**Noname manuscript No.**
(will be inserted by the editor)

# Modeling Influence of Action on Spatial Attention in Interactive Visual Environments

**Ali Borji · Dicky N. Sihite · Laurent Itti**

Received: date / Accepted: date

**Abstract** A large number of studies have been reported on top-down influences of visual attention. However, less progress have been made in understanding and modeling its mechanisms. In this paper, we propose an approach for learning spatial attention taking into account influences of physical actions on top-down attention. For this purpose, we focus on interactive visual environments (video games) which are modest real-world simulations, where a player has to attend to certain dimensions of visual stimuli and perform actions to achieve a goal. The basic idea is to learn a mapping from current mental state of the game player, represented by past actions and observations, to its gaze fixation. A data-driven approach is followed where we train a model from the data of some players and test it over a new subject. In particular, two contributions this paper makes are 1) employing multi-modal information including mean eye position, gist of a scene, physical actions, bottom-up saliency, and tagged events for state representation; 2) analysis of different ways of combining bottom-up and top-down influences. Comparing with other top-down task-driven models and bottom-up spatio-temporal models, our approach shows higher NSS scores in predicting eye positions.

A. Borji · D. N. Sihite · L. Itti
Department of Computer Science, University of Southern California
Hedco Neuroscience Building, 3641 Watt Way, Los Angeles, CA 90089-2520, USA

A. Borji
E-mail: borji@usc.edu

D. N. Sihite
E-mail: sihite@usc.edu

L. Itti
Neuroscience Graduate Program and Department of Psychology, University of Southern California
E-mail: itti@usc.edu

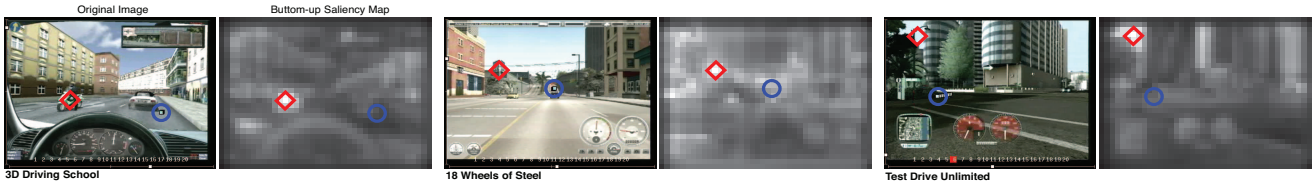


**Fig. 1** Sample frames from our game stimuli and their corresponding BU saliency maps [16]. Red diamond shows the maximum of the saliency map and blue circle is the actual eye position. Left and bottom bars in frames are the pedal and wheel positions, respectively. Actions are represented by numbers at the bottom. Attention is attracted to task-relevant regions which do not agree with the BU saliency.

## 1 Introduction

The concept of saliency has attracted a lot of attention over the past several years. Basically, it is a fast and low-cost way to select important image regions or objects to pass to higher level processes.

The main concern in modeling saliency is how, when, and based on what, to select salient image regions. It is often assumed that attention is attracted by salient stimuli or events in the visual array [1,2]. While this is the case, it is also known that a large portion of attentional behavior comes from ongoing task inferences which dynamically change and are dependent on the algorithm of the task. Understanding task influences on attention is conceptually hard to frame. The biggest challenge comes from the fact that we don't know much about how humans do different complex tasks which seem to be necessary for modeling top-down attentional influences. This has been at the focus of artificial intelligence (AI) and cognitive science research for past 50 years.

However, we know to some extent about algorithms and attentional behaviors of some laboratory-scale stimuli and tasks. One solution when dealing with complex problems is learning from data, experiences or history which could be gathered from the behavior of other humans especially when the goal is to explain human behavior.

There are already many button-up saliency models for static (still images) and spatio-temporal stimuli (videos). However, bottom-up models are inflexible (Figure 1) and can account for only a small fraction of the observed fixations in natural behavior [4–6]. Our goal in this study is to introduce a top-down spatial attention model which could automatically direct gaze based on task. Instead of trying to figure out an explicit algorithm for doing a task, (e.g. designing a state space and mapping its states to actions and attended locations), we are following a data-driven approach which could easily be applied to any task and situation.

### 1.1 Bottom-up (BU) models

The bottom-up saliency assumption is based on the hypothesis that certain features of the visual scene inherently attract gaze. That is, that vision is essentially reactive and stimulus driven. Typically, multiple low-level visual features such as intensity, color, orientation, texture, and motion are extracted from an image at multiple scales. A saliency map is computed for each feature and then they are normalized and combined in a linear or non-linear fashion into a master saliency map that represents the conspicuity of each pixel [16].

Our work in this paper falls in the category of saliency models based on machine learning approaches. Some models train a classifier to distinguish fixated patches from random patches. Facing a scene they assign a value to each patch that is the probability of that patch to be fixated. Kienzle *et al.* [7] learned a model of saliency directly from human eye movement data. Their model consists of a nonlinear mapping from a normalized image patch to a real value, trained to yield positive values on fixated patches, and negative values on randomly selected image patches. Judd *et al.* [8] used a SVM classifier for an attention model based on low-, mid and high-level features calculated by existing saliency methods. In modeling eye saccades of observers when looking for a pedestrian in a scene, Ehinger *et al.* [26] 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 single sources. Vig *et al.* [9] used 3D spatio-temporal volumes from video for spatiotemporal saliency modeling. Li *et al.* [10] proposed a multi tasking Bayesian approach for combining bottom and top-down saliency components. Kimura *et al.* [11] learned a dynamic Bayesian network (DBN) to predict the likelihood where humans typically focus on a video scene. Chikkerur *et al.* [27] presented a Bayesian model based on assumptions that the goal of the visual system is to say what is where and visual processing happens sequentially.

### 1.2 Top-down (TD) models

The other main component of attention comes from top-down demands such as knowledge of the task, emotions, expectations, predictions, etc. which are embedded in a temporally extended task. Modeling top-down attention is hard because 1) it is difficult to frame and conceptually define the problem 2) different tasks require different algorithms and 3) high inter-subject variability. In this paper, we take another step in modeling top-down spatial attention considering multi-modal information including physical actions.

Research on top-down attention dates back to the classic study by Yarbus [3] which showed that gaze patterns are dependent on the asked question when viewing a photo. Research on task-driven influences of gaze have been mostly at the analysis level. It has 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 [12]. 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 his/her eye movements for tasks like "block copying" [13] , "making tea" [4] , "driving" [14], etc.

In [15], Peters and Itti learned a mapping using gist of a scene to eye fixation from data of subjects playing a video game. In [17], using this model they showed that during the occurrence of an event (like hitting a target in shooting games or accident in driving games) bottom-up cues are more important than top-down cues. While this model is interesting, it does not benefit from the real potential of interactive environments which are interactions via physical actions. In our work, we follow a similar fashion by proposing a richer state representation and propose a new approach to combine bottom-up and top-down cues. In a related study, Navalpakkam and Itti [18] tried to build a top-down model in conjunction with the saliency model in situations where the algorithm for the task is at hand. Sprague and Ballard [19], proposed a method based on reinforcement learning for learning visio-motor behaviors and used their model to account for saccades in a side walking task.

### 1.3 Influence of action on attention

The integration between action and perception makes up one of the most important facets of everyday life. Many studies support the idea that perception affects action (e.g. [20]). It has also been proposed that changes due to actions lead to corresponding changes in perception [20,21]. A good example of interaction between actions and attentions is driving which also needs sophisticated attentional behavior. In [23], authors showed that preparation of a grasping movement affects detection and discrimination of visual stimuli. Our work also 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 [22].

### 1.4 Our approach and contributions

We aim to learn top-down spatial attention (where to look) from visual information and physical actions recorded from human subjects playing video games. The basic idea is to best estimate the mental state of the player and map it to an eye fixation. For state estimation, we merged all information including scene gist, physical actions, salient regions, and events. A classifier is learned from this data and used to predict the eye fixations of another subject.

A central open question in saliency modeling is "how the bottom-up salient and top-down task-driven stimuli are integrated in the course of a task"? We

tackled this question by evaluating different ways of integrating BU and TD attention components either in the decision space or at sensory level. Experiments were performed using a driving task which is a daunting task demanding high-level sensory-motor integration and attention/action coordination skills. It has also been subject to several behavioral and computational modeling studies (e.g. [14, 25]).

Our model 1) is easily applicable to interactive visual environments when subjects perform physical (motor) actions and visually attend; 2) has potential applications in interactive computer graphics environments ("virtual reality" or video games), flight and driving simulators (and assistants), as well as visual prosthetic devices.

## 2 Psychophysics and eye tracking

To set a basis and benchmark for future research and large-scale quantitative evaluation of studies on task-driven top-down saliency modeling, we have collected a large scale dataset of videos along with eye tracking data and actions. Accompanying code in C++ and Matlab will be available on the web to facilitate future research.

### 2.1 Data collection

Participants were 10 subjects between 18-25 years old with valid driving license and at least 2 years of driving experience. Experimental protocol was approved by the anonymous university Institutional Review Board. Subjects were compensated for their participation. Each subject played each of the 3 games: 3D Driving School (3DDS), 18 Wheels of Steel (18 WoS), Test Drive Unlimited (TDU) (see Figure 1). There was a 5-min training session for each game in which subjects were introduced to the goal of the game, rules, buttons, etc. After training, subjects played the game for another 5 minutes. At the beginning of the test session, eye tracker was calibrated using 9-point calibration. Training and testing phases were from the same game but different situations. Subject's distance from screen was 130 cm yielding field of view of $43^o \times 25^o$. The overall recording resulted in 2.5 hours of 156 GB video, 192,000 frames, 1,536,000 fixations, and 10,518 saccades.

Subjects played driving games on PC1 which had Windows XP running the games. An array of wheel, pedal and other actions (signal, mirror, etc) was logged with frequency of 62Hz. The frames were recorded on PC2 running Linux Mandriva OS. Game stimuli were shown to the subject at 30Hz. This machine sent a copy of each frame to LCD monitor and saved one copy to the hard disk. PC2 also instructed the eye tracker (PC3) for recording eye positions when watching the screen. PC2 had a dual- CPU processor and used `SCHED_FIFO` scheduling to ensure microsecond accurate timing. Each subject's right eye position was recorded at 240 Hz with a hardware-based eye-tracking

system (ISCAN Inc. RK-464). 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 in experiments), Force Feedback, connected via USB to the PC1.

### 2.2 Model-free analysis of the dataset

We were interested in two types of model-free analyses: 1) analysis of distribution of saccades and fixations in order to find out which locations attract subject's attention; and 2) correlation among eye fixation and actions to be used for eye movement prediction later.

Figure 2.a shows interesting (i.e task-relevant) locations for the 10 subjects (each dot is a saccade) over 3DDS. For illustration purposes, only saccades are shown since fixation maps are highly dense. Note that we are concerned with fixations rather than saccades in the prediction part of this paper (one fixation per frame). On 3DDS game, task-relevant regions are (see Figure 2.a rightmost panel and Figure 1 leftmost panel): an arrow sign at the top-left indicating direction, instruction command at top, instructor and rear-view mirror at the top-right, horizontal view and road (middle), red light slightly above road, and interior (speedometer) of the car at the bottom shown by blue ellipses. As it shows, there is a strong horizontal bias in this task similar to free-viewing and visual search tasks [8]. Profile of wheel vs. eye-y (image width) shows that subjects viewed all vertical line when wheel was released (value of 127) which is the case when they are driving straight or are stopped. There is a slight tendency to look at the bottom when turning left or right. Wheel vs. eye-x shows two main saccade directions 1) horizontal bias, and 2) diagonal which means wheel toward right then saccade right and wheel toward left saccade to the left. To further analysis the data we also tagged each frame of games based on different events that might happen in driving. Some games did not have all the events. Some events of each game could be found in Figure 4. For instance, in "going straight" event, there is a horizontal and vertical bias (wheel vs. eye) and subjects looked more at the center (center-bias) while looking around to get important information (eye-x vs. eye-y). For "turn right" event, there is a rightward shift of fixations based on wheel (similarly for turn left). For "red light" event, since task-driven influences are not much strong (stopping situation), then subjects have time to look everywhere (less task demand). Frequency of brake/gas is shown in right panels in Figure 2.b. There is a peak at the center meaning that most of the time, both pedals are released. When turning right, subjects pressed the gas more and pressed the brake more at the red light.

To learn about the temporal relationship between eye and wheel position, we plotted fixations (heatmaps) in steps of 32 of wheel position for all games in Figure 3. Green circle shows the mean fixation position and vertical red line is the normalized (linearly to its max) wheel position. It could be seen that

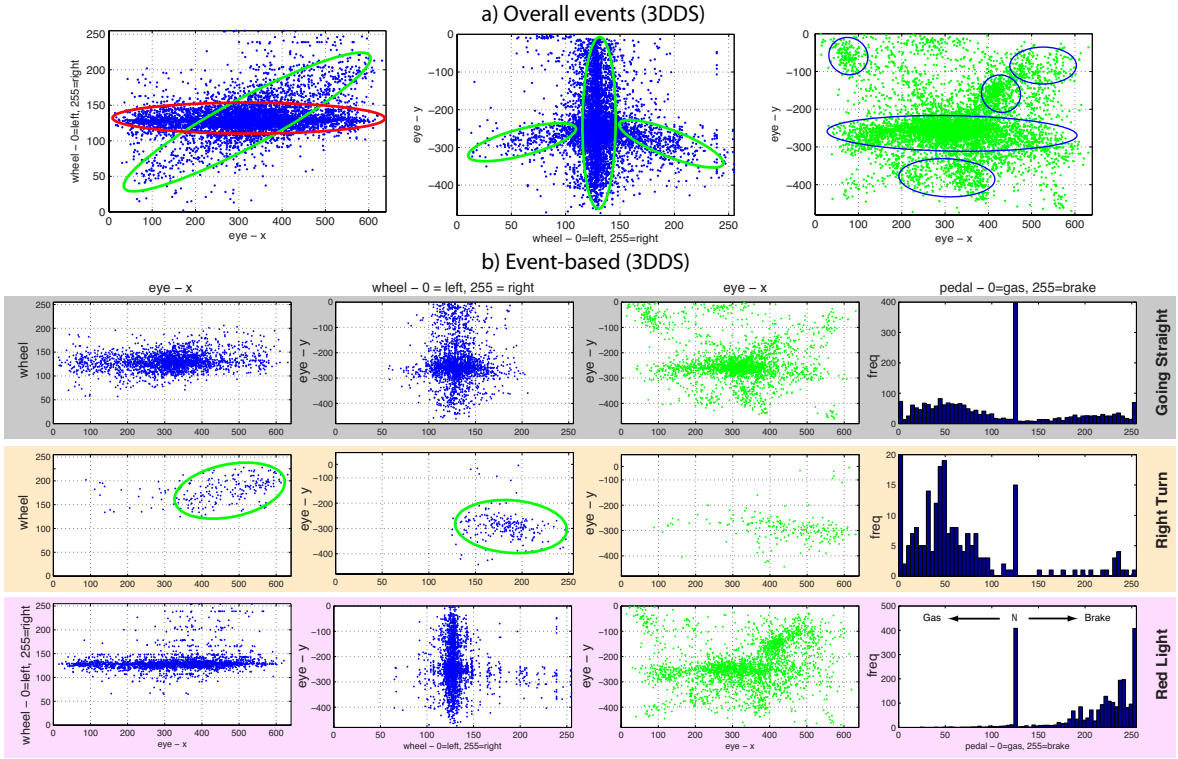


**Fig. 2** a) Correlation between wheel and eye-x and eye-y saccade coordinates (left two panels), and saccade positions (right panel) overall all events of 3DDS. b) Left 3 panels: same as (a) for sample events. Right panel shows the frequency of pedal positions.

a linear relationship with the eye fixation holds for wheel positions between 64 to 192, but for extreme values it seems that wheel position leads slightly. Temporal analysis along with tagging events, could help to build a fixation prediction model provided that the event could be predicted correctly.

Figure 4, shows the average number of saccades per frame for events of all three games. It shows more saccades happen in "red light" and "mistakes" events and less in "turning" and "going straight". This along with sparseness of saccades/fixations for an event indirectly is a measure of how demanding is a behavior (event) and could be used as a cue for weighting top-down and bottom-up saliency maps.

## 3 Learning task-dependent spatial map

In what follows, we explain our model for learning task-dependent, top-down influences on eye position. First, in the training phase, we compiled a training set containing feature vectors and eye positions corresponding to individual

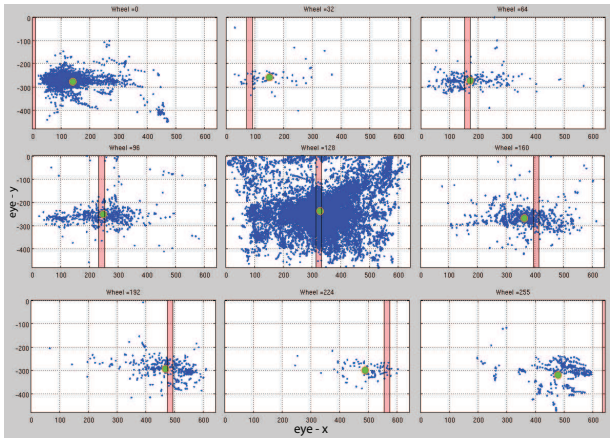


**Fig. 3** Mean fixation position vs. linearly estimated steering wheel position over all games and events.

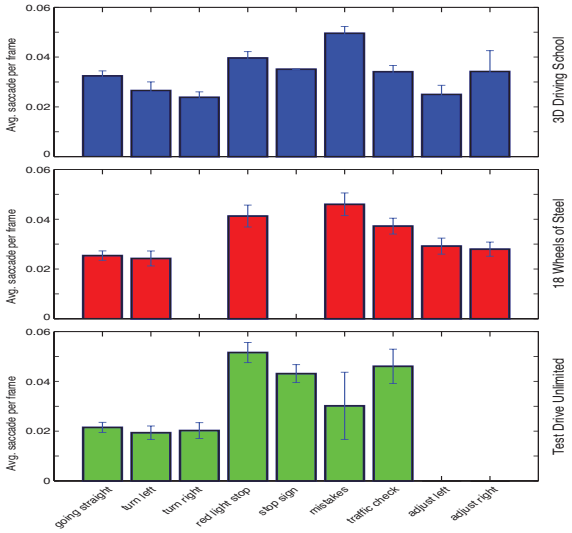


**Fig. 4** Average number of saccades per frame for each event. Error bars are standard deviations.

frames from several video game clips which were recorded while observers interactively played the games. Two approaches were followed. In the first approach called "decision combination method", individual predictors were learned by mapping a feature vector describing state or scene to corresponding fixation of that frame. We also tried to find a best predictor by combining the outputs of these predictors by adding or multiplying their outputs. In the second approach, "feature combination method", all feature vectors were

combined in a single vector in a hope that it might give a better scene/state description. Then a mapping is learned from this vector to fixations.

### 3.1 Training

Let $S_t$ be the state of a player at time $t$ defined as $S_t = [b, L_{t-m}, ...L_{t-1}, L_t]$ as a history of past scene representations where $L_t = [G_t, B_t, A_t, E_t]$ is the information at time $t$. $b$ is a scalar bias value, $G_t$ is the gist of the scene, $B_t$ is the raw bottom-up saliency map, $A_t$ is the associated action for frame $t$ and $E_t$ is the labeled event. $m$ is the depth of history. When a task is Markovian, all information regarding state is available at the current time (i.e $m = 0$). We are interested in finding a mapping $\mathcal{F}$ from $S_t$ to $P_{t+1} = [x_{t+1}, y_{t+1}]$, the eye position at time $t+1$. Over task $T$, assume $q$ subjects have done the task. We collected a dataset $D = \{M, N\}$ where $M$ is a $n \times |S|$ matrix of feature vectors and $N$ is the $n \times |P|$ matrix of eye positions. $n$ is the size of pairs of features to eye fixations from $q-1$ subjects. A classifier is learned from this data and is tested over the remaining $q-th$ subject in a leave-one-out approach.

In the decision combination approach, we combine decisions of predictions based on different features. That is, saliency map at time $t$, would be a function $\mathcal{G}$ of different $\mathcal{F}$ mappings, (each $\mathcal{F}$ could be considered as a single behavior). This arbitration mechanism, $\mathcal{G}$, itself is task dependent and tells how different top-down factors should be integrated. Here, we tried two simple integration functions: addition and multiplication. As it has been claimed that final behavior is a combination of pure bottom-up and top-down influences, we considered pure bottom-up map $B_t$ as an individual predictor as well.

In the feature combination approach, we started with a simple classifier $\mathcal{F}_0$ by only considering the bias term $b$ and gradually added more features to it to build more predictive classifiers $\{\mathcal{F}_1, \mathcal{F}_2, ...\}$.

Assuming a linear relationship between feature vectors and eye fixations, we solve the equation $M \times W = N$. Solution to this equation is: $W = M^+ \times N$, where $M^+$ is the pseudo inverse of matrix $M$. When feature vector is $b$, the solution (predicted map) is simply the average of all eye position vectors in $N$. This classifier is called mean eye position (MEP). This way, we are solving a linear regression classifier with the least squares method. We used SVD to find the pseudo inverse of matrix $M$. An important point here is that we set eigenvalues smaller than half of the biggest eigenvalue to zero to avoid numerical instability.

Vector $P$ which is eye position over the $640 \times 480$ image is downsampled to $20 \times 15$ and transformed into a $1 \times 300$ vector with a 1 at the actual eye position and zeros elsewhere. In testing phase, in order to predict the eye position for a new frame of a subject first, a feature vector (as above) is extracted and then a saliency map is generated by applying the learned mapping. Maximum of this map could be used to direct attention. In combination (addition and multiplication), first saliency maps are linearly normalized and combined to form a new saliency map.

### 3.2 Features

**Mean eye position**(MEP). MEP is the prediction when distribution of fixations is available. In dynamic environments used in this paper, since frames are generated dynamically and there are few fixations per frame, aligning frames (contrary to movies) is not possible. If a method could dynamically predict eye movements in a frame by frame basis, then achieving a higher accuracy than MEP is possible.

**Gist of the scene ($G$)**. Gist is a very rough representation of a scene and does not contain much details about individual objects or semantics but can provide sufficient information for coarse scene discrimination (e.g., indoor vs. outdoor). The pyramid-based feature vector (pfx) [24], relies on 34 feature pyramids from the bottom-up saliency model: 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 value of the entire feature map, values 1 to 4 are the average values of each $2 \times 2$ quadrant of the feature map and values 5 to 20 are the average value for each of the $4 \times 4$ grids of the feature map leading to overall of $34 \times 21 = 714$ elements.

**Bottom-up saliency map ($B$)**. 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 four oriented motion energies (up, down, left, right). After a center-surround difference operation and across scale competitions, a unique saliency map is created and subsampled to a $20 \times 15$ feature vector which is then linearized to a vector of $1 \times 300$ [16].

**Physical actions ($A$)**. In the driving experiment, actions are a 22D feature vector containing wheel positions, pedals (brake and gas), left and right signals, mirrors, left and right views, gear change, etc which are wheel buttons subjects used while playing.

**Labled events ($E$)**. Each frame of games was manually labeled belonging to one of different events such as left turn, right turn, going straight, adjusting left, adjusting right, stop sign, etc (Figure 4). Hence this is only a scalar feature.

## 4 Model-based results

To quantify how well model predictions matched observers' actual eye positions, we used the normalized scanpath saliency (NSS) metric, which is defined as the response value at the human eye position, $(x_h, y_h)$, in a model's predicted gaze density map that has been normalized to have zero mean and unit standard deviation.

In the first experiment, we trained the model over each separate game. Each game segment has 8,000 frames, therefore in a driving game training was done over $9 \times 8,000$ frames and tested over remaining frames of the test subject.

Figure 5.a shows NSS scores of models with single features and best answers for both combination approaches for each individual game. Over three games, decision combination approach resulted in higher NSS score. Saliency maps learned from gist features and the raw BU map were the most informative ones. Feature combination resulted in lower performance but still higher than all other single predictors. In agreement with previous results [15], the BU raw map resulted in the least performance (below 0.5) again indicating that BU saliency does not account for task-driven fixations. High NSS score for gist means that scene representation is a good predictor of state. However, our models in all three games were significantly above using only gist. Using only action features outperforms MEP and Gaussian models significantly indicating influence of action on prediction of top-down map. Predictor based on event feature was slightly lower than MEP but still better than Gaussian and BU raw map. All models were significantly above chance.

High NSS score for Gaussian indicates high center-bias in these tasks and could be further verified from mean heat maps (MEPs) in Figure 7. The fact that data is center-biased (by design) makes outperforming the MEP difficult (see Figure 7). The main reason for this is that a huge number of fixations happen in the center. Given the high center bias, a predictor can only improve its performance on the few samples that are off center.

In the second experiment, we trained the model over all games, each time over 29 subjects and tested over the remaining subject. Results are shown in Figure 5.b. Consistent with results in Figure 5.a, decision combination approaches led to higher NSS performance. NSS values for MEP, Gaussian, action, event and gist in order are: 2.68, 1.96, 2.72, 2.61, and 2.93. Our results for addition, multiplication and Feature Combination are 3.16, 3.08, and 2.96, respectively which are significantly higher (paired t-test, $p < 0.05$) across fixations. It shows that our approach have more prediction power compared with previous models [15,16]. Action features alone are significantly higher than MEP and Gaussian. In feature combination approach, adding action and event features to the state reprsentation improved the performance.

Figure 6 shows sample frames from three games along with the corresponding predicted saliency maps from the various models. Output of the bottom-up saliency map (BU raw) show spread activity with a weak maximum at the actual eye position. Predicted saliency maps by our models show dense activity at task relevant locations thereby narrowing attention and leading to higher NSS score. It seems that combined maps in general are more capable of finding the task-relevant regions. These maps change per frame as opposed to the static MEP and Gaussian.

## 5 Discussion and Conclusion

We analyzed the influence of action on driving task and proposed general methods for using it as an eye movement predictor. This approach performs

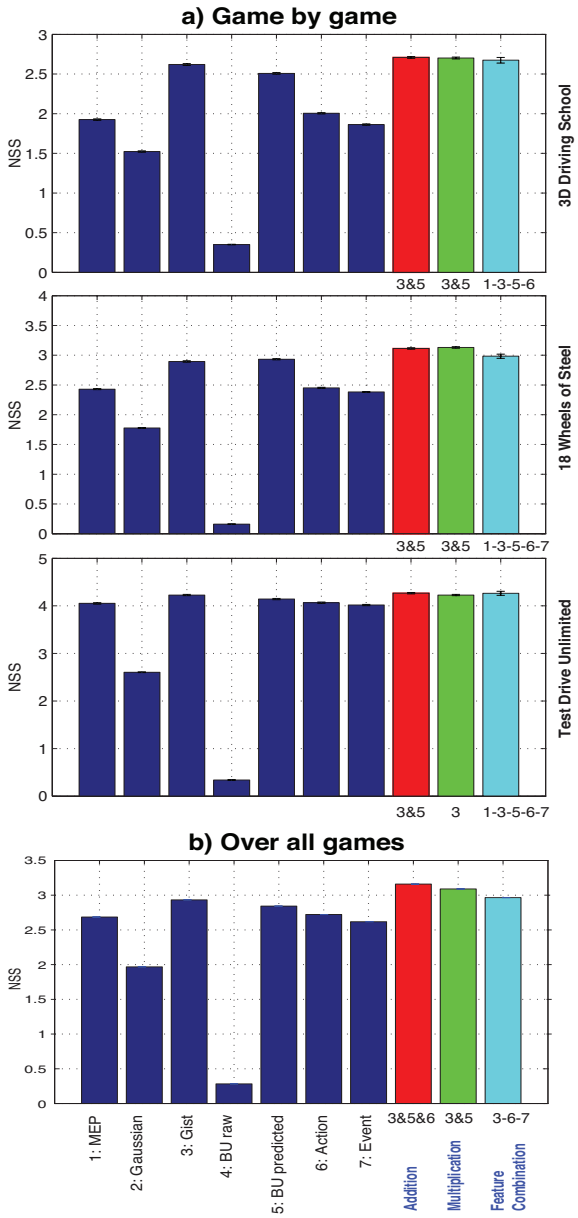


**Fig. 5** NSS scores of different models with actual recorded eye positions (single features and combinations). a) model tarined over each individual game, b) model tarined over all games. A larger NSS score means a better fit. Each bar represents mean  s.e.m. across all 80,000 fixations for each game. "&" sign means that final maps were combined (decision combination) and "-" sign indicated that features were combined (feature combination).

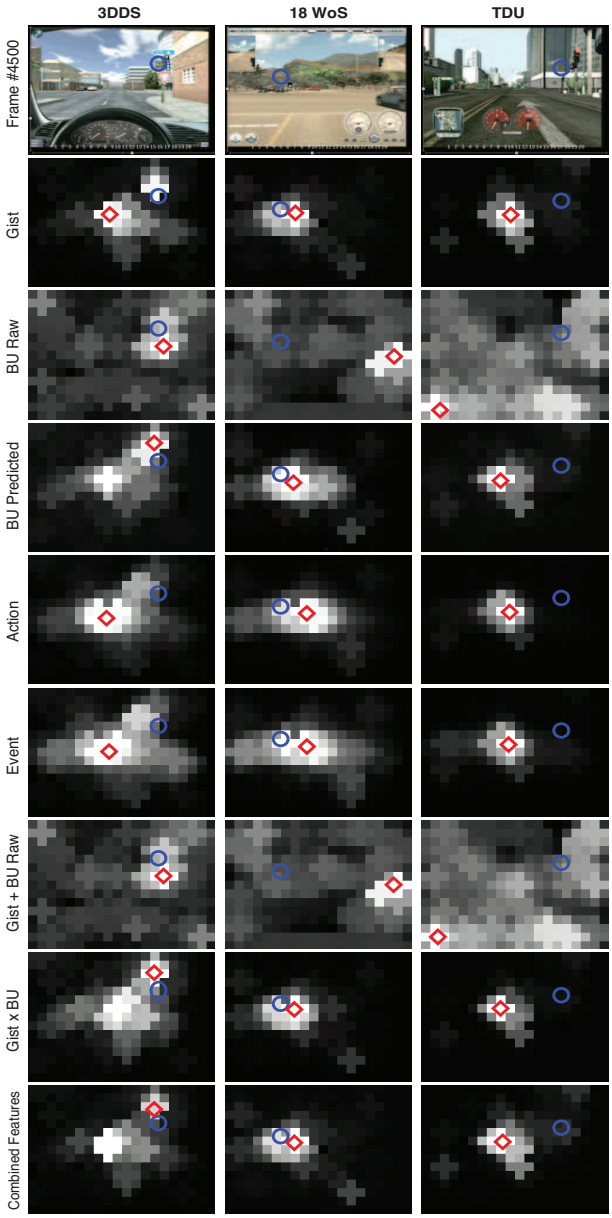


**Fig. 6** Sample frames along with predicted top-down maps of different models. Each row is the output of a different model. BU raw is the output of the purely bottom-up saliency model and BU predicted means predicted saliency map when raw BU features are used. Red diamond: maximum of each map, blue circle: actual eye position for that frame. Gist×BU is the point-wise product of BU predicted and gist models.

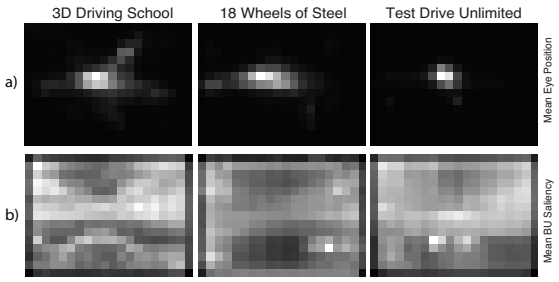


**Fig. 7** a) Mean eye position maps, b) mean bottom-up saliency maps for three games. BU saliency maps have stable white regions which creates false positives over many frames.

better when attention is more influenced by motor actions. Results show that combining decisions works better than combining features.

History depth ($m$) was set to 0 in the experiments, since higher history depths were not helpful. One reason might be that performance in eye fixation prediction in our data is not limited in gist classification, since gist over a single frame is already a good predictor. The performance at the other hand seems to be more limited by the correlation among subjects at looking at the same spots for a same scene. The higher the correlation, the better learning and prediction.

A big issue in saliency modeling (either BU or TD) is handling center-bias. Most of the available datasets are center-biased meaning that a large proportion of fixations happen to be in the center of the image. For example, available still images shows photographer bias when photographers intentionally put interesting eye catching objects in the center [8]. Similarly, game designers dynamically change the viewpoint in order to put the needed object (main character, road, etc) at the center. This contaminates the scores. Gathering a less center-biased dataset over movies or interactive natural setups would be very helpful for fair evaluation of top-down models. Another problem is being able to predict the eye fixation at the frame (millisecond) level. For instance, we know that subjects when approaching red light, might look at the red sign at one point but predicting the exact occurence time is very subjective and dependent on instantanous task demands. Therefore, some high level knowledge about the task seems to be necessary.

For future investigation, we are going to build more effective learning systems and classifiers that have more generalization capabilities (SVM, RBF networks, ...). Here, we also tried kNN classifiers but results were almost the same as regression. Also, recent video processing and analysis approaches (segmentation, action recognition, etc) might be interesting. Here we followed a data-driven approach. One promising extension would be trying to infer some high-level knowledge or behaviors from data similar to [19].

*Supported by the National Science Foundation (grant number BCS-0827764), and the Army Research Office (W911NF-08-1-0360 and W911NF-11-1-0046), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views ex-*

*pressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.*

## References


1. L. Itti and C. Koch. Computational modeling of visual attention. *Nat. Rev. Neurosci.*, 2(3):194-203, 2001.
2. 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.
3. A. Yarbus. Eye movements during perception of complex objects. L. Riggs, editor, Eye Movements and Vision, 1967.
4. M. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25-26):3559-3565, 2001.
5. M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cog. Sci.*, 9(4), 188-193. 2005.
6. C. Rothkopf, D. Ballard, and, M. Hayhoe. Task and scene context determines where you look. *Journal of Vision*, 7(14):16, 1-20, 2007.
7. W., Kienzle, A. F., Wichmann, B., Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. *NIPS*, 2007.
8. T. Judd, K. Ehinger, F. Durand and, A. Torralba. Learning to predict where humans look, *ICCV*, 2009.
9. E. Vig, M. Dorr, T. Martinetz, and, E. Barth. A learned saliency predictor for dynamic natural scenes, *ICANN*, LNCS, (6354): 52-61, 2010.
10. 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.
11. A. Kimura, D. Pang, T. Takeuchi, K. Miyazato, J. Yamato and K. Kashino, A stochastic model of human visual attention with a dynamic Bayesian network, *IEEE Transactions PAMI*. In Press.
12. M. Hayhoe. Advances in relating eye movements and cognition. *Infancy*, 6(2): 267-274, 2004.
13. D. Ballard, M. Hayhoe, and J. Pelz. Memory representations in natural tasks. *Journal of Cog. Neurosci.*, 7(1), 66-80, 1995.
14. M. F. Land and D. N. Lee. Where we look when we steer. *Nature*, 369: 742-744, 1994.
15. R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *CVPR*, 2007.
16. 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.
17. R.J. Peters and L. Itti, Congruence between model and human attention reveals unique signatures of critical visual events. *NIPS*, 2008
18. V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2): 205-231, 2005.
19. N. Sprague, D. H. Ballard, Eye Movements for Reward Maximization. *NIPS*, 2003.
20. H. Hecht, S. Vogt and W. Prinz, Motor learning enhances perceptual judgment: A case for action-perception transfer. *Psychological Research*, 65:3-14, 2001.
21. S. Schtz-Bosbach and W. Prinz, Perceptual resonance: Action-induced modulation of perception. *Trends in Cog. Sci.*, 11:349-355, 2007.
22. 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.
23. 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.
24. C. Siagian and L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions PAMI*, 29(2):300-312, 2007.
25. N. Pugeault and R. Bowden. Learning pre-attentive driving behavior from holistic visual features. *ECCV*, 2010.

26. K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: a combined source model of eye guidance. *Visual Cognition*, 17:945-978, 2009.
27. S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: a Bayesian inference theory of visual attention. *Vision Research*, 2010.