

# Directed Research

# End of Semester Report

Spring 2006

Under the guidance of

**Professor Laurent Itti**

at the iLAB, USC

3641 Watt Way
Hedco Neuroscience Building
University of Southern California
CA 90089-2520, USA
Tel: +1(213)740-3527

**Ankur Sahai**
Graduate Student
M.S. - Computer Science
University of Sothern California
CA 90089, USA
Tel: +1(213)747-3386

## Introduction

This is the end of semester report for the work done under Prof. Laurent Itti at the iLAB, USC. Throughout the semester, I undertook a study of publications in the area of *Vision;* majority of which were in-line with the type of research being carried out at the iLAB. I also went through some interesting publications in the areas like *Consciousness, Psychophysics etc.* In addition to my theoretical study of the research being carried out in the lab, I performed scientific data collection for the lab. My practical work was restricted to collecting centre-biased and non-centre-biased video clips that are to be used as stimuli for the subjects in the eye-tracking experiments to study their responses to these two categories of clips in ways discussed in *practical work* section of this report. I also wrote a simple C++ module for *calibration* of the eye-tracker readings to the actual positions of the stimulus on the screen; that is to be further developed to perform *smooth pursuit calibration*.

The underlying computational model of primate vision followed was the one proposed by **Itti, Koch & Niebur** [1] that was extended by **Itti, Dhavale & Pighin** [2] to include eye and head movement animations and later extended by **Navalpakkam & Itti** [3] to bring in modulations caused due to top-down task demands. I also got the opportunity to learn about the latest directions in research in Vision through the lab meetings and interactions with the members of the iLAB apart from the many interesting presentations made on different topics by experienced researchers from other labs and agencies. Especially, the interaction with the team from Klab, Caltech, headed by Christof Koch, was the most interesting and memorable one.

These opportunities gave me a rich experience, not only to learn about the latest research carried out and the novel directions followed in the field of *Vision*, but also to see how they are implemented practically. Of these general directions, I found the notion of top-down and bottom-up influences and how they interact to shape the vision in primates to be the most interesting one.

I also came up with a few ingenious theoretical ideas and practical techniques intended to further and extend the research carried out in the iLAB.

Subsequent sections of this report will summarize the theoretical study and practical work carried out and the relevant ideas that I came up with as a member of the lab during my stint as a directed research student here.

**Theoretical Study**

I primarily went through the **iLAB** and **Klab** publications. I also went through selected publications from the pioneering journals in Vision and Computational Neuroscience like *Vision Research, Journal of Vision, Nature, Visual Cognition, Journal of Cognitive Neuroscience, Neural Computation, Vision Sciences Society (VSS) Journal, computer Vision and Image Understanding, International journal of Computer Vision, Journal of Mathematical Imaging and Vision* etc. The following paragraphs summarize the general ideas and directions that I found to be the most interesting and stimulating while going through these publications.

As discussed earlier, I focused on the topic of top-down and bottom-up components and how they interact to shape the vision among the primates. There have been a lot of interesting studies done in this area. The seminal work concerning the notion of top-down and bottom-up influences was first produced by **Itti** in his Ph.D. thesis *Models of top-down and bottom-up visual attention* [4]. He explains clearly the role played by the bottom-up and top-down components in shaping the vision among the primates. Bottom-up component is constituted by those objects that have the ability to distinguish themselves from there surrounding in such a way as to capture our attention. To quote this idea from [4]:

> *"In a first approximation, focal visual attention acts as a rapidly shiftable "spotlight", which allows only the selected information to reach higher levels of processing and representation."* Itti [4]

It is based on the notion of a saliency map that is a 2-dimensional mapping encoding the degree of prominence of the different objects in a particular scene. *Winner-takes-all* is applied among the neurons in this map and the resulting location that emerges as the winner is attended next. Inhibition of return then comes into play by inhibiting this (last most salient location) and allowing the system to focus at the next most salient location.

Once the attention has been focused onto a filtered visual field; one may predict the future visual patterns to be simply shaped by only a feed-forward spatially selective filtering process; which is not the case as suggested by significant experimental evidence. Instead, there are some other guiding factors of which the most likely one seems to be some sort of feed-back and local modulation in a *top-down* manner.

The computational model of vision in primates proposed by Itti, Koch & Niebur [1] can be described as follows. Filtering is performed at eight spatial scales in the first pass to calculate the visual features. This is followed by calculation of center-surround differences to compute the local spatial contrast in each feature dimension. A lateral inhibition scheme is applied iteratively to model competition for conspicuity within a given feature-map. After this, these feature

maps are combined into one conspicuity map for each feature type. The seven conspicuity maps are then summed up into a unique topographic saliency map. This saliency map is implemented as a two-dimensional array of Integrate-and-Fire (I&F) neurons. The winner-takes-all is implemented using these I & F neurons; to detect the most salient location and the visual attention is redirected towards it. The inhibition-of-return mechanism then suppresses this location in the saliency map to redirect the attention to the next most salient location in the image.

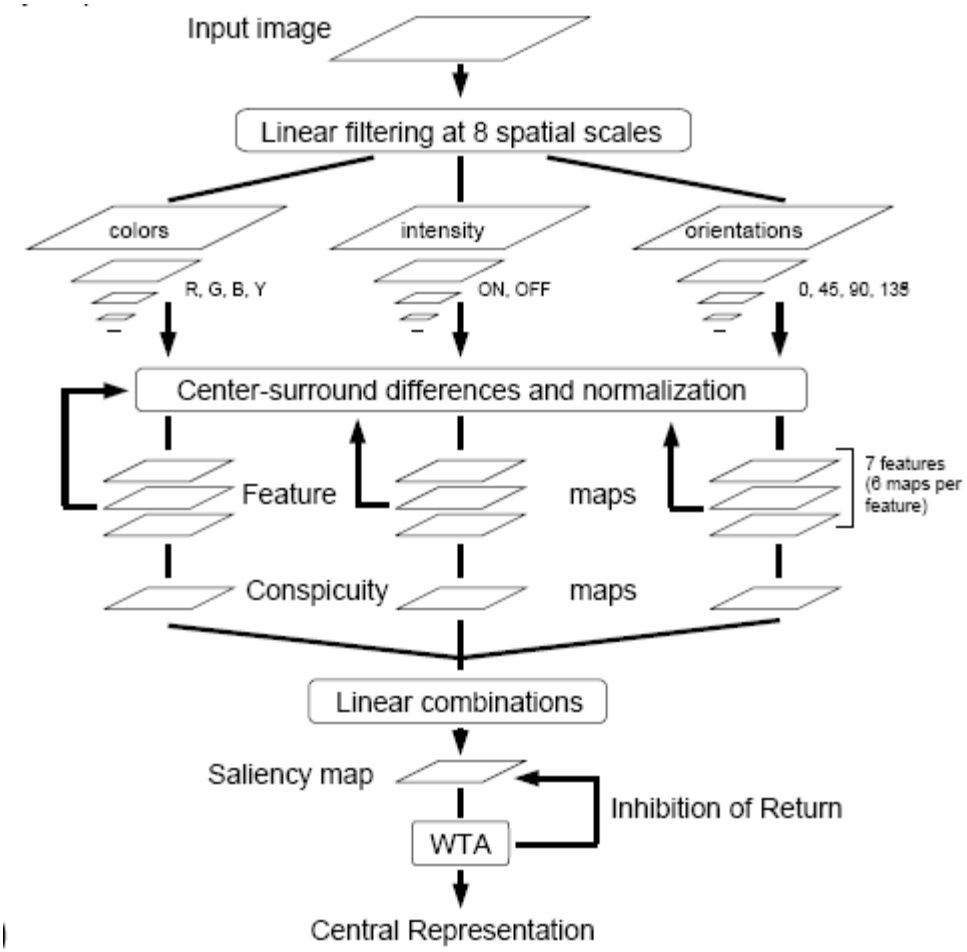


Fig.1. Computational model for vision in primates given by Itti, Koch & Niebur [1]

There have been some efforts to extend this model like the one by Navalpakkam & Itti [5]. This model intends to integrate the goal-driven top-down and the image-driven bottom-up components. Here the top-down component uses the previous knowledge to guide or tune the bottom-up maps in such a way so as to maximize target detection speed. This model is interesting because it proposes a possible model of interaction between the bottom-up and top-down components and gives us more to ponder about.

A study by Navalpakkam & Itti [6] provides direct experimental evidence that humans select visual cues to maximize the Signal-to-Noise-Ratio (SNR: ratio of useful to irrelevant information) between the desired target and its surroundings. The optimal cue selection strategy is selected by maximizing this SNR value. This optimal strategy successfully accounts for phenomena in visual search behavior like the effect of target-distracter discriminator, uncertainty in target's features, distracter heterogeneity, and linear separability.

I also studied selected publications on Consciousness. I found the mind-body problem to be a very interesting one. The **mind–body problem** is the one that involves defining the relationship between mind (or mental processes) and bodily states or processes. The perceptual experiences are said to be aroused from the stimuli received at the various sensory organs from the external world; these stimuli cause changes in the states of our brain which in turn leads to sensations (or feelings): either pleasant or unpleasant. For example, someone's desire for a slice of 'pizza' will tend to cause that person to move his body in a certain manner and direction in an effort to obtain the needful. The intriguing question here is that: how is it possible for conscious experiences to arise out of an inert lump of gray matter endowed with electrochemical properties? How does someone's desire cause that individual's neurons to fire and his muscles to contract in exactly the right manner (especially among babies)? These are some of the puzzles that have kept philosophers of mind interested for a long time.

There have traditionally been two schools of philosophies that follow two separate approaches to solve this problem: *Dualism* and *Monism.* The Dualist approach to the solution states that the mind and body are two separate entities and stresses that mind (separate from the brain) is non-physical substance with *consciousness* and *self-awareness.* The Monist approach to the solution, in particular the Physicalistic monism, states that there is only one fundamental substance and is physical in nature (i.e. mind ~ brain). The other concepts inspired by this problem are the notion of *Strong* Artificial Intelligence (AI) that aims at making computers with some form of consciousness, in contrast with *Weak* AI that solely aims at simulating mental states without stressing on consciousness. *Neuroimaging* procedures like fMRI (Functional Magnetic Resonance Imaging) have shed some light on this problem by helping us better study and understand the functioning of brain.

**Koch & Hepp** [7] explore the scope of *Quantum Mechanics* (and *quantum computation*) in place of the traditionally thought about enormous computational power an interaction among the neurons (explained purely in terms of neurobiological framework) as the basis for understand the higher level functions in the brain including *consciousness*.

Peters, Iyer, Itti & Koch [8] suggested that attentional guidance may also depend upon the interactions among the features instead of solely depending upon local visual features. Peters & Itti [9] combine bottom-up features with the

task-dependent top-down features to find large improvement in the predictions as compared with a purely bottom-up model. The task-dependent features are got by extracting a *gist* from each frame and comparing that with a database of eye position training frames to produce an eye position prediction map.

Peters and Itti [10] focused on designing heuristics that may be best suited for virtual agents, with human like visual attention, operating in the complex dynamic virtual environments like video games. They found that the heuristics which detect outliers from the global distribution of visual features as better predictors of gaze in humans than purely local ones. Further they also found that the heuristics sensitive to dynamic events performed best overall. Their findings also suggested simple neurally-inspired algorithms as better predictors of where humans may look while interacting with such environments.

**Rimey & Brown** [11] explore the use of Hidden Markov Models (HMM) in controlling the acquisition of visual information. HMMs can be used to model probabilistic sequences in the form of Markov chains. Studying the factors affecting the hidden transition probabilities in the HMMs looks like a very interesting topic.

**Practical Work**

The task assigned to me involved collecting both centre-biased and non-centre biased clips. These were divide into 5 sub-categories discussed below:

- *Stationary*: capturing the clip with the camera fixed at a location
- *Pan*: making an oscillatory motion of the form left → normal → right → normal with a symmetric span of $180^0$ at constant speed (6 deg/sec)
- *Tilt*: Up and down motion at constant speed; to overcome horizon bias
- *Follow*: Following different objects
- *Random*: arbitrary camera motion involving the above categories and zoom in/out function

The equipment used was the Sony DCR-HC21 Digital Video Camcorder and a standard tripod stand (approx. 5 feet in height) that had the following major specifications:

- Video Signal: NTSC color; EIA standards
- Usable cassette: Mini DV Cassette
- Recording: SP (high-quality): 60 minutes
  LP (long play): 90 minutes
- FF/Rewind time: 2 min 40s
- Image device: Approx 680 000 pixels
  Effective (movie): Approx. 340 000 pixels
- Power Requirements: DC 7.2 V (battery pack) / DC 8.4 V (AC Adaptor)
- Mass: 460 grams

The lab got a remote-controlled pan-tilt head in the mid-April that had the following specifications:

- Power consumption: alkaline battery AA*5 or 7.5V/0.5A DC adapter
- Pan total angle: 120 deg (60* each side from the normal)
- Tilt total angle: 20 deg (10* from the normal on each side)
- Pan speed: 6 deg/sec
- Tilt speed: 4 deg/sec
- Max remote controller distance: 10 meter

The main challenge was to collect non-centre biased clips to use them as stimuli in future experiments to test if the centre-biased eye-motion in humans (especially after jump-cuts) is a tendency because of some in-built impulse or because of the nature of the task itself.

Apart from that, it proved to be very difficult to get constant-speed pan and tilt clips manually made more difficult by the prevailing conditions outdoor like the

surface and wind. Further, the pan-tilt head and the remote had problems with functioning properly in open spaces, for example, the head would simply get stuck sometimes and not respond to the remote. On other occasions, the motion during the auto-pan was very shaky especially in windy conditions.

I collected clips from different locations in Los Angeles and San Diego. Natural, open spaces with more people were considered ideal to make the stimulus as good as possible in terms of lesser bias (i.e. avoiding large differences in saliency between the different objects in the scene). By choosing natural, open surroundings with more people it provided the subject with a wide range of tasks to select from a particular scene thus increasing the inter-subject variability. Major locations that I covered were:

- Coronado Islands, San Diego
- Santa Monica Beach, Santa Monica
- Cabrillo Monument, San Diego
- Redondo Beach
- Various Locations in Downtown, LA
- Sunset Cliffs, San Diego
- Various locations at USC (including a Football practice)
- Natural History Museum, LA
- Manhattan Beach
- Various locations in Malibu, CA
- Griffith Park, LA

I am very much interested in diversifying and enlarging this existing collection of clips in the future.

**Ideas for furthering the research in iLAB**

Instead of categorizing bottom-up and top-down influences as two discrete components, one can think of them as a *fuzzy logic* or a *continuum* of the form *[bottom-up, top-down]* i.e. instead of categorizing any particular eye movement (saccade, smooth pursuit/follow or fixation) as being caused solely due to any one type of influence (i.e. top-down or bottom-up) one can classify them into gradients or *degrees of truth* in the fuzzy-logic sense. So, instead of defining any one type of influence (bottom-up or top-down) as the sole basis for an eye-movement (like a saccade), a better question to ask would be: *how much of bottom-up and how much of top-down component was involved.*

I put forward a technique that can be used to study the top-down and bottom-up influences and their interaction. This involves showing a video-clip, preferably a thrilling movie clip, to at least 3 subjects. Then, blanking a scene (frame or set of frames) from that clip when shown to at least one and at most n-1 (n is the total number of subjects) preferably half (n/2) of the subjects. And then, comparing the eye-tracking data for all the subjects especially between these two categories of subjects (i.e. for those to whom the movie clips was shown as a whole and to those for whom the scene was blanked off) and also among the group of subjects for whom the particular scene was blanked. We must exclude the eye-tracking data collected in the interval when the scene was blanked when comparing these two categories of subjects and also among the all the subjects to whom the scene was blanked (because, comparing the visual response of the subjects during this interval would be misleading because blanking the scene is expected to produce arbitrary visual responses). Note that the scenes are blanked for maintaining the temporal relation between the stimulus; the other, more complex way around will be to simply cut-off the target scene for the respective subjects and then do some calibration to find the respective interval; this may also throw up the effects arising due to so called jump-cuts i.e. sudden transitions from one scene to another. The main idea here is to compare the eye-tracker data for these categories and plot the results using some metric like the DOH (difference of histograms) to observe interesting patterns after the occurrence of the blanked scene.

This can also be used as a measure (or heuristic) to compare the saliency across scenes i.e. to find the more salient scene/frame among the other frames of the movie. The target frame that causes maximum disagreement across the future eye-positions among all the subjects can be called as the more salient frame across the movie.

For illustration, let us consider a crime scene, with a possible suspects appearing in it, as the target frame(s) (i.e. the frame(s) to be blanked) from the movie. The subjects are expected to have higher agreement when this target frame is not blanked for any of them and the suspect appears in any future scene of the movie. Now, consider the case where this target frame was actually

blanked for some subjects, then the subjects, between these two groups or only from the group, in which the scene was blanked, are expected to have higher degree of disagreement when these suspects appear in any future scene.

To maximize this effect one could cut multiple scenes. One can also think of better ideas to improve this technique.

It may also let us study if the most salient or bottom-up scenes are the ones that affect the future top-down instincts of the observer the most or not. This could be done by controlling the stimulus after this blanked frame.

In a totally different context, I used my knowledge from *Artificial Intelligence* in general to propose the use of *machine learning* and *knowledge bases* to detect deadlocks in distributed systems in the term paper named *A Speculative Approach to Deadlock Handling* [12] as a part of the course: *Advanced Operating Systems* taken during the semester. The idea here is to use the Machine learning tools to learn the behavioral patterns of the distributed system over time and then this information can be modeled in form of either Resource Wait for Graphs (used to detect cycles suggesting deadlocks in the system) or as Knowledge Bases with required rules to predict deadlocks in the system.

Given the opportunity, I can contribute by using my Mathematical and Computational Science knowledge towards the theoretical research work involving development of Computational models for better understanding Vision. Further, I can use my strong C/C++ programming skills to develop and improve different types of programs used for research at the iLAB.

**References:**


(1) L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. **20**, No. 11, pp. 1254-1259, Nov 1998

(2) L. Itti, N. Dhavale, F. Pighin, Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention, **In:** *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology,* (B. Bosacchi, D. B. Fogel, J. C. Bezdek **Ed.**), Vol. **5200**, pp. 64-78, Bellingham, WA:SPIE Press, Aug 2003

(3) V. Navalpakkam, L. Itti, Modeling the influence of task on attention, *Vision Research,* Vol. **45**, No. 2, pp. 205-231, Jan 2005

(4) L. Itti, Models of Bottom-Up and Top-Down Visual Attention, California Institute of Technology, Jan 2000. *[Ph.D. Thesis]*

(5) V. Navalpakkam, L. Itti, An Integrated Model of Top-down and Bottom-up Attention for Optimal Object Detection, **In:** *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1-7, Jun 2006

(6) V. Navalpakkam, L. Itti, Optimal cue selection strategy, **In:** *Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005),* pp. 1-8, Cambridge, MA:MIT Press, 2006

(7) Koch C. and Hepp K. Quantum mechanics in the brain. *Nature* (2006) **440**, 611-612

(8) R. J. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images, *Vision Research,* Vol. **45**, No. 8, pp. 2397-2416, Aug 2005

(9) R. Peters, L. Itti, A computational model of task-dependent influences on eye position, **In:** *Proc. Vision Science Society Annual Meeting (VSS06),* May 2006

(10) R. J. Peters, L. Itti, Computational mechanisms for gaze direction in interactive visual environments, **In:** *Proc. ACM Eye Tracking Research and Applications,* pp. 1-6, 2006

(11) Rimey, R.D., Brown, C.M., Controlling Eye Movements with Hidden Markov Models, *International Journal of Computer Vision* (7), 1991, pp. 47-65

(12) Sahai, A., A Speculative approach to deadlock handling, Spring 2006 *(Term paper)* Advanced Operating Systems (CS 555), University of Southern California

**Appendix:**

(1) 30 sec ".wmv" clips in **/lab/beo1/drs06/video- camera/ Processed_WMVs/**

(2) 30 sec ".mpg" clips in **/lab/beo1/drs06/video-camera/Processed_MPGs/**

(3)C module for Calibration (finds optimal delay): **/lab/beo1/drs06/Callibration_Smooth_Pursuit/ Follow_Callibration.cc**