What/Where to Look Next? Modeling Top-down Visual Attention in Complex Interactive Environments

Ali Borji, *Member, IEEE,* Dicky N. Sihite, and Laurent Itti, *Member, IEEE,*

***Abstract*—Several visual attention models have been proposed for describing eye movements over simple stimuli and tasks such as free viewing or visual search. Yet to date, there exists no computational framework that can reliably mimic human gaze behavior in more complex environments and tasks such as urban driving. Additionally, benchmark datasets, scoring techniques, and top-down model architectures are not yet well understood. In this study, we describe new task-dependent approaches for modeling 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/recognition algorithms. 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 compared with the state-of-the-art.**

***Index Terms*—Visual attention, Bottom-up saliency, , top-down attention, Gaze prediction, eye movement prediction, Interactive environments, Task-driven attention, Complex natural scenes**

I. Introduction

SELECTIVE processing of scenes known as visual attention is a remarkable capability of human vision allowing subsequent complex processes (e.g. object recognition) feasible. Knowledge of the task is a crucial factor in this selection mechanism. A considerable amount of experimental and computational research have been conducted in the 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 of visual attention lacks principled computational top-down frameworks which are applicable to different task types. Aside from being an interesting yet challenging scientific problem, from an engineering perspective, there are numerous applications for attention modeling in computer vision and robotics, including video compression [36][51] and summarization [40], object recognition and detection [59][34][22], robot navigation and localization [39][41], interactive computer graphics (virtual reality or video games) [37], flight and driving simulators (e.g., driver assistant systems), and visual prosthetic devices [38].

It is widely agreed that visual attention operates in both bottom-up and top-down modes, where in the bottom-up mode, attention is driven by image-based conspicuities, while in the top-down mode, task, knowledge, memory, expectations, emotions, etc. guide gaze toward relevant/informative scene regions. Furthermore, visual attention can be either object-based, space-based or feature-based [2][3]. Thus, attention is a multi-faceted phenomenon engaging all of these mechanisms. A frequently referenced model for saccade prediction is the saliency map model of Itti *et al.* [20] built on top of the computational architecture of Koch and Ullman [71] and the feature integration theory (FIT) [4]. This bottom-up approach is based on contrasts of intrinsic image features such as color, orientation, intensity, flicker, motion, and others. Later implementations have added new feature channels to this model, including text and faces [62], symmetry [75], gist and horizontal lines [76] (See [89] for an interesting application of scene gist), optical flow [77], and these models have been able of accounting for an increasing fraction of human eye fixations. In addition, several other bottom-up models with significantly different inspirations – either biologically-inspired or purely computational – have been proposed, including: Bayesian models (e.g., surprise [54], SUN [56], discriminant saliency [46]), Information-theoretic models (e.g., Bruce and Tsotsos [43], Hou and Zhang [52], rarity model [53]), Spectral analysis models (e.g., PQFT [51], adaptive whitening saliency [49]), bottom-up Graphical models (e.g., GBVS [58], E-saliency [78], Pang *et al.* [15]), Classification-based approaches (e.g., Judd *et al.* [12], Kienzle *et al.* [80]). Please refer to [70] for a comprehensive review of the bottom-up saliency models. Although bottom-up models have been very successful in explaining fixations in free-viewing, they explain

Manuscript received Jan. 13, 2012; revised May 5, 2012. This paper updates and extends an earlier presentation of this research on British Machine Vision Conference, 2011 [1]. This work was supported by Defense Advanced Research Projects Agency (contract no. HR0011-10-C-0034), the National Science Foundation (CRCNS grant number BCS-0827764), the General Motors Corporation, and the Army Research Office (grant numbers W911NF-08-1-0360 and W911NF-11-1-0046). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

Authors are with the Department of Computer Science, University of Southern California, Los Angeles, CA, 90089.

L. Itti is also with the Departments of Neuroscience and Psychology, University of Southern California, Los Angeles, CA, 90089.

E-mails: {borji, sihite, itti}@usc.edu.

only a small portion of fixations in everyday tasks such as driving [79][8][5] .

A. Related Research on Top-down Attention Modeling

The second component of visual attention comes from top-down demands such as knowledge of the task, emotions, expectations, and predictions which are embedded in a temporally extended task. Modeling top-down attention is conceptually hard to frame since: 1) different tasks require different algorithms, and 2) there are often several factors (e.g., actions, objects, etc.) that need to be taken into account specially in the context of a long temporally extended task. Research on top-down attention dates back to the classic study by Yarbus [7] which showed that gaze patterns are dependent on the asked question when viewing a photo. It has also been shown that the vast majority of fixations are directed to task-relevant locations, and fixations are coupled in a tight temporal relationship with other task-related behaviors such as reaching and grasping [16]. Furthermore, eye movements often provide a clear window to the mind of an observer in a way that it is sometimes possible to infer how a subject solves a particular task from the pattern of her eye movements for tasks such as "block copying" [17] , "tea making" [8], and "driving" [18]. In [19], Peters and Itti learned a mapping from global context of a scene (a.k.a scene gist) to eye fixations using the data of subjects playing contemporary video games. The same authors [21], using this model evaluated the relative importance of bottom-up and top-down factors at the time of an event (e.g., hitting a target in shooting games or accident in driving games). Based on this, they built templates for different events and used them for event detection (thus combining stimulus and behavior information for event detection). Navalpakkam and Itti [22] proposed guidelines for top-down attention modeling in conjunction with the saliency model in situations where the algorithm for the task is at hand. Sprague and Ballard [23] proposed a method based on reinforcement learning for learning visio-motor behaviors and used their model to account for saccades in a sidewalk navigation task [24]. Hidden Markov Models (HMM) have been successfully applied to fixation prediction in reading (e.g., E-Z reader model [63]). In [81], Butko and Movellan have proposed a POMDP approach for visual search. Erez *et al.* [64] proposed a similar approach for a synthetic eye-hand coordination task. Rimey and Brown [65] modeled top-down attention with a Bayesian network for an object manipulation task. Cagli *et al.* [66] proposed a Bayesian approach for sensory-motor coordination in drawing tasks. Inspired by the visual routines theory (Ullman [67]), Yi and Ballard [68] programmed a DBN for recognizing the steps in a sandwich making task.

B. Integrated Bottom-up and Top-down Models

A central open question in modeling visual attention is "how the bottom-up salient and top-down task-driven stimuli are integrated in the course of a task?". Few attempts have been made to answer this question and existing models mainly apply to simple tasks such as visual search (e.g. [22]). An example application for integrated attentive systems is video surveillance where the aim is to detect goal-relevant targets like suspects while being aware of unexpected visual events such as gun shots or sudden explosions. Another example is robot navigation where top-down attention helps detection of landmarks and road signs while bottom-up attention detects unexpected obstacles and accidents. Some experimental studies have considered interaction of bottom-up and top-down attention. For instance, in modeling eye fixations of observers when looking for a pedestrian in a scene, Ehinger *et al.* [31] showed that a model of search guidance combining three sources: *low level saliency*, *target features*, and *scene context*, outperforms models based on any of these sources taken separately. Navalpakkam and Itti [22] proposed "optimal cue selection strategy" by tunning the gains of the basic saliency model [20] through maximizing the signal to noise ratio of the target object versus distractors (background) by considering target and distractors feature distributions. Peters and Itti [19] used "multiplication" of bottom-up saliency and their top-down fixation prediction. However, it is not clear how this simple mechanism will generalize to complex tasks. Overall, while several models have addressed understanding and modeling visual attention mechanisms separately, to date there exists no principled approach that combines these components in the context of a complex, interactive, and temporally-extended task such as those we are considering in this work.

C. Influence of Multi-modal Data on Attention

The interaction between attention and physical actions makes up one of the most important facets of our everyday life. Many studies support the idea that attention affects actions (e.g. [25]). It has also been proposed that changes due to actions lead to corresponding changes in attention and perception [25][26]. For instance, in [28] authors show that preparation of a grasping movement affects detection and discrimination of visual stimuli. A good example of interaction between action and attention is driving which needs sophisticated coordination between motor actions and eye movements. Our work here borrows from the ideas of sensory-motor integration: *The process by which the sensory and motor systems communicate and coordinate with each other (e.g. hand-eye coordination)*. The above statement is closely related to the premotor theory of spatial attention which argues that the major function of attentional selection is not only a reduction in the incoming information, but rather to select an appropriate action on the basis of a specific stimulus [27]. There are also interactions among other modalities such as auditory or emotion on attention. Here, we investigate the influence of physical actions on eye positions albeit our approach is scalable for using all other sources of information.

D. Our Contributions

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 have 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 (See for instance [85] and [86]). We have been inspired by the application of HMMs for analysis and segmentation of videos into semantic shot sequences (See [88] and [87]). These approaches are very helpful when sequential actions are linked, when documents are highly structured or organized such a tennis match which is composed of sets, games and points.

We introduce two types of models: Space-based and Object-based. Space-based models are discriminative models that estimate probability of the next attended object or spatial location over time directly from raw image and action features such as gist, bottom-up saliency, physical actions, and events. An advantage of space-based models over object-based models is that they are easily applicable to a wide range of interactive visual environments. Thus, we intend to take advantage of the sequential nature of everyday tasks for modeling visual attention and eye movement prediction. Although the whole framework is general (i.e., the equations do not change from one task to another) but it needs to be trained for each specific task. By being able to explain human attentional behavior, we hope that our framework could be used for engineering applications.

Since objects are essential building blocks in scenes, it is reasonable to assume that humans have instantaneous access to task-driven object-level variables, as opposed to only gist-like (scene-global) representations [82][76]. Our proposed object-based approach is a Bayesian framework developed to reason over objects. We compare this approach to several space-based 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 did an uncertainty (over objects) analysis when variables are fed from outputs of two highly successful object detection approaches (Boosting classifier [84]) and Deformable Part Model (DPM) by Felzenszwalb *et al.* [60].

II. Psychophysics and Data Gathering

To test our models, we have collected a large amount of multi-modal data from subjects playing video games. We share our data and accompany software to encourage follow-up research on modeling top-down attention[1].

A. Stimuli and Subjects

We chose video games since they resemble real-world interactive tasks in terms of having near-natural renderings, noise, and statistics. It is also easier to control data recording over video games versus real-world scenarios (for example in driving). Participants (variable number for each game, 18-25 years old) played PC video games under a protocol approved by the University of Southern California's Institutional Review Board (IRB). Subjects were compensated for their participation by cash or course credits. In the training session for each game, subjects were introduced to the goal of the game, rules, 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, thus, it is unlikely that the exact same image is rendered in multiple runs. After training, in a test session subjects played the game (but a different scenario) for several minutes.

[1]Our data is publicly available at: http://ilab.usc.edu/∼borji/Resources.html

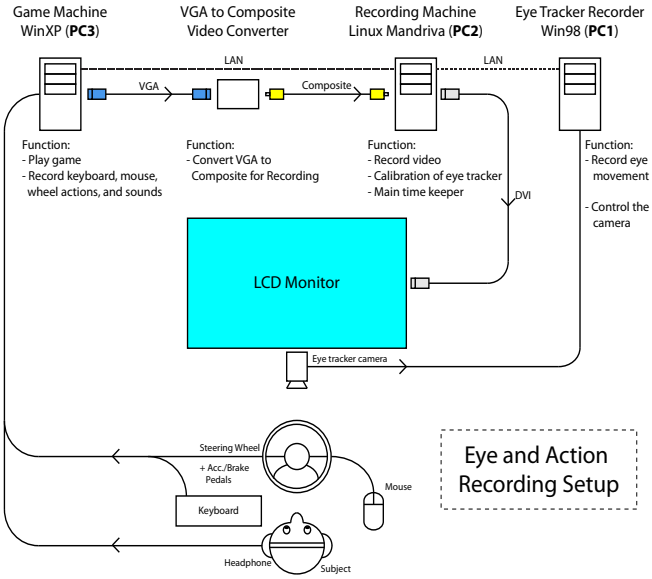

Fig. 1. Eye tracking and action recording setup. Subjects play video games using wheel, joystick, or keyboard with their actions being logged by a computer. Game stimuli are shown to the subjects and eye movements are recorded. Each data item has a time stamp which allows aligning frames, actions, and fixation data after recording.

B. Experimental Setup

Fig. 1 shows our eye and action recording setup. At the beginning of the test session, the eye tracker (PC1, Windows 95) was calibrated using 9-point calibration. Subjects were seated at a viewing distance of 130 cm (subtending a field of view of $43^\circ \times 25^\circ$). A chinrest (or headrest in driving games) 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 driving games, subjects drove using the Logitech Driving Force GT steering wheel, automatic transmission, brake and

gas pedals, 11-inch rubber-overmold rim, 900 degrees rotation (only 360 degrees; 180 left, 180 right; were used), force feedback, connected via USB to the PC3. In HDB and TG games, subjects used mouse and joystick for game playing, respectively. Multi-modal data including frames, audio (not processed here), physical actions, and eye positions were recorded.

C. Physical Actions and Eye Movement Data

Actions and fixations are tightly linked thus sometimes by knowing a performed action it is possible to tell where should be looked next. We recorded motor actions while humans were involved in game playing. We assumed that these actions correspond to some high-level events in the game (e.g., mouse click for shooting). We logged actions for driving games (e.g., wheel position, pedals (brake and gas), left and right signals, mirrors, left and right side views, and gear change), from which we only generated a 2D feature vector from wheel and pedal positions between 0 and 255. For other games, 2D mouse position and joystick buttons were used (further explained in Sec. III-A).

Fig. 2 shows a summary of our collected data over video games classified in three categories. Fig. 3 shows sample frames from each of the games. Part of our data has been previously collected by Peters and Itti [19][21].

III. Our Top-down Visual Attention Models

In contrast to the majority of previous models dealing with spatial attention, we aim to predict both the next object (what) and the next spatial location (where) that should be attended. Usually prediction of saccades (jumps in eye movements to bring the relevant object to the fovea)[2] has been considered by bottom-up models in free-viewing tasks. Here, we consider prediction of fixations for all frames (one fixation per frame) and saccades endpoints for those frames in which a saccade has happened.

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 the 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) **Memory-less**. The only difference is that in the memory-less mode, information of previous actions and gazes is not available. The memory-less mode has been mostly considered in spatio-temporal saliency modeling, where the input is a video and the task is to predict likely attended locations. However, in the context of sensori-motor interaction and online interactive tasks like those used here, memory-dependent prediction is a valid assumption with several applications when the goal is to predict gaze one step ahead.

In the rest, we explain the features that we use as indicators/predictors of eye fixations along with spatial and object based models using these features. We focus on three games for developing models and at the results section we analyze generalization power of our models over other games/tasks (See Fig. 4.a). Stimuli consist of: 1) a time-scheduling game known as **Hot Dog Bush (HDB)** in which subjects had to serve customers food and drinks; 2) a driving game called **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 known as **Top-Gun (TG)** where players had to control a simulated fighter plane with the goal of destroying specific enemy targets.

A. Features

As opposed to previous saliency models (e.g., [80][19][12]) that have only considered scene features for fixation prediction, we use features from both vision (features extracted from frames and visual events annotated for some games) and action modalities. Employed features include:

Gist. Gist is a light-weight yet highly discriminant representation of the whole scene and does not contain details about individual objects. We used the gist descriptor of [29] which relies on 34 feature pyramids from the bottom-up saliency model [20]: 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 $34 \times 21 = 714$ dimensions. In [19], the motion map has also been added for gist description. There are also other gist models which could be used here[3]. For a comparison of some gist models please see [83].

Bottom-up saliency map (BU). For the bottom-up component of our gaze-prediction model, we used the freely available implementation of the Itti-Koch saliency model [20][4]. This model includes 12 feature channels sensitive to color contrast (red/green and blue/yellow), temporal luminance flicker, luminance contrast, four orientations ($0°, 45°, 90°, 135°$), and 4 oriented motion energies (up, down, left, right). Then 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 (σ = 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. 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

[2]Saccades were defined by a velocity threshold of $20°/s$ and amplitude threshold of $2°$.

[3]For example the Gist descriptor by Oliva and Torralba: http://people.csail.mit.edu/torralba/code/spatialenvelope/

[4]http://ilab.usc.edu/toolkit/

Category	Name	Frames	Eyes	Actions	Play	Objects	Events	Meduim	Viewpoint	Comments
Hardcore	Super Mario Kart (SMK)	+	+	-	J	-	-	G	3rd	This is a racing game with various Nintendo character. Using a kart as vehicle, player is expected to finish as fast as possible by any means, including using items such as fireballs, bananas, etc to hinder other racers from finishing. In this stimuli, player races with 7 computer players. Player uses third-person view in this stimuli.
	Pc Man World (PMW)	+	+	-	J	-	-	G	3rd	A third-person adventure game involving Pacman, an iconic game character with yellow-coloured round body. Player controls him to follow a certain path to reach a destination by avoiding/killing enemies, avoiding obstacles, and solving puzzles.
	Mario Sunshine (MS)	+	+	-	J	-	-	G	3rd	This is a third-person adventure game involving Mario. Player controls his movement to retrieve an sun-like object called shine, a reward for completing a level. Each level requires player to either defeat a boss, killing several enemies, or finishing a puzzle to get the shine reward.
	Hulk	+	+	-	J	-	-	G	3rd	Hulk, a famous movie character, is a third-person adventure action game. Player controls him to run around bashing soldiers that tries to kill him.
	Wave Race (WR)	+	+	-	J	-	-	G	3rd	This is a jet-ski-themed racing. Player races on water to the finish line beating other racers and following path specified by water buoys. Player will be penalized by slowing the vehicle down for few seconds if he/she not following the correct path. In this stimuli, player race alone (time-attack mode) and in third-person view.
	James Bond Agent Under Fire (JBAUF)	+	+	-	J	-	+	G	1st	JBAUF is a first-person shooting starring James Bond. By controlling him, player must go to a specific point to finish a level. Along the path, player must kill enemies and use gadgets to solve puzzles. Bond is equipped with guns, multi-function cell phone, lazers, grappling hook, etc.
	Top Gun (TG)	+	+	-	J	-	+	G	1st	TG is a flight-combat simulator. Player controls a jet-fighter plane that can lock target and shoot missles, use afterburners to speed up, and do air manuvers. The main objective to complete the game is to completely destroy all targets on air and ground. Player uses first-person view in this stimuli.
Casual	Super Mario Bros (MB)	+	+	+	M	+	+	P	2D	SMB is a classic 2D-side-scrolling action game. Player controls Mario to a flagpole to finish the level. Mario can grow bigger if consume a mushroom and shoot fireballs if consume a flower. There are various enemies that can be killed by stomping on them or shooting fireballs. In this stimuli, player is expected not to take any means of shortcut such as running on ceiling (top of the level), teleport pipes, or warp points.
	Burger Shop (BS)	+	+	+	M	-	-	P	2D	BS is a 2D time-management game. Player serves customers under limited amount of time. They order specific foods, ie: burgers,fries, that can be assembled from a conveyor belt in the middle of the screen. The game ends when all customers are served.
	Hot Dog Bush (HDB)	+	+	+	M	+	-	P	2D	HDB is a 2D time-management game. Player serves customers hotdogs under limited amount of time. They will always order hotdogs either with or without ketchup. Later in the game, they will also order drinks. Player serves hotdogs by assembling the ingredients similar to BS with exception of the sausages, which need to be cooked properly.
Driving	Need for Speed Underground (NFSU)	+	+	-	J	+	-	G	1st/Bp	NFSU is a street-racing game with fancy and powerful cars to drive. This version of NFS series takes player into the night race in a city with various fancy city lightings such as buildings, street lights, etc. Player must cross the finish line before other racers to win a level. In this stimuli, player faces with other 3 computer players. Player uses first-person/bumper view in this stimuli.
	3D Driving school (3DDS)	+	+	+	W	+	+	P	1st/Db	3DDS is a driving simulator with simulated traffic conditions. Player must follow route defined by the game and european traffic rules (drive on right-side). An instructor will tell the player where to go by a text in a semi-translucent box above the screen and/or a small arrow on the corner top-left of the screen. Player uses automatic transmission to drive around the entire course, therefore player will focus only on the driving in a certain route and following the rules of this simulator. This stimuli has only dashboard view, an inside view from the driver-side towards the road.
	18 Wheels of Steel (18WOS)	+	+	+	W	-	+	P	1st/Db	18WoS is a semi/truck simulator with simulated traffic. In this game, player controls a big rig to a specific destination to retrieve money reward for delivering a trailer. Player must drive carefully as the truck cannot accelerate/brake suddenly due to its mass. In this stimuli, player is told to always make a left turn since there is no explicit instruction on the screen telling where to go. Player also uses first-person/bumper view.
	Test Drive Unlimited (TDU)	+	+	+	W	-	-	P	1st/Bp	TDU is a street-racing game with simulated traffic. Player controls a powerful exotic car and do various task such as racing with other cars, deliver goods/cars to a certain location, drive girls to a destination under specific time, etc. In this stimuli, player controls a ferrari and supposed to drive it to a certain location with a gps or arrow in the middle of the screen as navigation. Player is expected to maintain the car under certain speed to avoid destroying/crashing the car.
	Driver Test (DT)	+	+	+	W	-	-	P	1st/Bp	DT is a driving simulator similar to 3DDS. Player must follow certain routes and traffic rules. This simulator has more realistic rendering than 3DDS, however the instructions in this one is more difficult to follow. It only appear when the player is close to an intersection, making less time to react. This simulator also have a roundabout scenario, which is non-existant from 3DDS stimuli. This stimuli also uses dashboard view.
	Need for Speed Most Wanted (NFSMW)	+	+	+	W	-	-	P	1st/Bp	NFSU is a street-racing game with fancy and powerful cars to drive. This version of NFS series takes player into the countryside race. Player must cross the finish line before other racers to win a level. In this stimuli, player faces with only 1 computer player and no traffic. Player uses first-person/bumper view in this stimuli.

Play: J = joystick, **M** = mouse, **W** = Wheel and pedal ; **Meduim: G** = Nintendo gamecube, **P** = PC ;
Viewpoint: 3rd = 3rd person view, 1st = 1st person view, 1st/Bp = 1st with bumper view, 1st/Db = first-person with dashboard view

Fig. 2. Summary statistics of our video games classified in three categories: 1) Hardcore games demanding superb focus and have near natural visual renderings, 2) Casual games mostly consisting 2D video games with simple policies, and 3) Driving games including in-city and free-way driving routes. We intend to share our data for follow-up research on modeling top-down visual attention. Some games are simple 2D cartoon games.

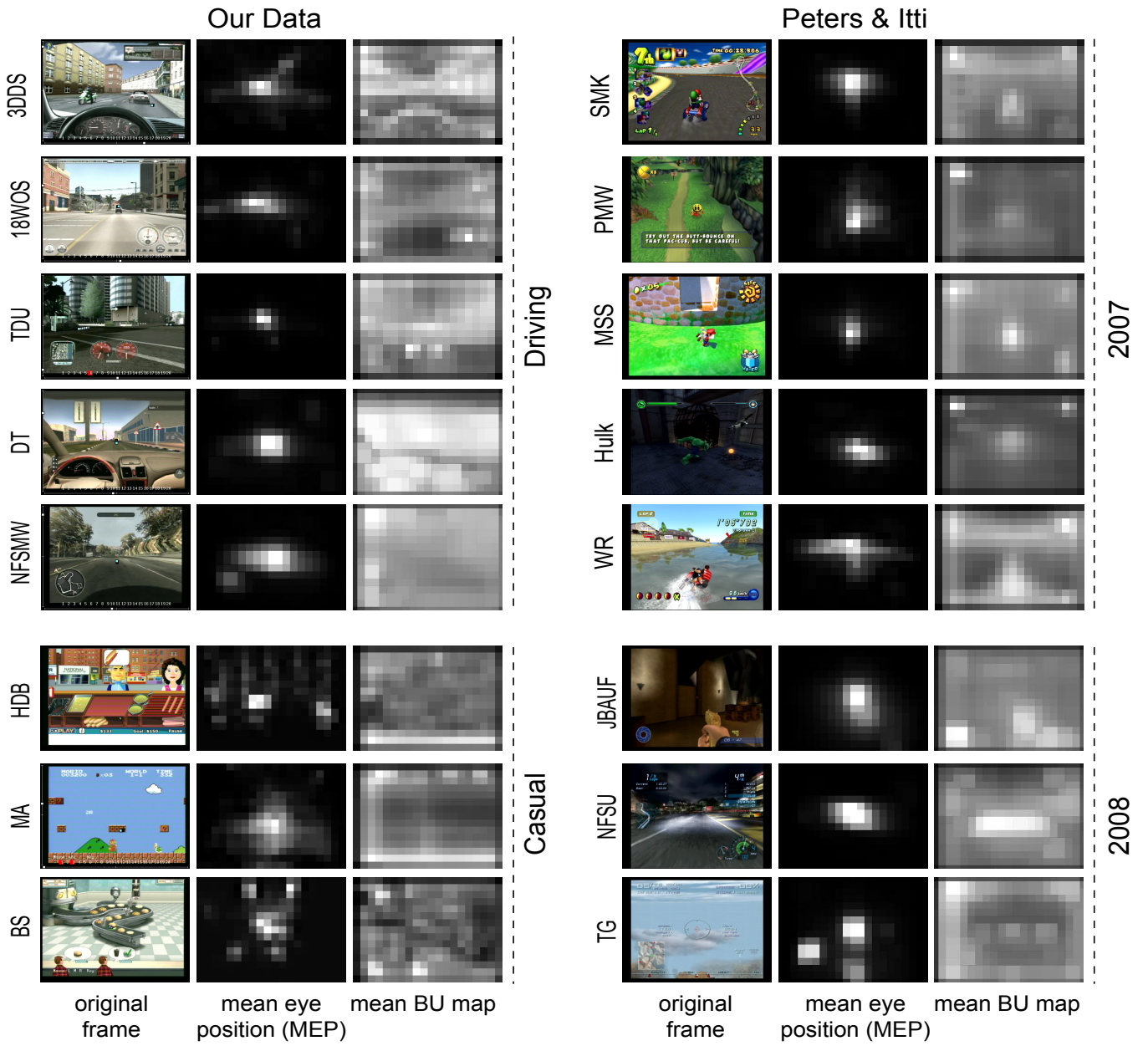

Fig. 3. Sample frames of the video games listed in Figure 2. Some of the data (right column) are collected in our lab by Peters and Itti [19][21]. Middle columns show the mean eye position of all subjects and the third columns show the mean bottom-up saliency map derived from Itti *et al.* [20] model showing the average bottom-up salient regions through the whole time course of a game.

finally contribute to a unique saliency map representing the conspicuity of each location in the visual field.

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 (rear, left), left and right side views for panning the current forward view to the left or right window (mainly its purpose is to do traffic check), and gear change}. Signals, mirrors and views are thus binary variables. 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, we don't have physical actions for the TG game.

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

Object features (F). This is a N dimensional vector of properties of objects ($F^{1:N}$) (e.g., here N=15 is the number

of objects and hence the cardinality of this vector) as will be further explained in section III-C.

B. Space-based Classifier Models

We first explain our models that learn a mapping from features to attended objects, eye fixations (or saccades) for each task. We developed several classifiers as well as brute-force control algorithms with the same input representations. The advantage of our space-based and classifier-based models is that they are easily applicable to every visual task and there is no need for object tagging. According to the Bayes theorem, these classifiers estimate:

$$P(R|M) = \frac{P(M|R)P(R)}{P(M)} \tag{1}$$

with R being either X or Y; and M being either the feature-based representation 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 the attended object from E, $P(Y|E)$) or using the marginal likelihood:

$$P(X|M) \propto P(X|Y) \times P(Y|M) \tag{2}$$

(i.e., a classifier first predicts attended object from features and then a second classifier maps the predicted attended object to the gaze position). The following linear and non-linear classifiers were developed:

Mean eye position (MEP). This family of predictors ignores feature vectors and simply uses the prior distribution over all fixations, saccades, or attended objects over entire training set. It is formally defined as:

$$MEP = \frac{1}{T}\sum_{j=1}^{T} R^j \tag{3}$$

where R is the spatial location (saccade or fixation) or attended object and T is the number of frames over the course of the game in the training phase for which a location or object was attended. 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 video frame) with probability $1/N$ where N is the number of remaining objects in the scene. For fixation 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 [44]), therefore a central Gaussian blob can score better than almost all saliency models when datasets are centrally biased. We thus also compare our results with this heuristic model, which is simply a Gaussian blob ($\sigma = 3$ pixels) at the image center.

Linear Regression (REG). This model does not take into account the temporal progress of a task and simply maps Gist of the scene to the eye position. Mathematically, the goal is to optimize the following objective function:

$$\begin{aligned} &\arg\min_{W} ||M \times W - X||^2 \\ &Subject\,to: W \geq 0. \end{aligned} \tag{4}$$

where M indicates the matrix of feature vectors (only Gist feature is used in [19]) and X is the matrix of eye positions (one fixation per frame or saccades for those frames that a saccade occurred). The least-squares solution of the above objective function is: $W = M^+ \times X$, where M^+ is the pseudo-inverse of matrix M through SVD decomposition. In our experiments, we only take the largest singular value of the SVD since this avoids numerical instability and results in higher accuracy. Given vector $E = (u, v)$ as the eye position over a 20×15 map (i.e., $w = 20, h = 15$) with $u \in [1, 20]$ and $v \in [1, 15]$, the gaze density map can then be represented by vector $X = [x_1, x_2, \ldots, x_{300}]$ with $x_i = 1$ for $i = u + (v-1) \times 20$ and $x_i = 0$ otherwise. Finally, for each test frame, we compute feature vector F and generate the predicted map $P = F \times W$ which is then reshaped to a 20×15 saliency map. The maximum of this map is used to direct spatial attention.

TABLE I
ALGORITHM 1: kNN ALGORITHM FOR GAZE PREDICTION

1: $Q_{w\times l} = U \times M^{'}$; (dot product similarity)
2: $Z_{w\times 300} = 0$ (initialization to zero)
3: **for** $j = 1 \cdots K$; (K numbers of neighbors)
4: $L_j = \arg\max_j Q$ (index of maximum column)
5: $Z = Z + X(L_j)$; updating the frequency of gaze at a location
6: $Q(L_j)$ = - *inf* ; to inhibit selection of maximum values in the next iteration
7: **end**
8: $Z = Z * LG$; convolution with ($LG_{1\times 300}$; a linearized Gaussian) to make the final map

kNN. Here, attention map for a test frame is constructed from the distribution of fixations of its most similar frames in the training set. Since kNN is usually slow, we developed a fast matrix implementation of kNN as shown in Table 1. Matrix Q denotes dot product similarity of all test frames U (for a subject) to all training frames M (other subjects). Number of neighbors (K) was 40, 5, and 5 for HDB, 3DDS, and TG, respectively. Parameter K was set to give high accuracy in a trial and error basis. Note that by performing matrix operations in this fashion, computational complexity of our kNN is less than when iterating through all test frames.

SVM. To ensure that SVM training did not overwhelm available computational resources, we first reduced the high-dimensional feature vectors (E) using PCA by preserving 95% of variance. Then a polynomial kernel multi-class SVM classifier was trained with p ($|Y| = 15$ objects or $|X| = 300$ eye positions) output classes. We used Libsvm [73], a publicly available Matlab version of SVM.

(a) Data Statistics

Game	Saccade	Fixations	Subjects	Total Duration	Train Time	Test Time	Types of Objects	Objects at screen	Objects / frame	Total frames	Actions	Events	Data size
3DDS	6,382	180,000	10	10 x 10 mins	5mins	10 mins	-	-	-	180,000	22D joystick	9	110 GB
TG	4,602	98,806	6 x 2 recordings	54.9 mins	until can play	until finish level	-	-	-	98,806	-	-	26 GB
HD	1,569 / 1,975	35,177 / 45,000	5	5 x 5 mins	finish 1st level or 5 mins	5 mins	15	26 max, 5 min	10.23	45,000	5D mouse	-	26.5 GB

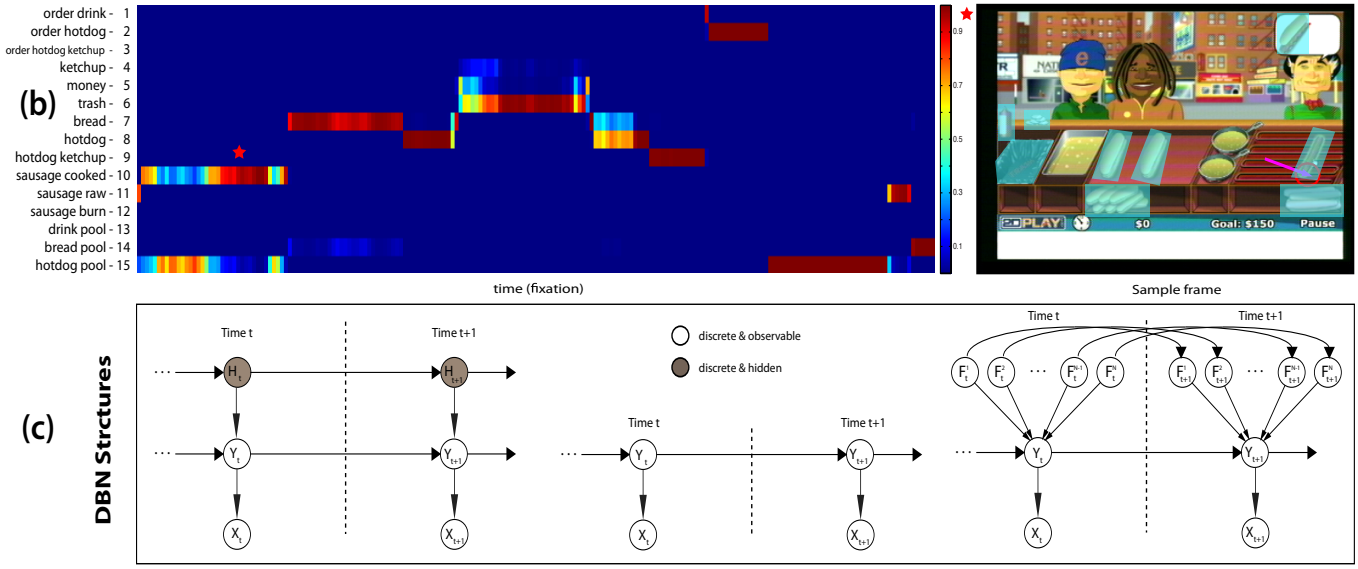

Fig. 4. (a) Summary statistics of three video games, (b) A time series plot of probability of objects being attended and a sample frame with tagged objects and eye fixation overlaid (some objects are available). Subjects could select an object by clicking. Jumps in fixation that pass a certain criteria are considered as saccades, and (c) Two time-slice representation of three DBNs used in the paper.

C. Object-based Bayesian Models

In order to be able to accurately predict which object or spatial location should be attended, a detailed understanding of a scene at the level of objects and their interactions is necessary (as opposed to models based on only global scene context). This object-based representation should be updated over time and the effect of physical actions on them should also be taken into account. We therefore argue that for learning a task, it is enough to learn which objects should be manipulated by which objects over time. By learning the task when can then learn task-driven visual attention. For that, we need an object-level representation of the scene which could be provided either by object annotations (segmenting objects by bounding boxes) of humans or automatic detection of objects using object detection approaches. We then compare performance of these models with space-based models mentioned in Sec. III-B.

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 saccade/fixation position X, i.e., $z(o^j) = 1/e^{\alpha d(X, C(o^j))}$ where $C(o^j)$ is the center of the object 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: $P(o^j) = z(o^j)/\sum_{i=1}^{N} z(o^i)$ where N is the total number of objects. Fig. 4.b 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 a selected object.

We followed 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 average over all n cross-validation runs. Object-based attention model is developed over HDB and classifier-based models are for all games.

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 (e.g., 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 selection as another object variable affecting the attended object. It is also possible to read out the next selected object (action in general) from DBN 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 shown in Fig. 4.c:

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.
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 HMM and third DBN, to be learned, reduces to:

$$\begin{aligned} &P(H_{1:T}, X_{1:T}, Y_{1:T}) = P(X_{1:T}, Y_{1:T}|H_{1:T})P(H_{1:T}) \\ &= P(X_{1:T}|Y_{1:T})P(Y_{1:T}|H_{1:T})P(H_{1:T}) \\ &= P(H_1)P(Y_1|H_1)P(X_1|Y_1) \times \prod_{t=2}^{T} P(H_t|H_{t-1}) \times \\ &\prod_{t=2}^{T} P(Y_t|H_t, Y_{t-1}) \times \prod_{t=2}^{T} P(X_t|Y_t) \end{aligned} \tag{5}$$

$$\begin{aligned} &P(X_{1:T}, Y_{1:T}, F_{1:T}^{1:N}) = P(X_{1:T}, Y_{1:T}|F_{1:T}^{1:N})P(F_{1:T}^{1:N}) \\ &= P(X_{1:T}|Y_{1:T})P(Y_{1:T}|F_{1:T}^{1:N})P(F_{1:T}^{1:N}) \\ &= \prod_{j=1}^{N} P(F_1^j)P(Y_1|F_1^j)P(X_1|Y_1) \\ &\times \prod_{t=2}^{T}\prod_{j=1}^{N} P(Y_t|F_t^j)P(Y_t|Y_{t-1}) \times \prod_{t=2}^{T} P(X_t|Y_t) \end{aligned} \tag{6}$$

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 (Fig. 4.c) that could generalize over many tasks. Therefore, the joint pdf in Eq.5 reduces to predicting next attended object thanks to independence assumptions. As an example we derive the formulation for the third case in Fig. 4.c:

$$\begin{aligned} &P(Y_{t+1}|F_{1:t+1}^{1:N}, Y_{1:t}, X_{1:t}) \ \%\ given\ all\ past\ info. \\ &= P(Y_{t+1}|F_{1:t+1}^{1:N}, Y_{1:t}) \ \% \quad Y_{t+1} \perp\!\!\!\perp X_{1:t} \\ &= P(Y_{t+1}|F_{t+1}^{1:N}, Y_t) \ \% \quad Y_{t+1} \perp\!\!\!\perp Y_{1:t-1} \\ &= \left(\Pi_{j=1}^{N} P(Y_{t+1}|F_{t+1}^j)\right) \times P(Y_{t+1}|Y_t) \\ &\%\quad F_{t+1}^i \perp\!\!\!\perp F_{t+1}^j,\ \forall\ i \neq j \end{aligned} \tag{7}$$

$P(Y)$ is initialized uniformly over the objects (time 0 and is equal to $P(o^j), j = 1 : 15$) and is updated over time. The HMM model (case 2) has one hidden variable 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. We used the Bayes Net Toolkit (BNT) [72] for learning parameters of DBN.

Since variables in our DBN take discrete values, while we have a pdf over the attended object Y_t, 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 a_t be the feature vector for the frame at time t, a tuple $< a_t, y_t, x_t >$ 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 (section 3.2) for a fair comparison with DBNs. Since DBN have access to the previous time information, a sample $< [a_t \quad y_{t-1}], y_t, x_t >$ is added to classifiers, where y_{t-1} and y_t are sampled from $J(Y_{t-1})$ and $J(Y_t)$, respectively (no y_{t-1} in memory-less mode).

Naive Bayes (NB). In the memory-less 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:

$$P(Y_{t+1}|F_{t+1}^{1:N}) = \frac{1}{Z}\prod_{i=1}^{N} P(F_{t+1}^i|Y_{t+1}) \tag{8}$$

where Z is a normalization constant. With no object information, this classifier reduces to priors $P(Y)$ and $P(X)$ which are equal to MEP. As in our DBN framework, here we also used validation strategy to avoid overfitting while training.

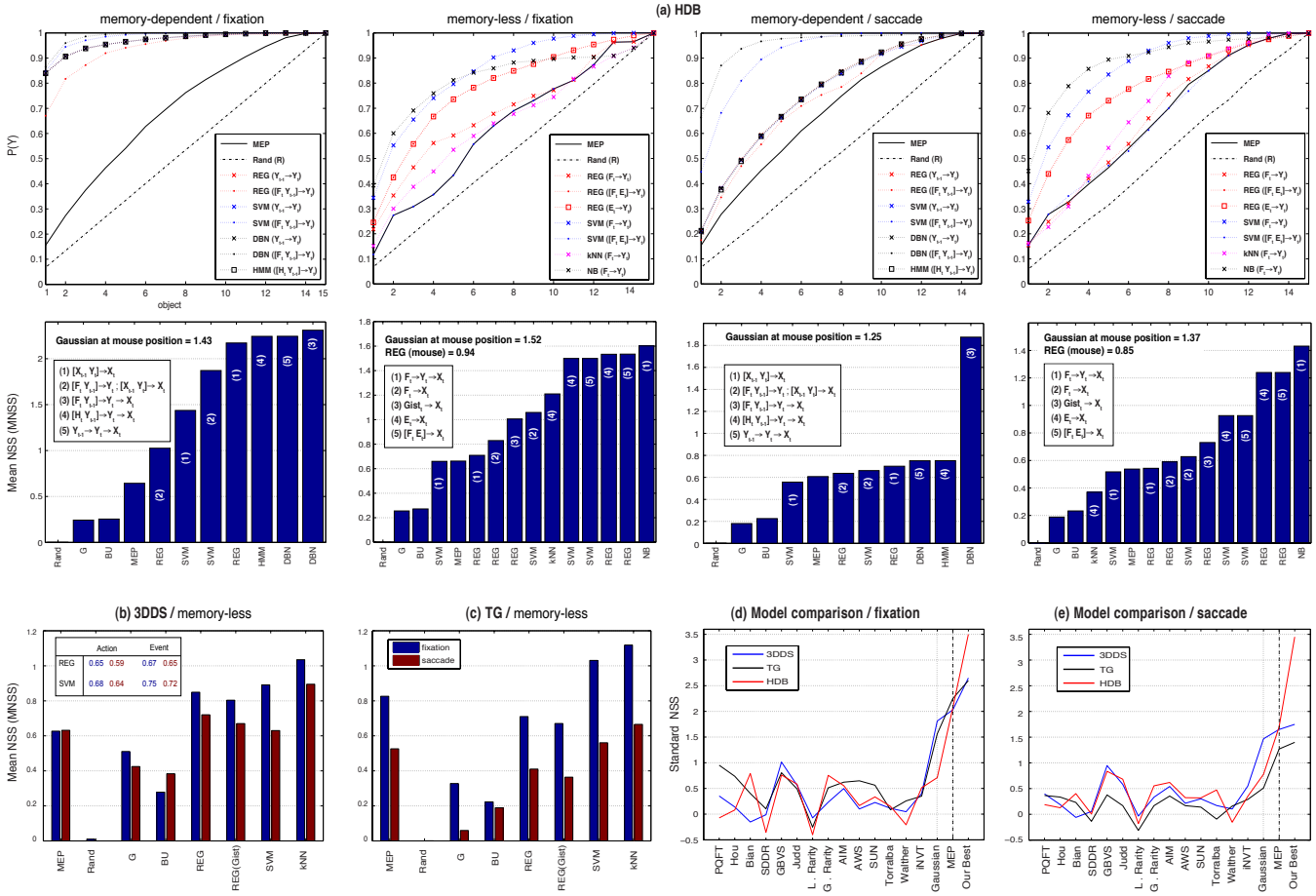

Fig. 5. Gaze prediction accuracies. a) probability of correctly attended object (first row) and MNSS scores for prediction of saccades and fixation positions (second row) for all models. White legends on bars show the mapping from feature types to gaze position X. For instance, REG ($F_t \rightarrow Y_t \rightarrow X_t$) 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 the scene or not (binary). b) and c) MNSS scores of our classifiers over 3DDS and TG games, d) and e) NSS scores (corresponding to $\gamma = 0$ in MNSS) of bottom-up models for saccade prediction over 3 games. Almost all of bottom-up models perform lower than MEP and Gaussian, while our models perform higher. Some models are worse than random (NSS < 0) since saccades are top-down driven instead of bottom-up.

IV. Experimental Results

Evaluation Metrics. Two scores were used to evaluate the accuracy of proposed models explained below:

1) *Normalized Scan-path Saliency (NSS)* [42]. NSS 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:

$$NSS = \frac{1}{\sigma_s}(s(x_h, y_h) - \mu_s) \tag{9}$$

$NSS = 1$ indicates that the subject's eye position fall in a region where predicted density is one standard deviation above average while $NSS = 0$ means that a model performs at chance. Due to high subject agreement (peaks in MEP; Fig. 13), Gaussian (when peak is in the center) and MEP 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 fixations 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" fixations, i.e., away from the central peak of MEP data. To summarize these scores, we defined Mean NSS (MNSS):

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

Along with MNSS, we also report the results using the original NSS score.

2) *Receiver Operating Characteristic (ROC)* [43]. ROC is a method used for evaluation of a binary classifier system with a variable threshold (Usually between two methods like saliency vs. random). Using this metric, the model's (or so-called "estimated") saliency map (ESM) is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels are classified as non-fixated. Human fixations are used as ground truth. By varying the threshold, the ROC curve is drawn as

the *false positive rate* vs. *true positive rate*, and the area under this curve indicates how well the saliency map predicts actual human eye fixations. Perfect prediction correspond to a score of 1. This metric has the desired characteristic of transformation invariance, in that area under the ROC curve does not change when applying any monotonically increasing function to the saliency measure.

To evaluate object-based models, for a frame, a hit is counted when the ground-truth attended object ($= \arg\max_j P(Y^{'})$) is in top maximums (accumulative i.e., $1:2, 1:3, \cdots, 1:15$) of the predicted object ($= \arg\max_{j=1\cdots15} P(Y)$). Hits are then averaged over all gazes for each j.

Gaze prediction. Fig. 5 shows prediction accuracies of models in all conditions (memory-dependent/memory-less, object/fixation/saccade, HDB/3DDS/TG). Bayesian models performed the best in predicting the attended object followed by SVM. All models performed significantly higher than random, MEP and Gaussian and a simple classifier from gist to eye position [19] using MNSS (same is true over NSS score (Figs. 5.d and 5.e), KL, and ROC scores; Fig. 6). Performances are higher in memory-dependent cases as we expected which shows that information from previous step is helpful. DBN model in the memory-dependent mode and Naive Bayes (NB) in the memory-less mode, scored the best MNSS over fixations and saccades (followed by HMM in memory-dependent and REG in memory-less 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/fixation case but not as good in the memory-less 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.

Over 3DDS and TG (Figs. 5.b and 5.c), with combination of all features, kNN achieved the best MNSS followed by SVM and Regression. Also, classifiers with event and action features performed higher than MEP and Gaussian.

Fig. 6 shows the ROC curves as well as the NSS scores for 3DDS and TG games in memory-less case and HDB game in memory-dependent case for both fixation and saccade location prediction. As it shows over 3DDS, SVM and kNN (using all features) score the highest using both AUC (area under ROC) and NSS sore for fixation prediction. Over saccades, kNN scores the highest. Over both saccades and fixations, kNN and SVM score higher than MEP. The same trend happens over the TG game with the exception that over saccades SVM scores higher than kNN.

For HDB game in memory-dependent case, there is large gap between prediction of our models and the MEP model using both AUC and NSS scores. The DBN approaches score

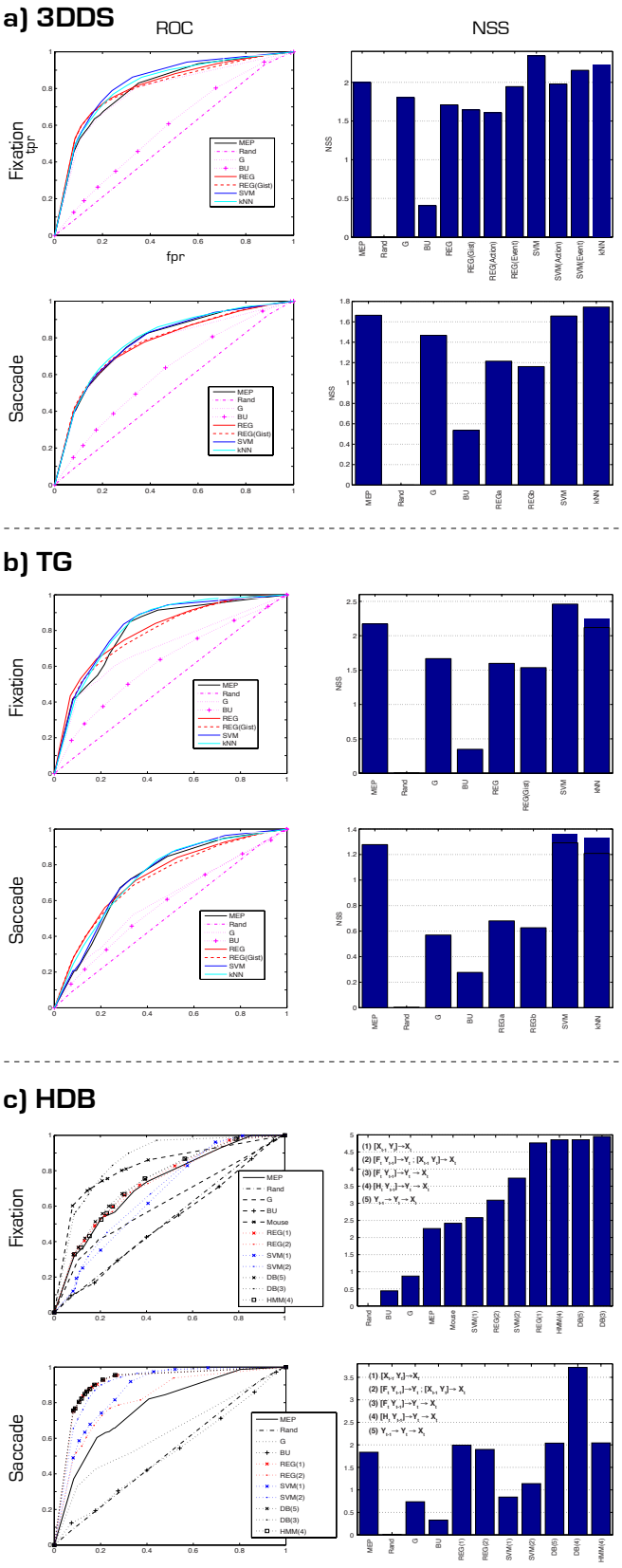

Fig. 6. ROC curve and NSS scores for fixation and saccade prediction, a) 3DDS, b) TG games in memory-less case and c) HDB game in memory-dependent case. In all cases, consistent with Fig. 5, our models score higher than MEP model. DBN models score the highest among classifiers and control models.

higher than all models for both fixation and saccade prediction.

Model comparison. We ran 14 state-of-the-art bottom-up saliency models[5] to compare saccade/fixation prediction accuracy over three games (cf. Figs. 5.d and 5.e). 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). Fig. 7 shows the prediction accuracy of Itti's bottom-up model [20] and a central Gaussian blob over all of our video games. As it shows a simple central Gaussian blob outperforms this model over all games using AUC score. Please note that these models do not use the motion channel which explains why some of them perform lower than chance. In [69], authors provide a comparative study of the state of the art bottom-up saliency models.

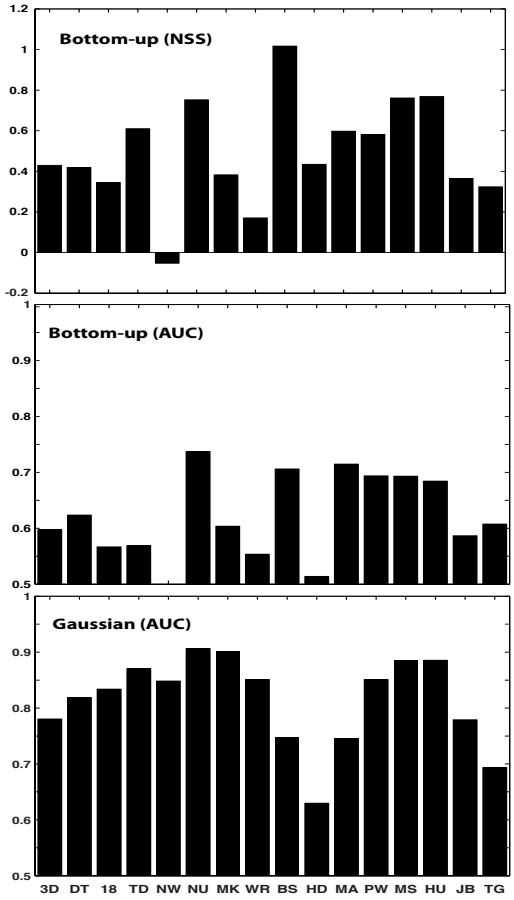

Fig. 7. Prediction accuracy (NSS and AUC scores) of bottom-up saliency model [20] and a central Gaussian blob for fixation prediction over over 16 video games.

[5]Compared bottom-up saliency models over our data include: iNVT [20], AIM [43], Hou *et al.* [55], Local and Global Rarity [53], PQFT [51], AWS [49], GBVS [58], Bian *et al.* [50], SDDR [48], Judd *et al.* [12], Torralba *et al.* [82], Walther *et al.* [59], and SUN [56].

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[6] and 2) the deformable part model (DPM)[7] [60] to automatically fill the variables instead of annotated data. These models were trained over a small set of cross validation data different from test frames. Fig. 8 shows a sample frame from the HDB game along with *hotdog pool* object template using DPM. Precision-recall curves of each of the 15 objects for both models are shown in Fig. 9. As opposed to Boosting model, DPM was very successful for detecting objects. Detection performance was very high for each object due to limited variation in object appearance. Therefore we used DPM for subsequent analysis.

We tested models again with variables filled with these data (instead of annotations). Accuracy of attended object prediction are shown in Fig. 10. As we expected, there is a graceful degradation in prediction of the attended object (in comparison with Fig. 5) but still performance of our DBN was higher than other models indicating partial robustness of our model (similar trend with MNSS score).

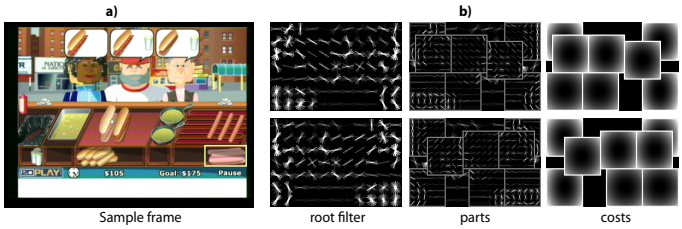

Fig. 8. A sample frame from Hot Dog Bush game (a) along with root and parts responses of the DPM model (b).

Analysis of Generalization Across Tasks. Here, we analyze the capability of the space-based classifier models for fixation prediction over other games. We trained a model from data of one game/task and applied it to another task. Fig. 11 shows the results. As is shows testing a model trained from a game on a different game results in a prediction accuracy still better than chance (AUC $>$ 0.5) and (NSS $>$ 0). This is partially because our models generate maps with high activation at the center thus applying this map to a new game has higher chance to predict fixations. Interestingly, this analysis showed higher prediction accuracy for similar games than dissimilar ones. For instance, training on one of the driving games (3D, DT, 18Wos, TD, ...) has better test accuracy over other driving games (See sub-clusters in Fig. 11, MEP and kNN models).

Analysis of Number of Subjects in Learning. Here, we investigate that how much adding new subjects can actually help learning. We started training on n subjects and tested over the remaining $P-n$ subjects (P is the whole number of subjects for a game). Fig. 12 shows the results for increasing n. As number of subjects are increased, fixation prediction accuracy also increases for our models over each of four games. This indicates that having more data can lead to better

[6]http://people.csail.mit.edu/torralba/shortCourseRLOC/boosting/boosting
[7]http://www.cs.brown.edu/∼pff/latent/

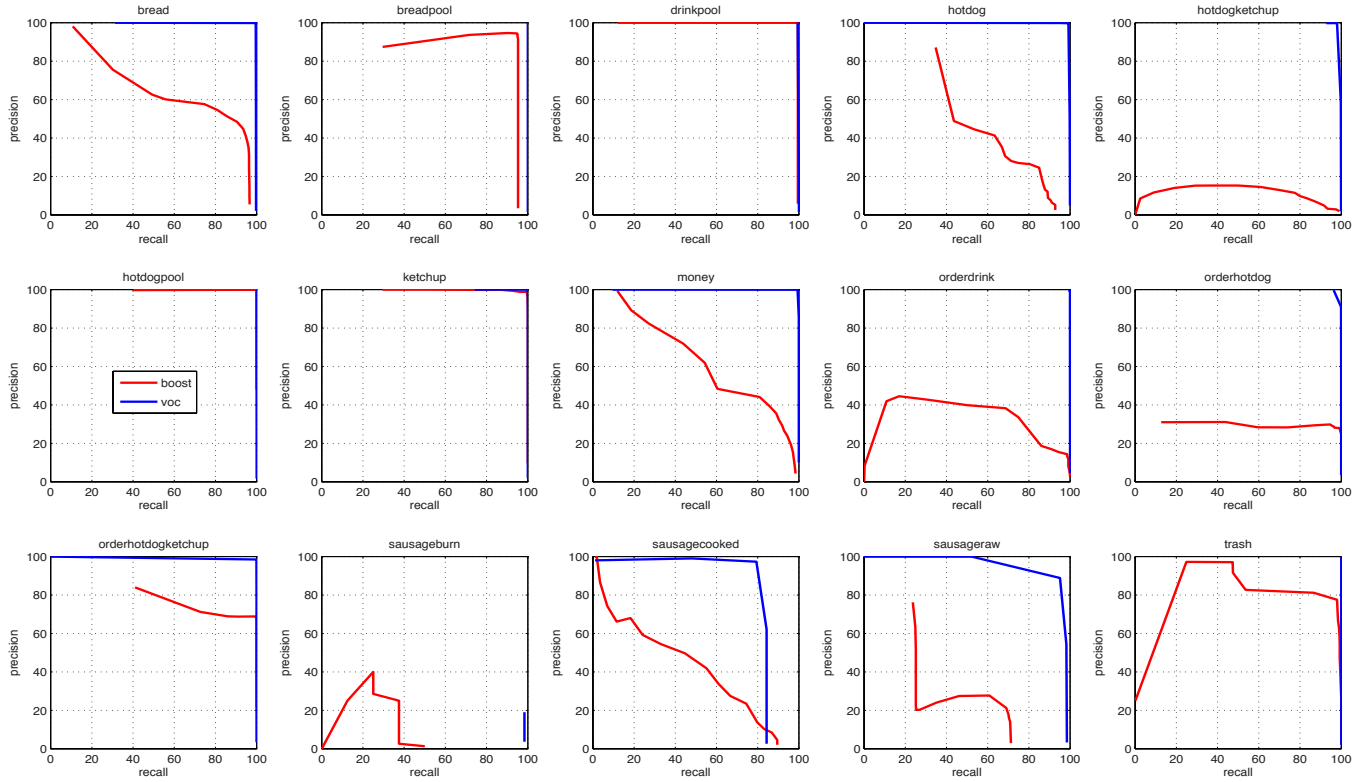

Fig. 9. Precision-recall curves of object detection using DPM [60] and boosting models (blue and red curves, respectively) for each object and frame of the HDB game. While booting model fails to detect many objects, DPM model works very well over almost all objects of the HDB game. For results of fixation and attended object prediction please refer to Fig. 10.

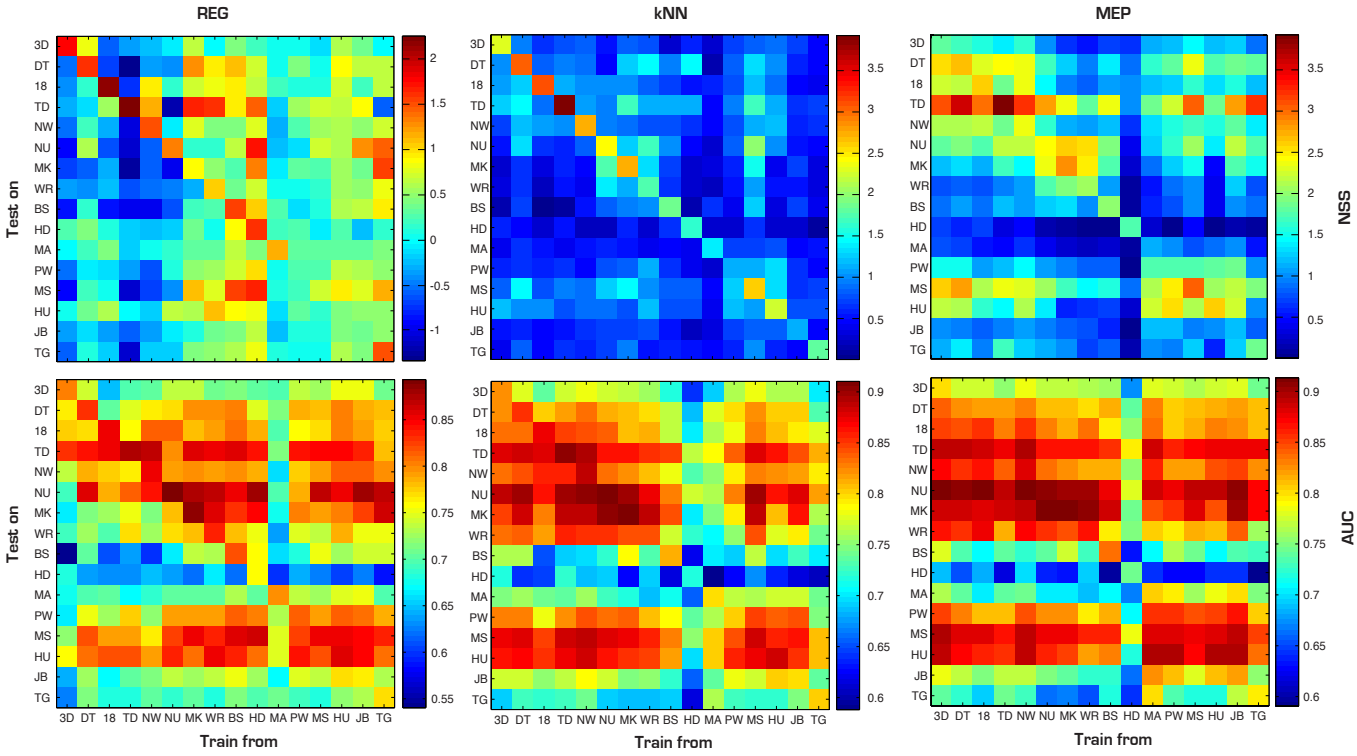

Fig. 11. Confusion matrices of applying a model trained over one game to other games. Three models were considered: Regression, kNN, and MEP. X-axis shows the trained games and Y-axis shows the tested ones. Top-row is the normalized scanpath saliency (NSS) score and the bottom row shows Area Under ROC Curve (AUC) score. Training over similar games leads to higher test performance.

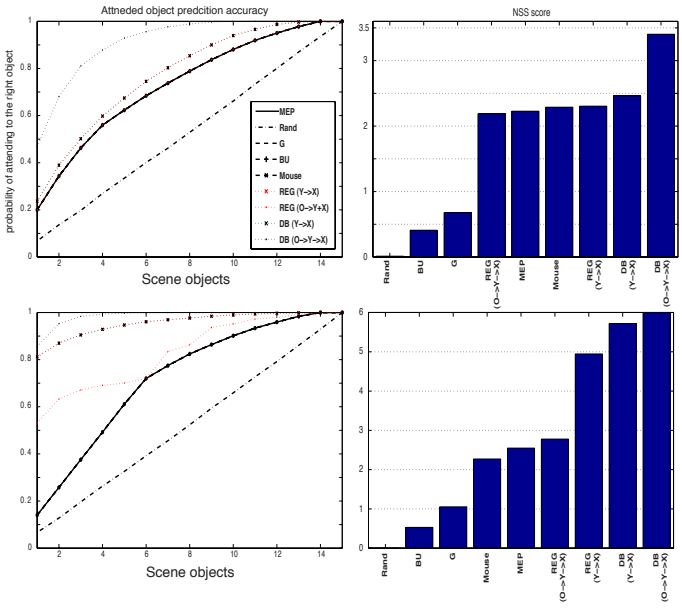

Fig. 10. Prediction of our Bayesian model in presence of noise (i.e., object detection results in left column and saccade prediction at right) over the HDB game. Top) for the memory-dependent saccade prediction, Bottom) for the memory-dependent fixation prediction cases.

results since subjects have different adventures in games and more data can capture the task demands better.

Fig. 13 shows sample predicted maps by our models for three games.

V. Discussions and Conclusions

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 where objects are processed sequentially in a spatio-temporal manner.

Using DPM model [60], we were able to automatically detect objects in HDB game with high detection accuracy, 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., relaxing conditional independence assumptions). One problem we experienced was learning the structure of DBN since to date, structure learning algorithms are limited to certain network structures and variable types.

Despite 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. Measuring and modeling covert attention in the context of top-down attention is a challenging topic for future research.

A more biologically plausible future extension would be using foveated representation of the scene similar to [74] where object features in the periphery are accessible with less confidence. Also, analysis of knowledge transfer would be a rewarding work. For instance, by training classifier-based models over a game and applying them over similar games,

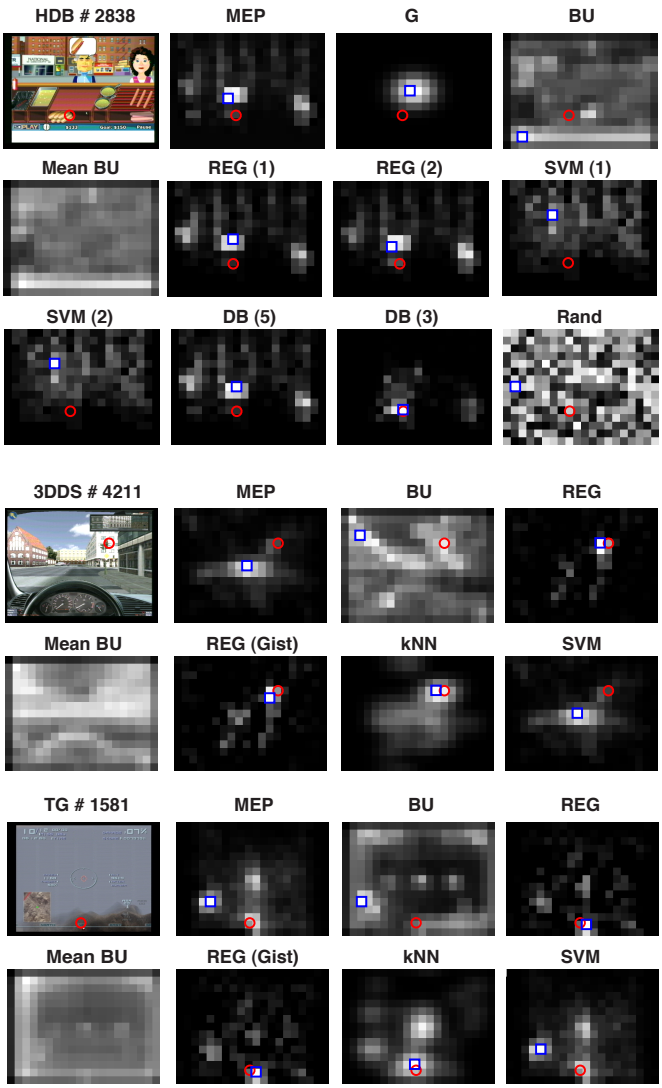

Fig. 13. Sample frames of video games and corresponding predicted maps of models. Red circle indicates the human fixation and blue square is the maximum point of each map. Smaller distance hence means better prediction. Currently we don't have action data for TG game.

we found that they scored better than chance implying that gist and action features to some extent capture the semantics directing gaze.

Here we compared several algorithms for modeling top-down attention over different tasks on very different stimuli. It would be also interesting to compare algorithms performing different tasks on the same visual stimulus close to the often-cited Yarbus experiment (similar to [90]).

We aim to build a top-down evaluation open challenge by sharing our datasets (some used here) and evaluation programs to make a fair comparison of models and raising interest in this field similar to PASCAL VOC challenge in object recognition literature.

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.

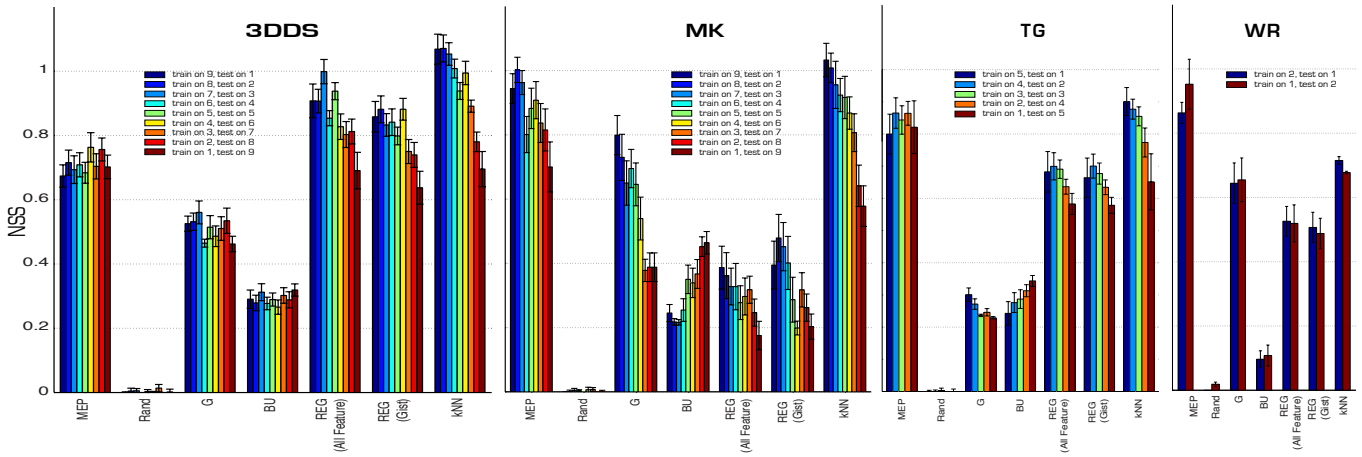

Fig. 12. Analyzing the effect of the number of training subjects on model accuracies over testing subjects using NSS score. Each bar is the mean over all possible selection of n training subjects (similarly for the remaining test subjects). The trend is the same for all compared models (except random) including: MEP, Gaussian, BU [20], Reg, and kNN. Results are for the memory-less fixation prediction case.

REFERENCES

[1] A. Borji. D.N. Sihite, and L. Itti. Computational Modeling of Top-down Visual Attention in Interactive Environments. *British Machine Vision Conference (BMVC)*, 2011.

[2] B.J. Scholl. Objects and attention: the state of the art. *Cognition*, 80(1/2):1-46, 2001.

[3] M.I. Posner. Orienting of attention. *Q. J. Exp. Psych*, 32, 1980.

[4] A.M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psych.*, 12:97-136, 1980.

[5] L. Itti and C. Koch. Computational modeling of visual attention. *Nat. Rev. Neurosci.*, 2(3):194-203, 2001.

[6] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.*, 5:1-7, 2004.

[7] A. Yarbus. Eye movements during perception of complex objects. L. Riggs, editor, Eye Movements and Vision, 1967.

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

[9] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cog. Sci.*, 9(4), 188-193. 2005.

[10] C. Rothkopf, D. Ballard, and, M. Hayhoe. Task and scene context determines where you look. *Journal of Vision*, 7(14):16, 1-20, 2007.

[11] W., Kienzle, A. F., Wichmann, B., Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. *NIPS*, 2007.

[12] T. Judd, K. Ehinger, F. Durand and, A. Torralba. Learning to predict where humans look. *ICCV*, 2009.

[13] E. Vig, M. Dorr, T. Martinetz, and, E. Barth. A learned saliency predictor for dynamic natural scenes. *ICANN*, LNCS, (6354): 52-61, 2010.

[14] J. Li, Y. Tian, T. Huang and W. Gao, Probabilistic multi-task learning for visual saliency estimation in video. *Int. Journal of Computer Vision*, (90)2:150-165, 2010.

[15] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino. A stochastic model of selective visual attention with a dynamic Bayesian network. ICME, 2008.

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

[17] D. Ballard, M. Hayhoe, and J. Pelz. Memory representations in natural tasks. *J. of Cog. Neurosci.*, 7(1), 66-80, 1995.

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

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

[20] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions PAMI*, 20(11):1254-1259, 1998.

[21] R.J. Peters and L. Itti. Congruence between model and human attention reveals unique signatures of critical visual events. *NIPS*, 2008

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

[23] N. Sprague, D. H. Ballard, Eye Movements for Reward Maximization. *NIPS*, 2003.

[24] N. Sprague, D. Ballard, and A. Robinson. Modeling Embodied Visual Behaviors. *ACM Transactions on Applied Perception*, 4(2), 2007.

[25] H. Hecht, S. Vogt and W. Prinz, Motor learning enhances perceptual judgment: A case for action-perception transfer. *Psychological Research*, 65:3-14, 2001.

[26] S. Schtz-Bosbach and W. Prinz, Perceptual resonance: Action-induced modulation of perception. *Trends in Cog. Sci.*, 11:349-355, 2007.

[27] G. Rizzolatti, L. Riggio, I. Dascola and, C. Umilt. Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25: 31-46, 1987.

[28] L. Craighero, L. Fadiga , G. Rizzolatti, and C. Umilt. Action for perception: a motor-visual attentional effect. *J Exp Psychol Hum Percept Perform*, 25(6):1673-92, 1999.

[29] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions PAMI*, 29(2):300-312, 2007.

[30] N. Pugeault and R. Bowden. Learning pre-attentive driving behavior from holistic visual features. *ECCV*, 2010.

[31] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes: a combined source model of eye guidance. *Visual Cognition*, 17:945-978, 2009.

[32] S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: a Bayesian inference theory of visual attention. *Vision Research*, 2010.

[33] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395-1407, 2006.

[34] S. Frintrop. VOCUS: A visual attention system for object detection and goal-directed search. PhD Thesis. Springer 2006.

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

[36] L. Itti. Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention. *IEEE Trans. Image Process.*, vol. 13, no. 10, 2004.

[37] M.S. El-Nasr, T. Vasilakos, C. Rao, and J. Zupko. Dynamic Intelligent Lighting for Directing Visual Attention in Interactive 3D Scenes. *IEEE Trans. on Comp. Intell. and AI in Games*, 2009.

[38] N. Parikh, L. Itti, and J. Weiland. Saliency-based Image Processing for Retinal Prostheses. *Journal of Neural Engineering*, 7(1), 2010.

[39] C. Siagian and L. Itti. Biologically Inspired Mobile Robot Vision Localization. *IEEE Transactions on Robotics*, 25(4):861-873, 2009.

[40] Y. Ma, X. Hua, L. Lu, and H. Zhang. A Generic Framework of User Attention Model and Its Application in Video Summarization. *IEEE transactions on multimedia*, 7(5), 2005.

[41] B., Mertsching, M., Bollmann, R., Hoischen, and S. Schmalz. The Neural Active Vision System. *In Handbook of Computer Vision and Applications, B. Jahne, H. Haussecke, and P. Geissler, Eds. vol. 3. Academic Press*, pp. 543568, 1999.

[42] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Res.*, 45, 2005.

[43] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. *NIPS*, 2005.

[44] B.W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions. *J. of Vision*. 14(7):1-17, 2007.
[45] X. Hou, J. Harel, and C. Koch. Image Signature: Highlighting Sparse Salient Regions. *IEEE Trans. PAMI*, in press.
[46] D. Gao, V. Mahadevan, and N. Vasconcelos. The Discriminant Center-surround Hypothesis for Bottom-up Saliency. *NIPS*, 2007.
[47] O. Le Meur, P. Le Callet and D. Barba. Predicting Visual Fixations on Video Based on Low-level Visual Features. *Vision Research*, 47/19:2483-2498, 2007.
[48] H.J. Seo and P. Milanfar. Static and Space-time Visual Saliency Detection by Self-Resemblance. *Journal of Vision*, 9(12):1-27, 2009.
[49] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Decorrelation and Distinctiveness Provide With Human-Like Saliency. *ACIVS*, 5807, 2009.
[50] P. Bian and L. Zhang. Biological Plausibility of Spectral Domain Approach for Spatiotemporal Visual Saliency. *LNCS*, 5506: 251-258, 2009.
[51] C. Guo and L. Zhang. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE Transactions on Image Processing*, 19(1):185-198, 2010.
[52] X. Hou and L. Zhang. Dynamic Visual Attention: Searching for Coding Length Increments. *NIPS*, 2008.
[53] M. Mancas. Computational Attention: Modelisation and Application to Audio and Image Processing. PhD. thesis, 2007.
[54] L. Itti and P. Baldi. Bayesian Surprise Attracts Human Attention. *NIPS*, 2006.
[55] X. Hou and L. Zhang. Saliency Detection: A Spectral Residual Approach. *CVPR*, 2007.
[56] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian Framework for Saliency Using Natural Statistics. *J. of Vision*, 8/7(32):1-20, 2008.
[57] D. Parkhurst, K. Law, and E. Niebur. Modeling the Role of Salience in The Allocation of Overt Visual Attention. *Vision Res*, 42(1):107-123, 2002.
[58] J. Harel, C. Koch, and P. Perona. Graph-based Visual Saliency. *NIPS*, 19:545-552, 2006.
[59] D. Walther and C. Koch. Modeling Attention to Salient Proto-objects. *Neural Networks*, 19(9):1395-1407, 2006.
[60] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *IEEE Trans. PAMI*, 32(9), 2010.
[61] http://people.csail.mit.edu/torralba/shortCourseRLOC/boosting/boosting
[62] M. Cerf, J. Harel, W. Einhuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *NIPS*, 2007.
[63] E. D. Reichle, K. Rayner, and A. Pollatsek. The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26:445-476, 2003.
[64] T. Erez, J. Tramper, B. Smart, and S. Gielen. A POMDP Model of Eye-Hand Coordination. *AAAI*, 2011.
[65] R.D. Rimey, C.M. Brown. Control of Selective Perception Using Bayes Nets and Decision Theory. International Journal of Computer Vision, 12(2/3):173-207, 1994.
[66] R.C. Cagli, P. Coraggio, P. Napoletano, O. Schwartz, M. Ferraro, and G. Boccignone. Visuomotor characterization of eye movements in a drawing task. *Vision Research*, 49, 2009.
[67] S. Ullman. Visual routines. *Cognition*, 18:97-159, 1984.
[68] W. Yi, and D. H. Ballard. Recognizing Behavior in Hand-Eye Coordination Patterns. *I. J. Humanoid Robotics*, 6(3): 337-359, 2009.
[69] A. Borji, D.N. Sihite, and L. Itti. Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study. *IEEE Trans. Image Processing.*, In Press.
[70] A. Borji and L. Itti. State-of-the-art in Visual Attention Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, In Press.
[71] C. Koch and S. Ullman. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, 4(4): 219-227, 1985.
[72] http://code.google.com/p/bnt/.
[73] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.
[74] J. Najemnik, and W. S. Geisler. Optimal Eye Movement Strategies in Visual Search. *Nature*, 434:387-391, 2005.
[75] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. *BMVC*, 2008.
[76] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001.
[77] M. Fukuchi, N. Tsuchiya, and C. Koch. The focus of expansion in optical flow fields acts as a strong cue for visual attention. *Journal of Vision*, 9(8), 2010.
[78] T. Avraham, M. Lindenbaum. Esaliency (Extended Saliency): Meaningful attention using stochastic image modeling. *PAMI*, 2010.
[79] J.M Henderson. Human gaze control during real-world scene perception. *Trends Cogn. Sci*, 7:498-504, 2003.
[80] W. Kienzle, A.F. Wichmann, B. Scholkopf, and M.O. Franz. A nonparametric approach to bottom-up visual saliency. *NIPS*,2007.
[81] N. J. Butko, and J. R. Movellan. Optimal Scanning for Faster Object Detection. *CVPR*, 2009.
[82] A. Torralba. Modeling Global Scene Factors in Attention. *Journal of Optical Society of America*, 20(7), 2003.
[83] C. Siagian, L. Itti, Comparison of gist models in rapid scene categorization tasks, *In: Proc. Vision Science Society Annual Meeting (VSS08)*, 2008.
[84] Y. Freund and R.E. Schapire. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *J. of Computer and System Sciences*, 1997.
[85] http://cocosci.berkeley.edu/tom/.
[86] http://web.mit.edu/cocosci/josh.html.
[87] J. Wang, C. Xu, and E. Chng. Automatic Sports Video Genre Classification Using Pseudo-2D-HMM, *ICPR*, 2006.
[88] J.H. Lai and S.Y. Chien. Baseball and Tennis Video Annotation with Temporal Structure Decomposition. *MMSP*, 2008.
[89] J. Hays and A.A. Efros. im2gps: estimating geographic information from a single image, *Computer Vision and Pattern Recognition (CVPR)*, 2008.
[90] J. Lim, Y. Liu, Y. Reinforcement Learning in Eye Movements: Modeling the Influences of Cyclic Top-down and Bottom-up Processes, *IEEE Transactions on Systems, Man, and Cybernetics*, 39(4), 706-714, 2009.

Ali Borji received the B.S. and M.S. degrees in computer engineering from Petroleum University of Technology, Tehran, Iran, 2001 and Shiraz University, Shiraz, Iran, 2004, respectively. He did his Ph.D. in cognitive neurosciences at Institute for Studies in Fundamental Sciences (IPM) in Tehran, Iran, 2009. He is currently a postdoctoral scholar at iLab, University of Southern California, Los Angeles, CA. His research interests include: visual attention, visual search, machine learning, robotics, neurosciences and biologically plausible vision.

Dicky N. Sihite received his B.S. degree in Electrical Engineering from Krida Wacana Christian University, Jakarta, Indonesia in 2007. He received M.S. in Electrical Engineering from University of Southern California in 2010 where he is currently persuing an M.S. degree in Computer Science. He is interested in visual attention, machine learning, robotics and computer vision.

Laurent Itti received his M.S. degree in Image Processing from the Ecole Nationale Superiere des Télecommunications in Paris in 1994 and his Ph.D. in Computation and Neural Systems from Caltech in 2000. He is now and associate professor of Computer Science, Psychology and Neurosciences at the University of Southern California. Dr Itti's research interests are in biologically-inspired vision, in particular in the domains of visual attention, gist, saliency, and surprise, with applications to video compression, target detection, and robotics.