Automatic Attention-Based Prioritization of Unconstrained Video for Compression

Laurent Itti

University of Southern California, Hedco Neuroscience Building
3641 Watt Way, Room 30A - Los Angeles, CA 90089-2520 - USA
itti@pollux.usc.edu

ABSTRACT

We apply a biologically-motivated algorithm that selects visually-salient regions of interest in video streams to multiply-foveated video compression. Regions of high encoding priority are selected based on nonlinear integration of low-level visual cues, mimicking processing in primate occipital and posterior parietal cortex. A dynamic foveation filter then blurs (foveates) every frame, increasingly with distance from high-priority regions. Two variants of the model (one with continuously-variable blur proportional to saliency at every pixel, and the other with blur proportional to distance from three independent foveation centers) are validated against eye fixations from 4-6 human observers on 50 video clips (synthetic stimuli, video games, outdoors day and night home video, television newscast, sports, talk-shows, etc). Significant overlap is found between human and algorithmic foveations on every clip with one variant, and on 48 out of 50 clips with the other. Substantial compressed file size reductions by a factor 0.5 on average are obtained for foveated compared to unfoveated clips. These results suggest a general-purpose usefulness of the algorithm in improving compression ratios of unconstrained video.

Keywords: Attention, saliency, priority encoding, region of interest, video compression

1. INTRODUCTION

An increasingly popular approach to reduce the size of compressed video streams consists of selecting a small number of interesting regions in each frame, and to encode these regions in priority. This spatial prioritization scheme relies on the highly non-uniform 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 fall-off with eccentricity.[1] Thus, the rationale is that it may not be necessary nor useful to encode each video frame with uniform quality, since human observers watching the compressed clips 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 be used to temporally sequence the delivery of contents (deliver priority regions first), or to continuously scale video quality depending on available transmission bandwidth (so that priority regions occupy the core of a compressed stream, while additional details outside the priority regions are 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 non-interactive video compression. Gaze-contingent video transmission typically uses an eye-tracking device to record eye fixations 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 spatial region around the viewer's current point of eye fixation, 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 that degradation is well matched to their visual system and viewing conditions. Furthermore, even in the absence of an eye-tracking device, this interactive approach has demonstrated usefulness, for example when there exists a set of fixed priority regions, or when the observer explicitly selects priority regions using a pointing device.[12] Further,

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 non-interactive 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. In this context, very high inter-observer variability precludes recording a single eye movement scanpath from a given observer, and using it to determine priority regions in the video clip of interest. Recording from several observers and using the union of their eye fixations partially overcomes this limitation,[14] but at a prohibitive cost: An eye-tracking setup, population of human subjects, and time-consuming recording 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 of approach has been particularly successful, and those properties which are well-defined (e.g., contrast sensitivity function, importance of motion, temporal masking effects) have already been widely implemented by 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 implement in a computational model (e.g., evaluating object size and shape requires that first object segmentation be solved in a general manner). An important contribution of the present study is to propose 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[22,4]), 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 that have been documented in early visual processing areas of monkeys and humans. Finally, computational algorithms have thus far typically been demonstrated on a small set of video clips, and often lack ground-truth validation. Another important aspect of our study is hence to validate our model against eye movements of human observers.

In the following sections, we start by describing our neurobiological model of human visual attention, which automatically selects regions of high saliency (conspicuity) in an unconstrained variety of video inputs, without requiring any per-clip tuning. We then validate the algorithm, for two settings of its parameters, 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 (each clip viewed by at least four subjects), 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.

2. ATTENTION MODEL AND FOVEATION

The model computes a topographic saliency map **(Fig. 1)**, which indicates how conspicuous every location in the input image is. Retinal input is processed in parallel by a number of multiscale low-level feature maps, which detect local spatial discontinuities using simulated center-surround neurons.[23,24] Twelve neuronal features are implemented, sensitive to color contrast (red/green and blue/yellow, separately), temporal flicker (onset and offset of light intensity, combined), intensity contrast (light-on-dark and dark-on-light, combined), four orientations (0°, 45°, 90°, 135°), and four oriented motion energies (up, down, left, right).[25,26] Center and surround scales are obtained using dyadic pyramids with 9 levels (from level 0, the original image, to level 8, reduced by a factor 256 horizontally and vertically). Center-surround differences are then computed as pointwise differences across pyramid levels, 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

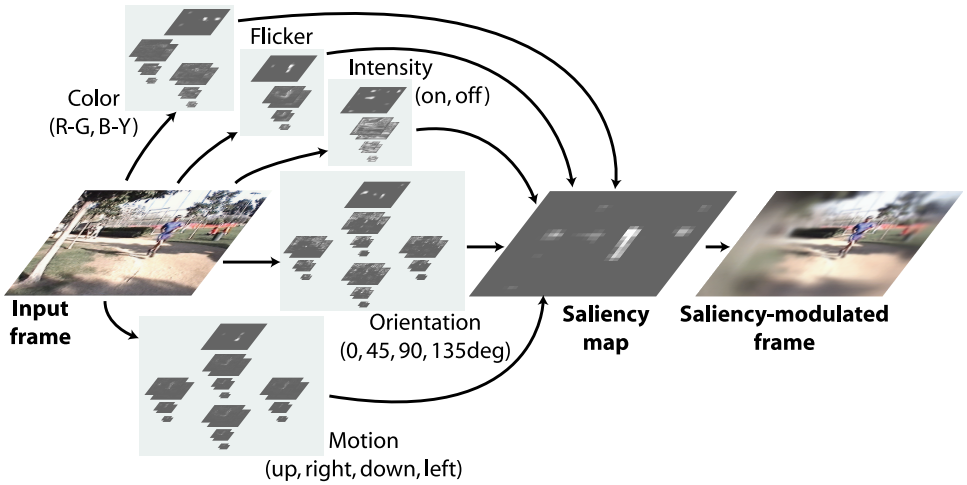

Figure 1. Overview of the model. Inputs are decomposed into multiscale analysis channels sensitive to low-level visual features (two color contrasts, temporal flicker, intensity contrast, four orientations, and four directional motion energies). After non-linear competition for salience, all channels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

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.[27] 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[27] 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 temporal smoothing and normalization by a squashing function **(Fig. 2.a)**. Alternatively, a small number of discrete virtual foveas endowed with mass/spring/friction dynamics attempt to track a collection of most salient objects, using proximity as well as feature similarity to establish correspondence,[28] over extended time periods, between salient locations and fovea centers. Interestingly, feature similarity between a salient location on a new frame and a fovea from the previous frame is based on the same multiscale low-level features that contribute to the saliency map: thus, at any location, a 72-component feature vector is constructed by reading out (with bilinear interpolation) values from the 72 feature maps at that location (note that those values depend on neighborhoods of sizes 4×4 to 256×256 pixels, depending on the spatial scale of the corresponding feature maps). Correspondence between the n most salient locations on a given frame and the positions of p foveas from the previous frame is then established through an exhaustive scoring of all $n \times p$ possible pairings between a new salient location $i \in \{1..n\}$ and an old foveation center $j \in \{1..p\}$ (typically, p is fixed and $n = p + 4$ to ensure robustness against varying saliency ordering from frame to frame).

The correspondence score combines four criteria: 1) Euclidean spatial distance between the locations of i and j; 2) Euclidean distance between feature vectors extracted at the locations of i and j; 3) a penalty term $|i - j|$ that discourages permuting previous pairings by encouraging a fixed ordered pairing; and 4) a tracking priority that increases with salience, enforcing strong tracking of only very salient objects, while foveas may easily disengage from less salient objects. Combined, these criteria tend to assign the most salient object to the first fovea, the second most salient object to the second fovea, etc. 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 next frame). Video compression priority at every location is then inversely related to the distance to the closest fovea center **(Fig. 2.b)**, as computed using

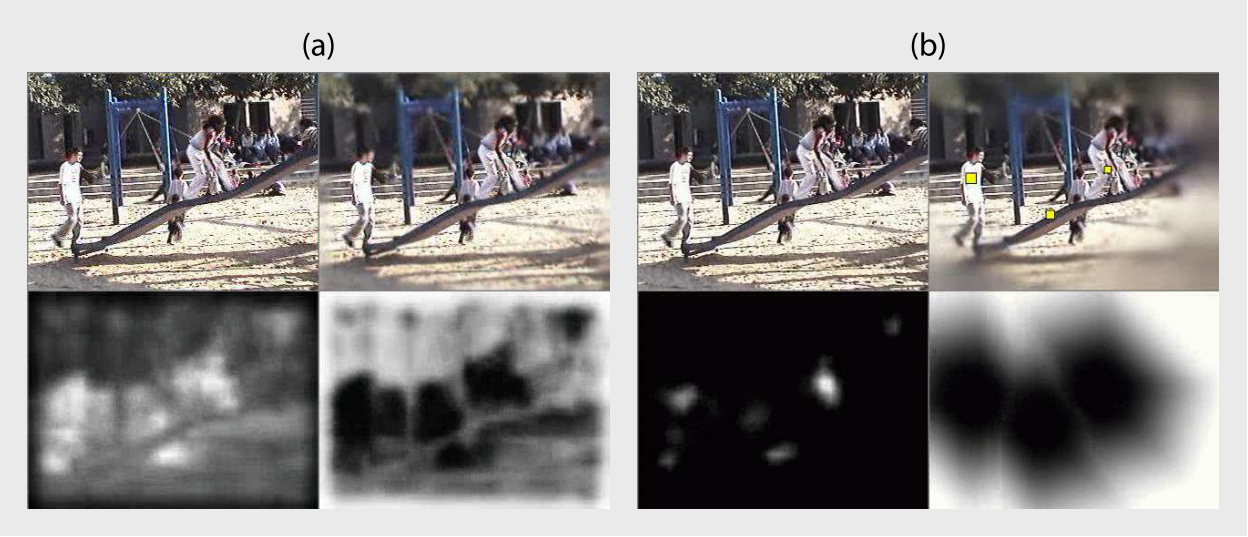

Figure 2. Examples of predicted priority maps, for two settings of our model: **(a)** Maxnorm feature normalization, continuous priority map (0 virtual foveas), foveation pyramid depth 4; **(b)** Fancynorm, 3 foveas, depth 6. **Top-left:** original frame; **top-right:** foveated frame (with fovea centers marked by yellow squares when using discrete foveas); **bottom-left:** saliency map (brighter for higher salience); **bottom-right:** priority map (darker for higher priority).

a 3/4 chamfer distance transform.[29] For further implementation details, we refer the reader to the source code of the algorithm, freely available online upon request.[30]

It is important to note that the dynamics of the virtual foveas do not attempt to emulate human saccadic (rapid ballistic) eye movements,[31] 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. It is also important to note that perfect tracking of a given set of objects is often not desirable for video compression purposes, which is why we have not used stronger object trackers like particle filters[32]; indeed, for efficient foveated video compression, it is very important to rapidly focus onto new salient objects, even though that often means losing track of some current object. Thus, the strategy implemented by our correspondence and tracking algorithm is a compromise between reliably tracking the few most salient objects, and cycling through a larger number of less salient objects. This maximizes coverage of less salient objects over long time periods, through time-sharing among a small number of virtual foveas. Finally, one key aspect of our approach is that is makes absolutely no direct assumption about 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.[26] Similarly, our foveation technique relies on spatial and feature-space distances rather than, for example, pre-defined object templates.

To evaluate the algorithm, we here simply use it as a front-end, applied before standard video compression algorithms (both MPEG-1 and MPEG-4 (DivX) encoding were tested): a spatially-variable blur is applied to the input frames, that is inversely related to the model-computed priority (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 the foveas move from frame to frame, the appearance of static objects far from the foveas will change, requiring continuous re-encoding of those changes while no re-encoding is necessary in the absence of foveation. Second, even when the foveas remain static for

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
```

Figure 3. Video grabbing and encoding

some time, peripheral moving objects will receive variable amounts of blur, depending on their distance to the closest fovea, hence defeating any motion compensation scheme in the encoder, and also requiring continuous re-encoding of their changing appearance. Specific video codecs have been proposed that 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, we here 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.

The encoder settings used in our study are shown in **Fig. 3**. For MPEG-1, we simply used the default settings of the `mpeg_encode` program.[33] For MPEG-4, the default settings of the `mencoder` program[34] are for constant-bitrate, variable-quality encoding, which is not appropriate for comparison between foveated and unfoveated clips (since the encoder would vary compression quality such as to produce a stream with approximately same filesize in both cases). 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.

3. 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 corrected or uncorrected 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 or text messages), but would instead help us evaluate

their general understanding of the contents of the clips. Thus our goal was to bias subjects towards regions of high cognitive interest in the video clips, as these would be the ones that should be encoded in priority. Whether these regions could be somewhat predictable by a simple bottom-up analysis of the image is a central question in this study: while instructions emphasized high-level concepts (actors and actions), our algorithm computes low-level properties of patches of pixels, and has no notion of object, actor, or action. While a dominant view on static scene understanding is that bottom-up saliency may only contribute to eye movements within the first half second when viewing a scene, with cognitive models of the world dominating the distribution of eye movements in a top-down manner afterwards,[35,36] here we test whether bottom-up analysis may have a more sustained influence on the selection of what could be the main actors and actions in dynamic scenes. The procedure was approved by USC's Internal Review Board, and informed consent was obtained from all subjects. A set of calibration points and clips not part of the experiment were shown to familiarize the subjects with the displays.

Stimuli were presented on a 22" computer monitor (LaCie Corp; 640×480, 60.27 Hz double-scan, mean screen luminance 30 cd/m^2, room 4 cd/m^2). Subjects were seated on an adjustable chair at a viewing distance of 80 cm ($52.5^\circ \times 40.5^\circ$ usable field-of-view) and rested on a chin-rest. A 9-point eye-tracker calibration was performed every five clips. Each calibration point consisted of fixating first a central cross, then a blinking dot at a random point on a 3×3 matrix. The experiment was self-paced and subjects could stretch before any 9-point calibration. Subjects fixated a central cross, pressed a key to start, at which point the eye-tracker was triggered, the cross blinked for 1206 ms, and the clip started. Stimuli were presented on a Linux computer, under SCHED_FIFO scheduling (process would keep 100% of the CPU as long as needed[37]). Each unfoveated clip (MPEG-1 encoded) was entirely pre-loaded 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[37] 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 33185 ± 2 μs.

Eye position was tracked using a 240 Hz infrared-video-based eye-tracker (ISCAN, Inc). Point of regard (POR) was estimated 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 ± 10 mm in our laboratory). Thus, no stricter restraint than a chin-rest was necessary, which is important as head restraint has been shown to alter eye movements.[38] All analysis was performed off-line. Linearity of the POR-to-stimulus coordinate mapping was excellent, as previously tested using a 7×5 calibration matrix, justifying the use of a 3×3 matrix here. The eye-tracker calibration traces were filtered for blinks, then automatically segmented into two fixation periods (the central cross, then the flashing point), or discarded if that segmentation failed a number of quality control criteria. 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 6 points remained after outlier elimination, recordings were discarded until the next calibration. A thin-plate-spline nonlinear warping algorithm was then applied to account for any small residual nonlinearity.[39] Data was discarded until the next calibration if residual errors greater than 20 pixels on any calibration point or 10 pixels overall remained. Eye traces for the five video clips following a calibration were remapped to screen coordinates, or discarded if they failed some quality control criteria (excessive eye-blinks, loss of tracking due to motion or excessive wetting of the eye, loss of corneal reflection due to excessive squinting). Calibrated eye traces were visually inspected when superimposed with the corresponding video clips, but none was discarded based upon that subjective inspection. Although we had no ground truth to further evaluate the accuracy of the recordings and calibrations, overall the quality of the recordings seemed remarkable with this eye-tracker (e.g., subjects tracking the 10-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 analog interlaced NTSC video sources **(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 can be viewed online[40] and 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;

Clip	frames	subj	Maxnorm, 0 fov, depth 4 %avg at eye	mpg1	divx	Fancynorm, 3 fov, depth 6 %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

Table 1. 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.

`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 algorithmic multi-foveation results and human eye movements may be examined online.[40] Clips had between 164 and 2814 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 1)**.

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 1**, 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** shows sample frames.

4. DISCUSSION

As previously mentioned, simply blurring the frames before compression is a worst-case scenario in terms of expected size gains. However, it is a useful exercise because of its compatibility with all existing video codecs. Further, 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. In this respect, the simple blur applied here was useful to visualize and evaluate our priority maps.

With both MPEG-1 and constant-quality MPEG-4 encoding, substantial size reductions were achieved on every clip by foveation. In **Table 1**, foveated clip size was approximately half of the unfoveated size, on average with our 50 test clips. Smallest size gains 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`, `beverly08`). Indeed, the amount of blur applied to the background would remain fairly unchanged in this case, but would vary greatly when discrete foveas focused on the moving 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). This can easily be understood by the fact that three discrete foveas will select at most three salient locations in the scene, while the continuous priority map may select many more, to varying degrees. Under conditions of complex full-field motion that defeats motion compensation in the encoder (e.g., pan, zoom), many salient locations were encoded crisply on each frame when in continuous mode (increasing file size), compared to only three when using discrete foveas (yielding smaller file size). Obviously, the better performance of the discrete foveation scheme with these clips came at the cost of of degrading possibly salient locations beyond the top three. A possible future extension of our algorithm thus consists of allowing the number of discrete foveas to vary from frame to frame.

An interesting feature of the results in **Table 1** 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 move, the variant with depth 4 yields overall more visually pleasant results. Indeed, the variant with depth 6 yields blurs that are often too strong for small details or fine text to remain distinguishable, and that most of the time are obvious to the observer. The variant with depth 4 and continuous blur is more subtle, and prioritizes a variable number

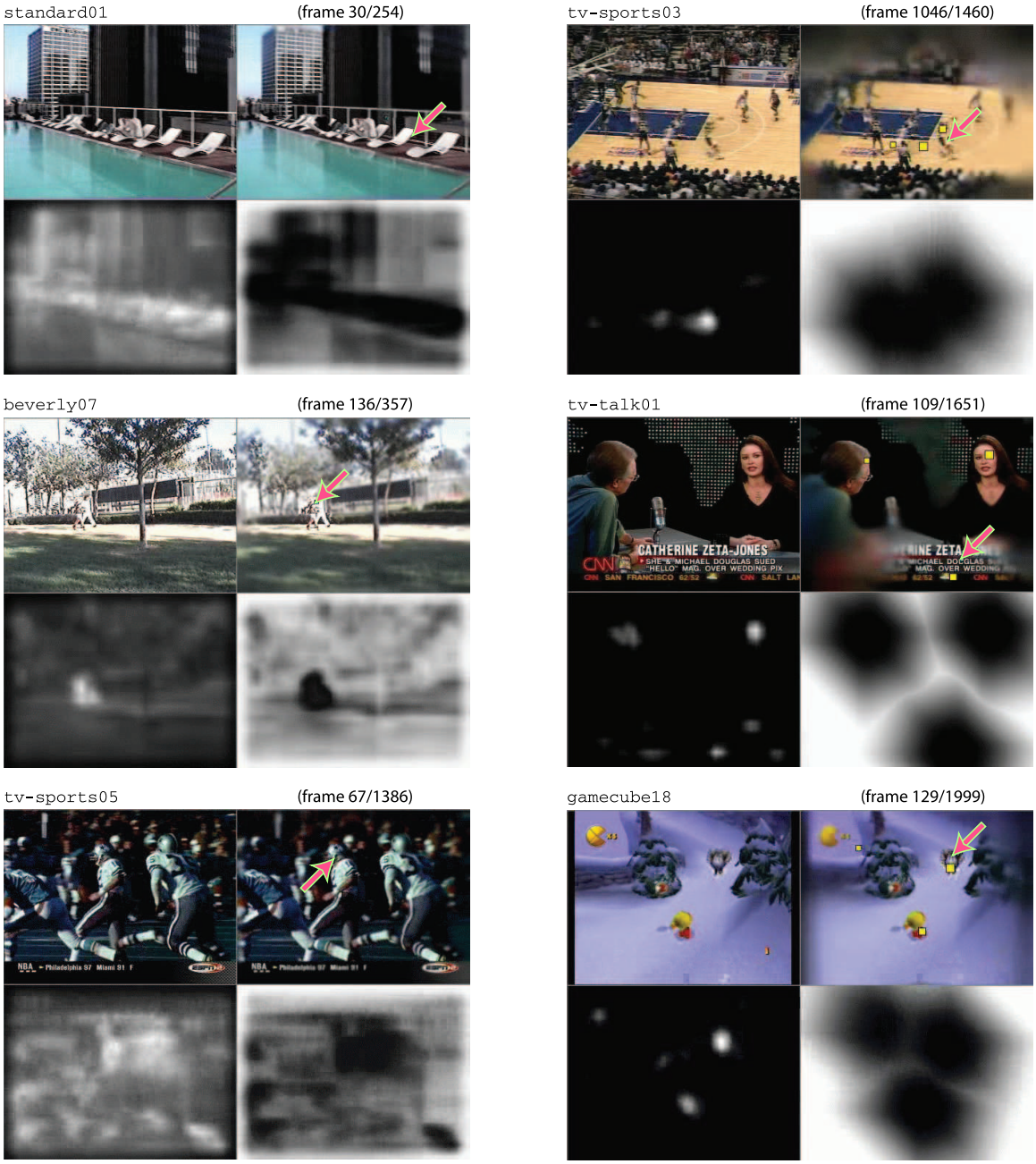

Figure 4. Examples of model-predicted priority maps, for the first (left column) and second (right column) variants of the model. Current eye position of one human observer watching the original unfoveated clip is shown (arrow) for comparison.

of locations to varying degrees in each frame. If fineprint text often is somewhat degraded with this variant, it is rare that an object of potential interest is so degraded as to become unrecognizable or unreadable. In contrast, any object far for the three discrete foveation centers will always be strongly degraded. Hence, our results argue somewhat against the traditional foveation paradigm and the binary selection of a small number of regions of interest.[12,4] 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.

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 was more pronounced with DivX encoding (as the default settings were to encode a keyframe only every 250 frames) than with MPEG-1 (where one intra-coded frame, compressed by itself and without reference to any other frame, was produced every 15 frames). Indeed, intra-coded frames would typically benefit from the reduction of texture complexity induced by foveation; in contrast, predictively-coded frames could suffer if large differences existed between current foveation mask and the foveation mask used for the last intra-coded frame, as these difference would defeat the prediction mechanism of the encoder. This is a clear limitation of the fairly simplistic blurring technique used here to visualize the predictions of the model. When using continuous blur, sometimes a similar problem existed, due to the fact that, inherently, the salience of an object depends on every other object in the scene. For example, consider a simple frame containing two equiluminant disks, one red and one blue, on a uniform black background. Both disks would be equally salient according to our model. Yet, the appearance of three additional red disks anywhere in the frame would reduce the salience of all red disks and increase that of the single blue disk (a so-called "pop-out" effect[27,26]). Similarly, in our more complex clips, the salience of a static object typically varied as other salient objects entered or left the field of view, even at distant locations in the image. With our simple testbed where blur was directly related to salience, this required re-encoding static objects to reflect changes in their appearance with varying blur. A fairly 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, for example in the context of continuously-rate-scalable encoding.[4] These limitations in the applicability of blur-based foveation to video compression may be reduced by the use of a bi-resolutional foveation blur instead of a continuously-variable blur.[11]

Overall, we found surprisingly strong agreement between our model's predictions and the scanpaths recorded from human observers. 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. This is a severe requirement, and visual inspection of the foveated clips suggests an even better agreement if such coincidence requirement was to be relaxed: indeed, often, given two objects present in the clip for an extended time, the model would foveate on the first object then the second, while a given human sometimes foveated the second and then the first. This resulted in high blur measures at both human fixations, while actually model and humans essentially were both interested in the same two objects. Our use of multiple foveas certainly reduced this trend, but did not entirely eliminate it.

Another difference which lowered the agreement between humans and model was that humans often fixated small details (e.g., fineprint text in the `tv-news` clips) which were not salient according to the model but were highly interesting and relevent to the human observers. This is perfectly reasonable, as we cannot expect human observers to only be drawn to salient image locations; instead, top-down influences (e.g., knowing that the anchor, although fairly bottom-up salient, will just speak, with a low probability of making interesting faces, and deciding that the text, although far less salient, might be more interesting) play a critical role in the spatiotemporal deployment of visual attention (in particular since our clips had no soundtrack).[41,26] Given this, it is remarkable that such good agreement was obtained for our wide variety of video clips. Our results indicate that throughout each video clip, bottom-up (image-based) influences remained very strong. 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 top-down guidance taking over afterwards.[36] 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 ignorant of what the main actors and actions in a video clip may be, its applicability to the analysis of extended video segments 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 was clear. This is a clear 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.

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.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, the National Eye Institute, the National Imagery and Mapping Agency, the Charles Lee Powell Foundation, and the Zumberge Faculty Research and Innovation Fund.

REFERENCES

1. B. Wandell, *Foundations of vision*, Sinauer Associates, Sunderland, MA, 1995.
2. J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Proc.* , pp. 3445–3462, 1993.
3. D. Taubman and A. Zakhor, "Multirate 3-d subband coding of video," *IEEE Trans. Image Proc.* **3**, pp. 572–588.
4. Z. Wang, L. Lu, and A. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Proc.* **12**(2), pp. 243–254, 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.* **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," in *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 and Systems for Video Tech.* **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 Proc.*, pp. 90–94, 1999.
9. U. Rauschenbach and H. Schumann, "Demand-driven image transmission with levels of detail and regions of interest," *Computers & Graphics* **23**(6), pp. 857–866, 1999.
10. E. M. Reingold and L. C. Loschky, "Saliency of peripheral targets in gaze-contingent multiresolutional displays," *Behav. Res. Meth., Instr. and Comp.* **34**(4), pp. 491–499, 2002.
11. D. J. Parkhurst and E. Niebur, "Variable-resolution displays: a theoretical, practical, and behavioral evaluation," *Hum. Factors* **44**(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," in *Proc. SPIE*, 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., Instr. and Comp.* **27**, 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," in *Proc. SPIE*, vol. 1453, pp. 147–152, 1991.
15. A. T. Duchowski and B. H. McCormick, "Pre-attentive considerations for gaze-contingent image processing," in *Proc. SPIE*, vol. 2411, pp. 128–139, 1995.

16. A. T. Duchowski and B. H. McCormick, "Modeling visual attention for gaze-contingent video processing," in *Ninth Image and Multidim.l Signal Proc. (IMDSP) Workshop*, pp. 130–131, 1996.
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 *Int. Conf. Patt. Recogn.*, pp. 701–704, Aug 1998.
18. X. Yang and K. Ramchandran, "A low-complexity region-based video coder using backward morphological motion field segmentation," *IEEE Trans. Image Proc.* **8**, pp. 332–345, Mar 1999.
19. F. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," in *Picture Coding Symp.*, pp. 101–104, Apr 2001.
20. W. Osberger, A. J. Maeder, and N. Bergmann, "Perceptually based quantization technique for mpeg encoding," in *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 *Data Compr. Conf.*, pp. 523–544, 2000.
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 Proc. (ICASSP2000)*, vol. IV, pp. 2345–2348, Jun 2000.
23. S. W. Kuffler, "Discharge patterns and functional organization of mammalian retina," *J. Neurophysiol.* **16**, pp. 37–68, 1953.
24. D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol. (London)* **160**, pp. 106–54, 1962.
25. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Patt. Anal. Mach. Intell.* **20**, pp. 1254–1259, Nov 1998.
26. L. Itti and C. Koch, "Computational modeling of visual attention," *Nat. Rev. Neurosci.* **2**, pp. 194–203, Mar 2001.
27. L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Elec. Imaging* **10**, pp. 161–169, Jan 2001.
28. S. Ullman, *High-level vision: Object recognition and visual cognition*, MIT Press, Cambridge, MA, 1996.
29. G. Borgefors, "Distance transformations in digital images," in *CVGIP: Image Understanding*, vol. 54, p. 301, 1991.
30. "Online: http://ilab.usc.edu/toolkit/."
31. D. L. Sparks, "The brainstem control of saccadic eye movements," *Nat. Rev. Neurosci.* **3**(12), pp. 952–964, 2002.
32. S. Maskell, M. Rollason, D. Salmond, and N. Gordon, "Efficient particle filtering for multiple target tracking with application to tracking in structured images," in *Proc. SPIE*, vol. 4728, 2002.
33. "Online: http://bmrc.berkeley.edu/research/mpeg/mpeg_encode.html."
34. "Online: http://www.mplayerhq.hu."
35. D. Noton and L. Stark, "Scanpaths in eye movements during pattern perception," *Science* **171**(968), pp. 308–11, 1971.
36. D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Res.* **42**(1), pp. 107–123, 2002.
37. S. A. Finney, "Real-time data collection in linux: A case study," *Behav. Res. Meth., Instr. and Comp.* **33**, pp. 167–173, 2001.
38. H. Collewijn, R. M. Steinman, C. J. Erkelens, Z. Pizlo, and J. van der Steen, "Effect of freeing the head on eye movement characteristics during three-dimensional shifts of gaze and tracking," in *The Head-Neck Sensory Motor System*, A. Berthoz, W. Graf, and P. P. Vidal, eds., Oxford University Press, 1992.
39. F. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Patt. Anal. Mach. Intell.* **11**, pp. 567–585, 1989.
40. "Online: http://ilab.usc.edu/bu/co/."
41. A. Yarbus, *Eye Movements and Vision*, Plenum Press, New York, 1967.