Probabilistic Learning of Task-Specific Visual Attention

Ali Borji, Dicky N. Sihite, and Laurent Itti
University of Southern California, Departments of Computer Science and Neuroscience

http://ilab.usc.edu/~borji/papers/cvpr-2012-top-downV8.pdf

Supported by the National Science Foundation, the General Motors Corporation, and the Army Research Office

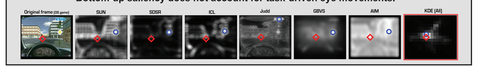
(1) Introduction

* Despite a considerable amount of previous work on bottom-up saliency modeling for predicting human fixations over static and dynamic stimuli, few studies have thus far attempted to model top-down and task-driven influences of visual attention.

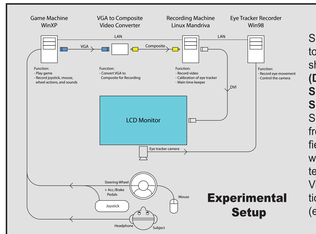
* Here, taking advantage of the sequential nature of real-world tasks [1], we propose a unified Bayesian approach for modeling task-driven visual attention. Several sources of information, including global context of a scene, previous attended locations, and previous motor actions, are integrated over time to predict the next attended location.

* Recording eye movements while subjects engage in 5 contemporary 2D and 3D video games, as modest counterparts of everyday tasks, we show that our approach is able to predict human attention and gaze better than the state-of-the-art, with a large margin (about 15% increase in prediction accuracy). The advantage of our approach is that it is automatic and applicable to arbitrary visual tasks.

Bottom-up saliency does not account for task-driven eye movements.



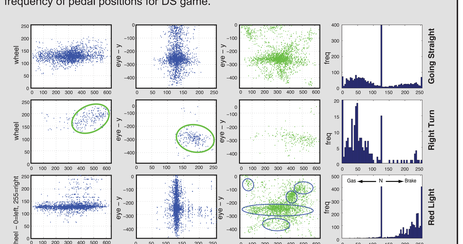
(2) Data Gathering



Experimental Setup

Subjects aged 20-30 were asked to play 5 games with the rig shown at left: **3D Driving School (DS), 18 Wheel of Steel (WS), Super Mario Bros (SM), Burger Shop (BS), and Top Gun (TG)**. Subjects were placed at 130cm from the screen subtending a field of view of 43° x 25°. There was a 5-min training before the test sessions for each game. Video frames [30Hz], Eye fixations [240Hz], and Actions [62Hz] (except TG) were recorded.

Correlation between actions and saccade positions. Rows indicate events (each frame was manually tagged based on its event type). Columns from left to right include: wheel vs. eye−x, eye−y vs. wheel, saccade coordinates during the game (eye−x vs. eye−y), and frequency of pedal positions for DS game.



Summary statistics of our data including overall number of saccades, subjects, durations per subject, frames, sizes in GB, and action types (J indicates joystick and M stands for mouse).

Game	# Sacc.	# Subj	Dur.	# Frames	Size	Action
DS	6382	10	10 min	180K	110	J
WS	4849	10	10 ″	180K	110	J
SM	1482	5	5 ″	45K	26	J
BS	1763	5	5 ″	45K	26	M
TG	4602	12	~4.5 ″	99K	57	N/A

(3) Features

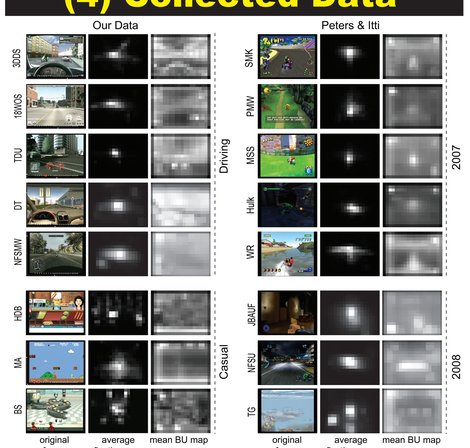
Global context (Gist, G). A quick summary of the quintessential characteristics of an image. We adopt the gist model of [2][1] as it is based on the bottom-up saliency model [3]. We consider 4 scales for each **orientation** pyramid, 6 scales for each **color** pyramid, and 6 scales for **intensity**. For each of the maps, average in each of the patches of grid sizes n × n (here n∈{1, 2, 4}) are calculated (thus 21 values). Overall the final gist vector will be the augmentation of (4×4+6×2+6×1)×21 = 714 values. We then exploit PCA to reduce the dimensionality of this vector. We also, investigate the ability of histogram of oriented gradient (HOG) [4] features to represent the global context of a scene[2].

Previous saccade location (X). A lot of everyday tasks need a number of perceptions and actions to be performed in a sequence (e.g., Sandwich making [1]). Therefore, knowing what object has been attended previously gives an evidence for the next attended object. We implement this idea over spatial locations. For instance, $P(X^{t+1}=b|X^t=a)$ indicates the probability of looking at location b in the next time step given that location a is currently fixated (e.g., looking at left first and then right when turning right).

Motor actions (A). Actions and fixations are tightly linked thus, by knowing a performed action, one can tell where to look next. We assume that these actions correspond to some high-level events in the game. We logged actions for driving games, from which we only generated a 2D feature vector from wheel and pedal positions. For other games, 2D mouse position and joystick buttons were used.

[1] http://ilab.usc.edu/siagian/Research/Gist/Gist.html
[2] http://pascal.inrialpes.fr/soft/olt/

(4) Collected Data



Sample frames of the video games. Second column shows the average fixation location over all subjects. Third column shows the mean bottom-up saliency map derived from the Itti et al.'s [3] model showing the average bottom-up salient regions through the whole time course of a game. Some of these data (right) has been collected in our lab by Peters and Itti [6]. Our data is freely available upon request.

(5) Benchmark Models

Linear Regression (REG). This model maps Gist of the scene to the eye position:

$$\arg\min_{W} ||M \times W - X_{sacc}||^2 \text{ Subject to } : W \geq 0.$$

where M indicates the matrix of feature vectors and X is the matrix of eye positions. The least-squares solution of the above objective function is: $W = M^* \times X$, where M^* is the pseudo-inverse of the matrix M through SVD decomposition. Given vector E = (u, v) as the eye position over a 20 × 15 map (i.e., w = 20, h = 15) with u∈ [1, 20] and v ∈ [1, 15], the gaze density map can then be represented by vector $X = [x_1, x_2, \ldots, x_{300}]$ with x_i=1 for i=u+(v-1)×20 and x_i=0 otherwise.

k Nearest Neighbor Classifier (kNN). The attention map for a test frame is built from the distribution of fixations of its k most similar frames in the training set:

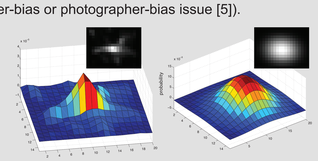
$$X^i = \frac{1}{k}\sum_{j=1}^{k} D(F^i, F^j)^{-1} X^j$$

where X^i is the fixation map of the j−th most similar frame to frame i which is weighted according to its similarity to frame i in feature space.

Average Fixation Map (AFM) is the average of all saccade positions during the time course of a task over all m training frames:

$$AFM = \frac{1}{m}\sum_{j=1}^{m} X^j$$

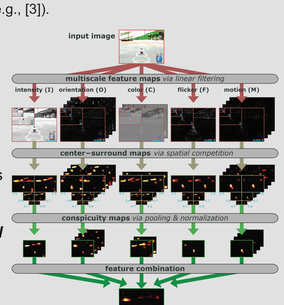
Central Gaussian filter (Gauss). The rationale behind using this model is that humans tend to look at the center of the screen when game playing (center-bias or photographer-bias issue [5]).



We fitted a 2D Bivariate Gaussian to the fixation data of each game using ML algorithm.

Bottom-up saliency models. We also compare the prediction power of our models against classic saliency models that predict fixations by detecting image outliers (e.g., [3]).

- input processed through as many as 5 multi-scale feature channels
- different versions of the model include different combinations of features
- ***salience is a global property***: outliers detected through coarse global competition



(6) Proposed Models

Case1 : Gist only

In this case, only global context information from all past and the current time is used. According to the Bayes theorem we have:

$$P(X_t|G_{1:t}) = P(X_t|G_t, G_{1:t-1}) = \frac{P(G_t|X_t)P(X_t|G_{1:t-1})}{P(G_t|G_{1:t-1})} = \mu P(G_t|X_t)P(X_t|G_{1:t-1})$$

Following Markov assumption, the current scene Gist (G_t) has all the necessary information for determining state and knowing the attended location.Thus X_t is independent of all previous gists: $P(X_t|G_{1:t-1}) = P(X_t)$. Therefore, we can write: $P(X_t|G_{1:t}) = \mu P(G_t|X_t)P(X_t)$ with $P(X_t)$ as the prior distribution over eye positions.

Case 2: Gist and previous saccade

In the second step, we add previous saccade locations to the formulation:

$$P(X_t|G_{1:t}, X_{1:t-1}) = P(X_t|G_t, G_{1:t-1}, X_{1:t-1}) = \frac{P(G_t|X_t)P(X_t|G_{1:t-1}, X_{1:t-1})}{P(G_t|G_{1:t-1}, X_{1:t-1})} = \mu_1 P(G_t|X_t)P(X_t|G_{1:t-1}, X_{1:t-1}) = \mu_1\mu_2 P(G_t|X_t)P(X_{t-1}|X_t)P(X_t|G_{1:t-1}, X_{1:t-2})$$

where μ_1 is equal to $P(G_t|G_{1:t-1}, X_{1:t-1})^{-1}$ and μ_2 is $P(X_{t-1}|G_{1:t-1}, X_{1:t-2})^{-1}$. Again, considering Markov assumption and defining $\mu = \mu_1\mu_2$, we have: $P(X_t|G_{1:t}, X_{1:t-1}) = \mu P(G_t|X_t)P(X_t|X_{t-1})P(X_t)$

Case 3: Gist, previous saccade, and motor actions

Finally, we combine all evidences in our Bayesian model. Following the steps in case 2 and simplifying we reach to:

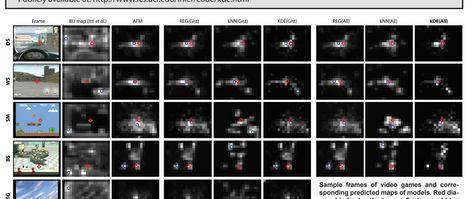
$$P(X_t|G_{1:t}, X_{1:t-1}, A^{j=1:n}_{1:t-1}) = \mu P(G_t|X_t)P(X_t|X_{t-1})P(X_t) \times \Pi_{j=1}^{n} P(A^j_{t-1}|X_t)$$

Above formula assumes that actions are independent of each other given the attended location (i.e., $A^k \perp A^l \mid X$). An important point here is whether actions influence saccades or vice-versa. Computing above requires estimation of $P(G_t|X_t)$ and similarly others. This can be done in several ways using non-parametric probability density estimation techniques such as generalized Gaussian model, histogram estimation or kNNs. We adapted the Kernel Density Estimation (KDE) approach. One pdf is calculated for each spatial location:

$$P(G|X=i) = \frac{1}{m}\sum_{j=1}^{m} K_h(G - G_{ij}) = \frac{1}{mh}\sum_{j=1}^{m} K\left(\frac{G - G_{ij}}{h}\right)$$

where K_h is a Gaussian kernel with smoothing parameter (sliding window or bandwidth) h and m is number of data points. We used a Matlab toolbox* for implementing KDE.

* Publicly available at: http://www.ics.uci.edu/ihler/code/kde.html



Sample frames of video games and corresponding predicted maps of models. Red diamond indicates the human fixation and blue circles is the maximum point of each map.

(7) Scoring and Results

Normalized Scanpath Saliency (NSS): the response value at the human eye position (x_h, y_h) in a model's predicted gaze density map that has been normalized to have zero mean and unit standard deviation:

$$NSS(t) = \frac{1}{\sigma_{s(x)}}\left(s(x(t)) - \mu_{s(t)}\right)$$

Area Under the Curve (AUC): A model's saliency map is treated as a binary classifier on every pixel; pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels are classified as non-fixated.

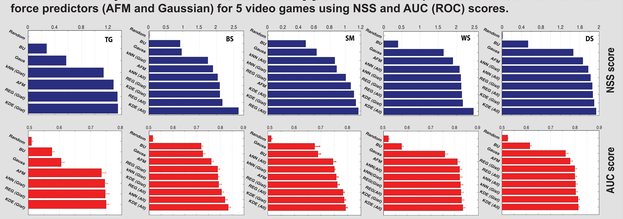
Game	ICL	SDSR	GBVS	AIM	SUN	Gauss	AFM	KDE(C=1)	KDE(C=2)	KDE(C=3)
DS	0.57	0.54	0.73	0.62	0.658	0.76	0.78	0.82	0.82	0.82
	0.19	0.05	0.948	0.54	0.30	1.47	1.66	1.9	1.91	1.95
WS	0.52	0.41	0.73	0.55	0.51	0.76	0.81	0.83	0.83	0.84
	0.27	-0.2	1.25	0.66	0.19	1.64	1.9	2.18	2.21	2.46
SM	0.61	0.69	0.72	0.67	0.62	0.67	0.75	0.78	0.79	0.79
	0.59	0.74	1.21	0.77	0. 33	0.62	1.07	1.13	1.21	1.11
BS	0.72	0.61	0.73	0.69	0.72	0.72	0.76	0.79	0.81	0.84
	1.04	0.54	1.1	0.80	1.2	0.96	1.89	2.1	2.2	2.7
TG	0.62	0.5	0.622	0.6	0.6	0.6	0.73	0.75	0.75	-
	0.58	0.01	0.55	0.51	0.29	0.57	1.28	1.36	1.34	-

AUC (1st rows) and **NSS** scores (2nd rows) of 5 state-of-the-art bottom-up saliency models, Gauss, AFM, and our models over 5 video games. The score of the best model in each row is shown in red. In almost all cases, while other models fall below Gaussian and AFM models, KDE (All) scores the best. In some cases, regression and KNN model score the best. C-x stands for Case x.

	Gist [2]		HOG [4]	
Game	kNN	REG	kNN	REG
DS	0.80 (1.77)	0.8 (1.86)	0.81 (1.88)	0.81 (2.05)
SM	0.75 (0.88)	0.76 (1.01)	0.74 (0.97)	0.79 (1.23)

Comparing AUC and NSS scores (in parenthesis) of the Gist model of Siagian et al. [2] and HOG features for saccade prediction using kNN and regression classifiers for 3D Driving School and Super Mario games.

Prediction accuracy of our KDE models, the Itti et al. [3], classifiers also implemented here, as well as brute-force predictors (AFM and Gaussian) for 5 video games using NSS and AUC (ROC) scores.



Conclusions & References

✓ **We proposed a unified Bayesian approach that is applicable to a large class of everyday tasks where global scene knowledge, the sequence of fixated locations, and physical actions, constrain future eye fixations.**

✓ **Applications: quantitative analysis of differences among populations of subjects (e.g., young vs. elderly or novices vs. experts) in complex tasks such as driving, assistant technologies for demanding tasks, prosthetic design, human computer interaction, context aware systems, and health care.**

✓ **It is still possible to gain higher performance by knowing more about the scene. For instance, by calculating the number or state of task-related objects.**

✓ **Extraction and addition of subjective factors such as fatigue, preference, and experience into our model is an interesting next step.**

[1] M. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? Vision Research, 2001.
[2] C. Siagian and L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention. PAMI, 2007.
[3] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. PAMI, 1998.
[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection.CVPR, 2005.
[5] B.W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions. Journal of Vision. 14(7):1-17, 2007.
[6] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. CVPR, 2007.