

GOAL DIRECTED VISUAL SEARCH BASED ON COLOR CUES: CO-OPERATIVE EFFECTS OF TOP-DOWN & BOTTOM-UP VISUAL ATTENTION

**ROGER S. GABORSKI,
VISHAL S. VAINGANKAR, ROXANNE CANOSA**

Laboratory for Applied Computing, Rochester Institute of Technology,
102 Lomb Memorial Drive, Rochester, NY 14623.

ABSTRACT

Focus of attention plays an important part in our perception of the world around us. Visual search is a combined effort of the top-down (cognitive cue) and bottom-up (low-level feature conspicuity) processes. Often during visual search our attention involuntarily gets directed to some irrelevant conspicuous objects, such as a bright object, regardless of the cued object. Objects that share similar characteristics with the cued object also influence our attention.

In this paper we analyze the mechanism of visual search based on color cues in natural images as well as computer generated images. We demonstrate the characteristics of the visual scene search by tracing the focus of attention path using data from RIT's human eye tracking system. The results are compared with the computer simulation results of our model in which we capture the interaction between the two systems and explain the behavioral dependencies. The top-down system is implemented using a neural network mimicking the working memory area of the brain and the bottom-up system is implemented using saliency maps.

1. INTRODUCTION

Focus of attention is a crucial mechanism we adopt while observing our surroundings. We as humans do not perceive every aspect of the surroundings, but instead focus on the interesting aspects and ignore the non-interesting ones. Thus attention acts as a gating mechanism to the higher level processing in the visual cortex, processing only the attended locations of the scene. Saccadic movements of the eye get the region of interest under the direct focus of the fovea, an area of maximal acuity in the retina. The filtering theory of attention (Broadbent, 1958) stated the need for attention for processing required information and ignoring the rest. We focus on certain aspects of the image with a higher priority than others. This prioritizing of objects leads us to the concept of saliency of low-level features (image based features) of objects. These low-level properties of objects are extracted by the bottom-up focus of attention mechanism that was initially described by Koch and Ullman (1985), where a saliency map reflects the relative saliency of objects from their surrounding. A top-down cognitive process modulates this early visual attention. For instance, given a visual search task, the knowledge about the target object's features biases the perception of the scene. High-level knowledge about the target object can be its shape, color or motion information. The guided search model (Wolfe, 1994) captures visual search in a computational paradigm. Their model created an activation map, which was a weighted summation of activity in the pre-attentive feature maps. Feature maps with the target object's features

are assigned higher weights. The activation map represented the task based relevance of objects in the scene. Psychology based research in color visual search by D’Zmura (1991) explains the serial/parallel nature of color search. Above-mentioned research in color search has been confined to synthetic images where the research goal has been to find relationship between target and distracter colors. We report results for color search in natural images.

2. MODEL

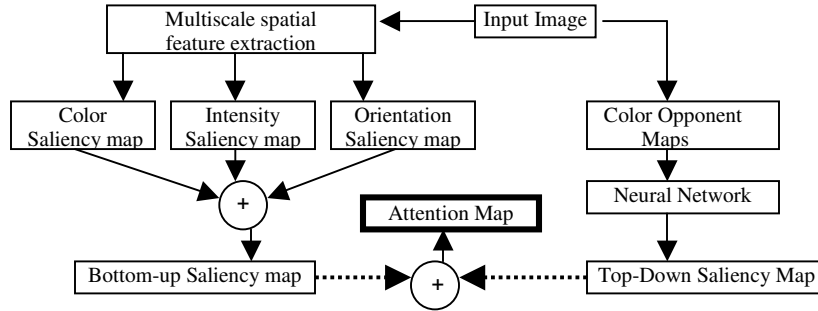


Figure 1: Block diagram of the attention system

2.1 Bottom-Up attention

Conspicuous sensory features of a scene are instrumental in directing our attention. A bright source of light in a dark room instantly catches our attention due to the light’s intensity contrast with the dark surrounding. Computational modeling of early visual mechanisms is implemented in intensity, color and orientation feature channels. Multiresolution feature extracting spatial filters are inspired from the biological vision principles simulating the function of the retina, lateral geniculate nucleus and the primary visual cortex. Intensity contrast in the input image is calculated using Difference of Gaussian filters that simulate the center-surround receptive field structure of the retinal ganglion cells. Color opponent cells in the lateral geniculate nucleus are modeled using the center-surround receptive field for color, where an on-center red, off-surround green filter extracts the red regions of the image and an on-center green and off-surround red extracts the green regions of the image. Blue-yellow color opponent filter extracts blue and yellow regions of the image. Color opponent cells that are excited by red and inhibited by green are denoted as having R+G- response. Cells with opposite response characteristics are G+R-. Similar behavior is associated with the blue-yellow color opponent cells. Processing in the color channel produces four different color opponent feature maps, which are as follows, red excitatory/green inhibitory (R+G-), red inhibitory/green excitatory (R-G+), (B+Y-), (B-Y+).

Orientation information is extracted using directional Gabor wavelets of 0°, 45°, 90° and 135° orientations. Gabor filters are convolved with the input image to extract edges of above-mentioned orientations. Convolution of spatial filters with the input image produces a set of topographical feature maps of the input image. Since the center-surround mechanism is used for extraction of the features, the topographical maps obtained are called feature salience maps since they contribute to the spatial competition within a map during feature extraction. Previous approaches (Itti and Koch, 2001)

constructed feature pyramids for each feature channel and applied spatial competition using Difference of Gaussian filters on the feature maps to generate saliency maps.

Multi resolution feature saliency maps are summed together within the respective feature channels to generate the final three feature saliency maps. The three feature saliency maps are linearly summed to form a final saliency map called the bottom-up saliency map as shown in Fig.1. This inter-feature channel summation contributes to the inter-feature competition in the bottom-up saliency map. The bottom-up saliency map is representative of the combined effects of the three feature channels with varying strength from each channel.

2.2 Top-Down attention (Neural Network Architecture)

During a 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. The neural network in our system mimics the working memory area of the brain. The inputs to neural network are the color opponent cell output maps (R+G-, R-G+, B+Y-, B-Y+). Four different neural networks are trained on the four prototype colors (Red, Yellow, Green, Blue). 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.

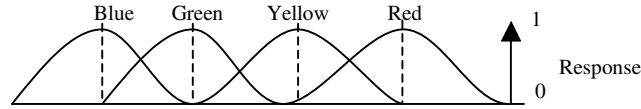


Figure 2: Color opponent cell responses.

As observed from the response curve Fig. 2, 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 color opponent cells. This behavior can be simulated using a neural network. The interaction between the responses of different color opponent cells gives the highest response for the prototype 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 prototype color. Saliency values of the output map close to 1 signify high target color and values closer to 0 lack the target color properties.

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 prototype colors.

Another neural network architecture proposed by De Valois et al. (1966) explained the interaction between the three cone types (short, medium and long wavelength) to explain the opponent responses of the color opponent cells that produce a myriad of color perception. Using the neural network as a model for capturing the interaction between the color opponent cells we are able to efficiently predict the regions of the input image with task-relevant saliency.

2.3 Interaction between Top-down & Bottom-up attention

Research conducted by Parkhurst et al. (2002) explains the effects of bottom-up and top-down attention on perception of a scene. The bottom-up and the top-down saliency maps in our system are summed to generate a final attention map. 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 1 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. The results section of this paper addresses this topic.

3. EXPERIMENTAL METHODS

Ten Rochester Institute of Technology students were selected as subjects for conducting the eye tracking experiments. Since color was the basis for the visual search, only subjects with normal color vision were chosen. Subjects were naive about the nature of experiments. Subjects were seated at a distance of 38" from the 50" Pioneer plasma display on which the images were displayed. Experiment images consisted of indoor scenes, bookshelf scenes and computer-generated scenes.

3.1 Procedure

The first part of the experiment displayed images for free viewing. Subjects were instructed to freely view images displayed on the screen, and proceed to the next image by pressing the space bar key. This part of the analysis gives us an understanding of the non-goal directed perception of a scene. The next part displayed images from a different image set for the search task. Search task consisted of searching for objects of a particular color in the image and subjects were instructed to proceed onto the next image after locating all the target colored objects.

3.2 Eye tracking

Rochester Institute of Technology's Visual Perception Laboratory uses an Applied Science Laboratory model 501 head-mounted eyetracker to monitor subject's eye movements while viewing a scene. Recent research (Pelz and Canosa, 2001) conducted on the eyetracker shows the scope of the eyetracker for predicting human behavior in different visual perception tasks. The eyetracker monitors the subject's pupil and the first corneal reflection of an infrared illuminator. The fixation coordinates and fixation duration of the scanpath are collected for data analysis. Fixation duration is the amount of time spent at the fixation points.

4. RESULTS

In order to estimate how well the fixations correspond to various saliency maps in the system, a correlation measure was calculated. The saliency map mean is indicative of the strength of individual maps. The scanpath is overlaid on the saliency map and the saliency values at the location of fixations are extracted. The mean of all the extracted saliency values is termed the fixation mean (fix_mean). A ratio of fix_mean to the map mean is calculated. If this ratio is close to 1, it indicates that the fixations were randomly distributed around the map since the fix_mean and map mean are almost similar. A ratio

significantly higher than 1 indicates the 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. Fixation duration also can be used for calculating the correlations values by multiplying the extracted fixation saliency with the corresponding fixation duration. The ratio of fix_mean_duration and map mean generates the correlation value based on duration.

4.1 Relationship between the target and distracter prototype colors in visual search.

The correlation for various prototype colors is calculated by overlaying the search scanpath onto the R+G-, R-G+, B+Y- and B-Y+ maps. Pattern bars in the graphs of Fig.3 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 normalized to the highest of the four correlations.

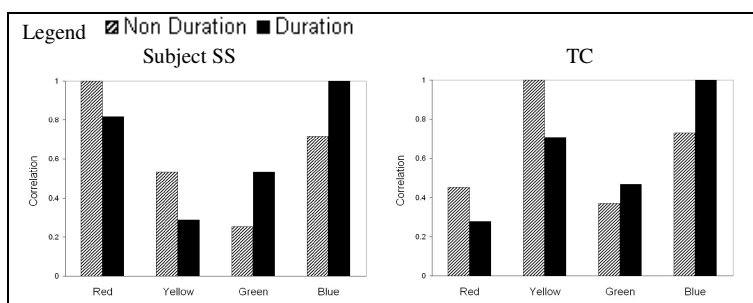


Figure 3. The target color is Blue. Graphs for two subjects are shown

In Fig. 3 the target color (blue) has a consistently high correlation value in the non-duration analysis, but other distracter colors tend to have comparatively high correlation values too. This behavior is because the subjects fixated the distracter colors while searching for the target colors. For subject SS, red distracter attains highest correlation in non-duration analysis and for subject TC, yellow was a prominent distracter. When applying fixation durations as weighting factors, the target color emerges with the highest correlation value, and the distracter colors attain more decorrelation with the target color. The graphs suggest that the fixation durations reflect important information about the nature of search.

4.2 Graphs explaining the relative effects of top-down and bottom-up attention on eye movements during visual search.

Analyses of our results show that bottom-up factors do affect the visual search task. The graphs in Fig. 4 are obtained by overlaying the search scanpath on the top-down and bottom-up saliency map. The y-axis shows the normalized correlation values, and the x-axis the sequential fixation numbers of the scanpath for the search conducted by subjects. The correlation values are normalized with respect to the maximum correlation possible for the respective saliency maps. Maximum correlation is calculated by taking a ratio of 1 (global maxima of the map) to the map mean. Correlation values in top-down curve are normalized with the maximum correlation of the top-down map, obtained by dividing 1 by the mean of top-down map. The normalized correlation values lie between 0 and 1, with 1 indicating the maximum correlation. This normalization allows us to compare the fixation strengths of the two curves. The top-down and bottom-up correlation curves are depicted in the graphs.

When not focusing on a target object, the low-level conspicuous features catch the subject's attention, which are predicted by the bottom-up saliency map. This is evident in the graphs of Fig.4 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 subject JS, NB. The results indicate that search task is influenced by intermediate task irrelevant bottom-up fixations in process of fixating on search relevant cognitive cues (top-down features). Another interesting observation is, even after the target object is located and if the search is in progress, the final fixations are primarily on high salient regions of the bottom-up map evident for subject NB, TC.

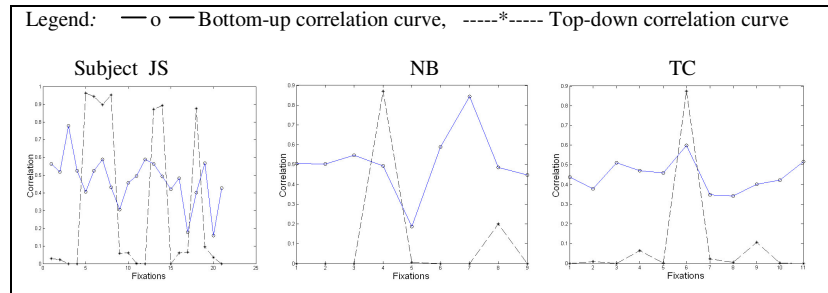


Figure 4: Search scanpath overlaid on top-down and bottom-up saliency maps.

5. CONCLUSION

As an engineering model the system is able to predict the regions of the scene that have a high probability of being fixated when confirmed with human data. This is evident from the high correlation values obtained from correlation curves in the graphs that are obtained from the eye movements of subjects. Neural network is efficient as a predictor for search task relevant regions in a scene. Color spectrum response to changing wavelength is captured using neural network. The system is not intended to be complete representation of the human visual search mechanism. Psychophysical factor like reduction in spatial resolution with increasing visual angle from the point of gaze has not been attempted.

ACKNOWLEDGEMENTS

We would like to thank Dr. Jeff B. Pelz, Center for Imaging Science, RIT, for granting us access to the eyetracker equipment in the Visual Perception Lab. We appreciate the support of all the students who participated as subjects in the experiments.

REFERENCES

- Broadbent D. E. (1958) *Perception And Communication*. London: Pergamon
- De Valois, Abramov, & Jacob, (1966). Analysis of response patterns in LGN cells. *Journal of the Optical Society of America*, 56, 966-977
- D'Zmura M. (1991) Color in Visual Search. *Vision Research*. 31(6) pp 951-966.
- Itti, L. and Koch, C. Computational Modeling of Visual Attention *Nature Neuroscience Review* (2001) 2, 194-204
- Koch C, Ullman S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*. 4, 219-227
- Parkhurst, Law, and Neibur (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1), 107-123.
- Pelz, J.B. & Canosa, R. (2001) Oculomotor Behavior and perceptual strategies in complex tasks. *Vision Research*. 41: 3587-3596.
- Wolfe, J.M (1994) Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review* 1(2), 202-238.