Computational Modeling of Top-down Visual Attention in Interactive Environments

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

http://ilab.usc.edu/publications/doc/Borji_etal11bmvc.pdf

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

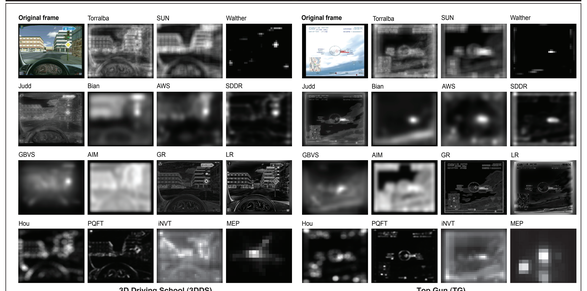
(1) Introduction & Motivation

- Visual attention is a key function for both machine and biological vision systems.

- Research efforts in computer vision have mostly been focused on modeling bottom-up saliency. Strong influences on attention and eye movements, however, come from instantaneous task demands.

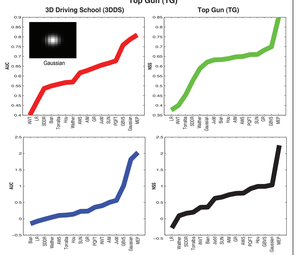
- Our models estimate the state of a human subject performing a task (here, playing video games), and map that state to an eye position. Factors influencing state come from scene gist, physical actions, events, and bottom-up saliency. Proposed models fall into two categories. In the first category, we use classical discriminative classifiers, including Regression, kNN and SVM. In the second category, we use Bayesian Networks to combine all the multi-modal factors in a unified framework. Our approaches significantly outperform 15 competing bottom-up and top-down attention models in predicting future eye fixations on 18,000 and 75,000 video frames and eye movement samples from a driving and a flight combat video game, respectively.

(2) Bottom-up Models



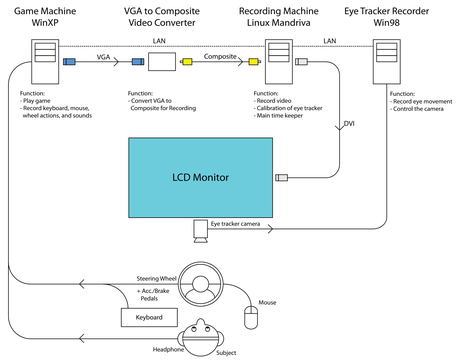
3DDS: 3D Driving School
TG: Top Gun
18WoS: 18 Wheels of Steel
TDU: Test Drive Unlimited

- Computational Models of Bottom-Up Sliency Can not Capture Task Demands



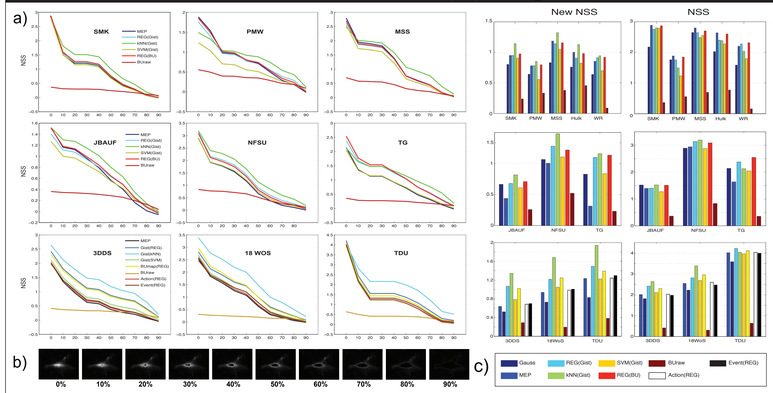
Top: Sample frames along with corresponding saliency maps of models. **Bottom:** AUC scores (chance level is 0.5, higher scores indicate better models) and NSS scores (chance level is 0.0, higher is better) of 14 saliency models over 3D Driving School and Top Gun games. Some models are able to detect the traffic light sign as salient, which happens to be task-related in the sample shown image. Overall performance of models is very poor compared to the inter-observer model.

(3) Data Gathering



Ten subjects aged 18-25 with valid driving license and at least 2 years driving experience were asked to play 3 games (3DDS, 18WoS, and TDU) with the rig above. Subjects were placed 130cm from the screen with field of view 43º x 25º. There was a 5-min training and 5-min test sessions for each game. Video frames (30Hz), Eye fixations (240Hz), and Actions (62Hz) were recorded resulting in total 2.5 hours recordings, 156GB of data, 192,000 frames, 1,536,000 fixations, and 10,518 saccades.

(4) Top-down Models



a) NSS scores over three video games for different amounts of data, **b)** Fixation maps with α% of data discarded and c) Average NSS over saliency levels (left) and NSS score over all fixations (i.e 0% case) for classifiers.

- Classifiers

Leave-one-out training across subjects.

Regression(REG): Assuming a linear relationship between feature vectors M and eye fixations N [3], we solve the equation M ×W = N. The solution is: W = M^+ ×N, where M^+ is the (least-squares) pseudo-inverse of matrix M. We used SVD to find the pseudo inverse of matrix M.

kNN: We look into training data and find similar neighborhoods to the current test frame and then make attention maps from the associated eye fixations. This resembles a local MEP model, where we make a map with 1's at fixated locations and zeros elsewhere. Then to generate an attention map, we convolve this map with a Gaussian filter.

SVM: To use SVM, we first reduced the high-dimensional feature vector using PCA to preserve 95% of variance. Then a linear multi-class SVM was trained from other subjects with 300 output classes. Experimenting over a subset of the data with low-resolution eye fixation maps (4 × 3 and 8 × 6 hence number of classes 12 and 48) and with polynomial and RBF kernels did not improve the results.

- Features

Mean eye position (MEP): Mean of the distribution of all human fixated locations

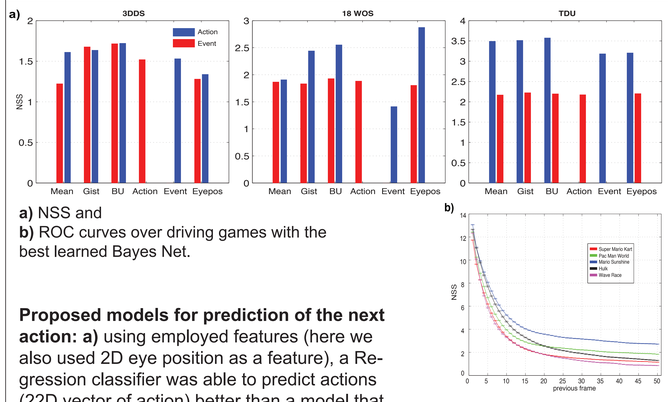
Gist (G): A very rough representation of a scene and does not contain much details about individual objects or semantics but can provide sufficient information for coarse scene discrimination (e.g., indoor vs. outdoor or category of the scene). The pyramid- based feature vector (pfx) [1] was used for scene representation.

Bottom-up saliency map (BU): We used the original bottom-up saliency map both as a signature of the scene and a saliency predictor [2].

Physical actions (A): In the driving experiment, action is a 22D feature vector containing wheel positions, pedals (brake and gas), left and right signals, mirrors and left and right views, gear change, etc which are wheels buttons that subjects used for driving.

Labeled events (E): Each frame of games was manually labeled as belonging to one of different events such as {left turn, right turn, going straight, red light, adjusting left, adjusting right, stop sign, traffic check and error frames due to unexpected events that terminate the games like hitting other cars}.

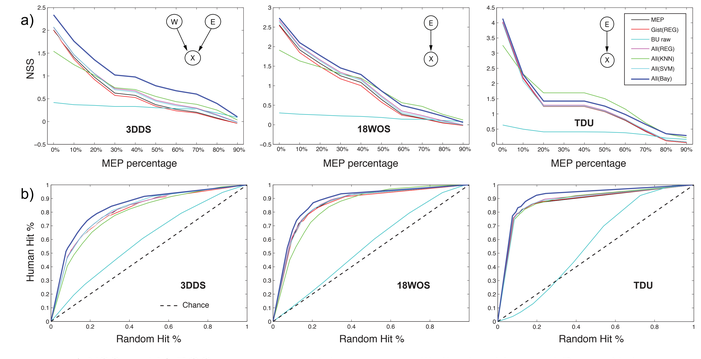
(5) Employing Actions



a) NSS and
b) ROC curves over driving games with the best learned Bayes Net.

Proposed models for prediction of the next action: a) using employed features (here we also used 2D eye position as a feature), a Regression classifier was able to predict actions (22D vector of action) better than a model that is the average of actions (similar to MEP for eye positions) in terms of NSS score. BU map and Gist scene descriptors performed better than other features. **b)** shows an upper bound on NSS score when fixations of previous frames were considered as predictors for the current frame (averaged across subjects for each game). This is the score of an optimal model that could consider subjectivity, noise and task demands, and it provides an interesting comparison point for our computational models.

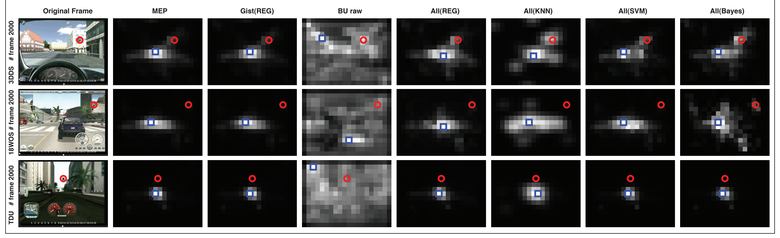
(6) Bayesian Implementation



a) NSS and **b)** ROC curves over driving games with the best learned Bayes Net.

We also propose a generative model based on Bayesian Networks to systematically learn relationships between variables and eye position. To accommodate features for use of Bayes Net, we clustered the high-dimensional Gist vector using k-means to r clusters (here r = 20). Continuous wheel and pedal positions were discretized to 8 values. Number of events were 9. Due to high complexity of these games a manually-designed Bayes Net is less likely to produce good results (We systematically experimented with several network topologies). Thus, we used a variant of Markov Chain Monte Carlo (MCMC) algorithm called Metropolis-Hastings (MH) to search the space of all DAGs in a network that has all variables (Gist (G), BU map (B), Wheel(W), Pedal(P) and Event (E)) connected to eye position (X). See above figure for results.

Sample Predicted Maps



Conclusions & References

- This study demonstrates that it is possible to develop computational models which are capable of estimating state and predicting task-dependent future eye movements and actions of humans engaged in complex interactive tasks.

- Scene context is an effective estimator of the subject's mental state and a good predictor of fixations.

- Taking advantage of temporal information leads to higher eye movement prediction accuracies and is a roadmap for future research.

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

[2] 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.

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