Modeling Selective Perception of Complex, Natural Scenes

Roxanne L. Canosa

Department of Computer Science
Rochester Institute of Technology
Rochester , NY 14623
rlc@cs.rit.edu

Abstract

Computational modeling of the human visual system is of current interest to developers of artificial vision systems, primarily because a biologically-inspired model can offer solutions to otherwise intractable image understanding problems. The purpose of this study is to present a biologically-inspired model of selective perception that augments a stimulus-driven approach with a high-level algorithm that takes into account particularly informative regions in the scene. The representation is compact and given in the form of a topographic map of relative perceptual conspicuity values. Other recent attempts at compact scene representation consider only low-level information that codes salient features such as color, edge, and luminance values. The previous attempts do not correlate well with subjects' fixation locations during viewing of complex images or natural scenes. This study uses high-level information in the form of figure/ground segmentation, potential object detection, and task-specific location bias. The results correlate well with the fixation densities of human viewers of natural scenes, and can be used as a pre-processing module for image understanding or intelligent surveillance applications.

Introduction

Visual perception is an inherently active and selective process with the purpose of serving the needs of the individual, as those needs arise. An essential component of active visual perception is a selective mechanism. Selective perception is the means by which the individual attends to a subset of the available information for further processing along the entire visual pathway, from the retina to the cortex. The advantage of selecting less information than is available is that the meaning of the scene can be represented compactly, making optimal use of limited neural resources. Recent studies on *change-blindness* (Rensink, O'Reagan, and Clark 1997) have shown that observers of complex, natural scenes are mostly unaware of large-scale changes in subsequent viewings of the same scene. These studies serve as an example of how efficient encoding may adversely effect visual recall.

A compact representation assumes that an attentional mechanism has somehow already selected the features to be encoded. The problem of how to describe an image in terms of the most visually conspicuous region usually takes the form of a 2D map of saliency values (Koch and Ullman 1985). In the saliency map, the value at a coordinate provides a measure of the contribution of the corresponding image pixel to the overall conspicuity of that image region.

The two most common methods of modeling the effects of saliency on viewing behavior are the bottom-up, or stimulus-driven approach, and the top-down, or task-dependent approach. Stimulus-driven models begin with a low-level description of the image in terms of feature vectors, and measure the response of image regions after convolution with filters designed to detect those features. Parkhurst, Law, and Niebur (2002) as well as Itti and Koch (2000) have used multi-resolution spatio-chromatic filters to detect color, luminance, and oriented edge features along separate, parallel channels. These models correlate well to actual fixation locations when the input image is non-representational and no explicit task has been imposed upon the viewer other than free-viewing, but do not correlate well to fixations on natural images of outdoor and indoor scenes.

Early studies on viewing behavior have found that the eyes do not fixate on random locations in the field, but rather on regions that rate high in information content, such as edges, lines and corners (Hebb 1949, and Kauffman and Richards 1969). These studies were primarily concerned with spontaneous fixation patterns during free viewing of scenes and largely ignored the high-level aspects of eye movement control, such as prior experience, motivation, and goal-oriented behavior.

Other early studies showed that high-level cognitive strategies are reflected in the patterns of eye movement traces (Yarbus 1967 and Buswell 1935). Distinctly different "signature" patterns of eye movement traces could be elicited from subjects when specific questions were posed to the subjects. More recently, Land, Mennie, and Rusted (1999) showed that eye movements monitor and guide virtually every action that is necessary to complete an over-learned task such as making tea. Turano, Geruschat, and Baker (2003) compared a bottom-up salience model with a top-down guided-search model in terms of the model's ability to predict the oculomotor strategies of subjects engaged in an active, natural task. The visual salience model was found to perform no better than a model based on random scanning of the scene. The top-down model, which incorporated geographic information in the form of expected location criteria performed better than the salience model. A model that used both salience information and geography performed best of all. Feature saliency may be a reliable indicator for determining which regions are fixated for free-viewing simple images, but not for oculomotor behavior that requires forming a plan of action (Pelz and Canosa 2001).

The purpose of the present study is to propose a biologically plausible model of selective visual attention that incorporates low-level feature information from the scene with high-level constructs and top-down, task-oriented constraints. The model takes the form of a topographic map of perceptual conspicuity values, and is called a *conspicuity map.* The value at a coordinate in the map is a measure of how conspicuous that coordinate is in terms of perception. The resulting model is shown to correlate well with the fixation densities of subjects who viewed natural scenes.

Model Description

This section describes in detail the steps that were taken to construct the conspicuity map. The conspicuity map consists of three essential modules – the pre-processing module that produces a color map and an intensity map, an edge module that produces an orientated edge map, and an object module that produces a proto-object map. The maps are merged together, and an object mask is applied to the result to inhibit areas that do not correspond to potential objects and enhance areas that do. Figure 1 shows a schematic of the processing modules and the resulting importance map.

Input Image Pre-processing

Before the low-level features of the conspicuity map can be computed, the input image must be pre-processed to represent the image in terms of the human early physiological response to stimuli. The pre-processing stage takes as input the original RGB formatted image and performs a non-linear transform of that image from the RGB color space to the CIE tristimulus values, X, Y, and Z. The tristimulus values take into account the spectral power properties of the display device (described in the next section) and the color-matching functions of the CIE Standard Colorimetric Observer.

The next step is to perform a linear transform of the tristimulus values into rod and cone responses, using the transformation matrices given in Pattanaik, *et al.,* (1998), as shown in Equations 1 and 2.

$$\begin{vmatrix} L \\ M \\ S \end{vmatrix} = \begin{vmatrix} 0.3897 & 0.6890 & -0.0787 \\ -0.2298 & 1.1834 & 0.0464 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} X \\ Y \\ Z \end{vmatrix} \qquad (1)$$

$$R = -0.702X + 1.039Y + 0.433Z \qquad (2)$$

The rod signal was derived from the tristimulus values as an approximation using a linear regression of the color matching functions and the CIE scotopic luminous efficiency function, V'(λ), as given in Wyszecki (1982). The cone responses are from the Hunt-Pointer-Estevez responsivities as given in Fairchild (1998). The final pre-processing step is to compute the two opponent color channels and the achromatic channel from the normalized rod and cone response signals. The opponent color channels detect chromaticity differences in the input image, and simulate the subjective experience of color resulting from the four chromatic primaries arranged in polar pairs – red/green and blue/yellow. The achromatic channel simulates the subjective experience of luminance along the black/white achromatic dimension.

The transformation from rod and cone responses into opponent signals, as shown in Equation 3, uses the matrices given in Pattanaik, *et al.,* (1998), which follow those of Hunt (1995), and are also used in the CIE color appearance model of CIECAM97 (Fairchild, 1998). In Equation 3, A refers to the achromatic channel, C1 refers to the R/G opponent channel, and C2 refers to the B/Y opponent channel. After the calculation of rod and cone signals, the low-level feature maps are computed.

$$\begin{vmatrix} A \\ C1 \\ C2 \end{vmatrix} = \begin{vmatrix} 2.0 & 1.0 & 0.05 \\ 1.0 & -1.09 & 0.09 \\ 0.11 & 0.11 & -0.22 \end{vmatrix} \begin{vmatrix} L \\ M \\ S \end{vmatrix} \qquad (3)$$

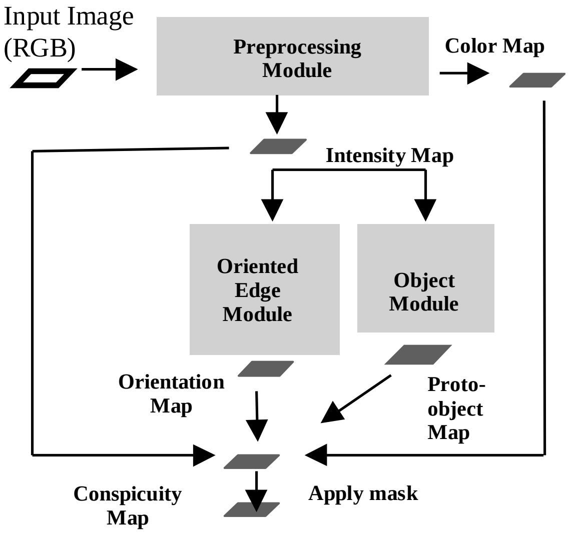

Figure 1. Construction of the conspicuity map.

The Low-level Saliency Map

The saliency map consists of three low-level feature maps – a color map, an intensity map, and an oriented edge map. Using these three maps to represent the low-level features of the saliency map follows Parkhurst, Law, and Niebur (2002), and Itti and Koch (2000), however the computational steps that derive the maps differ significantly from the earlier approaches. The color computation takes as input the two chromatic signals, C1 and C2. The resulting "colorfulness" of each pixel from the input image is expressed in color-opponent space as given in Equation 4. This value is calculated for every pixel, resulting in the color map.

$$\text{colorfulness} = \sqrt{(C1^2 + C2^2)} \qquad (4)$$

Combining the output from the rod signal with the output from the achromatic color-opponent channel creates the intensity map. Thus, the total achromatic signal, A_T, consists of information originating from the cone signals as well as from the rod signal. The rod signal is assumed to consist of only achromatic information. Pattanaik, *et al.* (1998) determined the differential weightings of the rod and cone signals that will result in an achromatic output which is monotonic with luminance, as given in Equation 5.

$$A_T = A + \text{rod} / 7 \qquad (5)$$

The oriented edge map also takes as input the intensity signal. To create the oriented edge map, the first stage of processing is the computation of a multi-resolution Gaussian pyramid (Burt and Adelson, 1983) from the intensity signal. To create the Gaussian pyramid, the intensity signal is sampled at seven spatial scales (1:1, 1:2, 1:4, 1:8, 1:16, 1:32, and 1:64) relative to the size of the original input image, 1280 x 768 pixels. Each level is up-scaled to the highest resolution level using bicubic interpolation.

The second stage of processing simulates the center-surround organization and lateral inhibition of simple cells in the early stages of the primate visual system by subtracting a lower resolution image from the next highest resolution image in the pyramid, and taking the absolute value of the result. The resulting six levels of difference images form a Laplacian cube. Each level of the Laplacian cube is a representation of the edge information from the original input image at a specific scale.

Since the human visual system has non-uniform sensitivity to spatial frequencies in an image, the levels of the Laplacian cube must be weighted by the contrast sensitivity function. Contrast sensitivity is modeled by finding the frequency response of a set of difference-of Gaussian convolution filters, and weighting each edge image of the Laplacian cube by the response. The filters alter the weight of each edge according to how sensitive the human visual system is to the frequency response of that particular edge.

The CSF weighting function begins by defining a Gaussian convolution kernel that is the same size as the kernel used for the bicubic interpolation described earlier, 5x5 pixels. Multiple kernels are derived from the original kernel by successively doubling the area. This simulates the effect of convolving a fixed size kernel with each level of the Gaussian pyramid. After all of the kernels have been normalized to 1, each kernel is transformed into the frequency domain using the Fast Fourier Transform algorithm.

Once the convolution kernels have been transformed into the frequency domain, they can be used to create bandpass filters that detect a specific range of frequencies in the input image. Subtracting one frequency domain kernel from another frequency domain kernel creates the bandpass filters. The range of frequencies that will be detected is determined by the frequency response of the filters.

After calibrating the frequency responses of the bandpass filters to correspond to degrees of visual angle, the contrast sensitivity function can be used to determine the visual response to a particular frequency in the Laplacian edge images. The visual responses are used as weights to be applied to the edge images, either enhancing the edge if the human visual system is particularly sensitive to that frequency, or inhibiting that edge if it is not. Equation 6 gives the contrast sensitivity function used to find the weights (from Manno and Sakrison 1974).

$$CSF = 2.6 * (0.0192 + 0.114 f)\, e^{-(0.114 f)\ \wedge\ 1.1} \qquad (6)$$

The final step in the creation of the oriented edge map is to represent the amount of edge information in the image at varying spatial orientations. This is done by convolving the edge image with Gabor filters at four orientations - 0°, 45°, 90°, and 135°. This simulates the structure of receptive fields in area V1 neurons that are tuned to particular orientations, as well as to specific spatial frequencies (Hubel and Wiesel 1968). Figure 2 is a graphical depiction of the basis functions used to model the receptive fields of these neurons.

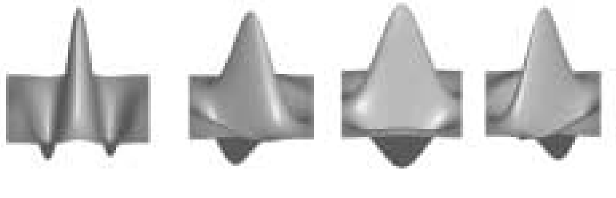

Figure 2. Basis functions of the Gabor filters used to model the tuning of receptive fields in area V1 of striate cortex. From left, 0°, 45°, 90°, and 135°.

Once the color map, intensity map, and the oriented edge map have been generated, they are linearly scaled from 0 to 1 and merged together to create a single low-level saliency map by adding the values from each map together on a pixel-by-pixel basis.

The High-level Proto-object Map

The proto-object map is constructed in parallel with the saliency map, and is used to identify potential objects in the image. The algorithm is based upon detecting texture from edge densities. The first stage involves segmenting an estimation of the background from the foreground of the image, using the intensity map of the image as input. The effect of this step is that regions of relatively uniform intensity in the image are localized, simulating the effect of figure/ground segmentation of perceptual organization (Rubin 1921).

For the next stage, a threshold is applied to create a binary representation of the foreground, which is subsequently used for edge detection. From the edge image, regions corresponding to potential objects in the image are grown using morphological operations. The result is called the proto-object map and represents the location of potential objects in the image. The proto-object map is used along with the color map, intensity map, and oriented edge map as an additional channel for the calculation of conspicuity. Once the four channels have been merged into a single map, the proto-object map is used once again as a mask to further inhibit regions in the image that are not likely to correspond to object locations, and enhance those regions that are. Figure 3 shows four example input images and the corresponding low-level and high-level maps for each image. The bright areas in the maps correspond to highly conspicuous regions in the example images. These areas are where the model predicts the viewer's visual attention is likely to be captured.

Verification of Model

In order to determine the correlation between the model and the fixation patterns of people viewing natural scenes, eye-tracking data was collected and analyzed. An ASL model 501 head-mounted eye-tracker was used to record gaze positions, which were calculated at a video field rate of 60 Hz providing a temporal resolution of 16.7 msec. A 50 inch Pioneer plasma display with a screen size of 1280 x 768 pixels and a screen resolution of 30 pixels/inch was used to display the images. The display area subtended a visual field of 60° horizontally and 35° vertically at a viewing distance of 38 inches. At this distance, approximately 21 pixels cover 1° of visual angle.

Data Collection

Eleven subjects participated in the eye-tracking experiment, all with normal or corrected to normal vision. A calibration procedure was performed for each subject prior to the beginning of the experiment, and checked at the end of the experiment. After offset and drift correction the average angular deviation from the calibration points was 0.73° ± 0.06° at the start of the experiment, and 0.56° ± 0.04° at the end of the experiment.

Each subject viewed a total of 164 color images divided over two sets of 82 images each. The two sets were labeled A and B, and were counter-balanced between observers.

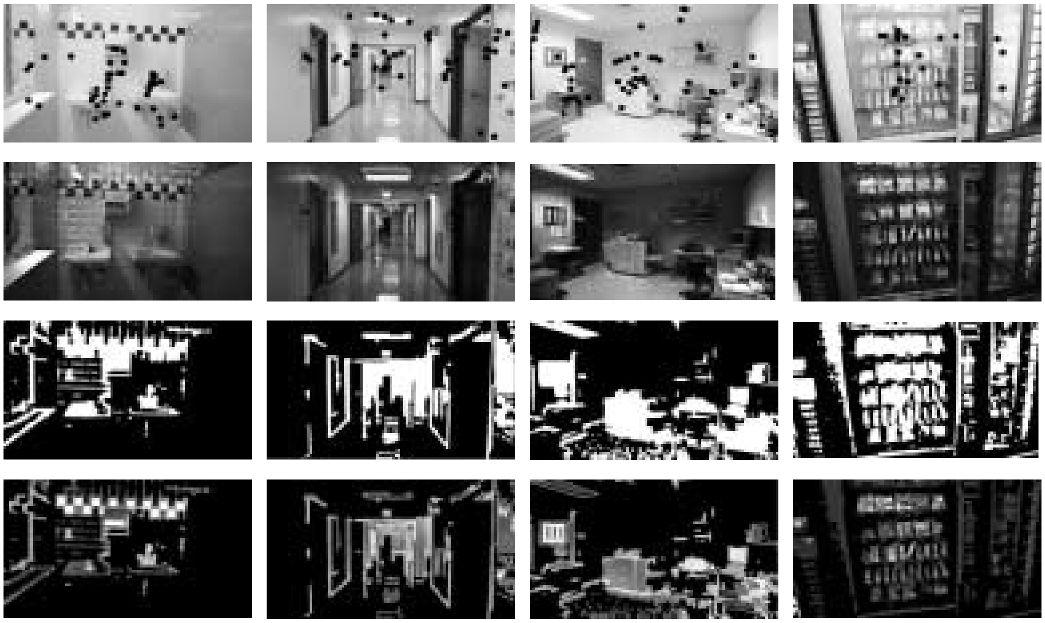

Figure 3. Input images with overlaid fixation plots (1st row) – from left, washroom, hallway, office, and vending. Low-level maps (2nd row), proto-object maps (3rd row), and final conspicuity maps (4th row).

The image database represented a wide variety of natural images, including indoor and outdoor scenes, landscapes, buildings, highways, water sports, scenes with people, and scenes without people. The experiment consisted of two parts – "free-view," where the subject was instructed to freely view each image as long as desired, and "multi-view," where the subject was given an explicit instruction before viewing the image. Free-view always preceded multi-view.

Results

A metric was developed to measure the correlation between the density of subjects' fixations on a particular image and the model predictions of fixation locations. The metric compares the conspicuity of the fixated regions to the average conspicuity of the map, and is referred to as the F/M ratio.

The mean conspicuity of fixations is defined as the average conspicuity value extracted from a map at the x,y-coordinates of the fixation locations, for all fixations on a particular image. The mean conspicuity of the map is the average value of the map generated from the model. The F/M ratio is the ratio between the mean conspicuity of fixations and the mean conspicuity of the map.

The F/M ratio is used to determine how well the model is able to predict fixation locations. If the F/M ratio is close to one, then the map generated from the model is not a good predictor of fixation locations, since the mean conspicuity of a feature map is the expected value at any random location in the map. If the F/M ratio is higher than one, then the map is a good predictor because the fixations tend to be on regions of the image that the model has computed as being highly conspicuous.

To compare the predictive power of the model using several different feature parameters, four maps were generated for the 164 images in the image database used for the eye-tracking experiment. The four maps generated are given in Equations 7, 8, 9, and 10:

$$\text{CIE map} = (C + I + E) / 3 \quad (7)$$

$$\text{P map} = P \quad (8)$$

$$\text{CIEP map} = ((C + I + E + P) / 4) \cdot P \quad (9)$$

$$\text{C\_Map} = (C\cdot w_1 + I\cdot w_2 + E\cdot w_3 + P\cdot w_4) \cdot P\cdot w_5 \quad (10)$$

C refers to the color feature map, I refers to the intensity feature map, E refers to the oriented edge feature map, P refers to the proto-object map, and C_Map refers to the conspicuity map. In Equation 10, w_1 through w_5 refer to weights applied to the features used to derive the conspicuity map. The weights were found using a genetic algorithm, where a near-optimal solution was found on a per-image basis. In essence, the genetic algorithm assigns a weight to each feature according to how well that feature is able to predict a fixation location. Thus, the C_Map uses knowledge about fixation locations to weight the individual features. This knowledge can be used in a future implementation of the map generation algorithm to classify images and assign weights based on the classification results. Figure 4 shows a comparison of the F/M ratios for each of the maps for the 152 images under the free-viewing condition.

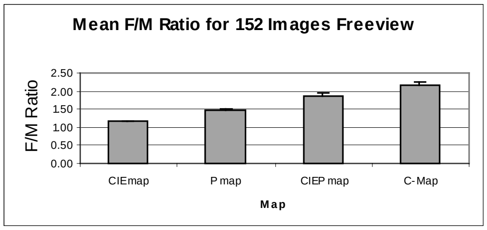

Figure 4. Mean F/M ratios for four maps generated from 152 images used for the eye-tracking study.

CIE is the conspicuity map consisting of only low-level image features such as color, luminance, and oriented edge information with equal weighting. The correlation of CIE conspicuity values to subjects' fixation locations is low, as the F/M ratio is close to 1 for nearly every image. Any random location on the map would produce nearly as high of a conspicuity value. P is the proto-object map used alone. This map shows a significantly higher correlation to fixation locations than the CIE map. CIEP uses the proto-object map as an added feature and as a binary mask to inhibit the features in the map that do not correspond to potential objects. The CIEP map has a higher correlation than either the CIE map or the P map alone does. The C_Map shows the highest correlation to subjects' fixation locations, which is expected because of prior knowledge.

The improvement in F/M ratio as the maps include more information about objects and potential objects shows that attention is more likely to be directed to objects in a scene, rather than to highly salient, non-object features. This is an indication that perceptual relevancy, rather than feature salience, guides fixation patterns. Ultimately, a map of perceptual conspicuity rather than of feature salience is likely to be a better predictor of fixation locations.

A positive feature of the object-oriented approach to predicting fixation locations is that a central bias in the image is preserved when the bias is warranted, and not preserved when the bias is not warranted. For example, an analysis of fixation distances from the center of each of the four example images found that most of the fixations were within ±10º of the center of three of the four images (washroom, hallway, and vending) even though the central area comprised less than 20% of the total image area. At a distance of ±10º, 63% of washroom fixations, 66% of hallway fixations, 44% of office fixations, and 52% of vending fixations were found. From this data it can be concluded that subjects preferentially fixated the center of these images; thus the fixations cannot be considered randomly distributed across the image space for any of the four example images.

To test the conspicuity map for a preserved central bias, a random fixation sequence was generated that constricted all fixations to a 1/16 image size window (14.5º x 8.7º). Figure 5 shows the ratios of fixation means to map means for the constrained fixations. The F/M ratio is high for both the proto-object and conspicuity maps when fixations are constrained to the centers of the washroom, hallway and vending images. This shows that the higher-level maps that incorporate object information correctly simulate the general tendency to look towards the center of an image, when that tendency is warranted. Other proposed models (Turano, Geruschat, and Baker 2003) include an imposed, explicit bias toward the center to improve performance, however the object-oriented model presented here preserves the central tendency without artificial means.

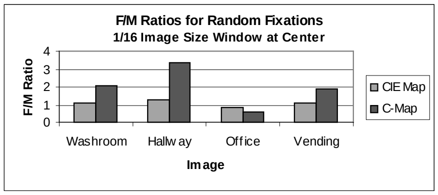

Figure 5. F/M ratios for random fixations restricted to 1/16 image size distance from center.

Conclusion

This study showed that locating highly conspicuous regions of an image must ultimately take into consideration the implicit semantics of the image – that is, the "meaningfulness" of the contents of the image for the viewer as exemplified by objects in that image. Objects as well as their locations in the scene play an important role in determining meaningfulness in natural, task-oriented scenes, especially when combined with action-implied imperatives. The low-level, bottom-up features of an image cannot be ignored, however, because it is those features that capture the attentional resources in the early stages of processing, sometimes in an involuntary way. Successfully predicting fixation densities in images requires computational algorithms that combine bottom-up processing with top-down constraints in a way that is task-relevant, goal-oriented and ultimately most meaningful for the viewer as well as for the particular image under consideration.

References

Burt, P. J., and Adelson, E.H. 1983. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, **31**(4):532-540.

Buswell, G.T. 1935. *How People Look at Pictures: A Study of the Psychology of Perception in Art*, Chicago:The University of Chicago Press.

Fairchild, M.D. 1998. *Color appearance models.* Reading, MA: Addison-Wesley.

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

Hebb, D.O. 1949. *The Organization of Behavior*, New York:John Wiley & Sons.

Hubel D.H., and Wiesel, T.N. 1968. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, **195**:215-243.

Hunt, R.W.G. 1995. *The reproduction of color.* 5thedition. Kingston-upon-Thames,England: Fountain Press.

Itti L., and Koch, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, **40**:1489-1506.

Kaufman L., and Richards, W. 1969. Spontaneous fixation tendencies for visual forms. *Perception and Psychophysics*, **5**(2), :85-88.

Hebb, D.O. 1949. *The Organization of Behavior*, New York:John Wiley & Sons.

Land, M.L., Mennie, N., and Rusted, J. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception*, **28**:1311-1328.

Manno, J.L., and Sakrison, D.J. 1974. The effects of a visual fidelity criterion on the encoding of images. *IEEE Transactions of Information Theory* **20**(4):525-535.

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

Pattanaik, S.N., Ferwerda, J.A., Fairchild, M.D., and Greenberg, D.P. 1998. A multi-scale model of adaptation and spatial vision for realistic image display. *Proceedings of the SIGGRAPH 98:*287-298.

Pelz, J.B. and Canosa, R. 2002. Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, **41**:3587-3596.

Rensink, R.A., O'Reagan, J.K., and Clark, J.J. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, **8**(5):368-373.

Rubin, E. 1921. *Visuell Wahrgenommene Figuren*, Glydenalske boghandel, Kobenhaven.

Turano, K.A., Geruschat, D.R., and Baker, F.H. 2003. Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, 43:333-346.

Wyszecki, G., and Stiles, W.S. 1982. *Color Science: Concepts and Methods, Quantitative Data and Formulae* (2nd edition). New York: Wiley.

Yarbus, A. 1967. *Eye Movements and Vision*, New York: Plenum Press.