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

Ali Borji **Dicky N. Sihite**
Department of Computer Science
University of Southern California
3641 Watt Way, Los Angeles, CA 90089-2520, USA
http://ilab.usc.edu

Laurent Itti
Department of Computer Science
Department of Psychology
Neuroscience graduate program
University of Southern California

Abstract

We introduce a new task-independent framework to model top-down overt visual attention based on graphical models for probabilistic inference and reasoning. We describe a Dynamic Bayesian Network (DBN) that infers probability distributions over attended objects and spatial locations directly from observed data. Probabilistic inference in our model is performed over object-related functions which are fed from manual annotations of objects in video scenes or by state-of-the-art object detection models. Evaluating over $\sim$3 hours (appx. $315,000$ eye fixations and $12,600$ saccades) of observers playing 3 video games (time-scheduling, driving, and flight combat), we show that our approach is significantly more predictive of eye fixations compared to: 1) simpler classifier-based models also developed here that map a signature of a scene (multi-modal information from gist, bottom-up saliency, physical actions, and events) to eye positions, 2) 14 state-of-the-art bottom-up saliency models, and 3) brute-force algorithms such as mean eye position. Our results show that the proposed model is more effective in employing and reasoning over spatio-temporal visual data.

Introduction and Background

To tackle information overload, biological vision systems have evolved a remarkable capability known as visual attention that gates relevant information to subsequent complex processes (e.g., object recognition). Knowledge of the task is a crucial factor in this selection mechanism. A considerable amount of experimental and computational research have been conducted in past decades to understand and model visual attention mechanisms, yet progress has been most rapid in modeling bottom-up attention and simple tasks such as visual search and free viewing. Furthermore, the field lacks principled computational top-down frameworks which are applicable independently of task type. Aside from being an interesting yet challenging scientific problem, from an engineering perspective, there are numerous applications for top-down attention modeling in computer vision and robotics, including video compression, object localization, scene understanding, interactive computer graphics, flight and driving simulators, and visual prosthetics (Toet 2011).

A rich qualitative understanding of gazing strategies for some tasks (e.g., tea and sandwich making, reading) is available (Land and Hayhoe 2001). For instance, Land and Lee (1994) proved that drivers track the tangent of the road when steering on a curvy cross-country road. In a block-copying task, Ballard, Hayhoe, and Pelz (1995) showed that the algorithm could be decoded from patterns of eye movements. Mennie, Hayhoe, and Sullivan (2007) explained the predictive nature of look-ahead fixations in walking. Land and Hayhoe (2001) classified eye fixations into four general categories: *locating a needed object* (e.g., milk in the fridge), *directing the hand* (grabbing something from shelf), *guiding* (lid onto kettle), and *checking* (water depth, spout), and proposed a schema for composing these so-called "object-related actions" to perform a task. Some computational models have been proposed to quantify these behaviors, though their generalizations across tasks remain limited. For instance, HMM models have been successfully applied to fixation prediction in reading (E-Z reader model by Reichle, Rayner, and Pollatsek (2003)). Renninger et al. (2005) suggested that observers fixate on locations that reduce local uncertainty in an object classification task. In a reward maximization framework, Sprague and Ballard (2003) defined three basic visual behaviors (litter collection, obstacle avoidance, and sidewalk following) for an avatar, and used reinforcement learning (RL) to coordinate these behaviors and perform a sidewalk navigation task. Butko and Movellan (2009) proposed a POMDP approach for visual search. Erez et al. (2011) proposed a similar approach for a synthetic eye-hand coordination task. In Navalpakkam and Itti (2005) some guidelines for top-down attention modeling are proposed when the task algorithm is known. Peters and Itti (2007) learned a spatial attention model by mapping a signature of scene (gist) to gaze fixation in navigation and exploration tasks. McCallum (1995) proposed the U-Tree algorithm for selective attention by discretizing the state-space of an RL agent to minimize the temporal difference error. Rimey and Brown (1994) modeled top-down attention with a Bayesian network for an object manipulation task. Cagli et al. (2009) proposed a Bayesian approach for sensory-motor coordination in drawing tasks. Inspired by the visual routines theory (Ullman 1984), Yi and Ballard (2009) programmed a DBN for recognizing the steps in a sandwich making task. Difficulties with models based on visual routines ideas are defining task modules, reward functions, and the use of very simple scene/object processings. Due to these obstacles, such models have rarely been applied to explain human saccades. In contrary, we follow data-driven approaches by learning models directly from human eye data for gaze prediction in novel situations.

Our primary goal is to present a general framework for interpreting human eye movement behavior that explicitly represents demands of many different tasks, perceptual uncertainty, and time. This approach allows us to model visuomotor sequences over long time scales, which has been typically ignored in vision sciences. For that, we employ graphical models which have been widely used in different domains, including biology, time series modeling, and video processing. Since objects are essential building blocks in scenes, it is reasonable to assume that humans have instantaneous access to task-driven object-level variables (e.g., (Einhauser, Spain, and Perona 2008)), as opposed to only gist-like (scene-global) representations (Torralba et al. 2006). Our proposed approach thus is object-based, in which a Bayesian framework is developed to reason over objects (it is also possible to add other variables such as gist, saliency map, actions, and events). We compare this approach to several spatial models that learn a mapping from scene signatures to gaze position. In this study, we use an object recognition oracle from manually tagged video data to carefully investigate the prediction power of our approach. For some of the tasks and visual environments tested here (older 2D video games), simple object recognition algorithms are capable of providing highly reliable object labels, but for more complex environments (modern 3D games) the best available algorithms still fall short of what a human annotator can recognize. Therefore, we also perform an uncertainty analysis when variables are fed from outputs of two highly successful object detection approaches (Boosting and Deformable Part Model (Felzenszwalb et al. 2010)).

Psychophysics and Data Collection

We chose video games because they resemble real-world interactive tasks in terms of having near-natural renderings, noise, and statistics. Participants (12 male, 9 female, 18-25 years old) played 3 PC video games under an IRB approved protocol and were compensated for their participation. Stimuli consisted of: 1) a time-scheduling game (**Hot Dog Bush (HDB)**), in which subjects had to serve customers food and drinks; 2) a driving game (**3D Driving School (3DDS)**) in which subjects were supposed to drive a car in an urban environment, following all traffic rules; and 3) a flight combat game (**Top-Gun (TG)**) where players control a simulated fighter plane with the goal of destroying specific enemy targets. In the training session for each game, subjects were introduced to the goal of the game, rules, and how to handle buttons, etc. All subjects were novice computer gamers and had no prior experience with our games, but some had limited experience with other games. Subjects had different adventures in games. After training, in a test session subjects played the game (but a different scenario) for several minutes. Table 1 shows summary statistics of our data.

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

Table 1: 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.

At the beginning of the test session, the eye tracker (PC1, Windows 95) was calibrated. Subjects were seated at a viewing distance of 130 cm corresponding to a field of view of $43° \times 25°$). A chin-rest (head-rest in 3DDS) was used to stabilize their heads. Stimuli were presented at 30 Hz on a 42" computer monitor at a resolution of 640×480 pixels and refresh rate of 60 Hz. Frames were captured at 30 Hz using a computer (PC2, Linux Mandriva OS) under SCHED-FIFO scheduling (to ensure microsecond accuracy) which sent a copy of each frame to the LCD monitor and saved one copy to the hard disk for subsequent processing. Finally, subjects' right eye positions were recorded at 240 Hz (ISCAN Inc. RK-464 eye tracker, PC1). Subjects played games on PC3 with Windows XP where all their joystick/steering/buttons actions were logged at 62 Hz. In 3DDS, subjects drove using the Logitech Driving Force GT steering wheel, automatic transmission, brake and gas pedals, 360 degrees rotation (180 left, 180 right), with force feedback. In HDB and TG games, subjects used mouse and joystick for game playing, respectively. Multi-modal data including frames, physical actions, and eye positions were recorded.

Modeling Top-down Visual Attention

We aim to predict both next object (what) and next spatial location (where) that should be attended under the influence of a task. We focus on predicting saccades which are jumps in eye movements to bring relevant objects to the fovea[1].

In its most general form, gaze prediction is to estimate $P(R_{t+1}|S_{t+1})$ where R_{t+1} is the next attended object Y_{t+1} or next attended spatial location X_{t+1}, and S_{t+1} is the subject's mental state. However, since it is not possible to directly access hidden (latent) variable S_{t+1}, we estimate $P(R_{t+1})$ directly from observable variables. Two modes for gaze prediction are possible: 1) **memory-dependent**, and 2) **memoryless**. The only difference is that in memoryless mode, information of previous actions and gazes is not available (only current time is used). In memory-dependent mode, goal is to predict gaze one step ahead.

Due to the noise in eye tracking, subjectivity in performing a task, and high-level gaze programming strategies, saccades do not always land on specific objects. One way to solve this problem is to ask humans to review the data, decide which object has been attended, and then take their average decisions. Instead, we followed a simpler and more objective approach by defining a function that assigns a probability to objects in the scene being attended, based on their inverse distance to the eye position X, i.e., $z(o^j) = 1/e^{\alpha d(X, C(o^j))}$ where $C(o^j)$ is the center of the ob-

[1] Saccades were defined by a velocity threshold of $20°/s$ and amplitude threshold of 2° similar to (Berg et al. 2009).

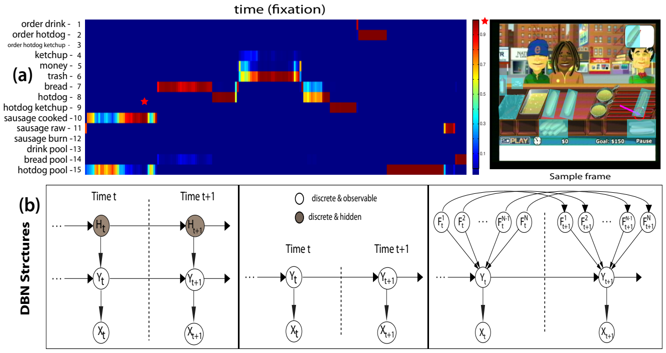

Figure 1: (a) A time series plot of probability of objects being attended and a sample frame with tagged objects and eye fixation overlaid. (b) Graphical representation of three DBNs unrolled over two time-slices. X_t is the current saccade position, Y_t is the currently attended object, and F_t^i is the function that describe object i at the current scene. H_t is the hidden variable in HMM which is learned using EM algorithm. All variables are discrete (see text).

ject o^j and d is the Euclidean distance. Parameter α controls the spatial decay with which an object is considered as attended for a given gaze location (here $\alpha = 0.1$). This way, closer objects to the gaze position will receive higher probabilities. These values are then normalized to generate a pdf: $P(o^j) = z(o^j)/\sum_{i=1}^N z(o^i)$; N is the total number of objects. Figure 1.a shows a sample time line of attended objects probabilities over HDB for $\sim 1,000$ frames along with a sample tagged frame. The object under the mouse position when clicking was considered as the selected object.

We follow a leave-one-out approach, training models from data of $n-1$ subjects and evaluating them over the remaining n-th one. The final score is the mean over n cross-validations. Object-based attention model is developed over HDB and classifier-based models are over all games.

Proposed Object-based Bayesian Framework

DBN is a generalized extension of Bayesian networks (BN) to the temporal dimension representing stationary and Markovian processes. For simplicity, we drop the index of subject in what follows. Let $O_t = [o_t^1, o_t^2, \cdots, o_t^N]$ be the vector of available objects in frame at time t. Usually some properties (features) of objects within the scene are important. Assuming that function $f(o)$ denotes such property, an object-level representation of this frame hence will be $F_t = \{f^i(o_t^j)\}$ where i is a particular property function and j is a particular object. In its simplest case, f could be just the number of instances of an object in the scene. More complex functions would take into account spatial relationships among objects or task-specific object features (For example, is ketchup empty or not?). Let $Y_{1:T} = [Y_1, Y_2, \cdots Y_T]$ be the sequence of attended objects, $X_{1:T} = [X_1, X_2, \cdots X_T]$ be the sequence of attended spatial locations, and $C_{1:T} = [C_1, C_2, \cdots C_T]$ be the selected objects by physical actions (e.g., by clicking, grabbing). Here, we treat the selected object as another object variable affecting the attended object. It is also possible to read out the next selected object (action in general) from our DBNs by slightly modifying the network structure, but here we are only interested in predicting the next attended object. Knowing the attended object, gaze location could be directly inferred from that.

We studied three types of general DBNs (Figure 1.b): 1) an HMM with a hidden variable (brain state H_t) connected directly to the attended object and from there to gaze position; 2) a DBN where the attended object is affected by the previously attended object (i.e., $P(Y_{t+1}|Y_t)$), hence prediction is only based on the sequence of attended objects; and 3) a DBN assuming that the attended object is influenced by properties of current objects in the scene as well as the previously attended object (i.e., $P(Y_{t+1}|Y_t, F_{t+1}^{1:N})$. Given the following conditional independence assumptions: 1) $X_t \perp\!\!\!\perp F_t^i \,|\, Y_t$, 2) $F_t^i \perp\!\!\!\perp F_t^j$ (due to general structure assumption), 3) $F_{t+1}^i \perp\!\!\!\perp F_t^i$ (happens when there is no uncertainty in case of having tagged data. It is not the case in general), and 4) $X_{t+1} \perp\!\!\!\perp X_t \,|\, Y_{t+1}$, then the full joint probability of the third DBN, to be learned, reduces to:

$$
\begin{aligned}
&P(X_{1:T}, Y_{1:T}, F_{1:T}^{1:N}) = P(X_{1:T}, Y_{1:T} | F_{1:T}^{1:N}) P(F_{1:T}^{1:N}) \\
&= P(X_{1:T} | Y_{1:T}) P(Y_{1:T} | F_{1:T}^{1:N}) P(F_{1:T}^{1:N}) = \qquad (1) \\
&\prod_{j=1}^{N} P(F_1^j) P(Y_1 | F_1^j) P(X_1 | Y_1) \prod_{t=2}^{T} \prod_{j=1}^{N} P(Y_t | F_t^j) P(Y_t | Y_{t-1}) \prod_{t=2}^{T} P(X_t | Y_t)
\end{aligned}
$$

where $F_{1:T}^{1:N} = [F_1^{1:N}, F_2^{1:N}, \cdots F_T^{1:N}]$ is the vector of functions representing object properties over time.

Inference and learning. Learning in a DBN is to find two sets of parameters $(m;\theta)$ where m represents the structure of the DBN (e.g., the number of hidden and observable variables, the number of states for each hidden variable, and the topology of the network) and θ includes the state transition matrix A $(P(S_t^i|Pa(S_t^i)))$, the observation matrix B $(P(O_t^i|Pa(O_t^i)))$, and a matrix π modeling the initial state distribution $(P(S_1^i))$ where $Pa(S_t^i)$ are the parents of S_t^i (similarly $Pa(O_t^i)$ for observations). Learning is hence to adjust the model parameters $V = (m;\theta)$ to maximize $P(O|V)$.

Since designing a different network for each task needs task-specific expert knowledge, to make the problem tractable, here we assume fixed structures (Figure 1.b) that could generalize over many tasks. Therefore, the joint pdf in Eq.1 reduces to predicting next attended object thanks to independence assumptions. As an example, we derive the formulation for the third case in Figure 1.b:

$$
\begin{aligned}
&P(Y_{t+1} | F_{1:t+1}^{1:N}, Y_{1:t}, X_{1:t}) \ \% \ given\ all\ information\ in\ the\ past \\
&= P(Y_{t+1} | F_{1:t+1}^{1:N}, Y_{1:t}) \ \% \ \ Y_{t+1} \perp\!\!\!\perp X_{1:t} \\
&= P(Y_{t+1} | F_{t+1}^{1:N}, Y_t) \ \% \ \ Y_{t+1} \perp\!\!\!\perp Y_{1:t-1} \qquad (2) \\
&= \left(\Pi_{j=1}^{N} P(Y_{t+1} | F_{t+1}^j)\right) \times P(Y_{t+1} | Y_t) \ \% \ \ F_{t+1}^i \perp\!\!\!\perp F_{t+1}^j, \forall\, i \neq j
\end{aligned}
$$

X_t is an integer between [1 300] (300 states). $P(Y)$ is initialized uniformly over the objects (time 0 and is equal to $P(o^j) = 1/N, j = 1:N, N = 15$) and is updated over time. The HMM model (case 1) has one hidden variable ([1 5]) and thus can be trained by exploiting the EM algorithm.

To avoid over-fitting parameters in conditional probability tables while training, train data was randomly split into k partitions, where DBN was trained over $k-1$ partitions and validated over the k-th partition. The model with best validation performance was applied to the test data.

Since variables in our DBN take discrete values, while we have a pdf over the attended object Y_t (ground-truth), we follow a stochastic sampling approach similar to the roulette-wheel algorithm. For a number of iterations, we loop through the training frames ($t = 1 \cdots T$) and generate more training sequences. Let m_t be the feature vector for the frame at time t, a tuple $< m_t, y_t, x_t >$ is added to the sequence ($< y_t, x_t >$ pair in the second DBN) where y_t is the index of an object sampled from $J(Y_t)$, the cumulative distribution of Y_t, and x_t is the eye fixation at that time (X_t). This way, objects with higher probability of being attended in a frame will generate more training samples. The same strategy is followed for classifier-based models (next section) for a fair comparison with DBNs. Since DBN has access to the previous time information, a sample $< [m_t \;\; y_{t-1}], y_t, x_t >$ is added to classifiers, where y_{t-1} and y_t are sampled from $J(Y_{t-1})$ and $J(Y_t)$, resp. (y_{t-1} is not added in memoryless mode).

Classifier-based Models

For a detailed assessment of our Bayesian approach, we developed several classifiers as well as brute-force control algorithms with the same input representations. According to the Bayes theorem, these classifiers estimate $P(R|M) = \frac{P(M|R)P(R)}{P(M)}$ (R being either X or Y; and M being either the feature-based representation of a scene E, or the object-based representation F, or a combination of both). Since calculating $P(M|R)$ and $P(M)$ is impractical due to high dimensionality of M, we follow a discriminative approach to estimate the posterior $P(R|M)$. Classifiers calculate either $P(X|E)$ (i.e., gaze directly from features; similarly predicting attended object from E, $P(Y|E)$) or $P(X|M) = \frac{P(M|Y)P(Y|X)}{P(M)} = \frac{P(Y|M)P(X|Y)}{P(X)}$ (i.e., a classifier first predicts attended object from features and then a second classifier maps the predicted attended object to gaze position). The following linear and non-linear classifiers were developed for predicting attended object and location:

Mean eye position (MEP). This family of predictors ignores feature vectors and simply uses the prior distribution over all saccade positions, or attended objects over all training data (Ak.a human inter-observer model). Note that while this model is easy to compute given human data, it is far from a trivial model, as it embodies human visual-cognitive processes which gave rise to the gaze.

Random predictor (R). At each time point, the next attended object is drawn from a uniform distribution (without replacement for the duration of the current frame) with probability $1/N$; N is the number of remaining objects in the scene. For saccade prediction, this is a random map.

Gaussian (G). It has been shown that subjects tend to look at the center of the screen (center-bias or photographer-bias issue (Tatler 2007)), therefore a central Gaussian blob can score better than almost all saliency models when datasets are centrally biased (Figures 2 and 3). We thus also compare our results with this heuristic model, which is simply a Gaussian blob ($\sigma = 3$ pixels) at the image center.

Regression (REG). Assuming a linear relationship between feature vectors M and R (X or Y), we solve the equation $M \times W = R$. The solution is: $W = M^{+} \times R$, where M^{+} is the (least-squares) pseudo-inverse of matrix M through SVD decomposition. Vector X is the eye position over 20×15 map which is downsampled from the 640×480 image. Given an eye position (u, v) with $1 \leq u \leq 20$ and $1 \leq v \leq 15$, the gaze density map would then be represented by $X = [x_1, x_2, \cdots, x_{300}]$ with $x_i = 1$ for $i = u + (v-1) \cdot 20$ and $x_i = 0$ otherwise. Similarly $Y = [y_1, y_2, \cdots, y_N]$ is the vector representing attended object ($y_j = 1$ for $j = \arg\max_j P(Y)$, and zeros elsewhere). Note that, if a vector of constant features (i.e., no scene information) is used as the M, the regression model generates the MEP. To predict eye positions/attended objects for test frames, feature vectors are first extracted, and attention maps are generated by applying the learned mapping which are then resized back to 20×15 2D array.

kNN. We also implemented a non-linear mapping from features to saccade locations. The attention map for a test frame is built from the distribution of fixations of its most similar frames in the training set. For each test frame, k most similar frames (using the Euclidean distance) were found and then the predicted map was the weighted average of the fixation locations of these frames (i.e., $X^i = \frac{1}{k} \sum_{j=1}^{k} D(M^i, M^j)^{-1} X^j$ 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 (i.e., $D(M^i, M^j)^{-1}$). We chose parameter k to be 10 which resulted in good performance over train data as well as reasonable speed.

SVM. To ensure that SVM training did not overwhelm available computational resources, we first reduced the high-dimensional feature vectors (i.e., E) using PCA by preserving 95% of variance. Then a polynomial kernel multi-class SVM classifier was trained with p output classes. p is equal to $N = |Y| = 15$ objects or $|X| = 300$ eye positions. We used LibSVM, a publicly available Matlab version of SVM.

Naive Bayes (NB). In memoryless case when there is no time dependency between attended objects, our DBN reduces to a static Bayes model incorporating only objects at time $t+1$. Assuming $F_{t+1}^{i} \perp\!\!\!\perp F_{t+1}^{j} | Y_{t+1}$, this classifier models $P(Y_{t+1}|F_{t+1}^{1:N})$ (probability of attended object given the current scene information). Therefore, $P(Y_{t+1}|F_{t+1}^{1:N}) = \frac{1}{Z} \prod_{i=1}^{N} P(F_{t+1}^{i}|Y_{t+1})$ (Z is a normalization constant). With no object information, this classifier reduces to priors $P(Y)$ and $P(X)$ which are equal to MEP. Here, as in DBN we also used validation strategy to avoid overfitting while training.

Features

We employed features from visual and action modalities:

Gist. Gist is a light-weight yet highly discriminant representation of the whole scene and does not contain details about individual objects. We used gist descriptor of Siagian and Itti (2007) which relies on 34 feature pyramids from the bottom-up saliency model (Itti, Koch, and Niebur 1998): 6 intensity channels, 12 color channels (first 6 red/green and next 6 blue/yellow color opponency), and 16 orientations. For each feature map, there are 21 values that encompass average values of various spatial pyramids: value 0 is the average of the entire feature map, values 1 to 4 are the average values of each 2×2 quadrant of the feature map and values 5 to 20 are the average values of the 4×4 grids of the feature map leading to overall of $34 \times 21 = 714$ D values.

Bottom-up saliency map (BU). This model includes 12 feature channels sensitive to color contrast (red/green and blue/yellow), temporal luminance flicker, luminance contrast, four orientations ($0°, 45°, 90°$, and $135°$), and four oriented motion energies ($up, down, left$, and $right$) in 9 scales. After center-surround difference operations in each scale and across scale competitions, a unique saliency map is created and subsampled to a 20×15 map which is then linearized to a 300 vector. We used the BU map both as a signature of the scene and as a saccade predictor.

Physical actions (A). In the 3DDS game, this is a 22D feature vector derived from wheel and buttons while subjects were driving. The main elements of this vector include: {wheel position, pedals (brake and gas), left and right signals, mirrors, left and right side views, and gear change}. Other action vector components are: {wipers, light indicators, horn, GPS, start-engine, radio volume and channel, show-menu, look-back view, and view change}. Subjects were encouraged not to use these latter buttons. In the HDB game, actions were {mouse position (x, y), left, middle, and right mouse clicks} by which subjects handled orders. Currently, there are no actions for the TG game.

Labeled events (L). Frames of 3DDS game were manually labeled as belonging to one of different events: {left turn, right turn, going straight, red light, adjusting left, adjusting right, stop sign, traffic check, and error frames due to mistakes that terminate the game like hitting other cars or passing the red light}. Hence this is only a scalar feature.

Object features (F). This is a 15D vector of properties of objects as discussed in previous sections ($F^{1:N}$).

Experiments and results

Scores. We used the Normalized Scan-path Saliency (NSS) score (Peters et al. 2005) which is the response value at the human eye position (x_h, y_h), in a model's predicted gaze map (s) that has been normalized to have zero mean and unit standard deviation: $NSS = \frac{1}{\sigma_s}(s(x_h, y_h) - \mu_s)$. $NSS = 1$ indicates that the subject's eye position fall in a region where predicted density is one standard deviation above average while $NSS \leq 0$ means that a model performs no better than chance. Due to high subject agreement (peaks in MEP), MEP and Gaussian (when peak is in the center) models generate many true positives which lead to high scores for them. Since the chance of making false positives is thus small, there is less opportunity for models to show their superiority over MEP or Gaussian. To stretch the differences between sophisticated and brute-force models, each time, we discarded those saccades that were in top $\gamma\%, \gamma \in \{0, 10, ..90\}$ of the MEP map. This gives an idea of how well models predicted "non-trivial" saccades, i.e., away from the central peak of MEP data. To summarize these scores, we defined Mean NSS ($MNSS = \frac{1}{10}\sum_{\gamma=0}^{90} NSS(\gamma)$) as a representative score of a model. To evaluate object-based models, for a frame, a hit is counted when the ground-truth attended object (i.e., $\arg\max_j P(Y')$) is in top k maximums (accumulative i.e., $1:2$, $1:3$, $\cdots$, $1:15$) of the predicted object pdf (i.e., $\arg\max_j P(Y)$). Hits are then averaged over all gazes for each k.

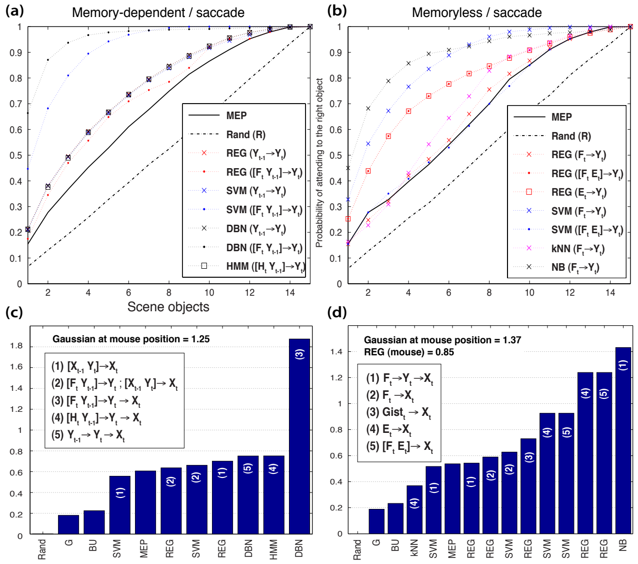

Figure 2: Gaze prediction accuracies for HDB game. a) probability of correctly attended object in memory-dependent/saccade mode, b) memoryless/saccade mode. $Q[F_t\ \ Y_{t-1}]$ means that model Q uses both objects and previous attended object for prediction. c) and d) MNSS scores for prediction of saccade 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).

Gaze prediction. Figure 2 shows prediction accuracies of models in all conditions (memory-dependent/memoryless; object/saccade position) for HDB game. Bayesian models performed the best in predicting the attended object followed by SVM. All models performed significantly higher than random, MEP, Gaussian, and a simple regression classifier from gist to eye position (Peters and Itti 2007) using MNSS score. Performances are higher in memory-dependent cases as we expected which shows that information from previous step is helpful. DBN model in memory-dependent mode

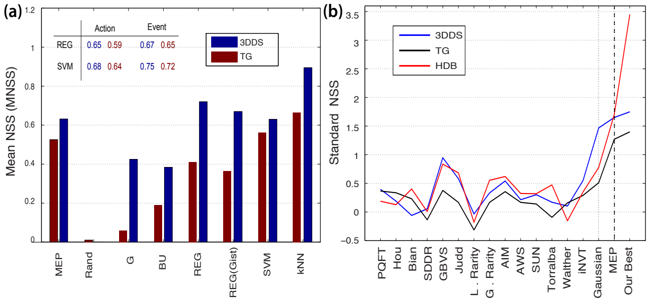

Figure 3: a) MNSS scores of our classifiers over 3DDS and TG games, b) NSS scores (corresponding to $\gamma = 0$ in MNSS) of BU models for saccade prediction over 3 games. Almost all BU models perform lower than MEP and Gaussian, while our models perform higher (same results using MNSS). Some models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

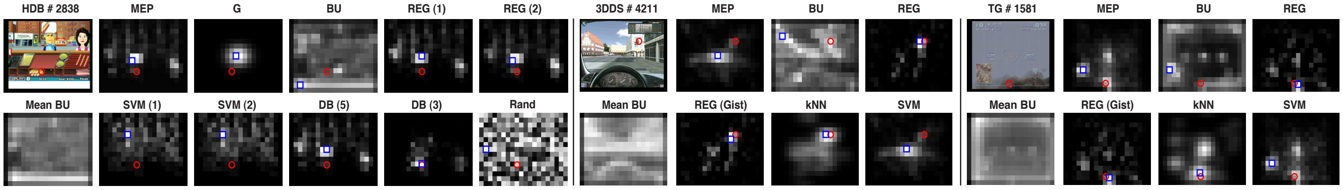

Figure 5: Sample predicted saccade maps by explained models. Each red circle indicates the observer's eye position superimposed with each map's peak location (blue squares). Smaller distance indicates better prediction.

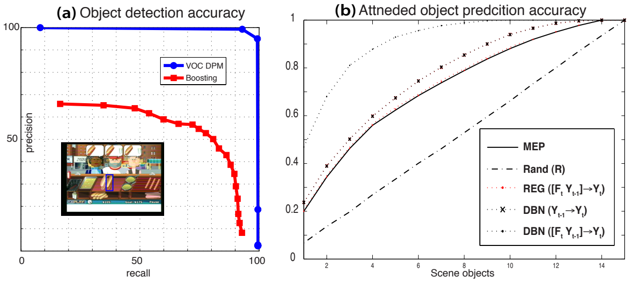

Figure 4: Analysis of uncertainty over HDB game. a) Average precision-recall curve over all 15 objects; red for boosting and blue for DPM, b) Accuracy of correctly predicting the attended object.

and Naive Bayes (NB) in memoryless mode, scored the best MNSS over saccades (followed by HMM in memory-dependent and REG in memoryless modes). Results show that inferring attended object first and using it to predict gaze position is more effective than directly mapping features to gaze position (DBN and NB). HMM model scored high on memory-dependent case but not as good in memoryless case. A similar HMM with added connection between object F_t and hidden variables H_t raised the MNSS to 1.5 in memory-dependent/saccade case. Best performance was achieved with 5 states for hidden variables in HMM. To test to what degree gaze follows mouse in HDB, we implemented two other algorithms: 1) by placing a Gaussian blob at mouse position, and 2) learning a regression classifier from mouse actions to eye position. These models scored high but still lower than Bayesian models.

Performance over 3DDS and TG games. Figures 3.a shows results over 3DDS and TG games using all features. kNN classifier achieved the best MNSS followed by SVM and Regression. Also, classifiers with event and action features performed higher than MEP and Gaussian indicating informativeness of these features for fixation prediction.

Model comparison. We ran 14 state-of-the-art BU models[2] to compare saccade prediction accuracy over three games (Figure 3.b). These models were the only ones that are readily applicable to our data compared to top-down models which thus far have been specific each to a particular task. Our models scored the best results compared with all bottom-up models. These results highlight the poor prediction power of bottom-up saliency models when humans are actively engaged in a task (notice the big difference between bottom-up, MEP, Gaussian, and our models).

Uncertainty Analysis. To analyze the degree to which our model is dependent on the uncertainty of the variables, we trained two object detection models: 1) Boosting model (BoostingToolbox) and 2) the Deformable Part Model (DPM) (Felzenszwalb et al. 2010) to automatically fill the variables instead of annotated data. Models were trained over a small set of cross validation data different from test frames. Average precision-recall curves of both models over 15 objects are shown in Figure 4.a. As opposed to Boosting, DPM was very successful to learn the objects thus we only used DPM. Detection performance was very high for each object due to limited variation in object appearance. As we expected, there was a graceful degradation in prediction of the attended object but still performance of our DBN was higher than the other models which indicates partial robustness of our model (Figure 4.b).

Sample gaze maps of all models are shown in Figure 5.

Discussion and conclusion

Results show the superiority of the generative Bayesian object-based approach to predict the next attended object/gaze position over 3 different complex tasks and large amount of data. This approach is applicable to many tasks when objects are processed sequentially in a spatio-temporal manner. Despite the promising results, there are some open questions for future research. Current analysis focuses on overt attention, however some parts of the scene are processed by subjects without direct gaze, e.g., by covert attention, which cannot be measured with an eye-tracker. A more biologically plausible future extension would be using foveated representation of the scene similar to (Najemnik and Geisler 2005) and (Larochelle and Hinton 2010) where object features in the periphery are accessible with less confidence or lower resolution. Using deformable part model, we were able to automatically detect objects in HDB with high detection rates (i.e., precision-recall for each object) yet there are still uncertainties in object variables. Having a causality structure over object variables could eventually give more evidence regarding the attended object (i.e., releasing conditional independence assumptions). One prob-

[2] To compare bottom-up saliency models over our data, we contacted model creators for codes, including: iNVT (Itti, Koch, and Niebur 1998), AIM (Bruce and Tsotsos 2005), Hou *et al.* (Hou and Zhang 2008), Local and Global Rarity (Mancas 2007), PQFT (Guo and Zhang 2010), AWS (Garcia-Diaz et al. 2009), GBVS (Harel, Koch, and Perona 2006), Bian *et al.* (Bian and Zhang 2009), SDDR (Seo and Milanfar 2009), Judd *et al.* (Judd, Ehinger, and Durand 2009), Torralba *et al.* (Torralba et al. 2006), Walther *et al.* (SaliencyToolbox), and SUN (Zhang et al. 2008).

lem we experienced, was learning the structure of DBN since to date structure learning algorithms are limited to certain network structures and variable types. Another area is investigation of generalization of proposed approaches. By training classifiers over a game and applying to similar games, we found that they scored better than chance implying that gist and action features to some extent capture the semantics directing gaze. We aim to build a website for sharing our codes, datasets (some used here) and evaluation programs for raising interest in this field. Finally, current work shows a promising direction to tackle this very complex problem, and helps designing experiments that can further shed light on mechanisms of top-down attention.

References

Ballard, D.; Hayhoe., M.; and Pelz., J. 1995. Memory representations in natural tasks. *Journal of Cognitive Neuroscience* 7(1).

Berg, D.; Boehnke, S.; Marino, R.; Munoz, D.; and Itti, L. 2009. Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision* 9(5):1–15.

Bian, P., and Zhang, L. 2009. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. *ICONIP'08, LNCS.*

BoostingToolbox. http://people.csail.mit.edu/torralba/shortcourserloc/boosting/boosting.

Bruce, N. D. B., and Tsotsos, J. K. 2005. Saliency based on information maximization. *Neural Information Processing Systems (NIPS).*

Butko, N. J., and Movellan, J. R. 2009. Optimal scanning for faster object detection. *CVPR.*

Cagli, R. C.; Coraggio, P.; Napoletano, P.; Schwartz, O.; Ferraro, M.; and Boccignone, G. 2009. Visuomotor characterization of eye movements in a drawing task. *Vision Research* 49.

Einhauser, W.; Spain, M.; and Perona, P. 2008. Objects predict fixations better than early saliency. *Journal of Vision.* 8(14).

Erez, T.; Tramper, J.; Smart, B.; and Gielen, S. 2011. A pomdp model of eye-hand coordination. *AAAI.*

Felzenszwalb, P.; Girshick, R.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI).* 32(9).

Garcia-Diaz, A.; Fdez-Vidal, X. R.; Pardo, X. M.; and Dosil, R. 2009. Decorrelation and distinctiveness provide with human-like saliency. *Advanced Concepts for Intelligent Vision Systems (ACIVS).* (5807).

Guo, C., and Zhang, L. 2010. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. on Image Processing.* 19.

Harel, J.; Koch, C.; and Perona, P. 2006. Graph-based visual saliency. *Neural Information Processing Systems (NIPS).* 19:545–552.

Hou, X., and Zhang, L. 2008. Dynamic visual attention: Searching for coding length increments. *Neural Information Processing Systems (NIPS).*

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI).*

Judd, T.; Ehinger, K.; and Durand, F., a. T. A. 2009. Learning to predict where humans look. *International Conference on Computer Vision (ICCV).*

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

Land, M. F., and Lee, D. N. 1994. Where we look when we steer. *Nature* 369:742–744.

Larochelle, H., and Hinton, G. E. 2010. Learning to combine foveal glimpses with a third-order boltzmann machine. *NIPS.*

LibSVM. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

Mancas, M. 2007. *Computational Attention: Modelisation and Application to Audio and Image Processing.*

McCallum, A. K. 1995. *Reinforcement Learning with Selective Perception and Hidden State.*

Mennie, N.; Hayhoe, M.; and Sullivan, B. 2007. Look-ahead fixations: Anticipatory eye movements in natural tasks. *Experimental Brain Research* 179(3):427–442.

Najemnik, J., and Geisler, W. S. 2005. Optimal eye movement strategies in visual search. *Nature.* 434:387–391.

Navalpakkam, V., and Itti, L. 2005. Modeling the influence of task on attention. *Vision Research* 45:205–231.

Peters, R. J., and Itti, L. 2007. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *CVPR.*

Peters, R.; Iyer, A.; Itti, L.; and Koch, C. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research* 45(8).

Reichle, E. D.; Rayner, K.; and Pollatsek, A. 2003. The e-z reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences* 26:445–476.

Renninger, L. W.; Coughlan, J. M.; Verghese, P.; and Malik, J. 2005. An information maximization model of eye movements. *Neural Information Processing Systems (NIPS).*

Rimey, R. D., and Brown, C. M. 1994. Control of selective perception using bayes nets and decision theory. *International Journal of Computer Vision* 12(2/3):173–207.

SaliencyToolbox. http://www.saliencytoolbox.net/.

Seo, H., and Milanfar, P. 2009. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision.* 9(12).

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

Sprague, N., and Ballard, D. H. 2003. Eye movements for reward maximization. *Neural Information Processing Systems (NIPS).*

Tatler, B. W. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision.* 14(7):1–17.

Toet, A. 2011. Computational versus psychophysical image saliency: a comparative evaluation study. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI).* 33(11):2131–2146.

Torralba, A.; Oliva, A.; Castelhano, M.; and Henderson. 2006. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review.*

Ullman, S. 1984. Visual routines. *Cognition* 18:97–159.

Yi, W., and Ballard, D. H. 2009. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics.*

Zhang, L.; Tong, M. H.; Marks, T. K.; Shan, H.; and Cottrell, G. W. 2008. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision.* 8(7).