Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention

Robert J. Peters* and Laurent Itti†
Departments of Computer Science*,†, Neuroscience† and Psychology†
University of Southern California, Los Angeles, CA 90089
http://ilab.usc.edu/rjpeters/

Abstract

*A critical function in both machine vision and biological vision systems is attentional selection of scene regions worthy of further analysis by higher-level processes such as object recognition. Here we present the first model of spatial attention that (1) can be applied to arbitrary static and dynamic image sequences with interactive tasks and (2) combines a general computational implementation of both bottom-up (BU) saliency and dynamic top-down (TD) task relevance; the claimed novelty lies in the combination of these elements and in the fully computational nature of the model. The BU component computes a saliency map from 12 low-level multi-scale visual features. The TD component computes a low-level signature of the entire image, and learns to associate different classes of signatures with the different gaze patterns recorded from human subjects performing a task of interest. We measured the ability of this model to predict the eye movements of people playing contemporary video games. We found that the TD model alone predicts where humans look about twice as well as does the BU model alone; in addition, a combined BU*TD model performs significantly better than either individual component. Qualitatively, the combined model predicts some easy-to-describe but hard-to-compute aspects of attentional selection, such as shifting attention leftward when approaching a left turn along a racing track. Thus, our study demonstrates the advantages of integrating BU factors derived from a saliency map and TD factors learned from image and task contexts in predicting where humans look while performing complex visually-guided behavior.*

1. Introduction

How do we decide where to look? Directing spatial attention is a key function for both machine and biological vision systems, and our goal in this study was to develop a fully automated gaze prediction system incorporating both task-independent and task-dependent influences. Models of spatial attention can be characterized by several attributes, including: the type of visual stimulus, the type of visual task being modeled, and the level of detail desired in the model. In that context, the present study introduces a novel combination of a dynamic natural *stimulus* (contemporary three-dimensional video games) and a natural interactive *task* (playing the video game) with a fully computational *model* (Figure 1) that captures both bottom-up/task-independent and top-down/task-dependent influences on eye position. In contrast, prior studies using natural interactive tasks have often offered only descriptive models of gaze behavior, while prior studies using fully computational models have applied the models only to non-interactive stimuli, without accounting for top-down or task-dependent influences (see Section 2, Related work). Our proposed model combines separate bottom-up and top-down modules, each of which generates a predicted gaze density map that highlights likely gaze targets. The bottom-up component is based on the Itti-Koch saliency model [12], which predicts interesting locations based on low-level visual features such as luminance contrast, color contrast, orientation, and motion. The novel top-down component is based on the idea of "gist," which in psychophysical terms is the ability of people to roughly describe the type and overall layout of an image after only a very brief presentation [15], and to use this information to guide subsequent target searches [32]. Our model (Figure 1) decomposes each video frame into a low-level image signature intended to capture some of the properties of "gist" [30], and learns to pair the low-level signatures from a series of video clips with the corresponding eye positions; once trained, it generates predicted gaze density maps from the gist signatures of previously unseen video frames. To test these bottom-up and top-down components, we compared their predicted gaze density maps with the actual eye positions recorded while people interactively played video games. Finally, we also tested a combined model including both components. With this combined model, we find: (1) qualitatively, the model can now mimic some aspects of

1-4244-1180-7/07/$25.00 ©2007 IEEE

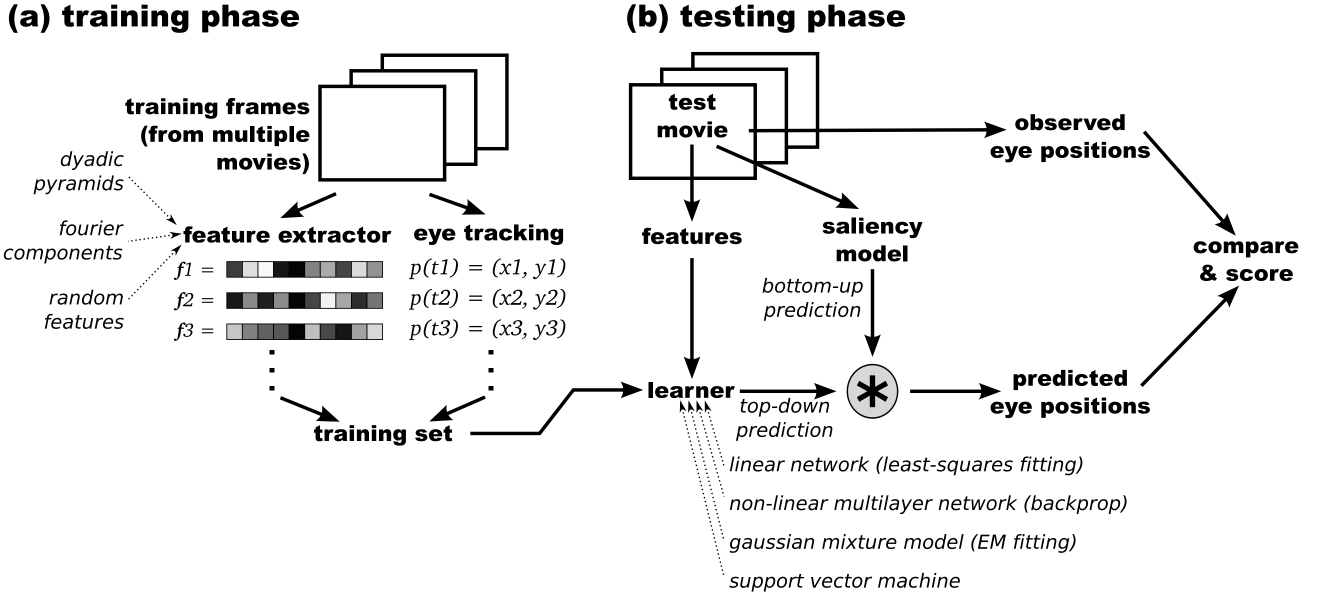

Figure 1. Schematic illustration of our model for learning task-dependent, top-down influences on eye position. First, in **(a)** the training phase, we compile a training set containing feature vectors and eye positions corresponding to individual frames from several video game clips which were recorded while observers interactively played the games. The feature vectors may be derived from either: the Fourier transform of the image luminance; or, dyadic pyramids for luminance, color, and orientation; or, as a control condition, a random distribution. The training set is then passed to a machine learning algorithm to learn a mapping between feature vectors and eye positions. Then, in **(b)** the testing phase, we use a different video game clip to test the model. Frames from the test clip are passed in parallel to a bottom-up saliency model, as well as to the top-down feature extractor, which generates a feature vector that is used to generate a top-down eye position prediction map. Finally, the bottom-up and top-down prediction maps can be combined via point-wise multiplication, and the individual and combined maps can be compared against the actual observed eye position.

high-level scene understanding, and (2) quantitatively, the model is significantly better than either individual component at predicting human gaze targets.

2. Related work

Studies of gaze direction can be considered according to the three attributes mentioned previously: **(1) Visual stimulus type**. Categories here include artificial psychophysical laboratory stimuli (*e.g.*, Gabor patches, simple search arrays) [34, 27, 18], static natural scenes (photographs, paintings) [37, 16, 28, 36, 26, 20, 32, 4, 19, 23, 25], and dynamic natural scenes (movies, cartoons, video games) [33, 14, 13, 6, 9, 22, 3]. One effect of ever-increasing computer power is that “artificial” stimuli can be rendered to appear very naturalistic, as in three-dimensional immersive virtual reality settings. **(2) Task type**. Possible task types include passive viewing, active viewing (visual search, scene comprehension, reading), and interactive viewing (video game playing, driving, web browsing). **(3) Model type**. Models of gaze prediction can be divided into two broad categories: (a) those that predict gaze from the scene’s semantic content, and (b) those that predict gaze from the scene’s raw image pixels alone. The first approach relies on an external source (typically, the experimenter) to provide some semantic pre-processing (*e.g.*, labeling scene fragments as “on-road” or “off-road” in a driving scene), so it is not a viable strategy for automated gaze-prediction. In turn, the second approach is fully automated, yet is typically limited to simple bottom-up features such as “saliency,” and misses important high-level regularities in human gaze behavior that are captured by the first approach (*e.g.*, people spend more time looking at the road when they are driving a car than when they are simply riding in a car as a passenger). Several previous computational studies combining bottom-up and top-down influences have worked within a visual search paradigm, taking the approach of biasing bottom-up features according to the properties of known properties of target and distractor items [4, 19]. In contrast, our approach here does not tune bottom-up processing for any particular target, but rather builds a separate top-down map that simply highlights task-relevant locations irrespective of the content at those locations.

When the task and stimulus are simple—for example, “find the unusual item” in a visual search array with a “pop-out” target—it is possible to make very accurate predictions of eye movements. In fact, those predictions can be made precise enough to be implemented in a machine-vision system that mimics human vision: it receives the same “retinal image,” processes that image computationally, and yields

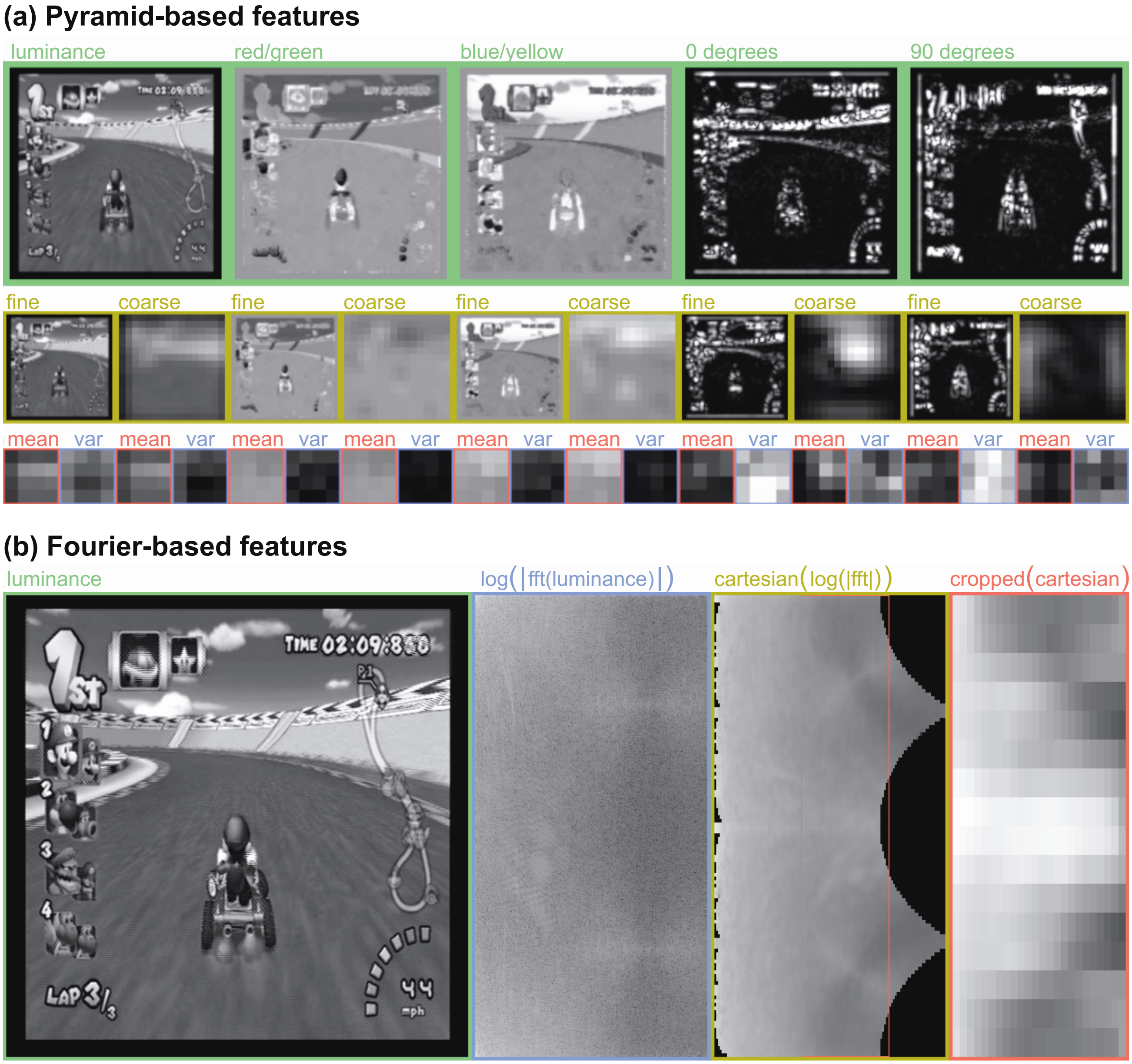

Figure 2. Two methods for extracting feature vectors for use in learning associations between "gist" and eye position, illustrated here for a single sample frame. **(a) Pyramid-based features**. Here, we use 7 of the 12 feature pyramids from the bottom-up model (green boxes): luminance, red/green and blue/yellow color contrast, and four orientations (though only two are shown in this figure). From each pyramid we extract a coarse and a fine scale (yellow boxes), and from each scale we compute the local mean (red boxes) and variance (blue boxes) within each patch on a 4×4 grid; these mean and variance values become the feature vector. **(b) Fourier-based features**. Here, we use the image luminance (green box) and first compute the log-magnitude of the FFT of the luminance (blue box). We transform the FFT into a Cartesian representation with θ and ω on orthogonal axes (yellow box) and select a subregion (outlined in red) from this representation. Within this subregion, we subsample the values down to 24 spatial-frequency bands and 16 orientation bands, and these values become the feature vector (red box).

simulated behavior like that of human observers. For example, human gaze is preferentially directed towards regions with: multiple superimposed orientations (corners or crosses) [38, 26]; above-average spatial contrast (variance of pixel intensities) [28], entropy [26], and texture contrast [21]; and above-average "saliency" [20, 23].

On the other hand, when the task and stimulus become less artificial and more complex—for example, driving a car through city traffic—such computational systems often fail to predict important aspects of eye movement behavior. In those cases, with current technology it is no longer possible to fully predict eye movements in the form of an algorithm operating on the retinal image; nevertheless, there are often very precise relationships between stimulus and behavior, as recent behavioral studies have shown [5]. For example, Yarbus [37] showed how gaze patterns depend on the task performed while viewing people in a painting: the observer's gaze fell on faces when estimating the people's age, but fell on clothing when estimating the people's material wealth. Other studies have used naturalistic interactive environments to describe how eye movements are guided by high-level task-relevant information, such as objects, agents, "gist," and short-term memory [2, 8, 29, 13, 31, 7, 1].

3. Methods

3.1. Videogame psychophysics and eye tracking

Five subjects (three male, two female) participated under a protocol approved by the University of Southern California Institutional Review Board. Each subject played four or five five-minute segments of standard Nintendo GameCube games (Mario Kart, Wave Race, Super Mario Sunshine, Hulk, and Pac Man World), using a standard GameCube controller to interact with the game. Stimuli were presented on a 22" computer monitor (640×480 pixels, 75 Hz refresh). Subjects rested on a chin-rest and were seated at a viewing distance of 80 cm, giving a usable field-of-view of 28°× 21°. To allow later processing with our computational models, the video game frames were captured, displayed and simultaneously saved on a dual-CPU Linux computer under `SCHED_FIFO` scheduling. Each of the 24 video game playing sessions led to 9,000 video frames giving 124GB of raw video data; for the analyses reported here we excluded the first and last 500 video frames (to avoid non-game frames such as navigation menus) from each clip and considered only the remaining 8,000 video frames and the corresponding eye position samples. Each subject's right eye position was recorded at 240Hz with a hardware-based eye-tracking system (ISCAN, Inc.) giving about 1.7 million total eye position samples.

3.2. Bottom-up saliency model

For the bottom-up component of our gaze-prediction model, we used the freely available implementation[1] of the Itti-Koch saliency model [12, 10]. Briefly, this model includes twelve feature channels sensitive to color contrast (red/green and blue/yellow), temporal luminance flicker, luminance contrast, four orientations (0°, 45°, 90°, 135°), and four oriented motion energies (up, down, left, right). These features detect spatial outliers in image space, using a center-surround architecture inspired from biological receptive fields. Center and surround scales are obtained from dyadic pyramids with 9 scales, from scale 0 (the original image) to scale 8 (the image reduced by a factor of $2^8 = 256$ in both the horizontal and vertical dimensions). Six center-surround difference maps are then computed as point-wise differences across pyramid scales, for combinations of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($\delta = \{3, 4\}$). Each feature map is additionally endowed with internal dynamics that provide a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition [11]. In this way, initially noisy feature maps can be reduced to sparse representations of only outlier locations which stand out from their surroundings. All feature maps finally contribute to a unique saliency map representing the conspicuity of each location in the visual field.

3.3. Top-down task-relevance model

The top-down component of our gaze-prediction model is designed to learn to associate the "gist" of an image with likely task-relevant locations under the current task. The model proceeds in two stages. First, in the training stage (Figure 1a), we build a training set using a leave-one-out approach—when one video game clip is used as the test clip, the training set is formed from the remaining 23 clips, and this procedure is repeated for each of the 24 clips. From each frame in the training clips we collect the recorded eye position of the observer who played the game as the clip was recorded, and we also compute from the entire image a low-dimensional feature vector that is intended to be diagnostic of the image's "gist." Two approaches for generating such feature vectors are described below. Second, in the testing stage (Figure 1b), we pass the set of observed eye positions and corresponding feature vectors to a learning algorithm, which, after training, can take feature vectors extracted from new test frames and generate eye position prediction maps.

Pyramid features. The pyramid-based feature vector (see Figure 2a), similar to [30], relies on 7 of the 12 feature pyramids from the bottom-up model: luminance, red/green and blue/yellow color opponency, and four orientations. From each of those pyramids we extract a fine scale (pyramid level 2, reduced from the original by a factor of $2^2 = 4$ in both the x and y dimensions) and a coarse scale (pyramid level 5, reduced by a factor of $2^5 = 32$). We divide each of those pyramid scales into 16 patches, on a 4×4 grid, and finally compute the within-patch mean and the within-patch variance for each patch. The within-patch means and variances become the elements of the feature vector, with a total of $7 \cdot 2 \cdot (16 + 16) = 448$ elements (7 pyramids, 2 scales per pyramid, 16 means plus 16 variances per scale).

Fourier features. The Fourier-based feature vector (see

[1] `http://ilab.usc.edu/toolkit/`

Figure 2b) uses Fourier energy from different orientations and spatial frequencies to form a "gist" descriptor, similar in spirit to [32]. Specifically, for each image we computed the logarithm of the magnitude of the FFT of the image luminance. We resampled the resulting array so that the cardinal axes represent orientation and spatial frequency, rather than the x and y directions. Finally, the feature vector consisted of $384 = 24 \cdot 16$ elements, each representing the average energy within one patch from among 24 spatial-frequency bands and 16 orientation bands.

Random features. As a control, we tested feature vectors with values drawn from a random distribution, thus testing how much the learner can learn from the eye positions alone, when the features are meaningless. In practice, and as expected, this gave identical results to those obtained from the "mean eye-position control" (see below), so for brevity we exclude this condition from further discussion.

Learning. To learn an association between eye positions and feature vectors, a number of potential machine-vision approaches could be applied, including multi-layer neural networks, support-vector machines, and the estimation/maximization (EM) algorithm. However, we started with an even simpler approach, which was to simply find a linear least-squares best fit. In order for the eye position data to be amenable to such a solution, we first coded the eye positions into coarse 300-element gaze density maps, with each position in the vector representing one of a 20×15 coarse array of eye positions. Thus, given an eye position (x,y) with $1 \leq x \leq 20$ and $1 \leq y \leq 15$, the gaze density map would be represented by a vector $\boldsymbol{p} = [p_1, p_2, \cdots, p_{300}]$ with $p_i = 1$ for $i = x + (y-1)\cdot 20$ and $p_i = 0$ otherwise.

Given T, the number of samples in a training set; M, the number of elements in each feature vector; N, the number of elements in each gaze density vector (always $20\cdot15 = 300$); $\boldsymbol{F}$, a $T\,\text{rows} \times M\,\text{columns}$ matrix of feature vectors; $\boldsymbol{P}$, a $T\times N$ matrix of gaze density vectors; and $\boldsymbol{W}$, a $M\times N$ matrix for which we wish to solve; then, the learning problem is represented by the matrix equation (1), and its linear least-squares best fit solution is given by inversion with the pseudo-inverse $\boldsymbol{F}^+$ of $\boldsymbol{F}$ (where $\boldsymbol{F}^+ \times \boldsymbol{F} = \boldsymbol{I}$) (2):

$$\boldsymbol{W} = \boldsymbol{F}^+ \times \boldsymbol{P} \quad (1)$$

$$\boldsymbol{F} \times \boldsymbol{W} = \boldsymbol{P}. \quad (2)$$

In practice, we computed the pseudo-inverse in terms of the singular value decomposition (SVD). To avoid numerical instability, eigenvectors whose eigenvalue was less than half of the largest eigenvalue were discarded during computation of the pseudo-inverse. Finally, for each frame during the test phase, we computed the test frame's feature vector $\boldsymbol{f}$, and generated a predicted gaze density map $\boldsymbol{p}$ as $\boldsymbol{p} = \boldsymbol{f} \times \boldsymbol{W}$. For visualization, the 300-element vector $\boldsymbol{p}$ can be unpacked back into a 20×15 two-dimensional array, examples of which are shown in Figure 3.

Mean eye position control. To test how effectively the pyramid-based and Fourier-based feature vectors actually captured task-relevant gist information, we used a control "learner" module which simply ignores the feature vectors, and instead just learns the mean gaze density vector, $\bar{\boldsymbol{p}}$, across all eye position samples in the entire training set (consisting of all 24 clips except for the current test clip). When this control model is asked to generate an eye position prediction corresponding to the feature vector from a new test frame, it again ignores the feature vector and just returns $\bar{\boldsymbol{p}}$. If the feature vectors carry no information that can be related to eye position, then the models based on such feature vectors will be expected to perform no better than the mean eye position control.

3.4. Normalized scanpath saliency (NSS)

To quantify how well the models' predictions matched observers' actual eye positions, we used the normalized scanpath saliency (NSS) [23], which is defined as the response value at the current eye position, $(x_{\text{human}}, y_{\text{human}})$, in a model's predicted gaze density map that has been normalized to have zero mean and unit standard deviation:

$$\text{NSS} = \frac{1}{\sigma_S}\left(S\left(x_{\text{human}}, y_{\text{human}}\right) - \mu_S\right), \quad (3)$$

where μ_S and σ_S^2 are the mean and variance of S, the unnormalized predicted gaze density map.

By definition, an NSS value of zero suggests that the model's prediction matches the observer's eye position no better than it would a random eye position; the observer's eye position would have fallen on an "average" location in that case. An NSS value of unity means that the observer's eye position fell on a location whose predicted salience was one standard deviation above average.

4. Results

Figure 3 shows several sample frames from the 192,000 frames that were analyzed, along with the corresponding gaze prediction maps from the various models. Considering first the predictions of the bottom-up saliency model (BU), we see that many times although the peak of the saliency map is distant from the actual eye position, there is a weaker local maximum very near to the eye position. In that context, the role of a top-down signal might be to narrow down the number of candidate locations that were highlighted by the initial bottom-up signal. Indeed, this is what we observe when the bottom-up maps are combined with one of the two top-down signals that we tested: either the static mean eye position (MEP), or the full dynamic top-down model based on pyramid features (TD) (maps from the top-down model based on Fourier features are not illustrated in the figure, but gave similar results both qualitatively and quantitatively; see Figure 4). Many times the top-down model

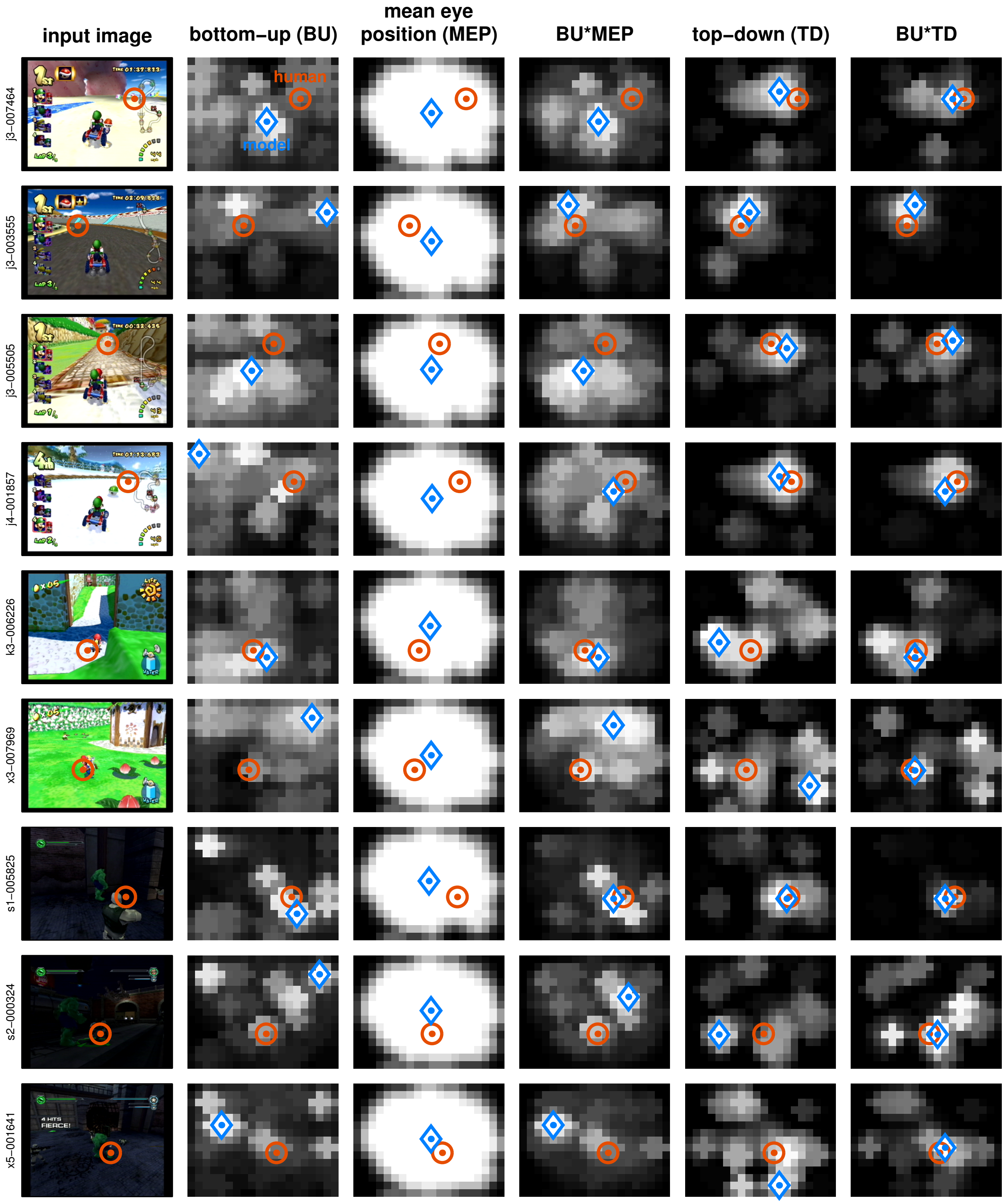

Figure 3. Each row shows a sample video game frame along with the predicted eye position maps generated by several of the computational models that we tested: the purely bottom-up saliency model (BU), the mean eye position prediction (MEP), the point-wise product of BU $*$ MEP, the top-down model based on pyramid features (TD), and the point-wise product of BU $*$ TD. Each orange circle indicates the observer's actual eye position from that frame; note that this is fixed within each row. Superimposed on the model prediction maps are blue diamonds which indicate the location of each map's peak location; a smaller distance between the orange circle and blue diamond suggests a better eye position prediction by the model.

alone already gives a good prediction of the eye position; however, several frames (rows 6, 8, and 9) illustrate cases where the bottom-up and top-down models each missed the target individually, but in combination came much closer to the actual eye position.

These qualitative trends are reflected quantitatively in the normalized scanpath saliency (NSS) scores across all 192,000 frames (Figure 4). The score for the bottom-up (BU) model alone was 0.58 ± 0.08, somewhat less than the NSS score of 0.69 ± 0.03 that was previously reported for the Itti-Koch saliency model applied to static photographs of natural outdoor scenes [23]. The mean eye position (MEP) control alone gave a significantly higher (paired t-test, $p < 0.05$) NSS score (0.76 ± 0.002); this reflects that a basic feature of observers' eye positions during the video game clips is that they tended to cluster near the center of the display, so that a trivial model that predicts a weak center bias will be accurate more often than not. Previous studies have demonstrated that there is often a bias toward central locations in subjects' eye positions [20, 23], which in this study is compounded by the fact that video games are typically designed to keep the important game elements at the center of the screen. Nevertheless, significantly better ($p < 0.05$) scores were obtained from the full top-down models, whether based on pyramid features (PFX: 1.07 ± 0.10) or Fourier features (FFX: 1.09 ± 0.13). Finally, when we combined the bottom-up and top-down components using a simple point-wise multiplication, we found significantly improved ($p < 0.05$) scores again for BU*MEP (1.10 ± 0.07), BU*PFX (1.19 ± 0.09), and BU*FFX (1.22 ± 0.11), with the best model overall (BU*FFX) scoring more than twice as well as the bottom-up model alone. Although these differences are statistically significant they are small in magnitude; again, this may simply reflect the central bias in the recorded eye position data, such that MEP scores are higher than would be expected if the eye positions were more uniformly distributed.

5. Discussion and Conclusion

The current model relies on simple algorithms in several places where more powerful or biologically-plausible computations could be substituted. For example, the learning stage is currently implemented by a linear least-squares best fit; one possible more sophisticated approach might involve a clustering network of radial-basis functions [24], where each node represents one canonical image "gist," and test frames are classified according to a k-nearest neighbor scheme. Likewise, the bottom-up/top-down combination stage currently involves only a simple point-wise product of the instantaneous bottom-up and top-down maps; a more sophisticated approach there might draw on neurophysiological [17] and psychophysical [35] studies of the millisecond-scale interactions between stimulus-driven and goal-driven influences. One possible mapping between neurobiology and the components of our model might be for the bottom-up map to represent a covert attention map containing many candidate locations of interest, with the top-down map acting as a gating mechanism which chooses the best task-relevant location from among those candidates. A current limitation of our proposed task-relevance model is that although generic enough to apply to any task or stimulus, its behavior in practice depends of course on its training set; it remains to be seen how diverse of a training set is needed to support broad generalization.

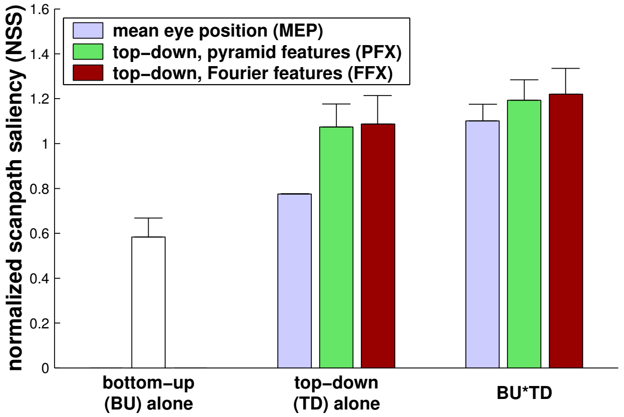

Figure 4. Normalized scanpath saliency (NSS) scores from comparing different models' predictions with actual eye positions recorded from human observers playing video games; a larger NSS score means a better fit, and an NSS score of zero would mean the model was no better than chance at predicting eye position. Each bar represents mean ± s.e.m. across 24 video game clips; each clip represents 8,000 individual NSS scores, one from each frame in the clip. The bottom-up (BU) model alone scores significantly above chance, but the scores are improved significantly when a top-down (TD) component is included in the model. The TD models alone score better than BU alone, and the combined BU*TD models score better than either type of model alone. Among the TD models, the full models based on pyramid features and Fourier features score significantly better than the simple mean eye position model. For significance scores, see Section 4, Results.

The main contribution of this study is its novel combination of three elements: a dynamic, naturalistic visual *stimulus*, an interactive *task*, and a fully computational gaze-prediction *model* that includes both bottom-up and top-down components. Our combined bottom-up/top-down model currently remains limited to processing of low-level features, and as such it is unable to reflect eye movement influences that depend on higher-level visual features (such as objects). Despite this limitation it is able in some cases to mimic such high-level behavior: for example, it often predicts that the observer's gaze will follow the direction of an upcoming turn, or will follow the location where enemy characters are likely to appear. This behavior occurs without any explicit high-level representations for objects and agents, but rather emerges from the fact that task-relevant relationships among such objects and agents may often be reflected in statistical regularities among low-level visual features. Our results suggest that these regularities can form the foundation of a powerful tool for predicting gaze direction during natural, interactive vision.

Supported by the NGA and the IC Postdoctoral Research Fellowship Program.

References

[1] J. Bailenson and N. Yee. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, 16:814–819, 2005.

[2] D. Ballard, M. Hayhoe, and J. Pelz. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80, Winter 1995.

[3] R. Carmi and L. Itti. The role of memory in guiding attention during natural vision. *Journal of Vision*, 6(9):898–914, Aug 2006.

[4] S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Pattern Recognition, Proceedings; in: Lecture Notes in Computer Science*, volume 3663, pages 117–124, 2005.

[5] M. Hayhoe. Advances in relating eye movements and cognition. *Infancy*, 6(2):267–274, 2004.

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

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

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

[9] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 631–637, San Siego, CA, Jun 2005.

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

[11] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, January 2001.

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

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

[14] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999.

[15] F. Li, R. VanRullen, C. Koch, and P. Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14):9596–9601, July 2002.

[16] S. Mannan, K. Ruddock, and D. Wooding. Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26(8):1059–1072, 1997.

[17] D. Munoz and S. Everling. Look away: the anti-saccade task and the voluntary control of eye movement. *Nature Reviews Neuroscience*, 5(3):218–228, March 2004.

[18] J. Najemnik and W. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, March 2005.

[19] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, January 2005.

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

[21] D. Parkhurst and E. Niebur. Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, 19(3):783–789, February 2004.

[22] V. Peli, T. Goldstein, and R. Woods. Scanpaths of motion sequences: where people look when watching movies. In *Proceedings of the Fourth Starkfest Conference on Vision and Movement in Man and Machines*, pages 18–21, Berkeley, CA, 2005. School of Optometry, UC Berkeley.

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

[24] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247(4945):978–982, 1990.

[25] M. Pomplun. Saccadic selectivity in complex visual search displays. *Vision Research*, 46(12):1886–1900, June 2005.

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

[27] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard. Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463, May 2002.

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

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

[30] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, February 2007.

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

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

[33] V. Tosi, L. Mecacci, and E. Pasquali. Scanning eye movements made when viewing film: Preliminary observations. *International Journal of Neuroscience*, 92(1-2):47–52, 1997.

[34] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[35] W. van Zoest, M. Donk, and J. Theeuwes. The role of stimulus-driven and goal-driven control in saccadic visual selection. *Journal of Experimental Psychology—Human Perception and Performance*, 30(4):746–759, August 2004.

[36] S. Voge. Looking at paintings: Patterns of eye movements in artistically naive and sophisticated subjects. *Leonardo*, 32(4):325–325, 1999.

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

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