Computational models for predicting gaze direction in interactive visual environments

Robert J. Peters and Laurent Itti
University of Southern California, Computer Science, Neuroscience

Computational models of bottom-up attention can perform significantly above chance at predicting eye positions of observers passively viewing static or dynamic images. Nevertheless, much of eye movement behavior (50% or more) is unexplained by purely bottom-up models, and is typically attributed to top-down, inter-observer, task-dependent, or random effects. Other studies have described such high-level effects in naturalistic interactive visual tasks (e.g., while driving, how often do people fixate other cars, or the road, or road signs); yet the underlying neurocomputational mechanisms are still unknown. Here, we introduce a simple computational model of task-related eye position influences in interactive tasks with dynamic stimuli. This model (Figure 1) extracts a low-dimensional feature signature ("gist") from each frame, compares that signature with a training database, and produces an eye position prediction map. Finally, we combine the task-related and bottom-up maps, and compare the combined maps with observers' actual eye positions across 216000 frames from 24 five-minute videogame-playing sessions. For analysis, each map was rescaled to have zero mean and unit standard deviation; the average predicted value at human eye position locations was 0.61 +/- 0.1 in the purely bottom-up maps, and 2.42 +/- 0.07 in the combined maps (a random model gives an average value of 0). Thus, this straightforward model of task-dependent effects offers some of the strongest purely computational general-purpose eye movement predictions to date, going significantly beyond what is explained by purely bottom-up effects; yet it relies only on simple visual features, without requiring any high-level semantic scene description.

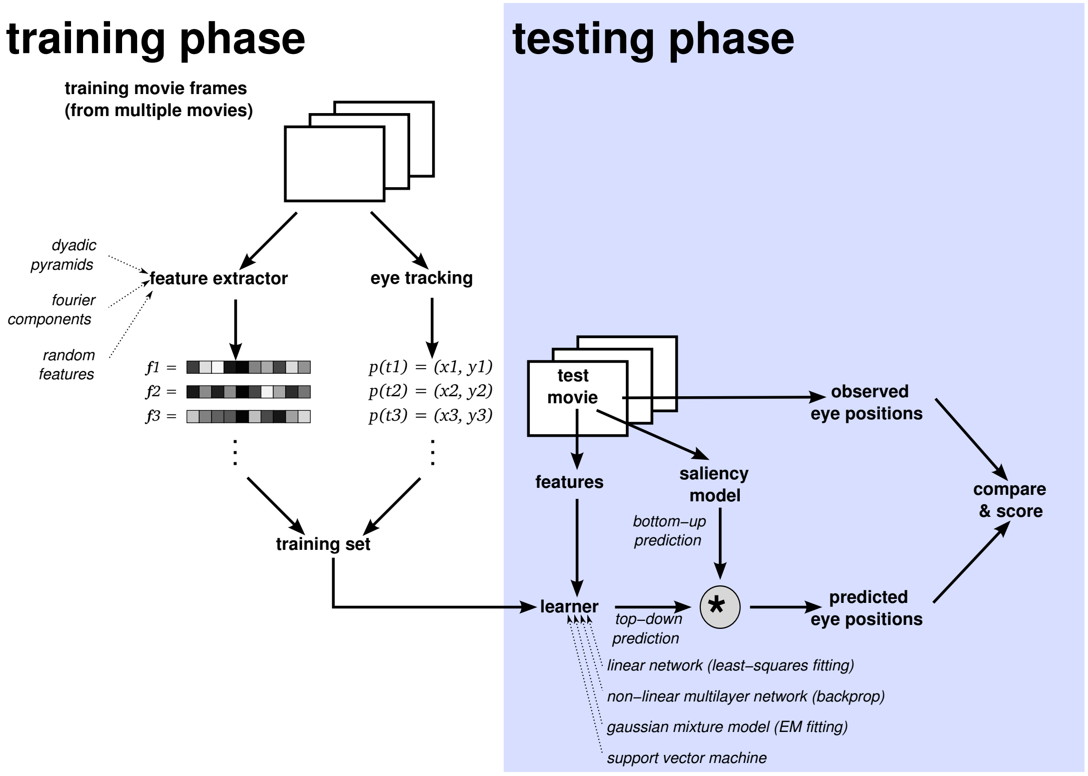

Figure 1: Schematic diagram of a model for learning top-down, task-dependent influences on eye position during interactive viewing of dynamic stimuli.