–2348.
[23] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 970–982, Sept. 2000.
[24] W. M. Osberger and A. M. Rohaly, "Automatic identification of regions of interest in complex video sequences," *Proc. SPIE*, vol. 4299, pp. 361–372, 2001.
[25] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vis. Res.*, vol. 42, no. 1, pp. 107–123, 2002.
[26] A. Yarbus, *Eye Movements and Vision*. New York: Plenum, 1967.
[27] J. M. Henderson and A. Hollingworth, "High-level scene perception," *Annu. Rev. Psychol.*, vol. 50, pp. 243–271, 1999.
[28] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cognit.*, vol. 7, pp. 17–42, 2000.
[29] L. Itti and C. Koch, "Computational modeling of visual attention," *Nat. Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.
[30] L. Itti, "Modeling primate visual attention," in *Computational Neuroscience: A Comprehensive Approach*, J. Feng, Ed. Boca Raton, FL: CRC, 2003, pp. 607–626.
[31] M. C. Potter and E. I. Levy, "Recognition memory for a rapid sequence of pictures," *J. Exp. Psychol.*, vol. 81, no. 1, pp. 10–15, July 1969.
[32] A. Friedman, "Framing pictures: The role of knowledge in automatized encoding and memory for gist," *J. Exp. Psychol. Gen.*, vol. 108, pp. 316–355, Sept. 1979.
[33] I. Biederman, R. C. Teitelbaum, and R. J. Mezzanotte, "Scene perception: A failure to find a benefit from prior expectancy or familiarity," *J. Exp. Psychol. Learning Memory Cognit.*, vol. 9, no. 3, pp. 411–429, July 1983.
[34] A. Oliva and P. G. Schyns, "Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli," *Cognit. Psychol.*, vol. 34, no. 1, pp. 72–107, Oct. 1997.
[35] A. Hollingworth and J. M. Henderson, "Does consistent scene context facilitate object perception?," *J. Exp. Psychol. Gen.*, vol. 127, no. 4, pp. 398–415, Dec. 1998.
[36] A. Torralba, "Modeling global scene factors in attention," *J. Opt. Soc. Amer. A*, vol. 20, no. 7, pp. 1407–1418, July 2003.
[37] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science*, vol. 229, no. 4715, pp. 782–4, 1985.
[38] B. Motter, "Neural correlates of attentive selection for color or luminance in extrastriate area V4," *J. Neurosci.*, vol. 14, no. 4, pp. 2178–89, 1994.
[39] Y. Yeshurun and M. Carrasco, "Attention improves or impairs visual performance by enhancing spatial resolution," *Nature*, vol. 396, no. 6706, pp. 72–75, Nov. 1998.
[40] S. Treue and J. C. M. Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature*, vol. 399, no. 6736, pp. 575–579, Jun 1999.
[41] M. Ito and C. D. Gilbert, "Attention modulates contextual influences in the primary visual cortex of alert monkeys," *Neuron*, vol. 22, no. 3, pp. 593–604, Mar. 1999.
[42] J. H. Reynolds, T. Pasternak, and R. Desimone, "Attention increases sensitivity of V4 neurons," *Neuron*, vol. 26, no. 3, pp. 703–714, June 2000.
[43] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 15, no. 3, pp. 419–33, 1989.
[44] J. M. Wolfe, "Visual memory: What do you know about what you saw?," *Curr. Biol.*, vol. 8, no. 9, pp. R303–4, 1998.
[45] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy, "Integration of visual and linguistic information in spoken language comprehension," *Science*, vol. 268, no. 5217, pp. 1632–1634, June 1995.
[46] C. F. Nodine and E. A. Krupinski, "Perceptual skill, radiology expertise, and visual test performance with nina and waldo," *Acad. Radiol.*, vol. 5, no. 9, pp. 603–612, Sept. 1998.
[47] F. J. Moreno, R. Reina, V. Luis, and R. Sabido, "Visual search strategies in experienced and inexperienced gymnastic coaches," *Percept. Mot. Skills*, pt. 1, vol. 95, no. 3, pp. 901–902, Dec. 2002.
[48] G. J. Savelsbergh, A. M. Williams, J. V. der Kamp, and P. Ward, "Visual search, anticipation and expertise in soccer goalkeepers," *J. Sports Sci.*, vol. 20, no. 3, pp. 279–287, Mar. 2002.
[49] G. E. Legge, T. A. Hooven, T. S. Klitz, J. S. S. Mansfield, and B. S. Tjan, "Mr. Chips 2002: New insights from an ideal-observer model of reading," *Vis. Res.*, vol. 42, no. 18, pp. 2219–2234, Aug. 2002.
[50] D. Peebles and P. C. Cheng, "Modeling the effect of task and graphical representation on response latency in a graph reading task," *Human Factors*, vol. 45, no. 1, pp. 28–46, 2003.
[51] A. Hollingworth, C. C. Williams, and J. M. Henderson, "To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes," *Psychon. Bull. Rev.*, vol. 8, no. 4, pp. 761–768, Dec. 2001.
[52] J. M. Henderson and A. Hollingworth, "Global transsaccadic change blindness during scene perception," *Psychol. Sci.*, vol. 14, no. 5, pp. 493–497, Sept. 2003.
[53] D. Noton and L. Stark, "Scanpaths in eye movements during pattern perception," *Science*, vol. 171, no. 968, pp. 308–11, 1971.
[54] M. Carreiras and C. Clifton, "Another word on parsing relative clauses: Eyetracking evidence from spanish and english," *Mem. Cognit.*, vol. 27, no. 5, pp. 826–833, Sept. 1999.
[55] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1254–1259, Nov. 1998.
[56] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Res.*, vol. 40, no. 10–12, pp. 1489–1506, May 2000.
[57] ——, "Feature combination strategies for saliency-based visual attention systems," *J. Electron. Imag.*, vol. 10, no. 1, pp. 161–169, Jan. 2001.
[58] C. Koch and S. Ullman, "Shifts in selective visual attention: Toward the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–27, 1985.
[59] S. W. Kuffler, "Discharge patterns and functional organization of mammalian retina," *J. Neurophysiol.*, vol. 16, pp. 37–68, 1953.
[60] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol. (London)*, vol. 160, pp. 106–54, 1962.
[61] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Proc. SPIE*, vol. 5200, pp. 64–78, Aug. 2003.
[62] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.
[63] G. Borgefors, "Distance transformations in digital images," *CVGIP: Image Understanding*, vol. 54, no. 2, p. 301, 1991.
[64] [Online]. Available: http://iLab.usc.edu/toolkit/
[65] D. L. Sparks, "The brainstem control of saccadic eye movements," *Nat. Rev. Neurosci.*, vol. 3, no. 12, pp. 952–964, 2002.
[66] S. Maskell, M. Rollason, D. Salmond, and N. Gordon, "Efficient particle filtering for multiple target tracking with application to tracking in structured images," *Proc. SPIE*, vol. 4728, 2002.
[67] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition – A gentle way," *Lecture Notes Comput. Sci.*, vol. 2525, pp. 472–479, Nov. 2002.
[68] [Online]. Available: http://bmrc.berkeley.edu/research/mpeg/mpeg_encode.html
[69] [Online]. Available: http://www.mplayerhq.hu
[70] [Online]. Available: http://iLab.usc.edu/bu/co/
[71] D. M. Stampe, "Heuristic filtering and reliable calibration methods for video based pupil tracking systems," *Behavior Res. Meth., Instrum. Comput.*, vol. 25, no. 2, pp. 137–142, 1993.
[72] S. A. Finney, "Real-time data collection in linux: A case study," *Behavior Res. Meth., Instrum. Comput.*, vol. 33, pp. 167–173, 2001.
[73] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 567–585, June 1989.
[74] V. Navalpakkam and L. Itti, "A goal oriented attention guidance model," *Lecture Notes Comput. Sci.*, vol. 2525, pp. 453–461, Nov. 2002.

Laurent Itti received the M.S. degree in image processing from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1994 and the Ph.D. degree in computation and neural systems from Caltech, Pasadena, CA, in 2000.

In September 2000, he became Assistant Professor of computer science, psychology, and neuroscience at the University of Southern California, Los Angeles. His primary research interest is in biologically-plausible computational brain modeling and the comparison of model simulations to empirical measurements from living systems. Of particular interest in his laboratory is the development of computational models of biological vision, with applications to machine vision.