2004

Goal Directed Visual Search Based on Color Cues: Co-operative Effectes of Top-Down & Bottom-Up Visual Attention

Vishal S. Vaingankar

Goal Directed Visual Search Based on Color Cues: Co-operative Effects of Top-Down & Bottom-Up Visual Attention

Thesis report submitted in partial fulfillment of the requirements of the degree

Master of Science in Computer Science

May 2004

Vishal S. Vaingankar

Advisor: Dr. Roger S. Gaborski 11 May 2004

Reader: Dr. Roxanne L. Canosa 5/11/04

Observer: Dr. Ankur M. Teredesai 5th May 2004

Thesis/Dissertation Author Permission Statement

Title of thesis or dissertation: Goal Directed Visual search Based on Color Cues : Co-operative effects of Top-Down & Bottom-Up Visual Attention

Name of author: VISHAL VAINGANKAR
Degree: Master of Science
Program: Computer Science
College: Golisano College of Computing and Information Sciences

I understand that I must submit a print copy of my thesis or dissertation to the RIT Archives, per current RIT guidelines for the completion of my degree. I hereby grant to the Rochester Institute of Technology and its agents the non-exclusive license to archive and make accessible my thesis or dissertation in whole or in part in all forms of media in perpetuity. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Print Reproduction Permission Granted:

I, Vishal Vaingankar, hereby **grant permission** to the Rochester Institute Technology to reproduce my print thesis or dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Signature of Author: Date: 5/13/04

Print Reproduction Permission Denied:

I, _____, hereby **deny permission** to the RIT Library of the Rochester Institute of Technology to reproduce my print thesis or dissertation in whole or in part.

Signature of Author: _____ Date: _____

Acknowledgements

I would like to thank all the people who have supported me through my master's education. A special thanks to Dr. Roger Gaborski for providing me guidance and support during the thesis process, and Dr. Roxanne Canosa for her significant contribution with this thesis work. I would like to thank Dr. Gaborski, my mentor, for inspiring me with his insightful and refreshing approach to conducting research. I would like to thank Dr. Canosa for training me to use the eye-tracker and providing me with the eye-tracking software. Her patience and constant guidance helped me conduct the experiments with utmost ease. In addition I must thank Dr. Ankur Teredesai for all the co-authored papers, and for guiding me through technical writing skills during the paper writing process. Thanks to Dr. Jeff Pelz, Center for Imaging Science, RIT for granting me the access to the eye-tracker equipment in the Visual Perception Lab. Also I appreciate the support of my comrades from the computer vision and data mining group in the Laboratory for Applied Computing.

Abstract

Focus of Attention plays an important role in perception of the visual environment. Certain objects stand out in the scene irrespective of observers' goals. This form of attention capture, in which stimulus feature saliency captures our attention, is of a bottom-up nature. Often prior knowledge about objects and scenes can influence our attention. This form of attention capture, which is influenced by higher level knowledge about the objects, is called top-down attention. Top-down attention acts as a feedback mechanism for the feed-forward bottom-up attention. Visual search is a result of a combined effort of the top-down (cognitive cue) system and bottom-up (low level feature saliency) system.

In my thesis I investigate the process of goal directed visual search based on color cue, which is a process of searching for objects of a certain color. The computational model generates saliency maps that predict the locations of interest during a visual search. Comparison between the model-generated saliency maps and the results of psychophysical human eye-tracking experiments was conducted. The analysis provides a measure of how well the human eye movements correspond with the predicted locations of the saliency maps. Eye tracking equipment in the Visual Perceptual Laboratory in the Center for Imaging Science was used to conduct the experiments.

Table of Contents

List of figures

Chapter 1: Introduction

1.1 Focus of Attention and Visual Search

Why do the red and yellow street signs stand out in our field of view when we are driving? We generally pay attention to only the important objects while driving, such as pedestrians, oncoming traffic, street signs etc. and generally ignore the non-important objects such as stones lying on the road side. Humans observe their surroundings by focusing on the interesting aspects and ignoring the non-interesting ones. We achieve this by making rapid (saccadic) eye movements that guide the fovea (region of maximal acuity in the retina) to the region of interest. Thus focus of attention acts as a gating mechanism to the higher level visual processes such as object recognition, categorization etc. The filtering theory of attention (Broadbent, 1958) stated the need for attention for processing required information and ignoring the rest. Focus of attention can be classified into two types: Bottom-Up and Top-Down focus of attention. An example of bottom-up focus of attention is a bright light in a dark room, which tends to stand out due to its contrast with the surrounding darkness. A red patch on a green area stands out by virtue of its color contrast. Objects that pop-out solely based on their stimulus features saliency or low-level (intensity contrast, color contrast) feature conspicuity capture attention in a bottom-up manner also known as exogenous control of attention.

Another kind of attention called top-down attention (endogenous control) guides attention based on subject's intentions and expectations. An instance of endogenous control is evident during visual search. In visual search the prior knowledge about the target object's features guides the attention. Thus visual search is a mechanism of searching for an object based on the previous knowledge about its features such as shape, color, motion etc. Often during visual search, our attention involuntarily is directed to some task irrelevant conspicuous objects, such as a bright object, regardless of the search for the cued object. Objects that share similar characteristics with the cued object also influence our attention. An example of this is searching for a person wearing a red shirt in a crowd. Attention is diverted to objects similar in color to red, such as red hat, red tie etc. Visual search can be said to be a combined effort of the top-down (cognitive cue) and bottom-up (low-level feature conspicuity) processes.

Studies of eye movements during perception of the environment can lead us to important information regarding strategies used for observing the environment. The eye-movements of a subject during a psychophysical experiment is an external indication of the subject's selective visual attention. A sequence of eye-movements can be an efficient indicator of the strategies used by a subject. These sequences of eye-movements (scan-paths) are different for different scenes. Analysis of the scan-paths can result in interesting theories regarding scene interpretations.

1.2 Problem definition

In my thesis I am proposing a computational model of color visual search. The developed framework highlights the interaction between bottom-up and top-down focus of attention mechanism. Bottom-up attention is computationally modeled using bottom-up saliency mechanism (topographic saliency maps) and top-down attention is modeled using a neural network trained on target color. The visual space is represented in the form of a saliency map that topographically codes the spatial locations based on feature saliency. A bottom-up saliency map represents saliency of stimulus features based on low-level feature competition. A top-down saliency map represents saliency of the same locations based their relevance to the current task at hand.

The goal of the thesis is to show the relative influence of the two attention systems on visual search. To prove the plausibility of the model, the analysis phase will show a comparative study of the model predictions and eye-tracking generated observations. This will highlight how well the computationally predicted spatial locations correlate with the eye-tracking fixation data collected using human subjects.

1.3 Thesis outline

The thesis is outlined as follows: Chapter 2 provides background on the human visual system, computational modeling of neuron properties, focus of attention, psychophysics of color, and visual search. Background material is provided on the related work in this domain. I will also provide a brief introduction on the neural correlates of these theories based on the literature which approaches it from a neurobiological perspective. This part will explain the importance of attention for implementing visual functions such as feature binding etc. A review of other

computational models on bottom-up focus of attention and color visual search is provided, which have inspired me to approach this problem from a different perspective. A description of some of the most influential computational models will lay the ground work for explaining my contribution in chapter 3.

Chapter 3 will highlight the computational modeling of visual search based on the two attention systems. The chapter provides a model diagram highlighting the various steps in the model. The chapter shows examples of the resulting bottom-up, top-down and the final attention maps.

Chapter 4 discusses the human eye tracking experiments conducted to provide a psychophysical evidence for the proposed computational model. An Applied Science Laboratory eye tracker is used to monitor eye movements of subjects during scene perception. This chapter will provide a detailed description of the experimental methods and procedure carried out to conduct the experiments. A brief introduction of the eye tracking equipment is provided.

Chapter 5 describes the results of the analysis conducted to correlate the saliency map and eye tracking data. A correlation measure between the computationally generated attention maps and the human eye-tracking data will provide a comparative analysis. Graphs will quantify the relative influence of the two attention systems on visual search.

Chapter 6 will provide the conclusions and recommendations for future work.

Chapter 2: Background

2.1 Human Visual System

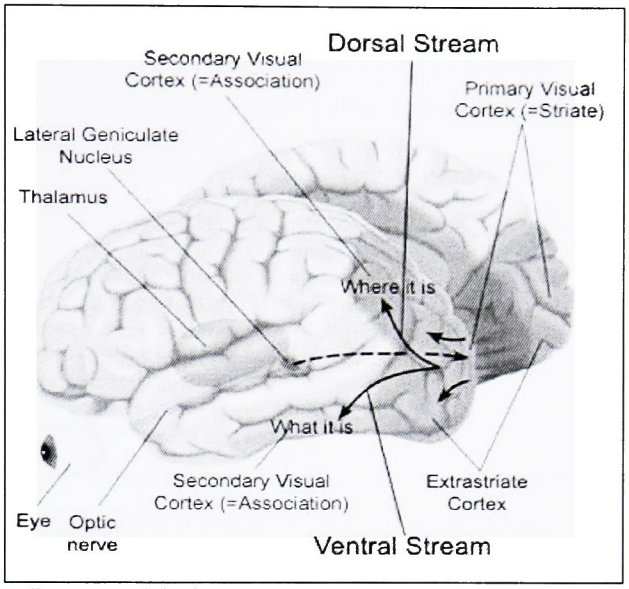


Figure 2.1: Diagram of the human brain emphasizing the visual cortex area. (Adapted from http://www.colorado.edu/epob/)

The visual information entering the eye passes through several visual areas before any sense is made about the semantics or objects present in the scene. This visual information undergoes transformations in a pathway called the visual pathway. Figure 2.1 shows the regions of the human brain related to visual perception. The visual pathway starts at the retina and extends all the way to the higher visual cortical areas. Light entering the eyes excites light absorbing photoreceptors on the retina. Photoreceptors are of two types: rods and cones (Brindley, 1970). Cones are responsible for color vision and high visual acuity. Rods are used during low-level illumination such as in the dark. Cones of three types have been identified, short, medium and long wavelength receptors. Light entering the eye on its own is not colored, rather the photoreceptors that absorb the light of various wavelengths help in color perception. Short wavelength cones absorb light of wavelengths that have low wavelength on the visible spectrum. Similarly the medium and long cones absorb medium and long wavelengths on the spectrum respectively. Combination of the three photoreceptors produces a myriad of color perception (Palmer, 1999). The input from these photoreceptors feeds into the retinal ganglion cells also present in the retina.

Retinal ganglion cells with their center-surround organization of concentric receptive fields, receive inputs form the receptors. The center-surround nature of neuronal receptive field can be seen at multiple layers in the visual cortex. Receptive field is a term for cells sensory pattern which receives impulses from a set of retinal receptors in an excitatory and inhibitory fashion. During firing of a neuron, light is focused on the receptors which produces a rise in the activity, these receptors form the excitatory part of the neuron's receptive field. Similarly light focused on receptors that produce a reduction in activity of the neuron under study, form the inhibitory part of the neuron's receptive field. This mapping of the excitatory and inhibitory receptors forms a receptive field of a neuron. Cells with on-center/off-surround receptive field organization get excited when light shines in the center of receptive field and it is dark around. Cells that get excited when it is dark in the center and light shines around are called off-center/on-surround cells. Receptive field acts as a transformation between the retinal image and the neuron's response. The visual information then passes via an optic nerve to the optic chiasm. The optic nerve is a group of nerves that connects the eye to the brain. From the optic chiasm the optic tract leads to the Lateral Geniculate Nucleus (LGN). From the LGN the visual information flows to the back of brain where the visual cortical areas are located.

The visual cortex is divided into several areas, each specialized for processing specific visual information. The areas are hierarchically arranged starting with the primary visual cortex (V1, V2) leading into V4, infero-temporal regions, parietal regions etc. The visual areas V1 and V2 have been studied extensively in the past (Hubel & Wiesel, 1962). Their findings showed that neurons in V1 are tuned to simple oriented line gratings. The V1 cells were classified as being simple or complex cells. Simple cells increased in activity when an oriented edge falls in the excitatory part of the receptive field. The complex cells tuned for the same orientation as the simple cell, gets excited regardless of the position of the bar in its receptive field. Neurons in the striate cortex are tuned to spatial properties like edges and spatiotemporal motion. The feature complexity (specificity) increases as the visual information passes through to the higher visual areas (Poggio & Riesenhuber, 2002). Visual cortex neurons increase in their specificity and feature complexity in the higher visual areas. Humans are known to achieve object recognition and categorization in 300 ms from the time the light enters the eye. Such high speed processing is possible because of the functional specialization of the various brain regions.

2.2 Computational modeling of the neuron properties

The electrophysiological study of the cat brain has led to insights into the structure of the visual cortical neurons, brain regions etc. Microelectrode recordings of the neurons have discovered the receptive field structure of these cells. Through this information various mathematical approximations of the receptive fields have been proposed.

Ganglion cells responses exhibit a contrast sensitivity response curve as shown in the Figure 2.2. The cells give the highest response to sinusoidal grating of 4-5 cycles/degree and the response falls off gradually with lower or higher frequency gratings. The peak in the contrast sensitivity curve indicates the grating's perfect tuning with the cells receptive field. Mathematically the basis function of the receptive field structure can be modeled using a first derivative of Gaussian or difference between two Gaussians. Young (1985) explains the physiological fit of the Gaussian derivative model to the physiological records of the simple cells spatiotemporal receptive field structure.

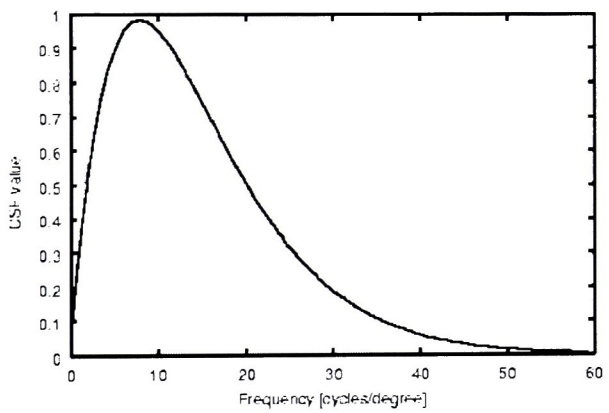


Figure 2.2: Contrast Sensitivity Function (adapted from Krešimir Matković thesis dissertation)

Simple and complex cells in the V1 area of the primary visual cortex are tuned to respond to oriented edges and bars. Jones & Palmer (1987) showed that 2D gabor filters can be effectively used to model the receptive field of a simple cell. Their model results were compared with the physiological data to produce this approximation. A gabor basis function is a product of a sine or cosine wave with a Gaussian function. A sine gabor wavelet is designed to detect oriented edge and a cosine gabor wavelet detects an oriented bar. (Olshausen & Field, 1996) give a complete

overview of using basis functions as a means for representing the receptive field structures for natural scenes.

Figure 2.3 and 2.4 illustrate the activity of simple and complex cell when a slit is arranged on the receptive field. Figure 2.3 shows the response of an on-center simple cell. The stimulus line at the bottom of the figure is the duration for which the stimulus was turned on and then turned off. The top most recording with a slit of perfect (optimum) size, orientation evokes a high activity indicated by the bars on the activity scale. Second arrangement, the slit lies on the inhibitory part of the on-center cell, this produces no response and only producing an off discharge once the slit is turned off. Third arrangement, the slit covers a part of the excitatory and the inhibitory part, thereby producing no cell activity. In the fourth experiment the entire receptive field is illuminated, no response is produced.

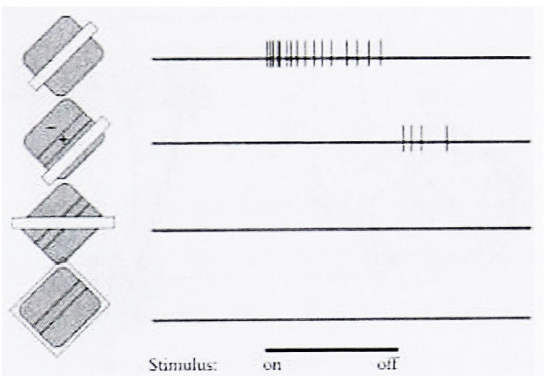


Figure 2.3: Simple cell activity for slit of different arrangements in the cell receptive field. (Adapted from "Eye, Brain, and Vision" by Hubel (1988))

Figure 2.4 shows the complex cell response to the same slit placed within the receptive field. The first three records show that the cell exhibits activity for a slit of optimum orientation regardless of where it lies in the receptive field. The final record indicates that a slit with non-optimal orientation produces no response. The complex cells are comparatively larger in size that the simple cells.

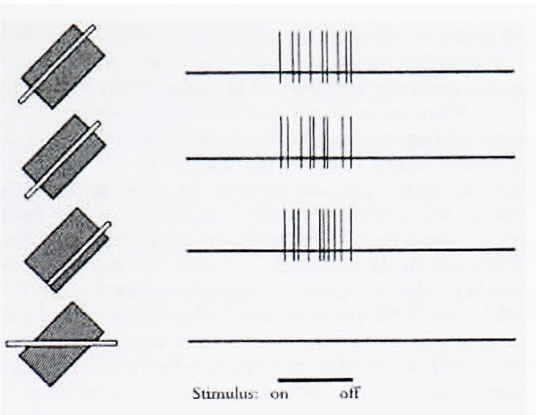


Figure 2.4: Complex cell activity for slit of different arrangements in the cell receptive field. (Adapted from "Eye, Brain, and Vision" by Hubel (1988))

Color opponent cells in the retina and LGN are responsible for color perception. The color opponent theory (Hering, 1964) stated that colors like red and green or blue and yellow cannot be perceived at the same location at the same time. This phenomenon explained what the tri-chromaticity theory (Helmholtz, 1925) failed to account for. According to tri-chromaticity theory the combination of the three primary colors (red, green, blue) gives the myriad of color combinations. The overlap between the response of the three photoreceptors accounts for color perception. Color opponent cells in the LGN have the concentric color center surround receptive fields. Red-center-on/Green-surround-off, Red-center-off/Green-surround-on, Green-center-on/Red-surround-off, Green-center-on/Red-surround-off are the types of color opponent cells.

2.3 Eye movements in Visual Search

Moving the eyes to the objects of interest enables us to process the object in detail. High resolution vision is only available at the central region of the retina (fovea) and the resolution progressively drops off further away from the fovea. Thus we make eye movements to process the information at full resolution at the location of interest. Our eye movements are ballistic with a series of fixations and saccades. Fixation is a momentary pause at a location during the scanning of a scene. A saccade is a movement from one fixation to the other. Recent research has emphasized the use of eye movements in inferring underlying cognitive processing (Pelz et al. 2000). The eyes cannot perceive during a saccade and the retinal image appears as a blur. It is only during the fixation that a stable image of the scene is formed. This succession of a stable percept followed by blurs and then again a different stable image raised the questions about how

the stable perception of the world emerges from this chaotic process (Rayner, 1978). What information is integrated across fixations? What part of the scene is memorized in the previous fixation and integrated with the information obtained in the successive fixations? Research by Palmer and Ames (1992) showed that subjects were able to correctly discriminate between the different shapes of two objects when the objects were viewed in completely different fixations. Another observation by Hayhoe, Lachter and Feldman (1991) indicated that the spatial information about points is held in a map-like representation which is formed over a number of fixations. Their experiments showed that subjects could correctly identify whether points viewed in different fixations formed the percept of a triangle. Such a map like representation is similar to the topographic representation techniques widely used to represent the scene.

Information integration across fixations sheds light on the topic of memory during visual search. Visual search experiments with search items trading spaces every 110ms in the display by Kristjansson (2000) proved the presence of memory. The search times were slower when a target item was relocated to a location previously taken by a distracter item. This behavior indicates that the already visited distracter locations are remembered and are least likely to be revisited. Contrary to this understanding, Horowitz and Wolfe (1998) showed that visual search lacks any memory. Their experimental methods also used a scene in which the items were randomly relocated regularly during the search so that it would be difficult to track search items based on their location and appearance during the search. But their results did not show a drop in search efficiency despite the relocation, indicating that visual search does make explicit use of memory. According to the previous beliefs (Klein, 1988), during search the already visited locations are tagged so that they are inhibited from being visited again. For a scene with constant relocation of items, this approach of tagging would result in the degradation of search efficiency.

During a fixation, which approximately lasts for 150 to 350 ms, the next fixation location is decided. Attending to objects without eye-movements is called covert attention. Focusing of attention by eye movement or head movement is called overt attention. Liversedge & Findlay (2000) analyzed how this covert attention interacts with the overt eye movements. Could multiple covert visual scans precede an overt eye movement? The rate of redeployment of covert attention is observed to be 30 ms per item. This would allow several of these covert movements

during fixation duration of 300 millisecond. The proximity of the target to the current fixation location vastly affects the search times. It is proved that the probability of locating the target is higher if the target lies close to the current fixation (Findlay et al, 2001). This emphasizes the possible role of covert attention for correctly locating the target if it lies within close proximity than locating a target which is further away. From the review of the literature on visual search, the factors like attention, memory across fixations, semantic information affect the search times.

The study of the patterns of eye movements while viewing a scene has lead to some interesting observations. The finding by Parkhurst et al.(2002) showed that the early fixations of a scan are influenced more by the bottom-up features saliency in the scene, but as the scene scan time increases, the later fixations are focused more on the semantically meaningful regions that is mediated by the top-down processes. Their research explains this effect based on studies using human subjects observing natural, synthetic, fractal scenes.

2.4 Focus of Attention

Given the enormous amount of information flowing down the optic nerve, it is not possible that every bit of this information can be processed by the visual system in a short amount of time. The visual system has to decide on the locations to selectively attend to based on stimulus feature saliency or volitional control. Selective attention has a significant importance in scene understand and interpretation. Change blindness studies have proved the importance of selective attention for detecting changes in a scene (Rensink, 2002). Change blindness experiments explain the inability to detect major changes in a scene if the changing object is not selectively attended. Without attention, subjects take a long time before they can detect the changes in the scene. Change detection is widely used as a testing paradigm in the psychophysics experiments. A typical change detection experiment shows the original image and then after a brief delay shows the same image with an object disappeared. The goal is to spot this disappearance of the object. Subjects who haven't focused on this changed object have difficulty spotting this change.

Physiologically the visual cortex is divided into two pathways, the dorsal and ventral visual pathway. The dorsal pathway also known as "Where" pathway is responsible for spatially locating the object. The ventral pathway called the "what" pathway is responsible for object

recognition. Neural correlates of visual attention have been studied extensively in the past (Desimone & Duncan, 1995). Their findings suggest that attention causes an increase in sensitivity of the neurons of the V4 and IT (inferotemporal) cortex area. Visual attention modulates cells in almost all visual areas including the primary visual cortex. Selective attention at the cell level is influenced by the competition between stimuli lying within the cell receptive field. Their study indicates that a single stimulus within the receptive field increases the firing rate. With two stimuli within the receptive field, the cell will respond differently depending on its selectivity. In case of a good and poor stimulus, if attention is directed to good, the response to the combination will increase to a level comparable to the single stimulus response. The same cell's response will decrease if the attention is focused on the poor stimulus. Here the good and poor state of stimuli is with respect to the selectivity of the cell.

Computational modeling of selective attention has been successfully approached in the past. Work of Koch & Ullman (1985) represents the relative importance or conspicuity of objects based on their surroundings in topographical map called the saliency map. A winner take all network then selects the location with the highest saliency in the saliency map. The basic theory behind this implementation is that the competition between the different saliency maps based on lateral inhibition eventually comes up with a winning location. This theory is entirely bottom-up in nature without any higher level knowledge feedback. This bottom-up saliency mechanism was later computationally modeled by Itti & Koch (2001). Their model represents the input scene with multi-resolution feature maps obtained from spatial filtering in parallel for each feature. The feature channels are intensity contrast, color opponent feature maps, and orientation feature maps. These multi-resolution feature maps are normalized within the respective feature channels and added to form three final saliency maps. These three saliency maps are then normalized and added together, this final saliency map represents the salient locations of the scene. The normalization process simulates the long range inhibition with each map and assigns non-linear weights to the individual maps. The map with large amount of information get the lowest weights and maps with sparse but highly localized regions are assigned highest weights. This kind of non-linear summation assures that the features that are active in only a few maps do not get subdued due to summation and features which are present in multiple maps do not get enhanced. An inhibition-of-return network then sequentially selects the winning location for

focusing. This winning location is then processed for higher level visual tasks. The inhibition-of-return network first selects the most salient region form the saliency map. Once selected, this location is inhibited and the next highest salient region is selected for focusing. This procedure acts as memory during scanning the scene, because by inhibiting a region and then focusing on other salient regions is representative of saying that the location of last focused region is remembered and not attended again. Their model captures the characteristics of an exogenous attention based system.

The theory of focus of attention cannot be complete without mentioning the covert and overt attention mechanisms. The importance of covert and overt attention is evident in psychophysical tasks using dual task paradigm and RSVP (rapid serial visual presentation). In dual task paradigm, the subjects are required to perform an attention demanding task in the central region of the display such as to find if all the letters are the same or not. At the same time while focusing on the central task, perform another peripheral task such as scene categorization etc. The subjects perform the peripheral task using covert attention while overtly fixating at the central task. In rapid serial visual presentation task a sequence of images are flashed one after the other on the screen, each image being visible only for a few milliseconds. The subjects are required to categorize the displayed images into natural or indoor images. With such short durations, subjects can only infer the gist of the scene by covertly attending to certain locations. The dual task paradigm analyses the thresholds of attention required when complete attention is available to a task and when the attention is divided between two tasks.

One of the inherent problems in object recognition has been that of feature binding. When there are multiple objects present in the visual field, the problem of associating or binding the features of individual objects is called as the binding problem (Schoenfeld, 2003). Selectively attending to the objects can help in solving this problem. This issue leads us to the domain of feature integration during visual search.

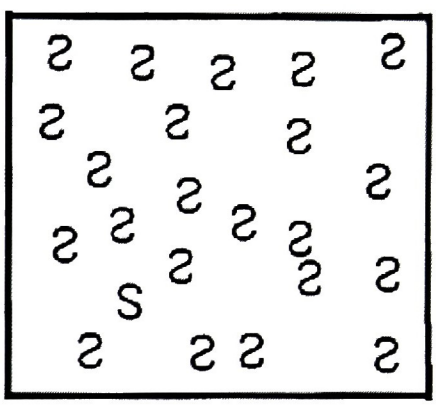
2.5 Visual Search

One of the earliest theories of feature integration was proposed by Triesman & Gelade (1980). Their work categorized visual search into two types; single feature search (disjunctive search) and feature conjunction search. In a disjunctive search the visual search requires search for only one feature, like a search for yellow in distracters of red and green. Conjunctive search involves search for a combination of features, example color/motion or color/orientation objects. Feature searches are fairly simple and occur in parallel, i.e. the target features pop out of the scene. The time taken to search the target object amongst the distracters is called the reaction time. In parallel searches the reaction time is independent of the number of distracters present because the target object tends to pop out of the scene in a pre-attentive manner. In serial searches the reaction time linearly increases with an increase in the number of distracters in the scene. According to Triesman & Gelade (1980), search is serial for target defined by the conjunction of features from different feature dimensions. Serial searches require the analysis of the search array to spot the target. Their model also predicts the use of topographic maps to focus attention on the target objects.

Buracas & Albright (1999) conducted comparative psychophysical studies on humans and macaque monkeys to compare the performance of human and non-human primates in disjunctive and conjunctive visual search tasks. The experiment consisted of briefly presenting a sample stimulus followed by the search array. The subjects were expected to spot the target stimulus from the search array. The stimuli presented were a combination of color, orientation and motion targets. As expected, the human and rhesus monkeys both gave serial and parallel searches as stated before. The feature integration model has met with some criticism lately regarding the serial searches for conjunction of features. Nakayama and Silverman (1986) proved that parallel searches are possible even in the case of feature conjunctions of color/stereoscopic disparity or motion/stereoscopic disparity. For combination of color/motion feature their experiments did give serial searches similar to that predicted by Triesman & Gelade (1980).

Another computational model of visual search that captures the interaction between the bottom-up and top-down attention systems is described in the Guided Visual Search model (Wolfe, 1994). The model extracts a set of stimulus features (color, orientation, motion) in the different feature channels in parallel. Stimulus features are represented in a bottom-up map. The bottom-

up map enhances the features which are unique compared to the surrounding features. Given a scene with one horizontal line between multiple vertical lines, the horizontal feature map will get the maximum bottom-up activation compared to the vertical. If this difference in activation is large enough, feature pop-out effect will take place resulting in a parallel search. The top-down map activates only the target object features. This map topographically represents the input space with high activation for target features. The bottom-up and top-down topographic maps are then summed together to form an activation map. The activation map has regions of high activity that will be focused on during the visual search. Each locus on the activation map is visited sequentially to detect if the target is present. The selection of features for top-down activation in their system is rule based. The rule selects features that give maximum discrimination between the target and distracter rather than just the feature which gives the maximum response. The technique used for each channel is to take the difference between average responses of the channel to the target from the channel's response to the distracter. In my contribution in this thesis, a response from the neural network which has learnt the features of the target is used to create the top-down map.



2.5: search for "inverted S" in a search image containing "S".

Figure 2.5 shows an example of serial search for target "S" amongst the distracters of inverted S. Since the target and distracters share some similar properties, deploying the bottom-up attention will produce high activation all over the visual field. Only with the help of slow top-down attention will the target object be located.

Color Visual Search:

Work by D'Zmura (1991), explains the serial/parallel nature of color visual search when searching for target colors amongst distracter colors. The research reports on the reaction time of human subjects to discriminate between target and distracter objects of different colors in the scene. Color visual search can be classified as being serial or parallel, depending on the amount of color difference between the target and the distracter object. Search in which the target color pops out of the scene due to large color differences between the target and distracter color is known as parallel search. Parallel searches are independent of the number of distracters present in the scene. In serial search the reaction time increases with an increase in the number of distracter objects in the scene due to small color difference between the target and distracter colors.

Their contribution discusses in length the various factors such as color saturation etc. that can affect the search. Nagy & Sanchez (1990) reported results suggesting that longer search times are observed if the color difference is small and shorter search times are seen with larger color difference between target and distracter. The search times were found to increase linearly with the number of distracters when the color differences were small, whereas the search times remained constant when color differences were large.

Work by (D'Zmura, 1991) showed that search for intermediate colors, such as orange can result in a parallel search amongst distracter made up of red and green. Their result emphasized that in this case, the observers used a mechanism tuned to yellow events in the visual field. Similarly orange targets popped out in the presence of yellow and blue distracters. The observers relied on the redness component to detect this orange in parallel. The reaction time graphs showed no increase in reaction time for the above two test cases. Serial searches were observed for orange targets amongst red and yellow distracters. Here since the color difference between the target and distracter is very small, serial scanning of visual field would be required. For the above parallel search cases, the color saturation of the target color was changed to observe if different saturations did affect the search. For three levels of saturations the search for orange amongst red/green and yellow/blue still remained parallel.

2.6 Interaction between bottom-up and top-down attention

Model by Hamker (2000) proposed the use of feedback between successive stages of three types of neurons, object selective cells, feature selective cells and location selective cells. The feedback connections from the higher levels to lower level cells optimizes the weights and makes the cells in the lower level selective to the features of the objects. The distributed competition between the IT (inferotemporal) cortical neurons enable parallel search even with feature conjunctions. Given the fact that the visual cortical areas are influenced by the reciprocal connections between two cortical layers, this model is physiologically plausible.

An iconic based visual search model developed by Rao et al (2002) represents the visual scene using oriented spatio-chromatic filters at multiple scales. The target object is also represented with a set of these filters. The search process proceeds from coarse to fine, comparing the target objects' feature template at coarse resolution with the scene. On finding a match at a coarse level, the search is terminated otherwise the next fine level is used for matching. The task relevant locations are represented in form of a saliency map, which then guides the saccades. The spatiochromatic basis functions are constructed at three scales. Each scale contains nine filters which are the first, second and third order derivative of 2D Gaussian. The feature templates of the target are used to measure the correlation with individual objects in the scene. Since the saliency map represents the locations based on the task relevance, the region with highest saliency is used as a location for shifting the gaze. The model's approach of computing the saliency maps by first using the lower resolution spatial filters, finding the correlation and proceeding to higher resolution, concurs with the studies that the coarse resolution have priority over higher resolution during the search (Schyns & Olivia, 1994).

Human attention varies considerably depending on nature of the task as shown by Yarbus (1967). The locations focused upon in a scene can vary based on the task at hand. One such model that captures this task relevance guiding of attention is described in Navalpakkam & Itti (2002). Their architecture uses four components, a visual brain, long term memory, working memory and agent. The working memory stores a task graph of all the entities that are relevant to the present task. The semantic information of the scene is represented as a graph in which the entities form the leaves and the edges form the relationships between these entities. Long term

memory contains the knowledge of the real world entities that are part of the given scene. The agent relays all this information between the above modules. The relevance of fixations is decided by traversing through the task graph. The working memory checks if a path exists between the current fixation entity to the other entities in the graph. If a path exists, the fixation relevance value is calculated based on the distance between the two entities in the graph. The relevance value decreases as a function of the distance. The final attention guidance map is a pixelwise product of the bottom-up salience map and the task relevance map. This map topographically represents the salience and relevance of each object based on its relevance to the task at hand.

Top-down knowledge about a scene can be a prior semantic knowledge about the scene. For example when asked to search for a printer in an office, the locations like office table or some stable surface are the regions of attention. Whereas the ceiling of the room is highly unlikely to be searched for a printer. Thus here the prior semantic knowledge that printers can be located on tables and not on ceilings can guide search process. The attention guidance model proposed by (Oliva et al, 2003) modifies attention based on the overall scene configurations. Observations from the human eye movements correlate well with the regions predicted by top-down model while search for people in a given scene.

Chapter 3: Computational Modeling

An image processing system that attempts to simulate the mechanisms of a visual system should be able to model the various stages of transformations an image undergoes in the visual pathway before any meaning is extracted about its structure. These transformations should take into account the processing stage at the retina where luminosity and color information is extracted. The image representation at lateral geniculate nucleus (LGN) is where center-surround cells enhance processing of color perception. The next stage is the modeling of neural properties of the cells found in the visual cortex that extract the contours and builds a higher level object representation from this contour image. In the domain of computational modeling, image processing kernels have been proposed in the past that simulate the response functions of the neurons in the retina, LGN and visual cortex. Gabor orientation filters, difference of Gaussian filters, hermite functions are all mathematical approximations of the neural receptive fields and their response. The edge detecting filters are seen as principal components required to represent a natural image without inducing redundancy. The goal is to find as smaller number of units as possible to represent the input space without losing the important information.

3.1 Physiologically inspired kernels

Receptive fields in the visual cortex are localized and have frequency band pass properties. These cells exhibit spatial as well as frequency overlap with the neighboring neurons. This is evident from the orientation selectivity of neurons in a single hypercolumn of striate cortex. The adjacent neurons within a hypercolumn have overlapping orientation tuning curves. Neurons in the primary visual cortex sparsely code the image properties. Sparse coding states that the visual cortex neurons are selectively tuned to a particular stimulus and do not respond to every stimulus in its receptive field. Thus computationally modeling these selective neurons is the goal of image processing filters. Bell & Sejnowski (1997) showed that edge filters are independent components of natural images. The filters developed by their approach of independent component analysis show responses similar to the responses of orientation tuned simple cell neurons. Another approach for building the edge filters which resembles the sparse coding in the cortex is by Olshausen & Field (1996). They use a minimum entropy approach to build a set of small number of descriptors from the available large set. These descriptors are localized, oriented and band-

passed similar to the Gabor basis functions. Sparse coding reduces the redundancy during image representation. It is known that a population of active cells in the visual cortex selectively represents the stimulus. This population of active cells can represent the stimulus in a local, dense or sparse manner (Foldiak & Young, 1995), depending on the number of neurons that are active. Activity ratio is the number of neurons in a population of neurons that are active for a stimulus representation. In a local code, each stimulus is encoded by one unit or neuron. Local code has the lowest activity ratio. In a sparse code the number of active units is relatively low, allowing efficient coding and avoiding the problems of a local code. The Gabor wavelets and difference of Gaussian filters can be used as basis functions to encode the stimulus in sparse manner.

The filters described used in this focus of attention model are all designed at multiple resolutions. This is in agreement with the fact that receptive fields in the retina and cortex have multiple spatial resolutions. The retinal ganglion cells in and immediate surrounding of the fovea are of high resolution (smaller spatial size) and progressively reduce in resolution (larger spatial size) around the periphery. The model described here filters the input image uniformly all over the image with each filter, without attempting to process high resolution at the point of fixation and with lower resolutions around.

Receptive field of the retinal ganglion cells and the lateral geniculate nucleus exhibit a center surround behavior. These cells are of two types, on-center/off-surround and off-center/on-surround. The Difference of Gaussian filter (DoG) is used to model the receptive field properties shown in Figure 3.1. DoG filter is obtained by taking a difference between the response of large width Gaussian and Gaussian of a smaller width. The same receptive field structure can be obtained from a first derivative of Gaussian.

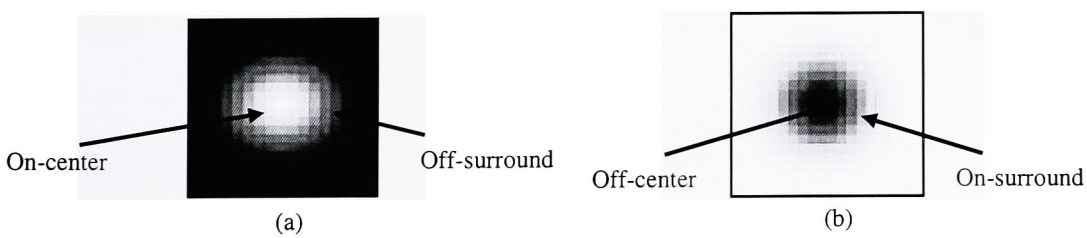
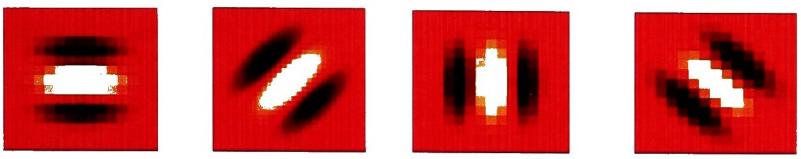


Figure 3.1: Center Surround Difference of Gaussian filters (a) On-center/Off-surround filter. (b) Off-center/On-surround filter.

Gabor filters are used to model the edge and bar detecting cells found in the visual cortex V1. Sine and cosine Gabor filters form a quadrature pair filters to detect orientations of edges and bars in the input scene. Figure 3.2 shows the sine Gabor wavelets.



0 degree. 45 degree. 90 degree. 135 degree.

Figure 3.2: sine Gabor orientation filters. (Bar detection filters)

The 0 degree Gabor filter when convolved with the input image extracts all the horizontal edges from the image. Similarly the 90 degree filter extracts vertical edges. The Gabor filters simulate the processing of the orientation tuned simple cells in the primary visual cortex. Image processed at the retina the LGN is the input to the visual cortex. Therefore the input image to the cortex is the intensity contrast image which is a result of the center surround retinal ganglion cells and LGN. The Gabor filters are convolved with the intensity contrast image.

Color opponent cells

The outputs of the short, medium and long wavelength cones are combined in an opponent fashion. This is based on the opponent theory by Hering (1878), that certain color combinations do not appear together. These color combinations tend to cancel each other. The cells in retina and LGN color are coded by color opponent cells. Figure 3.3 illustrate the different color opponent cells.

(a) (b) (c) (d)

Figure 3.3 : Color opponent cells (a) Red-center/Green-surround (b) Green-center/Red-surround (c) Blue-center/Yellow-surround (d) Yellow-center/Blue-surround

Red-center/Green-surround cells will produce maximum activity if the cell in excited by red in the center and no green around. The opposite is true for the green-center/red-surround cells that show activity for green regions of the scene. The same behavior is observed for Blue-Yellow center surround opponent cells.

3.2 Computational model

Figure 3.4 shows the block diagram of the model. The processing of the system is divided into two processing channels. First channel is called the bottom-up attention. Within this processing domain spatial filtering is applied on the image in the color, intensity and orientation channels. This forms the part of bottom-up attention. Another processing channel called the top-down attention channel processes image in a top-down fashion. Neural network is used to generate the task relevant salient image regions.

3.2.1 Bottom-Up Attention

The multi-resolution Gabor filters, difference of Gaussian and color opponent kernels are used for convolutions. The output feature maps within each channel encode the feature properties. The feature maps generated within each channel are all of the same spatial size but convey different information due to convolution with the multiresolution feature extraction filters. The normalized luminosity image is convolved with the difference of Gaussian filters to generate the intensity contrast feature maps. Luminosity image is calculated using the (r + g + b)/3 method. The on-center/off-surround and off-center/on-surround filters extract different regions of image. The filter gives a zero response to an image patch that has a uniform luminosity without any variation in luminosity. A change in intensity evokes an increased response from the filter. The intensity contrast map is shown in Figure 3.5 (b).

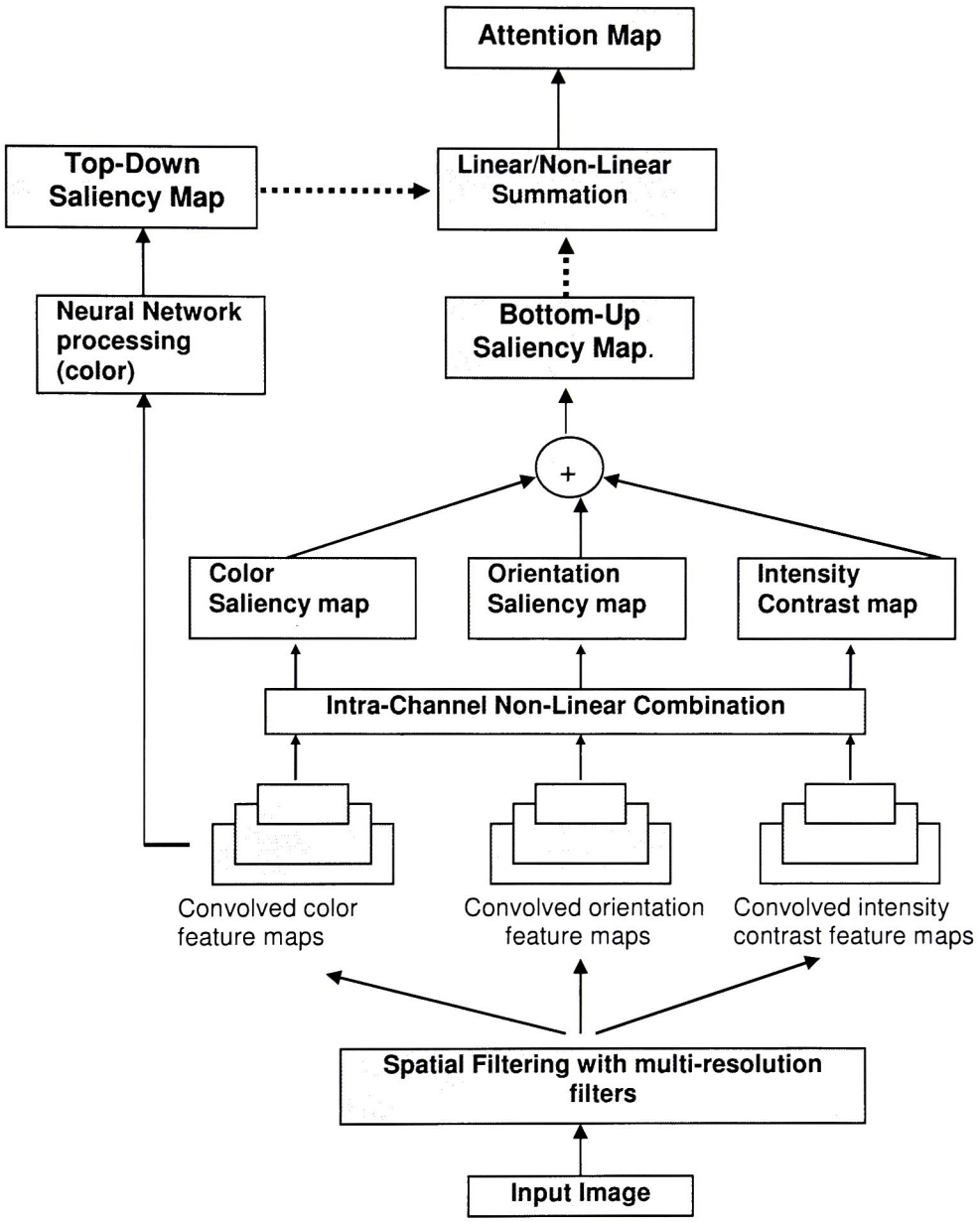


Figure 3.4: System block diagram.

The four color channels (r, g, b, y) are created by removing the luminosity component from the image. The color planes used are: Red (R) = r – (g + b)/2, Green (G) = g- (r + b)/2, Blue (B) = b-(r + g)/2, Yellow (Y) = r + g -2 (|r - g| + b) similar to the ones used in Itti & Koch (2001). r, g, b, y are the normalized color planes. The color opponent filters extract the centre region from

one color plane and surround region from the other opponent color plane. For example a Red-center-on/Green-surround-off filter convolves the center region from red plane (R) and surround region from the green plane (G). The amount of red in center-on forms the positive weights and the green in surround-off forms the negative weight this filter. If the positive and negative weights of the two regions are equal, then they cancel each other and result in zero response. But if there is no green in the surround and only red is present in the center, then this filter gives the maximum response. The color opponent maps are shown in Figure 3.5 (c-d).

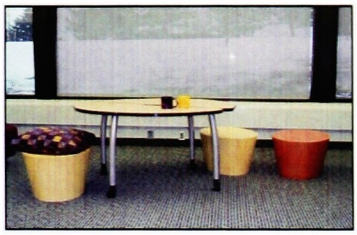
The orientation information is extracted by convolving the intensity contrast maps with the mutiresolution Gabor filters. This approach is different from the one used in Itti & Koch (2001), where the luminosity image is convolved with the Gabor filters. Our approach is more biologically plausible, since the orientation detection receptive fields exist in the primary visual cortex which receives the neuron inputs from the neurons of the LGN and retina (Usrey et al., 2000). Thus the image representation that reaches the V1 area is a center-surround filtered image from the retina and LGN. Thus convolving the Gabor filters with the center-surround intensity contrast image seems to be a plausible approach. The convolution generates 0^{o}, 45^{o}, 90^{o} and 135^{o} orientation feature maps all of same spatial size. The orientation maps are illustrated in Figure 3.5 (e-h)

Within each feature channel, the feature maps are normalized using non-linear weights and summed together within the channel to generate three feature saliency maps: color saliency map, intensity contrast saliency map, and orientation saliency map. These feature saliency maps in turn are combined to generate a final saliency map which is called the bottom-up saliency map. The result of this summation for an example scene is shown in Figure 3.6. This final bottom-up saliency map is representative of saliency across three feature dimensions. The regions in the bottom-up saliency map compete for attention based on the lateral inhibition mechanism.

The within-feature channel summation of the multiple feature maps to generate a single feature saliency map can be achieved using certain normalization techniques. The normalization technique helps in inhibiting maps that posses a lot of dense regions of information and enhances the maps that possess sparsely distributed locations of information. This technique allows for

intra channel competition in accordance with the concept that a highly localized and isolated region in a feature map stands out in the image more than many of these regions in close proximity. Based on the lateral inhibition technique this map with large number of salient regions generates low saliency overall. For example a salient object may invoke response in maybe a single or a few feature maps, while many other feature maps may not show response for this feature. If we apply a simple linear combination strategy, then this single isolated feature may get inhibited due to absence in a large number of maps. This incurs a lot of noise in the final combined map. We need to apply a non-linear combination strategy to sum the maps intra channel and inter channel. The "global non-linear combination" proposed in Itti & Koch (2001) was used to incorporate non-linear weighted summation of maps. This approach starts by normalizing (between 0 to 1) all the maps involved in the summation process so that uniformity is maintained in values of each map. In our case all the maps values were in the range 0 to 1. Then calculate the Global maximum (M) for the map and the mean of all the other local maximums (m) for the same map. Then multiply this map by $(M - m)^2$ factor. The maps with higher difference will get the higher weights than the maps with smaller differences. Then the weighted maps are summed together. This combination strategy is applied for intra-feature channel map combination. The three feature saliency maps: color saliency map, intensity contrast saliency and orientation saliency map obtained from the above procedure are then linearly summed. The combined map is called the bottom-up saliency map shown in Figure 3.6.

In order to confine the focus of attention regions to a few locally salient regions, an iterative lateral inhibition approach was experimented with. The lateral inhibition is applied to the bottom-up saliency map. This technique simulates the lateral inhibition amongst the neurons in visual cortical level. The difference of Gaussian filter with its center surround receptive field induces short range excitatory and long range inhibitory connections to the surrounding regions. Figure 3.7 shows the output saliency maps over a number of iterations of convolving the difference of Gaussian filter with the original saliency map. As can be noticed with the increasing number of iteration, the saliency map becomes sparse in the number of locations shown as salient.

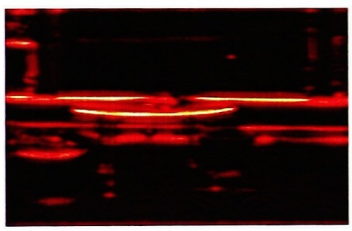


Input image (a)

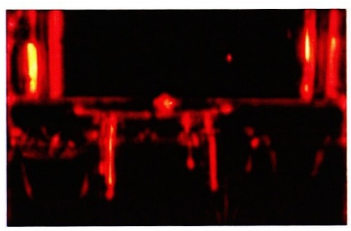
(b) Intensity contrast map

(c) Red green color opponent map

(d) Blue yellow color opponent map

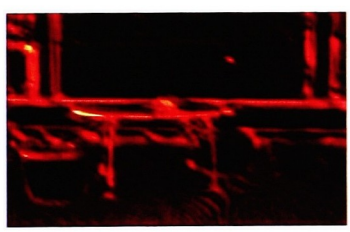


(e) 0 degree feature map



(f) 90 Degree feature map

(g) 45 degree feature map



(h) 135 degree feature map

Figure 3.5: (a) Input image, (b) intensity contrast image, (c-d) Color opponent maps (e-h) Orientation maps

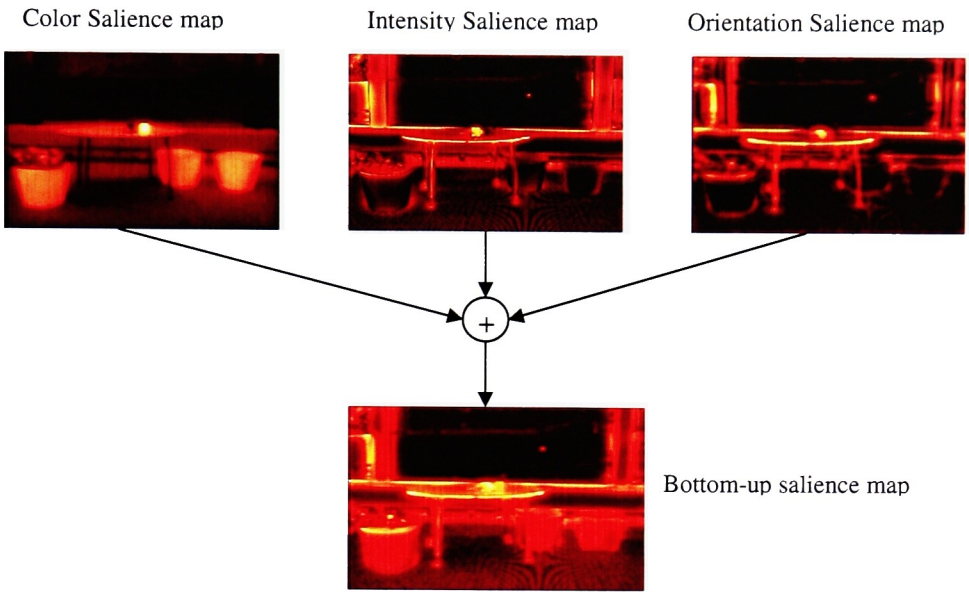


Figure 3.6: Summation of the feature salience maps to generate the bottom-up salience map

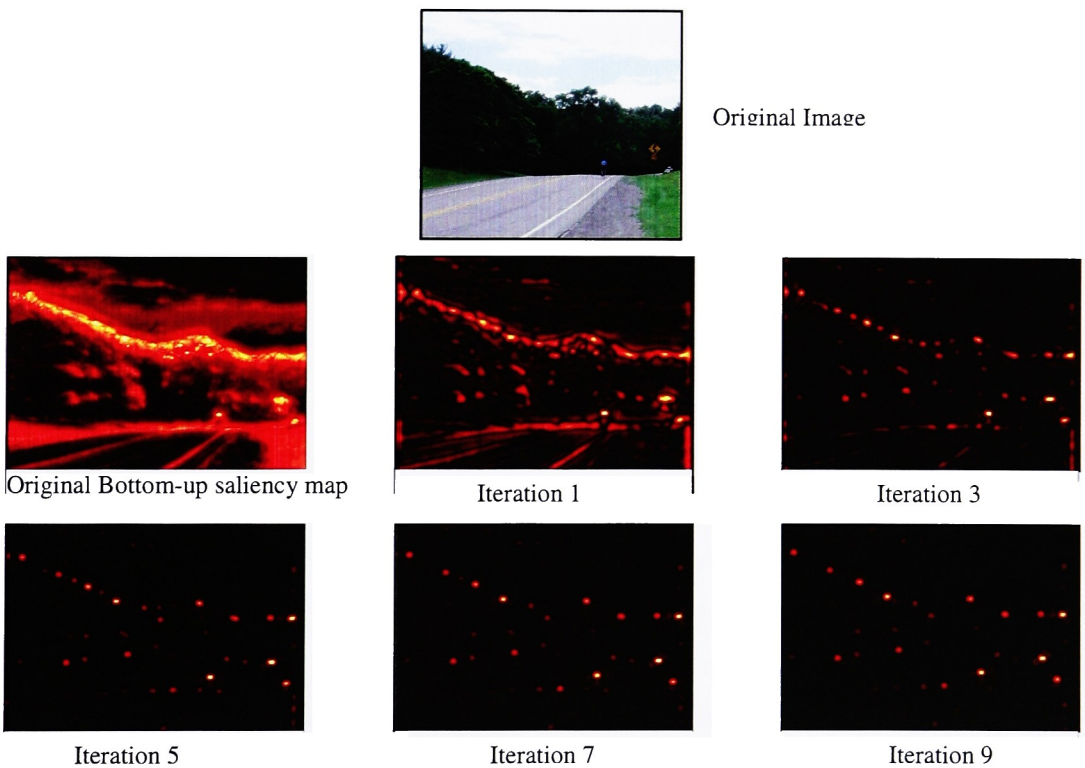


Figure 3.7: Original image and bottom-up saliency maps over multiple iterations of convolution with Difference of Gaussian filters

In Figure 3.7 the original saliency map shows saliency for the bicycle rider, street sign, approaching car and the tree edges. With increasing iterations the tree edges disappear and only a few isolated salient locations are left. The biker, street sign and car are still salient in the map after iteration nine, since these locations encounter least competition from the surround.

The bottom-up saliency map represents object saliency based only on the stimulus feature saliency. Attention is not always bottom-up, as discussed the top-down knowledge of a certain feature modulates the attention. The next section discusses the top-down attention model for a color cues.

3.2.2 Top-Down Attention

During visual search for objects of interest, high-level information about the searched-for object usually guides our focus of attention. We have experimented with color as a high-level feature for visual search. During search for color red the top-down map assigns high saliency to regions that are red and progressively reduced saliency to colors adjacent to red on color spectrum. The model uses neural network to assign relative saliency to regions that posses cue color properties. The output of the neural network is a top-down saliency map, where cue color regions are assigned higher saliency and lower saliency to distracter color regions. Primary colors are red, green, yellow and blue. Intermediate colors are orange, bluish-green etc.

The use of neural network in learning the weights for primary colors and intermediate colors could be challenged by an argument that the color opponent maps could be used as top-down maps. In this section I will provide an explanation in support of the use of neural network. The contradictory comment could be that if a neural network is used to find regions of image which have cue color properties then the color opponent maps could just as well be used to do the same. For example during search for color red, the top-down map invokes high saliency for the regions that are red or similar to red. For this purpose the Red-center-on/Green-surround-off (R+G-) color opponent map could be used as top-down map to locate regions of red. This is true in case of search for primary colors that the color opponent maps can be used as top-down map. But this approach fails if the target color is an intermediate color such as orange which is a combination of red and yellow. During search for orange, the top-down map should ideally give high saliency for orange and reduced saliency for red and yellow regions. The color opponent maps give a

lower saliency for the intermediate colors because they are tuned to detect the primary colors, thus would give orange lower saliency than red and yellow. In that case the top-down map will be a linearly summed map of red and yellow color opponent maps, thereby assuming that both red and yellow have equal weights while representing orange. The assumption that red and yellow have equal saturation in orange is not necessarily true. Some variations of orange have higher red component than a yellow. So a linear sum of the red and yellow color opponent maps does not adhere well with this theory. A solution to this problem of non-linear weighting of primary color during an intermediate color search is to use a tool that can learn the non-linear weights of primary color for representing the intermediate color and assign saliency accordingly. Neural networks provide a sophisticated technique to learn the weights thus is an efficient choice to represent the top-down maps in this model (Gaborski et al. 2003).

The inputs to neural network are the color opponent output maps (R+G-, R-G+, B+Y-, B-Y+). As shown in Figure 3.4 the convolved color feature maps are the inputs to the neural network. The neural network is trained on search (cue) color. Figure 3.8 shows the output saliency responses of four different neural networks trained on the four primary colors (Red, Yellow, Green, Blue). The saliency response curves are similar to the color opponent cell response for the entire visible color spectrum. Assuming that a neural network is trained on pure saturation of red, this network gives an output of 1 for regions containing a high value of red. The output response reduces with the reduction in amount of the trained color. Thus orange which is a combination of Red and Yellow, produces a lower response than a pure red does. Orange generates response in both the red and yellow trained neural networks.

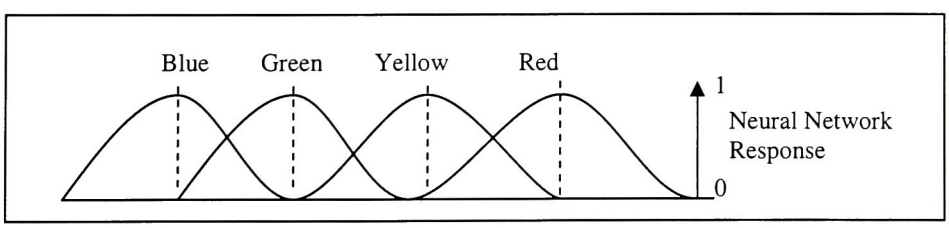
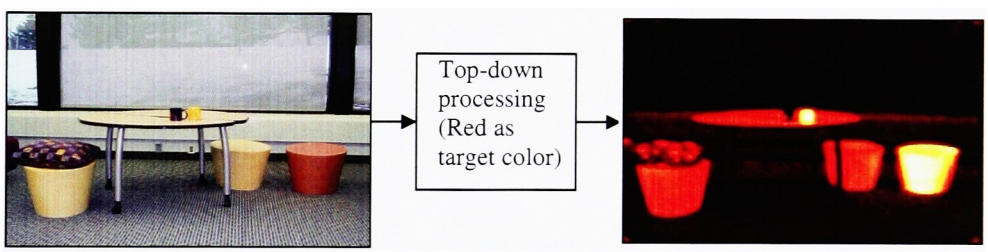


Figure 3.8: Color saliency response curve for the prototype cue colors.

As observed from the response curve Figure 3.8, there is no overlap between the red and green response curves, indicating the mutually exclusive nature of the two colors. Similar behavior is true for blue and yellow response curves. This behavior can be simulated using a neural network.

The interaction between the responses of different color opponent cells give the highest response for the primary colors, marked with vertical dotted lines on the response curve. A reduced response for the deviation colors (orange, violet) is indicated by the fall in the response (to the right and left of the vertical dotted lines). While training the neural network, the training image is an image containing red, yellow, green, and blue regions.

The target matrix is a binary image of the training image with the region of the color to be learned as 1 and the remaining three color regions as 0. Thus while training the neural network on red as the target color, the red region in the training image is 1 and distracter colors, yellow, green and blue regions as 0. The output of the neural network processing is an output map (top-down saliency map) with relative saliency values with respect to the trained color. Saliency values of the output map close to 1 signify high target color and values closer to 0 lack the target color properties. The top-down map for an example scene with red as a search color is shown in Figure 3.9. The small orange table and the orange coffee mug on the table are assigned relatively higher saliency than the yellow tables which gets lower saliency in the top down map. This saliency response of the top-down map concurs with the observation of Figure 3.8.



Input image Top-down salience map

Figure 3.9: Top-down salience map for red as search color

A two layer feed forward neural network is used, with the first layer using a hyperbolic tangent sigmoid transfer function and the second layer using a linear transfer function. The back propagation learning technique is employed in which a training vector and target vector are provided. A maximum of 200 epochs were allowed for the network while training. The number of epochs selected for the network gives expected outputs without over training the network. This is proved by the reduced response generated for deviation colors after being trained on the primary colors. De Valois et al. (1966) proposed a neural network architecture to capture the interaction between the three cone types (short, medium and long wavelength) to explain the

color opponent responses of the color opponent cells that produce a myriad of color perception. In our model the color opponent cells are used as inputs to the neural network but only to generate the saliency values for target and distracter colors. Using the neural network as a model for capturing the saliency of target and distracter colors can efficiently predict the regions of the input image with task-relevant saliency.

Figure 3.10 shows the simulated saliency response (dotted orange curve) for a neural network trained on orange. This saliency response for orange is superimposed on the responses of the primary distracter colors. As seen from the response this particular orange has a higher saturation of red than yellow. During search for this orange the red (R+G-) component gets higher weights than yellow channel (B-Y+). This network assigns non-linear weights to the color channels (R+G- and B-Y+) rather than equal weights to the two channels. The advantage of using a neural network becomes evident in case like the one mentioned here. The saliency response in the top-down saliency map for red and yellow regions is lower than the orange regions.

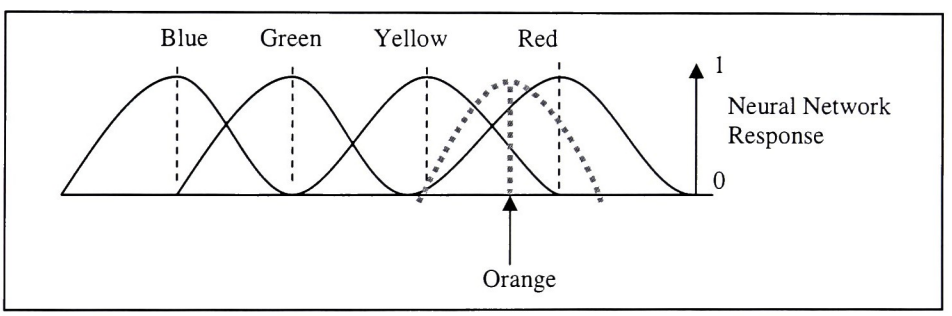


Figure 3.10: Saliency response curve for orange as cue color

3.2.3 Attention Map

The bottom-up and the top-down saliency maps are summed to generate a final attention map. Figure 3.11 shows that summation of the maps enhances the saliency values of the relevant search task regions, and simultaneously maintains the saliency of task irrelevant conspicuous features. Figure 3.4 of the model diagram shows the interaction between two attention systems. The salient regions in the attention map are the regions that have a high probability of being fixated when tested with human subjects. Research conducted by Parkhurst et al. (2002) explains

the effects of bottom-up and top-down attention on perception of a scene. The results section of this paper addresses this topic.

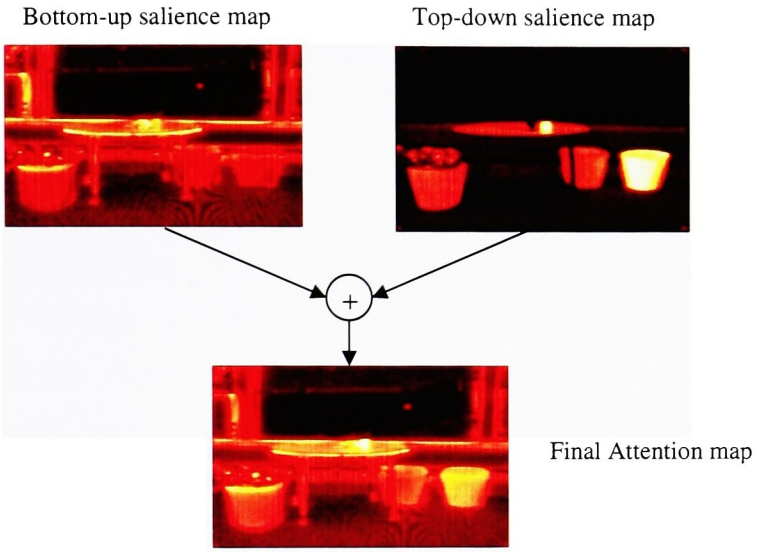


Figure 3.11: Summation of Bottom-up and Top-down salience map.

Chapter 4: Eye-tracking Psychophysical Experiments

4.1 Introduction

The goal of this thesis is to predict the strategies subjects use while searching a scene for cued color objects. Monitoring subjects' eye movements during a psychophysical experiment is an efficient indicator of the objects of interest in a scene. Eye-tracking experiments monitor the subject's point of gaze called fixation across locations in the scene. The eye-tracking theory is based on the fact that in order to perceive an object the eyes have to selectively attend to the object of interest. This is achieved by overtly focusing on that object. This overt focusing on the object is called a fixation point also known as gaze location. The eye-tracker records the point of fixations, the saccades from one fixation location to the other and the duration of fixations. Complex tasks such as image quality judgments, map reading, model building, and hand washing have been successfully analyzed (Pelz et al 2000). Their experiments on hand washing task showed that majority of the eye fixations were on objects that were of immediate use and only small number of fixations were on objects that would become relevant in the near future. Another set of experiment analyzed the fixation durations in a model building task. Reading, searching and manipulation formed the three subtasks in building a model rocket. Analysis of the fixation durations in the three subtasks showed that 'manipulation' task resulted in long fixation durations with a mean of about 450 ms. Reading and searching tasks showed shorter fixations with a mean of 275 msec. Eye-tracking studies enable one to predict the expectations and strategies used during scene perception.

This chapter provides a brief description of the eye-tracker equipment used in the Visual Perception Laboratory (VPL) at Rochester Institute of Technology. The next section describes the methods and procedures used to collect eye-tracking data from subjects. The later section shows the sample visual scanpaths correlated with test images used in the experiments.

4.2 Eye-tracker & Experiment Procedure

The Applied Science Laboratories (ASL) model 501 eye-tracker was used to monitor the eye movements. The eye-tracker headgear has the capability to integrate the effects of head movement during the experiments. The ASL model 501 provides the flexibility of slight head movements without the need to tightly fasten the headgear to the subject's head. Figure 4.1 shows the headgear of the ASL 501 eye-tracker. The raw data collection capability of the ASL system integrates the horizontal and vertical position of the eye with the position of the head in space (Canosa, 2003). The eye position is measured using the separation between the center of the pupil and the center of the corneal reflection. The change in this separation is proportional to the change in line of gaze. The eye-tracking equipment consists of a scene camera, eye camera, control unit, and visible wavelength laser. The eye camera captures the image of the pupil and the corneal reflection and shows it on the eye monitor. Some thresholding is necessary to acquire a distinguishable image of the pupil. The scene camera focuses an image of the currently viewed scene and creates a frame of reference for measurement of the line of gaze. The control unit is used to process the eye camera signal to find the center of the pupil and the center of the corneal reflection. The control unit receives video signals from the eye and scene camera and uses this to calculate a line of gaze and displays it as a cross hair on the video image of the scene. Further details regarding the ASL eye-tracker and its various components are given in Canosa (2003).

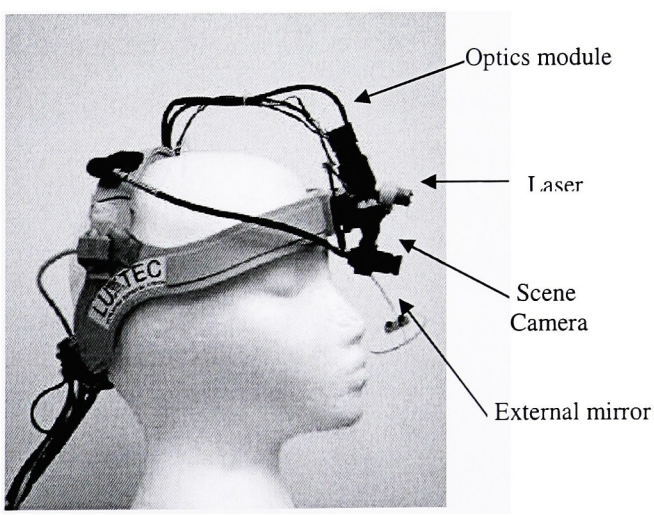


Figure 4.1: ASL model 501 eye-tracker
Adapted from Canosa (2003)

Figure 4.2 shows the eye-tracker headgear mounted on the subject's head. A typical eye-tracking experiment consists of first setting up and calibrating the equipment. The next step is to capture the subject's corneal reflection and the pupil image by adjusting the eye camera and mirror. During this step the subjects are requested to move their gaze to either side of the monitor, to ensure that eye movements do not result in loss of the center of pupil and corneal reflection. The next step is to calibrate the eye-tracker for each subject since each subject has different shaped corneas and corneal reflection properties. This is carried out before the experiment is started. The subjects are instructed to fixate (constant gaze) on a sequence of nine points arranged in a grid fashion. During the calibration step the subjects are required to keep their head steady. In order to ensure that the head movement is minimum driving calibration, a laser beam pointer projects to the centre of the display screen. Subjects use this point as a reference to keep their head steady during calibration. At each one of the nine fixation locations on the screen, the subject's fixation coordinates at the fixation point are recorded. Once the calibration step is complete subjects are free to move their head. After the calibration step the experiment is started using the test scenes for which the eye-tracking data is collected. At the end of the experiment a similar calibration step is conducted. This final calibration check helps in determining the change in recorded positions on initial calibration and final calibration check. The error or deviations in the two checks can be a result of movement of the headgear during the experiment. The accuracy of the eye-tracker after calibration is approximately 1 degree of visual angle.

The subjects sit in front of a 50 inch Pioneer Plasma monitor which is used to display images. The screen size was 1280 x 768 pixels with a screen resolution of 30 pixels/inch. The display subtends a visual field of 60 degree horizontally and 35 degree vertically at a viewing distance of 38 inches. Based on this configuration setting, approximately 22 pixels cover 1 degree of visual angle.

Figure 4.2 : Subject with an ASL 501 head mounted eye tracking gear during an experiment. Image acquired from the Visual Perception Laboratory RIT.

4.3 Experiment Design

A set of 20 test images were collected for experiments. The images consisted of indoor, outdoor scenes as well as computer generated images. The images were divided into two groups first called the "free view task" and second "search task". The choice of the two different task categories helps in inferring the visual perceptual strategies used during a particular task. Given an example scene a free view elicits a pattern of eye movements from a subject which will be different from a pattern during search task of the same image. During a free view task the subjects were required to freely view the images displayed on the screen, and proceed to the next image by pressing the space bar key. This part of the experiment gives an understanding of the non-goal directed perception of a scene. The next part displays images form a different image set for the search task. Search task required searching for objects of a particular color and proceed on to the next image after locating all the target colored objects in the given image. The subject was informed about the target color from text displayed on the screen prior to the appearance of the search image on the screen. This target color was the search color for the subsequent image.

During the experiment a subject first viewed the free view images and then the search task images. No images were repeated in both the free view and search. This was necessary because on first viewing of a particular image in free view the subject acquires

a general understanding about the images and on the second presentation of the same image the subject's perception would be influenced by the first view of the image. Therefore images viewed in the free view are different from the images viewed in the search task. Subjects chosen for the experiments were between the age of 20 to 30. Subjects had normal or corrected to normal vision. Since the experiment required correct search for color objects, subjects without colorblindness were chosen which was tested using the color vision chart. Each subject signed a consent form before the start of experiment. Experiments were conducted on ten subjects. Data collected for two subjects resulted into bad eye-tracks and hence their eye-tracks were not considered for further analysis.

4.4 Visual Scanpaths

The raw data collected during the eye-tracking session is used to analyze the data in a fixed coordinate system. The software written in the VPL is used to locate the fixation coordinates and fixation duration with respect to the viewed image. Before this raw eye-tracking data can be used for interpolation on the image, the data is corrected for drifts and offset errors which might have occurred during the experiment. This is an important phase of data interpretation because during an experiment the headgear may shift on the subject's head. The calibration tests carried out in the beginning and end of the experiment is used to correct this drift error. The software written in the Visual Perception Laboratory was used for drift and offset correction.

The corrected fixation data is then superimposed on the actual image to identify the fixation locations on the viewed scene. The algorithm for plotting the fixation locations is explained in (Canosa 2003). Figure 4.3 shows a test image and the scanpath of a subject's eye movement overlaid on the image. The blue line is the visual scanpath and the yellow * are the actual fixation points. The fixations are numbered in a sequential order from the first fixation to the last observed fixation. As observed from the scanpaths and fixation locations different subjects give different scanpaths but the general locations of fixations are overall similar. In the scene of Figure 4.3, the subjects tend to fixate on the coffee mug, small tables, and boxes on the window. The sequence in which the objects are fixated also varies. This observation concurs with the tenet that size and frequency of

fixations varies across subjects for a given complex natural scene (Andrews & Coppola, 1999).

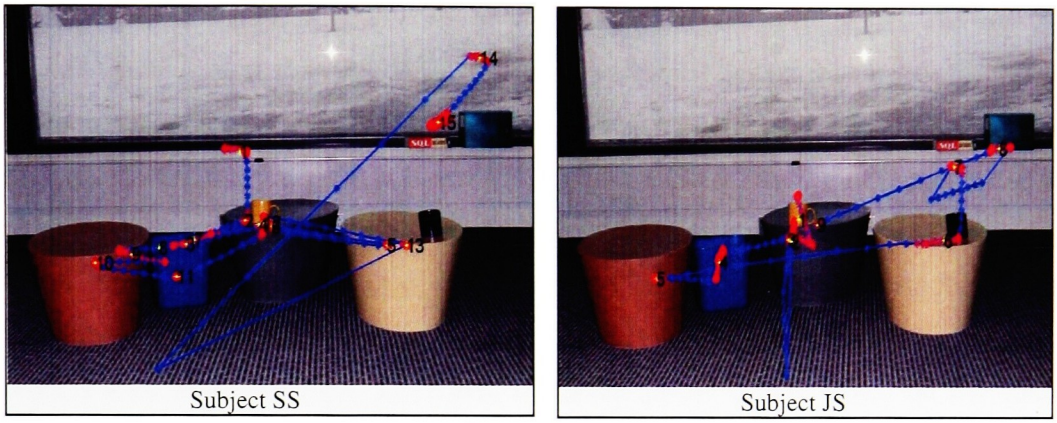


Figure 4.3: Freeview scanpaths overlaid on the image for two subjects. Blue track is the scanpath. Red cluster of points are the fixation points observed around a region. The yellow * is the centroid of each cluster of fixation points.

Figure 4.4 shows resulting scanpaths of a different set of subjects during search for red objects in the same scene as Figure 4.3. The difference in the scanpaths between the free view task and search task is evident in the two figures. The scanpaths in Figure 4.4 are influenced by the goal of searching for red color objects which in this case is the small orange table, and the red book on the window. It can be seen that distracter colored objects (blue, yellow) also capture subject's attention in search task suggesting the possibility of a bottom-up influence on the search.

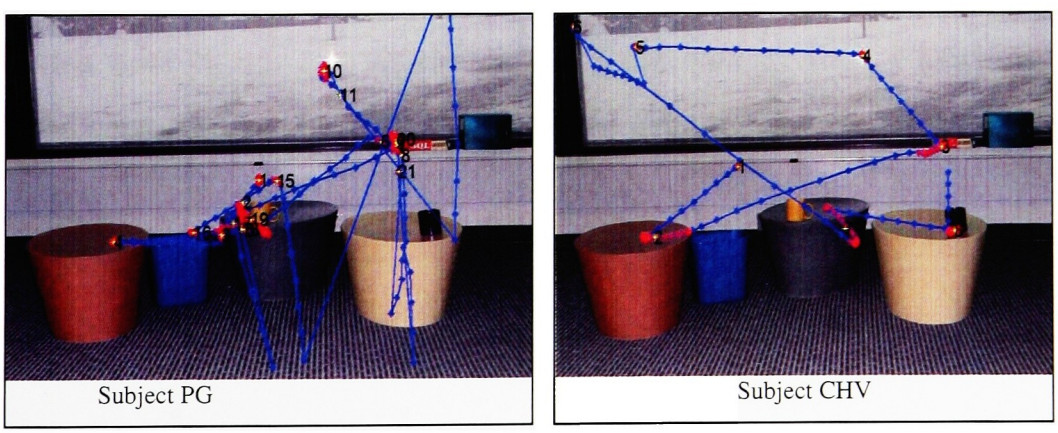


Figure 4.4: Search task scanpaths during search for red objects.

4.5 Top-Down influence on eye movements

The goal of conducting eye-tracking experiments for this thesis is to explain the effects of bottom-up and top-down attention on visual search. As stated before the top down knowledge of a scene can influence the eye movements. Given a search task to search for a paper stapler, the key locations of interest will be a table or a computer desk. Regions such as room ceiling or floor are not likely to be searched. Such prior expectations were observed in our experiments on natural scenes. Given an example scene of a laboratory, Figure 4.5 (a,b) shows the sample eye-tracks for free view and (c,d) shows the scanpaths for blue color search. In this scene the subjects tend to attend to semantically meaningful objects such as the monitor, keyboard, computer-mouse on the table. On the contrary during the search for blue objects the blue bin is attended instantaneously. The object that was initially not considered important during free view acquires importance in search task. This change in pattern of eye movement based on the search task proves that attention is influenced by the task at hand as well as the semantically meaningful objects in the scene.

The rapid extraction of scene gist helps in attending to the semantically meaningful objects in a given scene. Prior research by Parkhurst (2002) addresses this issue of eye movements being influenced by the top-down knowledge about the scene once the gist of the scene is generated. Their research suggest that initial fixations of non-goal directed free views of scenes are influenced by the bottom-up stimulus features but within short period of time top-down influences are observed. In Figure 4.5 the very familiar lab scene the top down influences set-in very early as visible from the fixation sequence. The semantically meaningful objects like computers capture attention. The next chapter shows quantified results for the correlation between the generated saliency maps and scanpaths.

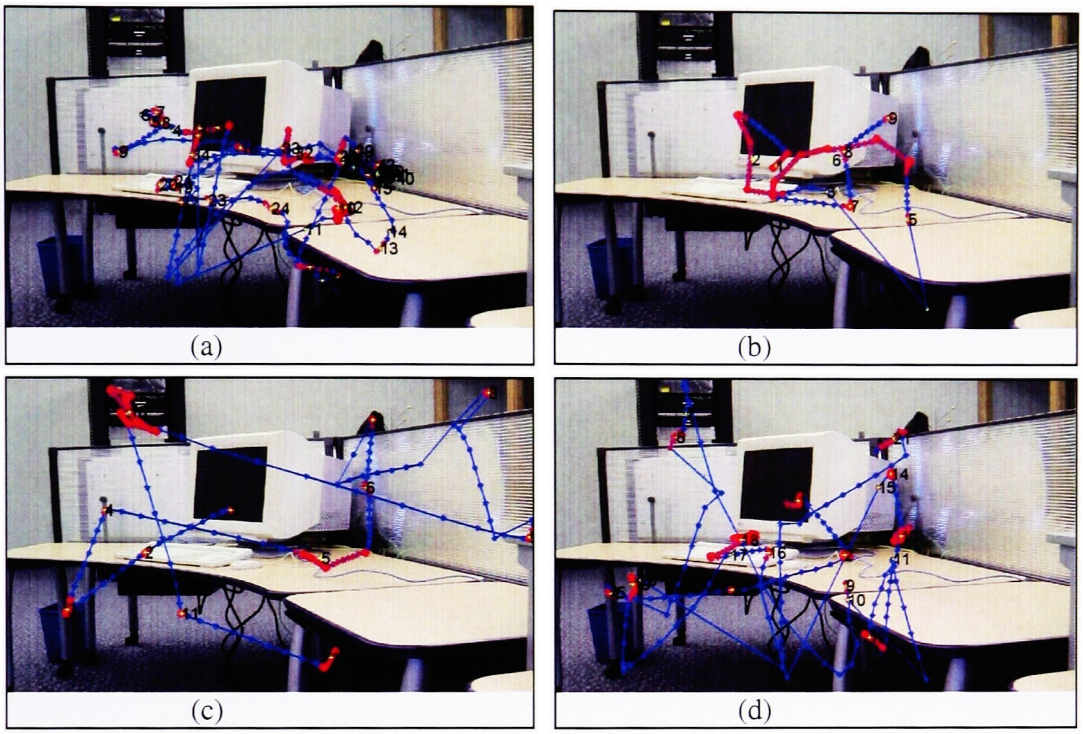


Figure 4.5: (a,b) free view scanpaths for two subjects. (c,d) search scanpaths for search of blue color in the scene.

Chapter 5: Analysis

5.1 Eye Movement and Saliency Map Correlation

The previous two chapters explained the generation of the saliency maps and the eye tracking studies conducted to collect eye movement data. This chapter provides analysis that shows the computationally predicted saliency maps correctly predict the local maxima regions of saliency that agree with the eye movement data. A correlation measure is calculated in order to estimate the correlation between the eye movement generated fixations and the saliency map predicted regions. The saliency map predicts local maxima regions of saliency. The mean saliency of a given saliency map is indicative of the strength of the map predictive power. In order to quantify the correlation between maps and eye movements the following procedure is used. The scanpath is overlaid on the saliency map and the saliency values at the fixation point corresponding to the (x, y) coordinate map location are extracted. In order to account for the possibility that the subjects actual fixation might be within 1 degree visual angle of the recorded fixation point, the maximum saliency value within 1 degree (21x21 pixel) region around the fixation point is chosen to be the saliency value for the fixation. The maximum saliency value is chosen within 1 degree region because the recorded fixation point might be offset by a certain pixels from the actual point of gaze within that region.

The mean of all the extracted saliency values is termed the fixation mean ($\mu_{fixation}$). The mean (μ_{map}) of the complete saliency map is taken to represent the mean saliency by chance alone. The mean saliency by chance represents the salience of the map if locations are chosen randomly across the entire saliency map. A ratio of $\mu_{fixation}$ to the μ_{map} is called the Fixation/Map ratio (F/M) (Canosa, 2003). This ratio is analogous to a correlation indicator of how well the saliency map predicts the eye fixations. If F/M ratio is close to 1, it indicates that the subject's fixations were randomly distributed across the map since the μ_{map} and $\mu_{fixation}$ are almost similar. A ratio significantly higher than 1 indicates that majority of fixations occurred on the high saliency regions. If the ratio is less than 1, it indicates the fixations tend to be focused on the low saliency regions.

The duration of individual fixation also can be used for calculating the F/M ratio by multiplying the saliency value at each fixation location with the corresponding fixation duration. The mean of these duration weighted fixation saliencies is called ($\mu_{fixation\_duration}$). Here the fixation duration

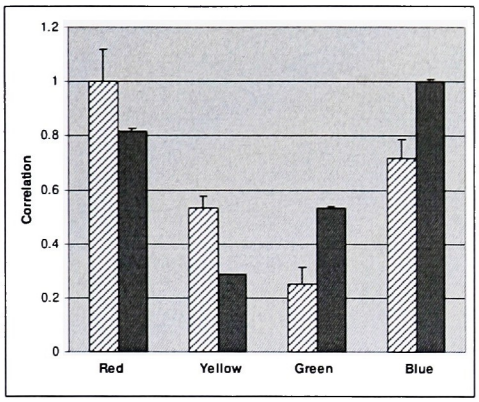
is the weighting factor for the individual fixations. The ratio of $\mu_{fixation\_duration}$ and μ_{map} generates the correlation value influenced by fixation durations. The freeview scanpaths were used to correlate with the saliency maps in order to understand the task independent top-down effects. The search scanpaths were used to understand the search task relevant changes in scene perception.

5.1.1 Effects of target and distractor prototype colors

The F/M ratio for the four prototype colors is calculated by overlaying the search scanpath on the R+G- (Red excitatory, green inhibitory), R-G+, B+Y-, B-Y+ maps. These maps are generated as part of the bottom-up processing of the input image in the color channel as explained in 3.2.1. Figure 5.1 shows graphs for search of target color blue and Figure 5.2 shows graphs for red as target color. Pattern bars in the graphs of Figure 5.1 and 5.2 are the correlation values without considering the fixation durations. Black bars are correlation values with fixation durations as a weighting factor. Y-axis represents correlation values (F/M ratio) normalized to the highest of the four F/M ratios.

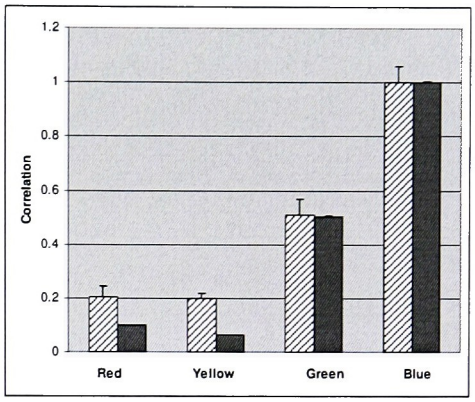
Legend ▨ Non Duration ■ Duration

Subject SS



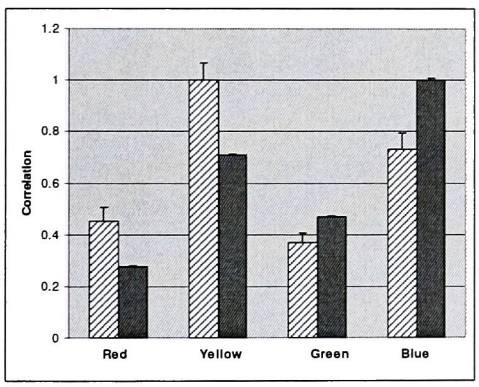
(a) Prominent distractor in non-duration graph is red

JS



(b) No prominent distractor in non-duration graph.

TC



(c) prominent distractor in non-duration graph is yellow.

Figure 5.1: Bookshelf image. The target color is Blue. Graphs for three subjects are shown.

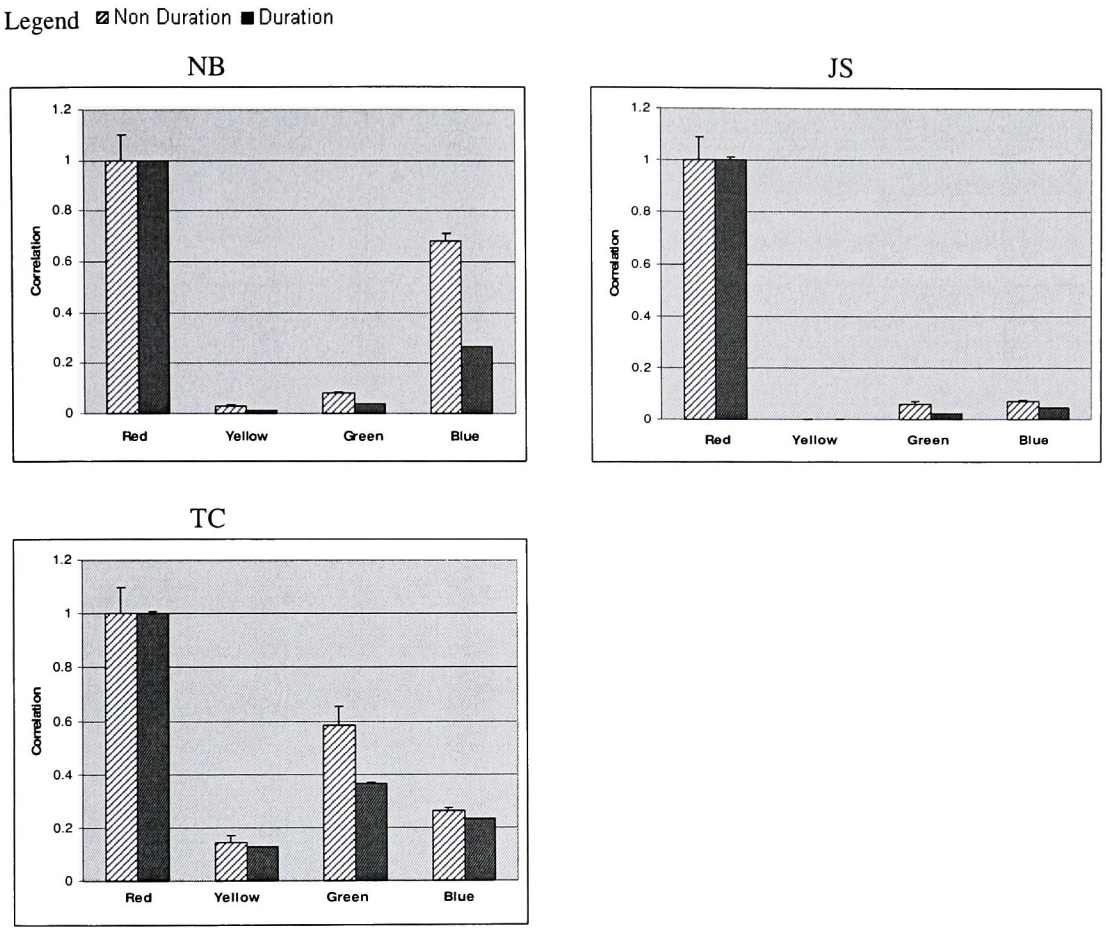


Figure 5.2: Computer generated scene. The target color is Red. Red stands out with highest correlation in each graph irrespective of using duration as weight.

Observation: In Figure 5.1 the target color (blue) has a consistently high correlation value in the non-duration analysis, but other distractor colors tend to have comparatively high correlation values too. This behavior is because the subjects fixated the distractor colors while searching for the target colors. In Figure 5.1 for subject SS, red distractor attains highest correlation in non-duration analysis and for subject TC, yellow was a prominent distractor. When applying fixation durations as weighting factors, the target color emerges with the highest correlation value, and the distractor colors attain more de-correlation with the target color. In graphs of Figure 5.2 the search yields highest correlation for the red target color map both in duration and non-duration tasks. The duration as a weighting factor does not affect the correlation of the red target whereas the blue targets in graphs of Figure 5.1 show increased correlation using duration as weight. This would possibly suggest that there is something inherently different about searching for red versus

searching for blue. A possible explanation for this could be that yellow and red stand out as prominent distractors in search for blue but this effect reduces when fixation duration is considered. This indicates that the fixation durations reflect important information about the nature of search. The reduction in correlation for distractor colors when fixation duration is used as a weighting factor is due to the fact that shorter duration fixations are observed when fixations occur on distractor colors. Whereas longer duration fixations occur for fixations on target colors. Therefore on using fixation duration as a weighting factor for correlation analysis, the target color's F/M is enhanced and disctrator color's F/M ratio is reduced. This enable target colors to appear with higher correlation than distractor colors. Work by Findlay et al. (2001) also suggests a similar observation that brief fixations occur when the eye lands on a distractor. The color experiment they conducted led to the conclusion that short-fixation durations occurred if the first saccade landed on a distractor of opposite color to the target color. To identify the cause for why fixation durations change with search requirements is beyond the scope of this thesis. Detailed experiments and analysis will be required to find the cause.

5.1.2 Effects of top-down and bottom-up attention

This part of the analysis shows graphs that indicate that bottom-up factors affect the visual search task. Figure 5.3 shows the subject's search scanpath superimposed on the bottom-up, top-down and final attention maps. The scanpath shows that majority of the fixations take place on the high saliency regions of the maps.

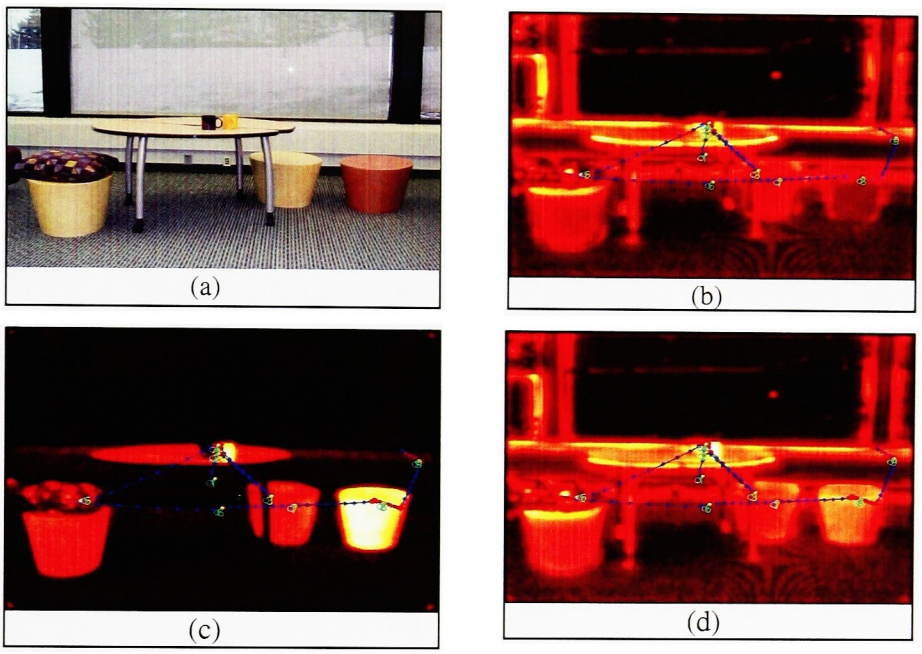
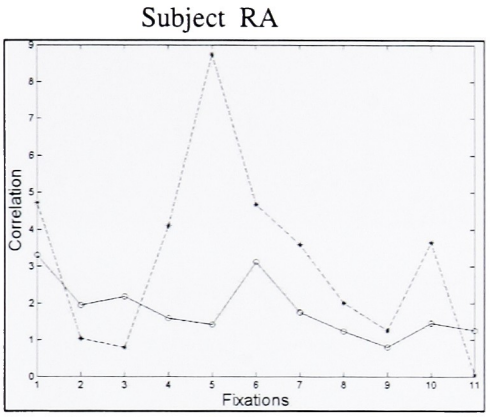


Figure 5.3 : Subject TrC's search scanpath overlaid on the (b) Bottom-up saliency map (c) top-down saliency map (d) Final attention map for the (a) input image with search for red objects.

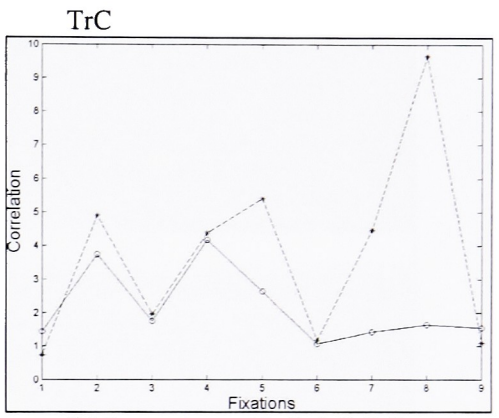
The graphs of Figure 5.4 and 5.6 show the top-down and bottom-up correlation curves obtained by overlaying the search scanpath on the top-down and bottom-up saliency map and calculating the correlation values (F/M ratio) of each fixation with the respective map. Graph of Figure 5.4 is based on the search for color red conducted on scene shown in Figure 5.3. The y-axis is the correlation value axis, and the x-axis the sequential fixation numbers of the scanpath for the search conducted by the subjects. The correlation value (F/M ratio) for each fixation is calculated by taking a ratio of individual fixation location's saliency value with the map mean (μ_{map}). The fixation correlation values indicate how well the individual fixation correlated with the predicted saliency map. This method allows us to compare the fixation strengths of the two curves. The top-down and bottom-up correlation curves are depicted in the Figure 5.4 and 5.6. Either the top-down or bottom-up correlation curve that yields the higher correlation for a given fixation can be

said to have influenced that fixation. For example in Figure 5.4 (a) the fixation numbers 2, 3 give higher correlation for bottom-up curve whereas fixations 1, 4, 5, 6, 7 etc are influenced by top-down curve.

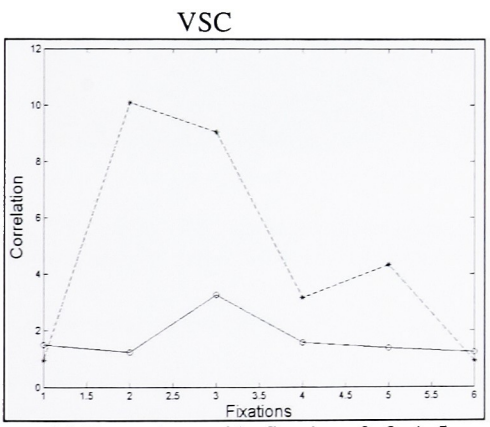
Legend —o— Bottom-up correlation curve, -----*----- Top-down correlation curve



(a) Target located in fixation number 1, 4, 5, 6, 7



(b) Targets located in 2, 4, 5, 7, 8.



(c) Target located in fixations 2, 3, 4, 5

Figure 5.4: (a-c) shows the correlation curve for the three subjects for an indoor scene. The search scanpath overlaid on top-down and bottom-up maps

Figure 5.5 (a) shows the scene for which the search required locating blue colored objects. A sample scanpath of subject TC is superimposed on the (b) bottom-up saliency map, (c) top-down map and, (d) the final attention map.

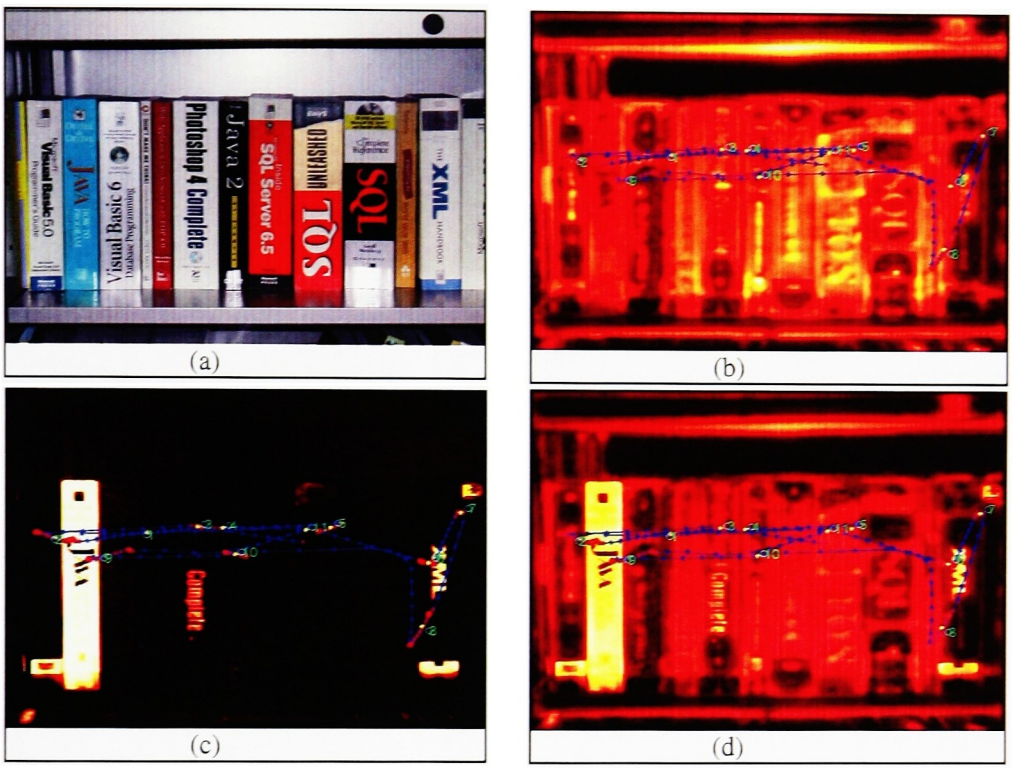


Figure 5.5 Subject TC search for blue objects in bookshelf scene (a). The search scanpath overlaid on the (b) Bottom-up saliency map (c) top-down saliency map (d) Final attention map

It can be seen from Figure 5.5 (b) that red regions and regions with text on the books tend to stand out as highly salient. The eye fixations on the bottom-up map are on the high saliency regions. In (c) the top-down map shows that target colored (blue) regions are fixated within the few fixations after the image onset. Figure 5.6 shows the graphs for correlation curves obtained from three subjects for search of blue in bookshelf scene shown in Figure 5.5.

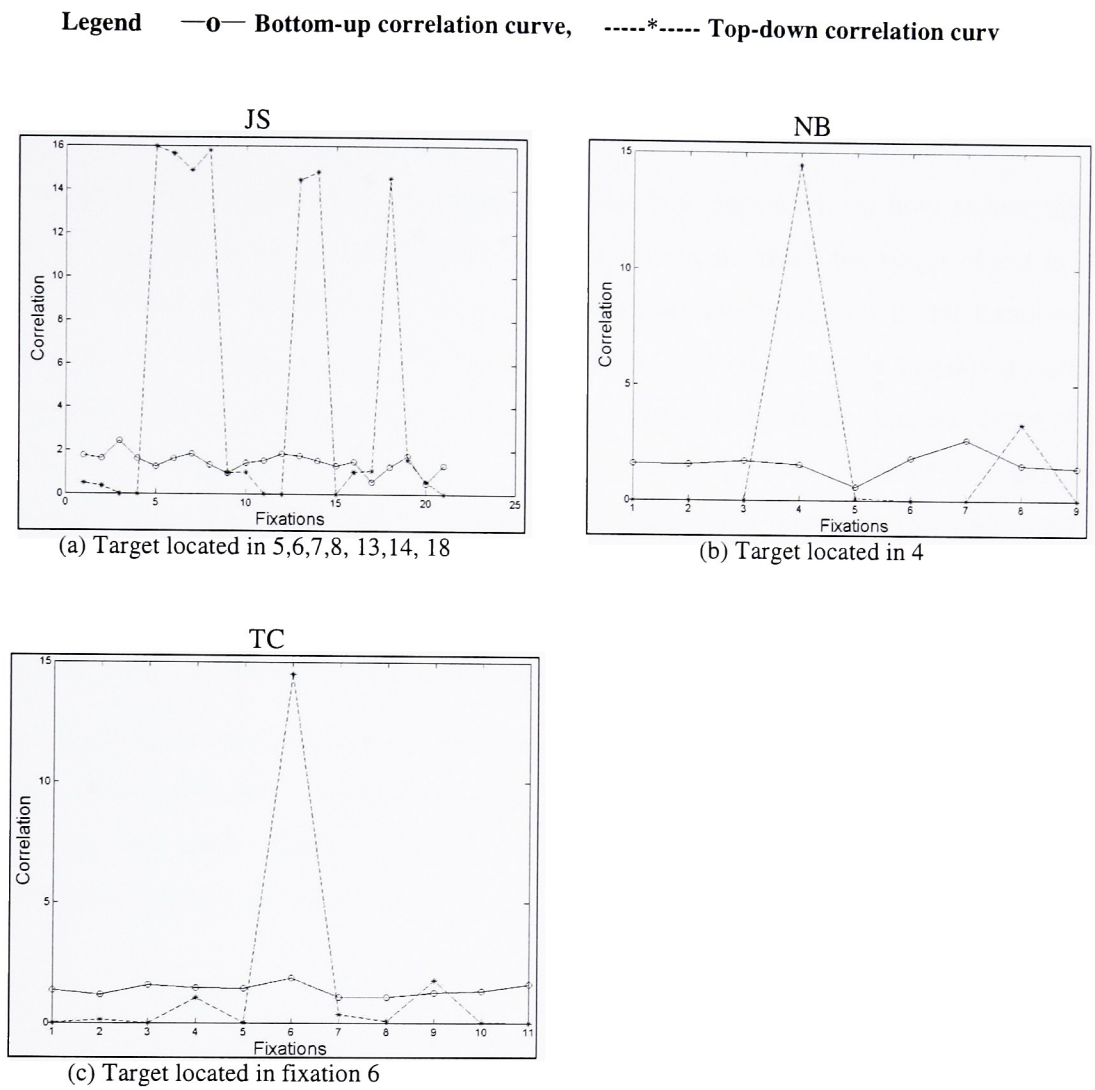


Figure 5.6: Top-down and bottom-up correlation curves for search scanpaths in bookshelf image. Search color is blue.

Observation: When not focusing on a target object, the low-level conspicuous features catch the subject's attention, which are predicted by the bottom-up attention map. This is evident in the graphs where the fixations with low top-down correlation values are compensated with high bottom-up correlation value, suggesting a potential bottom-up influence on the fixation seen in Figure 5.6 for subject RA, JS, NB. The results indicate that search task is influenced by intermediate task irrelevant bottom-up fixations in process of fixating on search relevant top-down features. Fixations 1,2,3,4, 10, 11 of subject JS in Figure 5.6 (a) show the above described behavior.

Another observation is that if the search continues even after locating the target object, the final fixations are primarily on high salient regions of the bottom-up map evident in Figure 5.6 (subject NB, TC).

The analysis of the graphs is scene dependent. Search is dependent on how salient the target object is compared to the distractors. For instance, scenes in which the target object is readily visible or stands out in the scene will be located immediately without initial fixations being focused on the distractor objects. Therefore in graphs of Figure 5.4 the top-down correlation curve yields high correlation even in early fixations of the scanpath because the target object is high in bottom-up saliency as well as top-down saliency. This behavior changes for the bookshelf scene of Figure 5.5 and its corresponding analysis in Figure 5.6. The blue target object is located in fixation 5 and thereafter for subject JS of Figure 5.6 (a). Target is located on fixation 4 for NB and fixation 6 for subject TC. Thus the properties of the highly salient distractors affect the search pattern based on the type of the scene.

From the analysis of the graphs we do understand that the system is efficiently able to predict the regions of the image that are high in saliency for both top-down and bottom-up visual attention processing. The search scanpaths are affected by combined effects of the bottom-up and top-down saliency of objects in the scene.

Chapter 6: Conclusion and Future Work

As an engineering model the system described in the thesis is able to predict the regions of the scene that have a high probability of being fixated when confirmed with human eye tracking data. The model generated bottom-up saliency map topographically represents the scene with salient regions based on low level stimulus saliency. The top-down saliency map represents the regions that are relevant to the current search task. The combination of the two saliency maps is called the attention map, represents regions of the scene which are a combination of the stimulus saliency (intensity contrast, color saliency and orientation saliency) and the search task relevant features. Neural network is efficiently used to capture the chromatic relations between the target and distractor colors for predicting search task relevant regions. In order to test the efficiency of the maps in predicting possible location of interest, human eye tracking data was used. The eye tracker generated scanpaths were compared with the saliency maps and a correlation measure was calculated. The chapter 5 explains all the results analysis for comparing the maps and scanpaths in form of graphs and analysis of variance tests. The system is not a complete representation of the actual human visual search mechanism. Various psychophysical factors have not been incorporated in the model. Some of these factors and recommendation for future work are described in this chapter.

There are instances when the low-level feature saliency is not enough to define the objects of interest as shown in example of section 4.5. In the laboratory scene the monitor, key board and mouse tend to be fixated regularly. The bottom-up saliency map for this scene does not capture keyboard and mouse as salient objects and thus the correlation between the eye movements and the saliency map suffers. Therefore in addition to the low level stimulus feature saliency, an object level representation will be required to create more plausible saliency maps. Subject's eye movements are not random but follow a strategic pattern for perceiving the scene. Subjects tend to look at meaningful information in the scene and are not always influenced by the stimulus saliency. Research by Canosa (2003) addresses this issue by creating a higher level proto object map which topographically represents the foreground objects in the scene. The proto object map combined with the low level saliency map forms the conspicuity map, is able to better predict the fixation locations for a subject's scanpaths.

As shown in the research by Parkhurst et al. (2002) that after the onset of an image, the initial fixations are influenced by the low level stimulus saliency and as time progresses the correlation between bottom-up saliency map and eye fixations reduces. This drop in correlation is because after a few fixations the top-down influence takes over. Humans are very efficient at extracting the gist of the scene within millseconds of showing the scene. Once the gist is extracted, the subject's attention is directed to semantically related objects for that scene and bottom-up effects may reduce. Their approach also creates the low level saliency map using technique described by Itti and Koch (2001). These experiments indicate that in order to fully capture the subject's locations of eye fixations a higher level map will be required, that accounts for subject's expectations and also assigns task relevant saliency to the objects in the scene. This research shows that subjects don't always focus on every object in the scene rather focus on task relevant objects.

In my thesis, color was chosen as high level feature for guiding visual search. The top-down map topographically represents regions based on the similarity to target color. The map assigns saliency according to the color opponent response curve described in section 3.2.2. The colors adjacent to the target color on the color opponent curve are assigned higher saliency than colors that directly inhibit the target color. This technique accounts for the fact that colors similar to the target color are highly probable to be searched for due to their similarity to the target color. Such a technique is able to account for the higher level knowledge that orange is perceptually similar to red during search for red. A more plausible approach for creating top-down maps could base the analysis on the color discrimination capabilities of a human visual system. Research by Fletcher & Voke (1985) shows that capability to discriminate between two colors varies considerably across the color wavelength spectrum. At the extremes of the color spectrum in the region of deep reds, hue discrimination is poor and a large difference in color wavelength is required to notice the color changes. Where as in the centre of the spectrum where color changes from blue to green and yellow regions, a little difference in wavelength is enough to perceive the color change. Adding a feature which modifies the top-down map to account for chromaticity discrimination will provide an efficient and a plausible method for representing color discrimination feature.

Summing the feature saliency maps (color saliency, intensity contrast saliency and orientation saliency) from different feature dimensions is an important aspect of stimulus saliency that needs to be addressed in order to build better computational models. Research by Nothdurft (2000) states that the saliency of a target object during visual search is dependent on how different the target object's features are with respect to the distractor objects in the scene. A horizontal bar between vertical bars is very salient because of the difference in orientation between the horizontal and vertical bars. But this same horizontal target will be less salient amongst bars that are nearly horizontal. Thus the orientation difference between the target and distractor defines the saliency of a target object. This difference term is defined as contrast in a feature dimension. The saliency of objects is related to the local differences in the stimulus dimensions. When objects are defined with a combination of stimulus dimension, (e.g. target defined by color and orientation) the stimulus salience from the different dimensions add up but not linearly. Targets which posses feature contrast in more than one dimension were more salient than ones with feature contrast in single dimension but often less salient than predicted by the sum of feature contrast from the different dimensions. This reduction in gain of combined saliency indicates that saliency between feature dimensions do not add up linearly rather there is some overlap between different feature saliency. Their research showed that combination of color and orientation contrast produced the strongest gain reduction about 90 % for color in orientation. Combination of color and motion contrast revealed about 50 % gain reduction than what would be predicted by the linear sum. This non-linear summation of feature dimensions could be incorporated in the model for generating bottom-up saliency maps. The model as it stands now, linearly adds the feature saliencies from three feature dimensions. Using the psychophysics data provided by Nothdurft (2000), more plausible models of visual search could be built.

Another technique that could help in non-linear summation of feature saliency maps is by using eye tracking data. Subject scanpaths display properties based on the strategies subjects use while observing a scene. In order to learn the feature weights for non-linear summation of feature saliency maps, the scanpaths collected from the subjects can be used for learning the role that different feature dimensions play in attention allocation. As describe in research by Parkhurst et al. (2002) the eye tracks can be used to find the weights for different feature dimensions for natural, fractal images and city scenes.

Correlation between the eye-movements and the saliency maps at present take into account only the fixation saliency mean and map mean. Thus if the entire scanpath consisted of 7 fixations, the fixation saliencies from the bottom-up map and top-down saliency map are collected. By taking the means of the fixation saliencies for finding the correlation, we ignore the variance of the saliency values collected by the scanpath for the two saliency maps. Thus if the entire scanpath consisted of 7 fixations, it will be important to find what was the distribution of saliency values of the scanpath. Suppose for a scanpath consisting of 7 fixations when overlaid on the bottom-up map gives fixation saliency values of [0.2 0.65 0.7 0.4 0.45 0.7 0.6] and standard deviation (σ_{BU}) of 0.186. The same scanpath when overlaid on a top-down map gives values of [0 0 0.8 0.6 0 0 0] and a higher standard deviation (σ_{TD}) of 0.346. The standard deviations of the data values of the two groups (Bottom-up and Top-down) can be used as weights to calculate the new F/M ratio in the following way.

$$F/M_{bottom\text{-}up} = (F/M_{bottom\text{-}up}) * (1 - \sigma_{BU})$$
$$= (F/M_{bottom\text{-}up}) * (1 - 0.186)$$
$$= (F/M_{bottom\text{-}up}) * 0.814$$
$$F/M_{top\text{-}down} = (F/M_{top\text{-}down}) * (1 - \sigma_{TD})$$
$$= (F/M_{top\text{-}down}) * (1 - 0.346)$$
$$= (F/M_{top\text{-}down}) * 0.654$$

Here the 1- σ factor acts as a weighting factor. This weighting will ensure that the scanpath that collected saliency values with higher variance (high σ) in the values will lower its F/M ratio due to a low weight. The scanpath with low variance (low σ) in fixation saliency values will be modulated with a higher weight. The correctness of the above weighting procedure with respect to eye-movement and map correlation will have to be researched into. At present lack of enough evidence and experimental analysis prevent us from applying the following method of correlation calculation. One possible explanation in support of the above method could be that a scanpath that with high σ, was distributed unevenly in the map and simultaneously collected high and low saliency values. A scanpath with low σ, consistently collected tightly grouped low or high fixation saliency values.

7: References

Aks, D.J., Zelinsky, G., & Sprott, C. (2002). Memory across eye-movements: 1/f dynamic in Visual Search. Journal of Non-linear Dynamics in Psychology & the Life Sciences.

Andrews T. J. and D.M. Coppola, (1999) "Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments," *Vision Research*, **39**, pp. 2947-2953.

Bell A. J. and T. J. Sejnowski, (1997) The independent components' of natural scenes are edge filters." *Vision Research 37(23), pp. 3327-3338.*

Brindley G. S.(1970) *Physiology of the Retina and Visual Pathway*, Edward Arnold (Publishers) Ltd.

Buracas, G. T., & Albright, T.D. (1999) Covert visual search: A comparison of performance by humans and macaques (Macaca mulatta). *Behavioral Neuroscience., 113,* 451-464

Canosa Roxanne (2003). Seeing, Sensing and Selection: Modeling Visual Perception in Complex Environments. *Phd Thesis, Rochester Institute of Technology*

Desimone, R., Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193-222.

de Valois R. L. and K.K. de Valois. (1988) *Spatial Vision*. Oxford Press.

D'Zmura, M. (1991). Color in Visual Search. *Vision Research* 31, 951-966.

Findlay, J.M., Brown, V., & Gilchrist, I.D. (2001). Saccade target selection in visual search: the effect of information from the previous fixation. *Vision Research,* 41, 87-95

Fletcher R., Voke J. (1985), Defective colour vision fundamentals, diagnosis and management.

Foldiak, P & Young, M (1995). *Sparse coding in the primate cortex*. The Handbook of Brain Theory and Neural Networks, 895--898. (MIT Press, Cambridge, MA).

Gaborski R. S., Vaingankar V. S.,Canosa R. L., (2003) "Goal Directed Visual Search Based on Color Cues: Co-operative Effects of Top-down & Bottom-up Visual Attention". *Proceedings of the Artificial Neural Networks in Engineering, Rolla, Missouri, 2003*

Hamker F. H., (1999) The role of feedback connections in task-driven visual search In: Connectionist Models in Cognitive Neuroscience, Proc. of the 5th Neural Computation and Psychology Workshop (NCPW'98). D. Heinke, G. W. Humphreys, A. Olson. (Ed.) London: Springer Verlag 1999, 252-261.

Hayhoe, M., Lachter, J., and Feldman, J. (1991). Integration of form across saccadic eye movements. *Perception, 20:*393--402.

Hering, E. (1878/1964). Outlines of a theory of the light sense, (L.M. Hurvich and D.J. Jameson, trans.) Cambridge, MA: Harvard University Press.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*(6693), 575-577.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106-154.

Hubel D (1988), Eye, Brain, and Vision. Scientific American Library Series. New York: WH Freeman

Jones J. P. and L.A. Palmer. (1987) An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1233-1258.

Klein R. (1988). Inhibitory tagging system facilitates visual search. *Nature, 334,* 430–431.

Koch C. and Ullman S.,(1985) ``Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, 4:219-227.

Kristjansson, A. (2000). In search of remembrance: Evidence for memory in visual search. *Psychological Science, 11*(4), 328-332.

Li F. F., R. VanRullen, C. Koch and P. Perona. (2002) Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. 99,* 8378 – 8383.

Liversedge, S.P., & Findlay, J.M. (2000). Eye movements reflect cognitive processes. *Trends in Cognitive Science,* 4, 6-14

Nagy A., and Sanchez R. Critical color differences determined with a visual search task. *Journal of the optical society of America*, A7(7): 1209-1217, 1990

Navalpakkam V., Itti L., (2002) A Goal Oriented Attention Guidance Model, *Lecture Notes in Computer Science,* Vol. **2525**, pp. 453-461.

Nakayama, K. and Silverman, G.H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature 320:* 264-265.

Nothdurft H. C., (2000) "Salience from feature contrast: additivity across dimensions", *Vision Research, 40,* pp. 1183-1201.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top down control of visual attention in object detection. *IEEE Proceedings of the International Conference on Image Processing, Vol I,* 253-256.

Olshausen BA, Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381:* 607-609.

Palmer, J. & Ames, C. T. (1992). Measuring the effect of multiple eye fixations on memory for visual attributes, *Perception & Psychophysics, 52,* 295-306.

Palmer S. E. (1999). *Vision Science: Photons to phenomenology.* Cambridge, Massachusetts: MIT Press

Parkhurst D, K. Law, and E. Niebur. (2002) Modeling the role of salience in the allocation of overt visual attention, *Vision Research, 42,* pp. 107–123.

Pelz, J.B., Canosa, R., and Babcock, J. (2000) "Extended Tasks Elicit Complex Eye Movement Patterns," *ACM SIGCHI Eye Tracking Research & Applications Symposium 2000.*

Philippe.G. Schyns., & Aude Oliva. (1994). From blobs to boundary edges : Evidence for time- and spatial scale dependent scene recognition. *Psychological Science, 5,* 195-200.

Rayner. K. (1978) Eye movements in reading and information processing *Psychological bulletin, 85,* 618-660

Rao R.P.N., Zelinsky G., Hayhoe M.M., & Ballard D.H. (2002). Eye movements in iconic visual search. *Vision Research*, 42, 1447-1463.

Rensink R A (2002), Change Detection. *Annual Review of Psychology,*53:245-277

Riesenhuber, M. and T. Poggio. (2002) Neural Mechanisms of Object Recognition, *Current Opinion in Neurobiology*, 12, 162-168.

Schoenfeld MA, Tempelmann C, Martinez A, Hopf JM, Sattler C, Heinze HJ, Hillyard SA. (2003) Dynamics of feature binding during object- selective attention. *Proc Natl Acad Sci U S A*

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, **12**, 97-136.

Usrey, W.M., Alonso, J.M., Reid, R.C. (2000). Synaptic interactions between thalamic inputs to simple cells in cat visual cortex. *Journal of Neuroscience* 20:5461-5467.

Wolfe, J.M (1994) Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review 1(2),* 202-238.

Wooding, D.S. (2002). Fixation maps: Quantifying Eye-Movement Traces. *In: Eye Tracking Research and Applications Symposium Proceedings (ETRA, New Orleans, LA, 2002), ACM.*

Yarbus, A. (1967). *Eye movements and vision.* L.A. Riggs, trans., New York: Plenum Press.

Young R. (1985). The Gaussian derivative theory of spatial vision: analysis of cortical cell receptive field weighting profiles. General Motors Research Report GMR 4920

Appendix

Visual Search software

Main.m-
This is the main script of the system. The script contains calls to other functions for calculating intensity contrast saliency maps, color opponent maps and the orientation saliency maps.

NewVDog8.m-
This script is used for creating the 8x8 Difference of Gaussian filters. The kernel simulates the center surround receptive field structure of the retinal ganglion cells.

MakeGrating2.m-
This script creates Gabor orientation kernels of 0, 45, 90, 135 degree orientation. The kernels simulate the structure of the oriented simple cells of the primary striate visual cortex. The kernels are of two types : sine Gabor filters and cosine Gabor filters.

Normalise.m-
This function is used to normalize maps to implement the global non-linear combination technique discussed in section 3.2.1. The function uses the map's global maximum and local maximums to calculate the weight for the map. Maps with multiple dense regions of saliency are assigned lower weights and maps with sparse localized salient regions are assigned higher weights. This technique is applied to maps before summing the maps within a feature channel or between feature channels.

ColorChannelsV1.m-
Computes the top level matrices of Red, Green, Blue, and Yellow for color opponent saliency calculations

ColorOpponentFilters.m-
This script extracts the center and surround regions of the difference of Gaussian regions. The extracted regions are used for convolution with the color maps in ColorFeature.m

ColorFeature.m-
Script for creating the BY and RG double color opponent feature maps.

Learn.m-
This script implements learning the color properties of the target color using a neural network. The Matlab neural network toolbox is used to implement learning of feature weights. During the learning process, the inputs to the network are the multi-resolution red, green, blue, yellow color opponent maps. The pixel values of the feature maps form the data matrix (input matrix). The target matrix is row wise representation of the binary map with 1 in the regions where the target color is present and 0 in the regions of the distractor colors. The network is allowed a maximum of 200 epochs for learning. This enables the network to learn the feature weights without over training.

Test.m-
Script for creating the top-down saliency map for a test image. Loads the neural network created in Learn.m. The output of the script is a top-down saliency map with values of the pixels indicating the saliency values with respect to the target color. Colors that are similar to target color are assigned higher saliency and colors that are distant from the target color on the color opponent curve are assigned lower saliency.

Eye-tracking scripts-
Eye-tracking software source code can be referred from Canosa (2003). The software code includes scripts for extracting eye movement scanpath and fixation data from the raw eye-tracking data. The software also includes scripts for extracting saliency values from the collected scanpaths.

Main.m

```
% Main Script of Focus of Attention

NewVDog8   % call script to create Difference of Gaussian 8x8 filter
NewVDog16  % Difference of Gaussian 16x16 filter
NewVDog32   % Difference of Gaussian 32x32 filter
NewVDog64  % Difference of Gaussian 64x64 filter

% Enter the filename to be processed
FileName= 'DSCN0717.jpg';

name= char(FileName);
image= imread(FileName);

image=imresize(image,[256 256],'bicubic');

[ro,co,n]= size(image)
image0= double(image)/255;
Iimage0=image0;
Iimage0= rgb2gray(image0);

%---------------------------------------------------------
% Calculating the Intensity Contrast Saliency map.
% Flipping the outer rows and cols of the image to avoid edge effects

xx16 = [[flipud(Iimage0(1:16,:));Iimage0];flipud(Iimage0(ro-15:ro,:))];
yy16 = [[fliplr(xx16(:,1:16)),xx16],fliplr(xx16(:,co-15:co))];

xx64= [[flipud(Iimage0(1:32,:));Iimage0];flipud(Iimage0(ro-31:ro,:))];
yy64 = [[fliplr(xx64(:,1:32)),xx64],fliplr(xx64(:,co-31:co))];


imConv_8=conv2(yy16,Dog8,'same');
imConv_8= imConv_8(17:ro+14,17:co+14);

imConv_16=conv2(yy16,Dog16,'same');
imConv_16= imConv_16(17:ro+14,17:co+14);

imConv_32=conv2(yy16,Dog32,'same');
imConv_32= imConv_32(17:ro+14,17:co+14);

imConv_64=conv2(yy64,Dog64,'same');
imConv_64= imConv_64(33:(ro+30),33:(co+30));

%New size of the images is 254

imConv_8new= abs(imConv_8);
imConv_16new= abs(imConv_16);
imConv_32new= abs(imConv_32);
imConv_64new= abs(imConv_64);

Inten_Sal= imConv_8new1 + imConv_16new1 + imConv_32new1 + imConv_64new1;
Inten_Sal= Inten_Sal / max(max(Inten_Sal)); % Final Intensity Saliency map


imConv_8new= imConv_8new/ max(max(imConv_8new));
```

```
imConv_16new= imConv_16new/ max(max(imConv_16new));
imConv_32new= imConv_32new/ max(max(imConv_32new));
imConv_64new= imConv_64new/ max(max(imConv_64new));

%-------------------------------------------------------

% Calculating the Color opponent maps.

Cimage0= mat2gray(image0);
Cimage0=imresize(Cimage0,[254 254],'bicubic');
Iimage0=imresize(Iimage0,[254 254],'bicubic');

ColorchannelsV1; Creates the red, green, yellow, blue planes of the input image.
ColorOppn; % executes the Color Opponency Script for Creating the DoG filters for Colors
ColorFeat; % Creates the Color Feature Maps

BY_final= BY_8 + BY_16 + BY_32 + BY_64;
BY_final= BY_final/ max(max(BY_final));

RG_final= RG_8 + RG_16 + RG_32 + RG_64;
RG_final= RG_final/ max(max(RG_final));

%% For intra-channel normalization apply normlise script on the RG_final and BY_final maps

% BY_final= normalise(BY_final);

% RG_final= normalise(RG_final);


ColSalient= BY_final + RG_final;
ColSalient= ColSalient/ max(max(ColSalient));

%-------------------------------------------------------
% Calculating the Orientation Saliency Map

% Create Gabor filters.
Gabor90 = makegrating2( 0, 10, 2 );
Gabor45 = makegrating2( 45, 10, 2 );
Gabor0 = makegrating2( 90, 10, 2 );
Gabor135 = makegrating2( 135, 10, 2 );

Gabor0_7 = imresize(Gabor0, [ 7 7 ], 'bicubic' );
Gabor45_7 = imresize(Gabor45, [ 7 7 ], 'bicubic' );
Gabor90_7 = imresize(Gabor90, [ 7 7 ], 'bicubic' );
Gabor135_7 = imresize(Gabor135, [ 7 7 ], 'bicubic' );

Gabor0_15 = imresize(Gabor0, [ 15 15 ], 'bicubic' );
Gabor45_15 = imresize(Gabor45, [ 15 15 ], 'bicubic' );
Gabor90_15 = imresize(Gabor90, [ 15 15 ], 'bicubic' );
Gabor135_15 = imresize(Gabor135, [ 15 15 ], 'bicubic' );

Gabor0_31 = imresize(Gabor0, [ 31 31 ], 'bicubic' );
Gabor45_31 = imresize(Gabor45, [ 31 31 ], 'bicubic' );
Gabor90_31 = imresize(Gabor90, [ 31 31 ], 'bicubic' );
Gabor135_31 = imresize(Gabor135, [ 31 31 ], 'bicubic' );

% Convolve the intensity contrast maps with the gabor filters.
```

```
Gab0(:,:,1)=(abs(conv2(imConv_8new,Gabor0_7,'same')));
Gab0(:,:,2)=(abs(conv2(imConv_8new,Gabor0_15,'same')));
Gab0(:,:,3)=(abs(conv2(imConv_8new,Gabor0_31,'same')));
Gab0(:,:,4)=(abs(conv2(imConv_16new,Gabor0_7,'same')));
Gab0(:,:,5)=(abs(conv2(imConv_16new,Gabor0_15,'same')));
Gab0(:,:,6)=(abs(conv2(imConv_16new,Gabor0_31,'same')));
Gab0(:,:,7)=(abs(conv2(imConv_32new,Gabor0_7,'same')));
Gab0(:,:,8)=(abs(conv2(imConv_32new,Gabor0_15,'same')));
Gab0(:,:,9)=(abs(conv2(imConv_32new,Gabor0_31,'same')));

Gab45(:,:,1)=(abs(conv2(imConv_8new,Gabor45_7,'same')));
Gab45(:,:,2)=(abs(conv2(imConv_8new,Gabor45_15,'same')));
Gab45(:,:,3)=(abs(conv2(imConv_8new,Gabor45_31,'same')));
Gab45(:,:,4)=(abs(conv2(imConv_16new,Gabor45_7,'same')));
Gab45(:,:,5)=(abs(conv2(imConv_16new,Gabor45_15,'same')));
Gab45(:,:,6)=(abs(conv2(imConv_16new,Gabor45_31,'same')));
Gab45(:,:,7)=(abs(conv2(imConv_32new,Gabor45_7,'same')));
Gab45(:,:,8)=(abs(conv2(imConv_32new,Gabor45_15,'same')));
Gab45(:,:,9)=(abs(conv2(imConv_32new,Gabor45_31,'same')));

Gab90(:,:,1)=(abs(conv2(imConv_8new,Gabor90_7,'same')));
Gab90(:,:,2)=(abs(conv2(imConv_8new,Gabor90_15,'same')));
Gab90(:,:,3)=(abs(conv2(imConv_8new,Gabor90_31,'same')));
Gab90(:,:,4)=(abs(conv2(imConv_16new,Gabor90_7,'same')));
Gab90(:,:,5)=(abs(conv2(imConv_16new,Gabor90_15,'same')));
Gab90(:,:,6)=(abs(conv2(imConv_16new,Gabor90_31,'same')));
Gab90(:,:,7)=(abs(conv2(imConv_32new,Gabor90_7,'same')));
Gab90(:,:,8)=(abs(conv2(imConv_32new,Gabor90_15,'same')));
Gab90(:,:,9)=(abs(conv2(imConv_32new,Gabor90_31,'same')));

Gab135(:,:,1)=(abs(conv2(imConv_8new,Gabor135_7,'same')));
Gab135(:,:,2)=(abs(conv2(imConv_8new,Gabor135_15,'same')));
Gab135(:,:,3)=(abs(conv2(imConv_8new,Gabor135_31,'same')));
Gab135(:,:,4)=(abs(conv2(imConv_16new,Gabor135_7,'same')));
Gab135(:,:,5)=(abs(conv2(imConv_16new,Gabor135_15,'same')));
Gab135(:,:,6)=(abs(conv2(imConv_16new,Gabor135_31,'same')));
Gab135(:,:,7)=(abs(conv2(imConv_32new,Gabor135_7,'same')));
Gab135(:,:,8)=(abs(conv2(imConv_32new,Gabor135_15,'same')));
Gab135(:,:,9)=(abs(conv2(imConv_32new,Gabor135_31,'same')));

% Feature Map for 0 Degree
[r,c]= size(Gab0(:,:,1));
sum0= zeros(r,c);
for i= 1:9
      sum0(:,:)=Gab0(:,:,i) + sum0(:,:);
end

% Feature Map for 45 Degree
sum45= zeros(r,c);
for i= 1:9
   sum45(:,:)=Gab45(:,:,i) + sum45(:,:);
end


% Feature Map for 90 Degree
```

```
sum90= zeros(r,c);
for i= 1:9
  sum90(:,:)= Gab90(:,:,i)+ sum90(:,:);
end

% Feature Map for 135 Degree
sum135= zeros(r,c);
for i= 1:9
   sum135(:,:)=Gab135(:,:,i) + sum135(:,:);
end


% For intra orientation channel normalization of feature maps (sum0, sum45, sum90, sum135)
% sum0 = normalize(sum0);
% sum90 = normalize(sum90);
% sum45 = normalize(sum45);
% sum135 = normalize(sum135);


Orient= sum0 + sum45 + sum90 + sum135;
Orient= Orient/ max(max(Orient)); % Orientation Saliency Map

% Interfeature normalization of saliency maps before summing the maps.
norm_color = normalise(Inten_Sal);
norm_orient = normalise(Orient);
norm_inten = normalise(ColSalient);

% Final Bottom-Up Saliency map
TotalSalient= norm_color + norm_orient + norm_inten;
```

NewVDog8.m

```
% Author : Roger Gaborski
% Generates 8x8 Difference of Gaussian filter

imageSize = 8; % Size of the filter
sigmaex = .04*imageSize; %  excitatory part of the filter.
sigmainh = .16*imageSize; % inhibitory part of the filter.

[x,y] = meshgrid(-3:4 -3:4);

%First Term

exp1 = exp( -1*(x.*x+y.*y)./(2*sigmaex*sigmaex));
FirstTerm = 9.95*exp1;
%%With this value 9.95 the convolution sum is -1.1 instead of +1

%Second Term

exp2 = exp( -1*(x.*x+y.*y)./(2*sigmainh*sigmainh));
SecondTerm = 1.*exp2;

Dog8 = .41*(FirstTerm-SecondTerm); % Difference of Gaussian filter
```

MakeGrating2.m

```
% Author : Roger Gaborski
% Generates Sine and Cosine Gabor filters
% Input arguments :
% orient= orientation of the filter to generate;
% numOfSamples
% numOfCycle : number of cycles in the filter.

function [Gabor_cos, Gabor] = MakeGrating2( orient, numOfSamples, numOfCycles )

%sd must be set up acorrding to size of the original image im
sd = 12;  %12

orient = 2*pi - ( orient*pi/180 );

% create the grating
step = 1/numOfSamples;
[ x,y ] = meshgrid( -pi:step:pi, -pi:step:pi );
ramp = (cos( orient ) * x)  -  (sin( orient ) * y);

im = sin(ramp*numOfCycles);
im_cos = cos(ramp*numOfCycles);

%Generate Gabor
filtSize =min(size(im));
x=linspace(-1,1,filtSize)*filtSize/2;

y=(1/sqrt(2*pi*sd)).*exp(-.5*((x/sd).^2));
filt=(y'*y);
filt=filt./max(filt(:));

Gabor_sine = im .* filt;        % Sine Gabor filter
Gabor_cos = im_cos .* filt;   % Cosine Gabor filter
```

Normalise.m

```
% Author : Vishal Vaingankar
% Function returns the normalized saliency map based on global non-linear combination technique
% Input arguments:
% map: The input saliency map
% scaledMap: output map result of the normalization process.

function [scaledMap]= normalise(map)

% Consider only local maxima regions which are within 70 percent tolerance range of the global maxima
tol=70;

if tol~=0;

        map= map/max(max(map)); % normalising the images betn 0 & 1
        M= max(max(map));

        range=(M*tol)/100; %  tolerance to global maximum
        thresh= M-range;
        locmax=[];

        filter=9;
        marg= (filter-1)/2;
        for i= 5:10:250
          for j= 5:10:250

            val= max(max(map(i-marg:i+marg, j-marg:j+marg)));
            if val ~= M
              if val > thresh
                locmax= [locmax;val];
              end
            end
          end
        end

        meanval= mean(locmax)
        scale= (M-meanval)^2;  % scale is the weight with which the map is modulated

        scaledMap= scale.* map;
else
        map= map/max(max(map)); % normalising the images betn 0 & 1
        scaledMap=map;
end
```

ColorChannelsV1.m

```
% The full size intensity image Iimage0
%   was already created in CarlVisualMainV4.m
%
% Computes the top level matrices of Red, Green, Blue, and Yellow
%   for color contrast calculations
%
% Created by Roger Gaborski
% Validity check for adequate brightness turned on by Carl Reynolds 10/2001

Imax = max(max(Iimage0));

% Cimage0 is the color image from Main.m

r = Cimage0(:,:,1);
g = Cimage0(:,:,2);
b = Cimage0(:,:,3);

Ivalid = Iimage0 > .1*Imax; % Pixels must be bright enough to show color

R0 = r - ( g + b ) / 2;
valid = R0 > 0;       % find values greater than zero
R0 = R0 .* valid;     % Level 0 in the Red color pyramid
R0 = R0 .* Ivalid;    % Must be a bright enough pixel

G0 = g - ( r + b ) / 2;
valid = G0 > 0;       % find values greater than zero
G0 = G0 .* valid;     % Level 0 in the Green color pyramid
G0 = G0 .* Ivalid;

B0 = b - ( r + g ) / 2;
valid = B0 > 0;
B0 = B0 .* valid;     % Level 0 in the Blue color pyramid
B0 = B0 .* Ivalid;

Y0 = (r + g) - 2 * ( abs( r - g ) + b );
valid = Y0 > 0;
Y0 = Y0 .* valid;     % Level 0 in the Yellow color pyramid
Y0 = Y0 .* Ivalid;

clear valid Ivalid;
```

ColorOpponentFilters.m

```
% This script extracts the center and surround regions of the difference of Gaussian regions

RG= R0- G0; % Red excite and Green Inhibitory
GR= G0- R0;  % Green excite and Red Inhibitory

BY= B0- Y0; % Blue excite and Yellow Inhibitory
YB= Y0- B0;  % Yellow excite and Blue Inhibitory

% For OnCenter(R-G) and OffSurround(G-R)
%-----------------------------------
% Creating Center Surround filters for Double Color Opponency
% extract the excitatory part of the multiresolution DoG filter
ex_8= Dog8>0;
newex_8= ex_8.* Dog8;

ex_16= Dog16>0;
newex_16= ex_16.* Dog16;

ex_32= Dog32>0;
newex_32= ex_32.* Dog32;

ex_64= Dog64>0;
newex_64= ex_64.* Dog64;

% extract the inhibitory part of the multiresolution DoG filters
inh_8= Dog8<0;
newinh_8=inh_8.* Dog8;

inh_16= Dog16<0;
newinh_16=inh_16.* Dog16;

inh_32= Dog32<0;
newinh_32=inh_32.* Dog32;

inh_64= Dog64<0;
newinh_64=inh_64.* Dog64;
```

ColorFeature.m

```
% Script for Creating the BY and RG  Color opponent feature maps with different scale filters.

im1=conv2(BY, newex_8, 'same'); % Convolve excitatory region with the map.
im2=conv2(YB, newinh_8, 'same'); % Convolve the inhibitory region with the map.
im3=conv2(1-BY, newex_8, 'same');
im4=conv2(1-YB, newinh_8, 'same');

BexcY_8= im1(:,:)+ im2(:,:); % Blue excitatory/ Yellow  inhibitory map
BinhY_8= im3(:,:)+ im4(:,:);  % Blue inhibitory/ Yellow excitatory map

BY_8= abs(im1(:,:)+ im2(:,:)); % Blue-Yellow double color opponent map

%-----------------------------------------------------------
im1=conv2(BY, newex_16, 'same');
im2=conv2(YB, newinh_16, 'same');
im3=conv2(1-BY, newex_16, 'same');
im4=conv2(1-YB, newinh_16, 'same');

BexcY_16= im1(:,:)+ im2(:,:);
BinhY_16= im3(:,:)+ im4(:,:);

BY_16= abs(im1(:,:)+ im2(:,:));

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
im1=conv2(BY, newex_32, 'same');
im2=conv2(YB, newinh_32, 'same');
im3=conv2(1-BY, newex_32, 'same');
im4=conv2(1-YB, newinh_32, 'same');

BexcY_32= im1(:,:)+ im2(:,:);
BinhY_32= im3(:,:)+ im4(:,:);

BY_32= abs(im1(:,:)+ im2(:,:));

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
im1=conv2(BY, newex_64, 'same');
im2=conv2(YB, newinh_64, 'same');
% im3=conv2(1-BY, newex_64, 'same');
% im4=conv2(1-YB, newinh_64, 'same');

BY_64= abs(im1(:,:)+ im2(:,:));
%===========================================
im1=conv2(RG, newex_8, 'same');
im2=conv2(GR, newinh_8, 'same');
im3=conv2(1-RG, newex_8, 'same');
im4=conv2(1-GR, newinh_8, 'same');

RexcG_8= im1(:,:)+ im2(:,:); Red excitatory/ Green inhibitory map
RinhG_8= im3(:,:)+ im4(:,:);  Red inhibitory/ Green excitatory map

RG_8= abs(im1(:,:)+ im2(:,:)); Red-Green double color opponent map
```

```
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
im1=conv2(RG, newex_16, 'same');
im2=conv2(GR, newinh_16, 'same');
im3=conv2(1-RG, newex_16, 'same');
im4=conv2(1-GR, newinh_16, 'same');

RexcG_16= im1(:,:)+ im2(:,:);
RinhG_16= im3(:,:)+ im4(:,:);

RG_16= abs(im1(:,:)+ im2(:,:));

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
im1=conv2(RG, newex_32, 'same');
im2=conv2(GR, newinh_32, 'same');
im3=conv2(1-RG, newex_32, 'same');
im4=conv2(1-GR, newinh_32, 'same');

RexcG_32= im1(:,:)+ im2(:,:);
RinhG_32= im3(:,:)+ im4(:,:);

RG_32= abs(im1(:,:)+ im2(:,:));

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
im1=conv2(RG, newex_64, 'same');
im2=conv2(GR, newinh_64, 'same');
% im3=conv2(1-RG, newex_64, 'same');
% im4=conv2(1-GR, newinh_64, 'same');

RG_64= abs(im1(:,:)+ im2(:,:));

clear  im1, im2;
```

Learn.m

```
% Script for Creating the Neural Network trained on the target color.

FileName= 'orange.jpg'; % Enter the filename for which the MAT file needs to be loaded for learning

% Loading the Mat file
[token, rem] = strtok(FileName, '.');
temp= ['load ' token ';'];
eval(temp);

[row,col]= size(RG_final);
no_of_feat= 12;
no_ip= row*col; % number of inputs

data= zeros(no_of_feat, no_ip); % Data matrix
output= zeros(1, no_ip); % Output Matrix

% extracting the Mask pixel values in the output matrix.
% orangeMask.jpg is the binary image of the input image (FileName) with the target color regions as 1 and
% distractor color regions as 0
maskimg= imread('orangeMask.jpg');
maskimg= imresize(maskimg,[row,col],'bicubic');
bin= rgb2gray(maskimg);
bin= double(bin)/255;

ip=0;
% Arranging the Matrix values in the Data and Output Vector
% RG is the Red Green Feature Map, BY is the Blue Yellow Feature Map which will be used from the
% current workspace

% copy the map pixel values in the data matrix
for i= 1:row
   for j=1:col
      ip= ip+1;
      feat=0;
      data(feat+1, ip)= RexcG_8(i,j);
      data(feat+2, ip)= RinhG_8(i,j);
      data(feat+3, ip)= RexcG_16(i,j);
      data(feat+4, ip)= RinhG_16(i,j);
      data(feat+5, ip)= RexcG_32(i,j);
      data(feat+6, ip)= RinhG_32(i,j);
      data(feat+7, ip)= BexcY_8(i,j);
      data(feat+8, ip)= BinhY_8(i,j);
      data(feat+9, ip)= BexcY_16(i,j);
      data(feat+10, ip)= BinhY_16(i,j);
      data(feat+11, ip)= BexcY_32(i,j);
      data(feat+12, ip)= BinhY_32(i,j);

      output(1,ip)= bin(i,j);    % Bin is the Binary Mask Of the Learning Image
   end
end

[r,q]= size(data);
[s2,q]= size(output);
```

```
% Creating the Neural Network
net= newff(minmax(data), [8, s2], {'tansig','purelin'}, 'trainrp');

net.trainParam.epochs= 100;
net= train(net, data, output); % Training the Neural Network.
```

Test.m

```
% Script for testing the Test Image with Trained Neural Net.
FileName= 'colimg.jpg'; % Enter the filename for which the MAT file needs to be loaded for testing

%load Olearnt_net; % load the stored neural network for the learnt color
% Loading the Mat file
[token, rem] = strtok(FileName, '.');
temp= ['load ' token ';'];
eval(temp);

[row,col]= size(RG_final);

no_of_feat= 12; % number of features
no_ip= row*col; % number of inputs is total number of pixels.

Testdata= zeros(no_of_feat, no_ip);

% copy the map pixel values in the Testdata matrix
feat=1;
ip=0;
for i= 1:row
   for j=1:col
      ip= ip+1;
      feat=0;
      Testdata(feat+1, ip)= RexcG_8(i,j);
      Testdata(feat+2, ip)= RinhG_8(i,j);
      Testdata(feat+3, ip)= RexcG_16(i,j);
      Testdata(feat+4, ip)= RinhG_16(i,j);
      Testdata(feat+5, ip)= RexcG_32(i,j);
      Testdata(feat+6, ip)= RinhG_32(i,j);
      Testdata(feat+7, ip)= BexcY_8(i,j);
      Testdata(feat+8, ip)= BinhY_8(i,j);
      Testdata(feat+9, ip)= BexcY_16(i,j);
      Testdata(feat+10, ip)= BinhY_16(i,j);
      Testdata(feat+11, ip)= BexcY_32(i,j);
      Testdata(feat+12, ip)= BinhY_32(i,j);
   end
end

net1= sim(net,Testdata);
out_mat= zeros(row,col); % Stores the results of the testing process
row_no=1;
for i= 1:col:row*col
   out_mat(row_no,:)= net1(1, i:i+col-1);
   row_no= row_no + 1;
end
%
mini= min(min(out_mat)); % since the top-down map is between 1 & -1 we need to normalise it to between  0 & 1
storeout_mat= out_mat;
out= out_mat > 0;
out= out .*out_mat;

if max(max(out))>1
   out= out./max(max(out));
end
```