Integrating low-level and high-level visual influences on eye movement behavior

Robert J. Peters and Laurent Itti

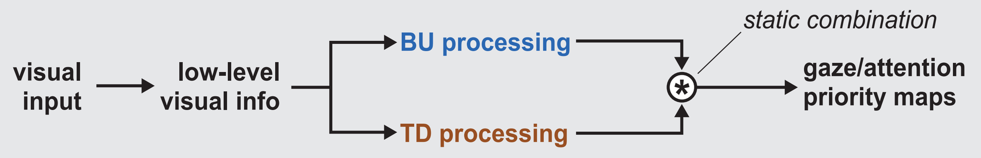
University of Southern California, Computer Science, Neuroscience

http://ilab.usc.edu/rjpeters/pubs/2007_VSS_Peters.pdf

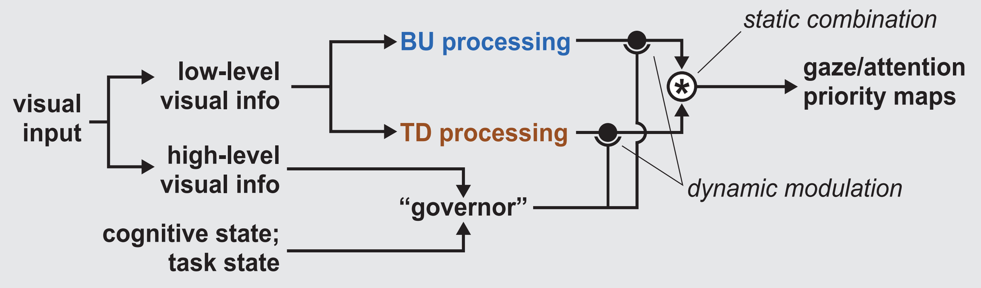
Supported by the National Geospatial-Intelligence Agency (NGA), Irvine Sensors Corporation, and the Intelligence Community (IC) Postdoctoral Research Fellowship Program

(1) Introduction & Conclusion

How does the brain integrate **bottom-up** (BU) and **top-down** (TD) influences on eye position? Often a simple **static combination** of factors is assumed:



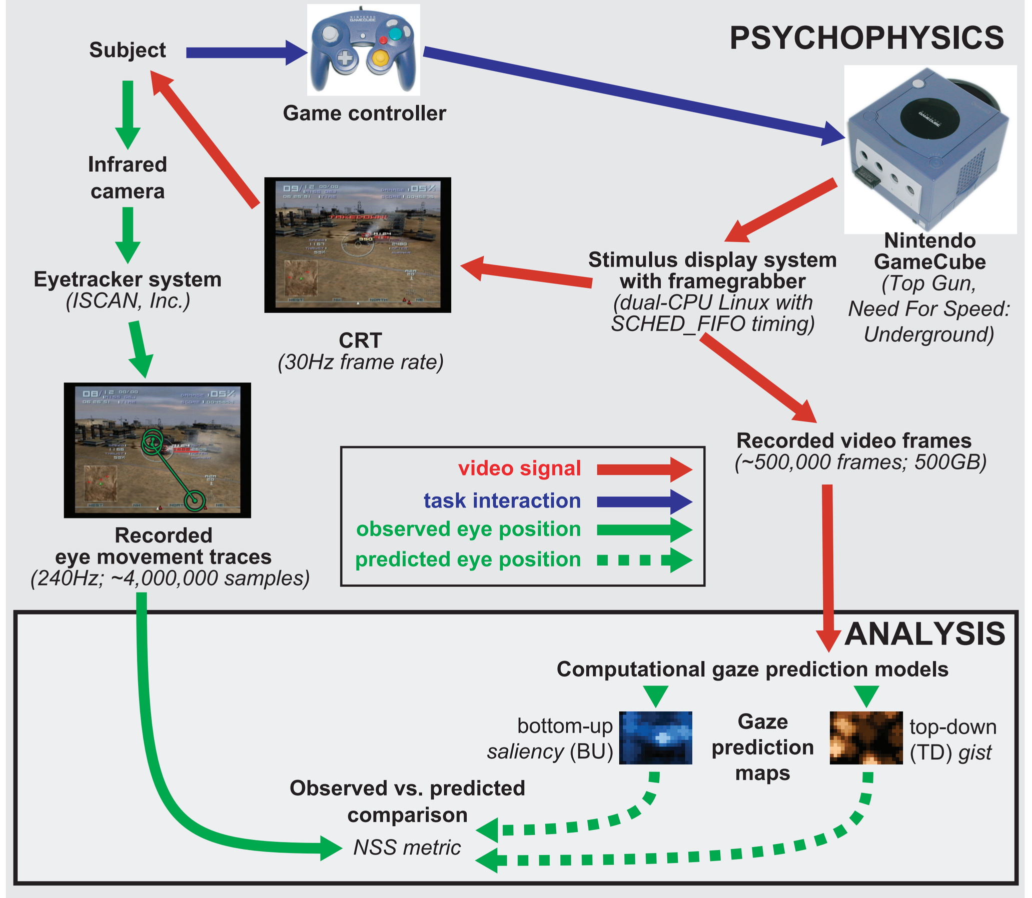
Here we propose a framework for **dynamic combination** of factors, in which a "governor" module receives both high-level visual information (such as object identities) as well as information about cognitive task state and goals, and uses this high-level information to selectively modulate the influence of BU and TD signals on eye position:



We report finding evidence consistent with such a system in the relationships between eye tracking recordings and the predictions of computational BU and TD models: we estimate cognitive task state by isolating important game events, and show that the predictive strength of BU and TD models changes significantly around these events.

(2) Overall method

- eye movements recorded while subjects play video games
- eye movements compared with model predictions



(8) Results: Top Gun (flight combat)

(9) Results: Need For Speed: Underground (car racing)

"target destroyed" events | *"missile fired" events* | *"damage taken" events* | *"start speed-up" events* | *"lose race position" events*

Game frames surrounding event at times *t*-δ, *t*, *t*+δ

shown here for a single example event

Game events extracted from a continuous signal across the full game session

shown here for a single example session

Eye position (in screen coordinates) recorded from observer playing the game

shown here for a single example session

Prediction strength of BU and TD models in predicting observers' eye position

shown here for a single example session

Event-locked prediction strength of BU and TD models, averaged across all events of a given type

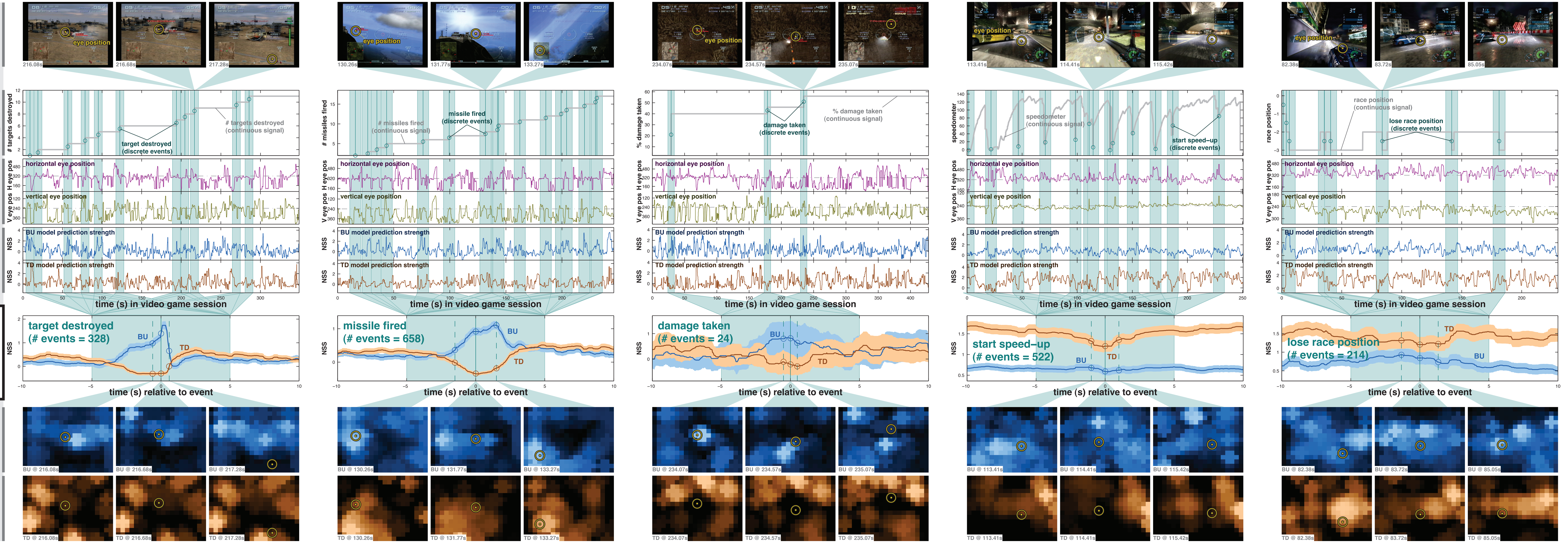
*shown here for **all events of a given type***
shaded area represents 98% confidence interval

BU prediction map of gaze position at times *t*-δ, *t*, *t*+δ, relative to an event time

shown here for a single example event

TD prediction map of gaze position at times *t*-δ, *t*, *t*+δ, relative to an event time

shown here for a single example event

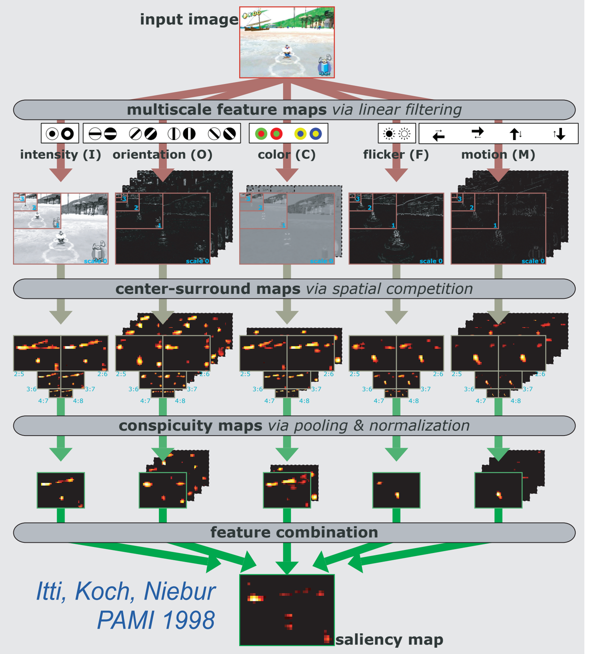


(3) Videogame stimuli

- Nintendo GameCube games "Top Gun" and "Need For Speed: Underground"
- 3 game levels per game
- 5 subjects
- Subjects practiced several hours on each game prior to eye movement recordings
- During eye tracking, subjects played each game level twice
- Each game level lasted ~3–9 minutes
- Data set spans ~4.5 hours in 60 sessions (5 subjects x 2 games x 3 levels/game x 2 sessions/level)
- Data set includes ~500,000 video game frames and ~4,000,000 eye position samples

(4) Bottom-up model

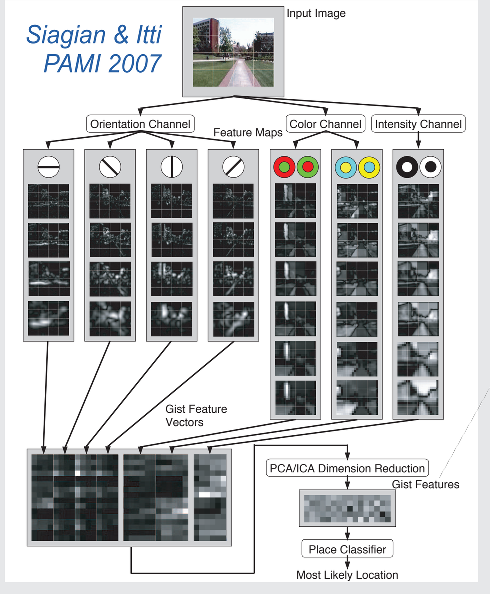
- **12 features**
  intensity
  orientation x 4
  color x 2
  flicker
  motion x 4
- **9 scales**
- **6 center-surround pairs**
- **combination across scales**
- **combination across features**



Itti, Koch, Niebur PAMI 1998

(5) Top-down/task-relevance model

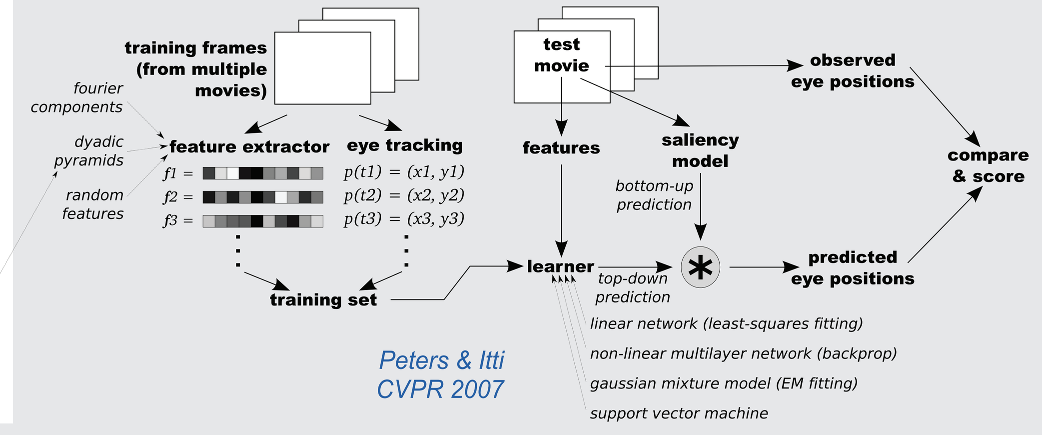
(a) feature extraction



Siagian & Itti PAMI 2007

- model learns to associate ***image "gist" signatures*** with corresponding ***gaze density maps***
- a leave-one-out approach is used for testing: for each test clip, the remaining 59 clips are used for training

(b) training phase **(c) testing phase**

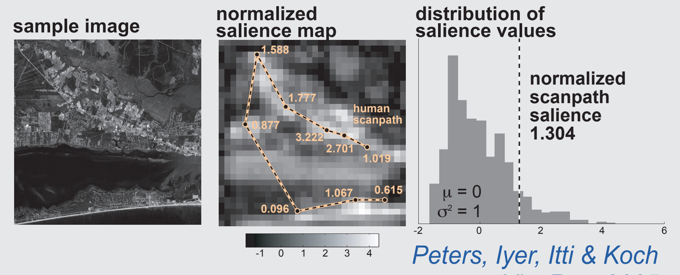


Peters & Itti CVPR 2007

(6) Model scoring

normalized scanpath salience (NSS)

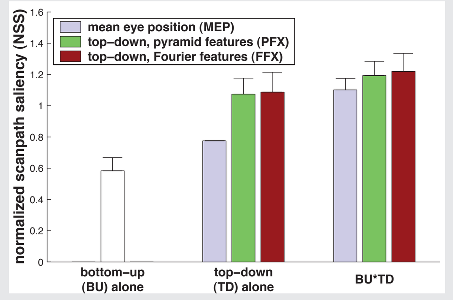
- normalize each salience map to mean=0 and stdev=1
- overlay human scanpath on normalized salience map
- extract normalized salience value at each eye position
- average these values to give the overall NSS value



Peters, Iyer, Itti & Koch Vis. Res. 2005

(7) Previous results

- ***individual BU and TD models*** each perform better than chance at predicting gaze
- ***combined BU*TD model*** performs better than either individual model, and better than a simple mean eye-position control model



Peters & Itti VSS 2006
Peters & Itti CVPR 2007