An Object-based Bayesian Framework for Top-down Visual Attention

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

http://ilab.usc.edu/~borji/papers/borjiAAAI.pdf

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

(1) Introduction

+ The idea is to predict the next attended object or saccade location when there is a task

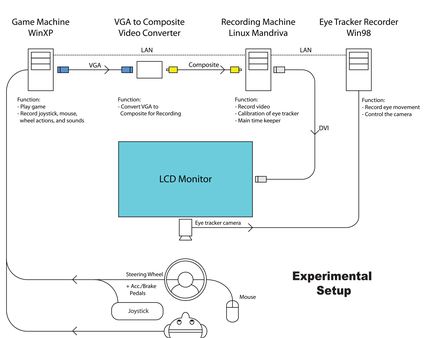
+ Other than global scene context, physical actions and sequential nature of everyday tasks provide rich information for gaze prediction

+ Different tasks demand different strategies, but many of them have common structures

+ Here we learn a Bayesian model from gaze data

Bottom-up saliency does not account for task-driven eye movements [5].

(2) Data Gathering



Experimental Setup

Subjects aged 20-30 were asked to play 3 games with the rig shown at the top: **Hot-dog Bush (HDB), 3D Driving School (DS), 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.

Game	# Sacc.	# Subj	Dur. (train-test)	# Frames (fixs)	Size	Action
HDB	1569	5	5-5min	35K	26.5GB	5D-Mouse
3DDS	6382	10	10-10	180K	110	2D-Joystick
TG	4602	12	5-5	45K	26	2D-Joystick

Summary statistics of our data including overall number of saccades, subjects, durations per subject, frames (and hence fixations, one to one relationship), sizes in GB, and action types.

(3) Features

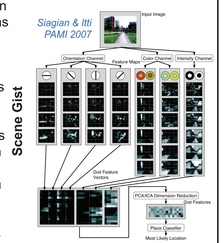
Global context (Gist, G). A quick summary of the quintessential characteristics of an image. We adopt the gist model of [2][1] as it is based on the bottom-up saliency model [3].

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.

Labeled Events (E). Frames of 3DDS game were manually labeled as belonging to one of different events: {left turn, right turn, going straight, ...}. Hence this is only a scalar feature.

Object Features (O). Properties of objects in the scene. At the simplest case could be the number of the instances from each object type of its presence or esence.

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



(4) Baseline Models

Linear Regression (REG). This model maps Gist of the scene to the eye position [4]:

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

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⁺ × 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, . . . , 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^j 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.

Mean Eye Position (MEP) is the average of all saccade positions during the time course of a task over all m training frames:

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

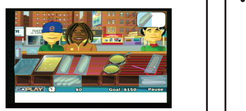
Central Gaussian filter (Gauss). The rationale behind using this model is that humans tend to look at the center of the screen when game playing (center-bias or photographer-bias issue [5]).

(5) Bayesian Object-based Model

+ Probability of an object being attended next and gaze position is calculated in a Bayesian framework from knowledge of all objects in the scene
+ Need detailed information about the scene and causal task structure
+ Are more descriptive than gist-only representations

Two cases of our model:
1) Memory-dependent: has access to previous time information
2) Memory-less: only uses the current time information

sample frame from the hot-dog bush game



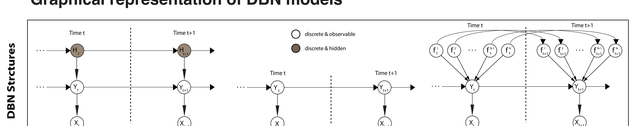
- **In general we are interested in:** $P(R_{t+1}|S_{t+1})$ (next attention unit | mental state)
  - No direct access to S_{t+1}, therefore estimate if from observables
- **Four modes for modeling attention:**
  - Memory-dependent (assumes access to previous gaze and action) vs. memory-less
  - We consider saccades here
- **A pdf over scene objects of being attended:**

Euclidean distance | fixation position | center of the j-th object | spatial decay

$$z(o^j) = 1/e^{\alpha d(X, C(o^j))}$$

$$P(o^j) = z(o^j)/\sum_{i=1}^{N} z(o^i)$$

Graphical representation of DBN models



$O_t = [o_t^1, o_t^2, \cdots, o_t^N]$ object-based representation of scene at time t

Define some functions encoding properties of objects: $F_t = \{f^i(o_t^j)\}$

$X_{1:T} = [X_1, X_2, \cdots X_T]$ Sequence of attended spatial locations
$Y_{1:T} = [Y_1, Y_2, \cdots Y_T]$ Sequence of attended objects
$C_{1:T} = [C_1, C_2, \cdots C_T]$ Sequence of selected actions
$F_{1:T}^{1:N} = [F_1^{1:N}, F_2^{1:N}, \cdots F_T^{1:N}]$ Sequence of object-level scene representations

Conditional independence assumptions:

1) $X_t \perp\!\!\!\perp F_t^i | Y_t$ gaze position is independent of scene given attended object
2) $F_t^i \perp\!\!\!\perp F_t^j$ No interaction between objects (assuming a general structure)
3) $F_{t+1}^i \perp\!\!\!\perp F_t^i$ Property of an object is independent of its property at previous time (given annotated data hence 100% accuracy in labeling)
4) $X_{t+1} \perp\!\!\!\perp X_t | Y_{t+1}$ gaze positions are independent through time given attended object

Reasoning:

$P(Y_{t+1}|F_{1:t+1}^{1:N}, X_{1:t}, Y_{1:t})$ % given all information in the past

$= P(Y_{t+1}|F_{1:t+1}^{1:N}, Y_{1:t})$ % $Y_{t+1} \perp\!\!\!\perp X_{1:t}$

$= P(Y_{t+1}|F_{t+1}^{1:N}, Y_t)$ % $Y_{t+1} \perp\!\!\!\perp Y_{1:t-1}$

$= \left(\prod_{j=1}^{N} P(Y_{t+1}|F_{t+1}^j)\right) \times P(Y_{t+1}|Y_t)$ % $F_{t+1}^i \perp\!\!\!\perp F_{t+1}^j, \forall i \neq j$

Further assuming that Y_{t+1} is independent of Y_t we end up to a Naive Bayes classifier (as a control for temporal model):

$$P(Y_{t+1}|F_{t+1}^{1:N}) = \frac{1}{Z}\prod_{i=1}^{N} P(F_{t+1}^i|Y_{t+1})$$

Xt is an integer between [1 300] (300 states). P(Y) is initialized uniformly over the objects (time 0 and is equal to P(oj)=1/N,j=1:N,N=15) and is updated over time.

(6) Scoring and Results

Scores

- Normalized scan-path saliency (NSS)

$$NSS = \frac{1}{\sigma_s}(s(x_h, y_h) - \mu_s)$$

- Mean NSS ($MNSS = \frac{1}{10}\sum_{\gamma=0}^{90} NSS(\gamma)$)

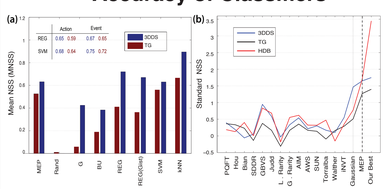
$$\gamma\%, \gamma \in \{0, 10, ..90\}$$

- Probability of correctly detecting the attended object

Ground-truth $(= \arg\max_j P(Y'))$

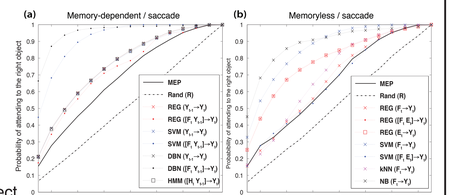
Predicted $(= \arg\max_{j=1\cdots 15} P(Y))$

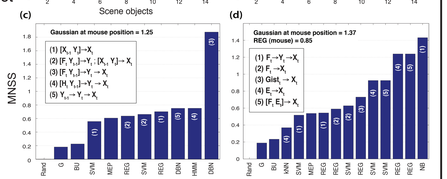
Accuracy of Classifiers



a) MNSS scores of our classifiers over 3DDS and TG games, b) NSS scores (corresponding to γ = 0 in MNSS) of BU models for saccade prediction over 3 games. Almost all BU models perform lower than MEP and Gaussian, while our models perform higher (same results using MNSS). Some models are worse than random (NSS ≤ 0) since saccades are top-down driven.

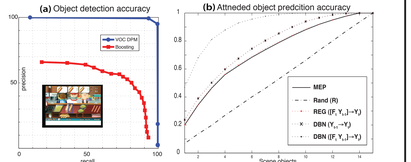
Main Results: Gaze Prediction





Gaze prediction accuracies for HDB game. a) probability of correctly attended object in memory-dependent/saccade mode, b) memoryless/saccade mode. $Q_{(F_t\ Y_{t-1})}$ means that model uses both objects and previous attended object for prediction. c) and d) MNSS scores for prediction of saccade position in memory-dependent and memoryless modes. White legends on bars show the mapping from feature types to gaze position X. For instance, REG ($F_t \rightarrow Y_t \rightarrow X_t$) maps object features to the attended object and then maps this prediction to the attended location using regression. Property functions f(.) in HDB indicate whether an object exists in a scene or not (binary).

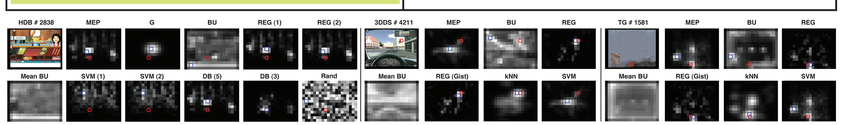
Uncertainty Analysis



Analysis of uncertainty over HDB game. a) Average precision-recall curve over all 15 objects; red for boosting and blue for DPM, b) Accuracy of correctly predicting the attended object.

Summary & Conclusions

✓ **We proposed a unified Bayesian approach that is applicable to a large class of everyday tasks where objects are attended sequentially.**
✓ **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.**
✓ **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.