

# MODELING SELECTIVE PERCEPTION OF COMPLEX, NATURAL SCENES

ROXANNE CANOSA<sup>†</sup>

*Department of Computer Science  
Rochester Institute of Technology  
Rochester, New York, USA*

*rlc@cs.rit.edu*

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.

*Keywords:* Machine perception; active vision; saliency map; attention.

## 1. 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 visual pathway, from

---

<sup>†</sup>Completed as part of dissertation at the Center for Imaging Science, Rochester Institute of Technology, Rochester, New York.

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*<sup>1</sup> 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 affect 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 regions usually takes the form of a 2D map of saliency values.<sup>2</sup> 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. Previous attempts<sup>3,4</sup> 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 or 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.<sup>5,6</sup> 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 patterns of eye movement traces.<sup>7,8</sup> Distinctly different “signature” patterns of eye movements could be elicited from subjects when specific questions were posed to them. More recently, studies have found that eye movements monitor and guide virtually every action that is necessary to complete an over-learned task such as making tea.<sup>9</sup> In a separate study, a bottom-up salience model was compared to a top-down guided-search model in terms of the model’s ability to predict oculomotor strategies of subjects engaged in an active, natural task.<sup>10</sup> 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 salience may be a reliable indicator for determining which regions will be fixated for free-viewing simple images, but not for predicting oculomotor behavior that requires forming a plan of action.<sup>11</sup>

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 a high-level potential object detector and a central bias, when that bias is warranted. The model takes the form of a topographic map of perceptual conspicuity values, and is called a *conspicuity map*. The value of a coordinate in the map is a measure of how perceptually conspicuous that particular coordinate is for the human visual system. The resulting model is shown to correlate well with the fixation densities of subjects who view natural scenes.

## 2. 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 oriented 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 probable object locations and enhance areas that do. Figure 1 shows a schematic of the processing modules and the resulting conspicuity map.

### 2.1. *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 transformation 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 properties of the display device and the color matching functions of the CIE Standard Colorimetric Observer.

The next step is to perform a linear transformation of the tristimulus values into rod and long-, medium-, and short-wavelength cone responses,

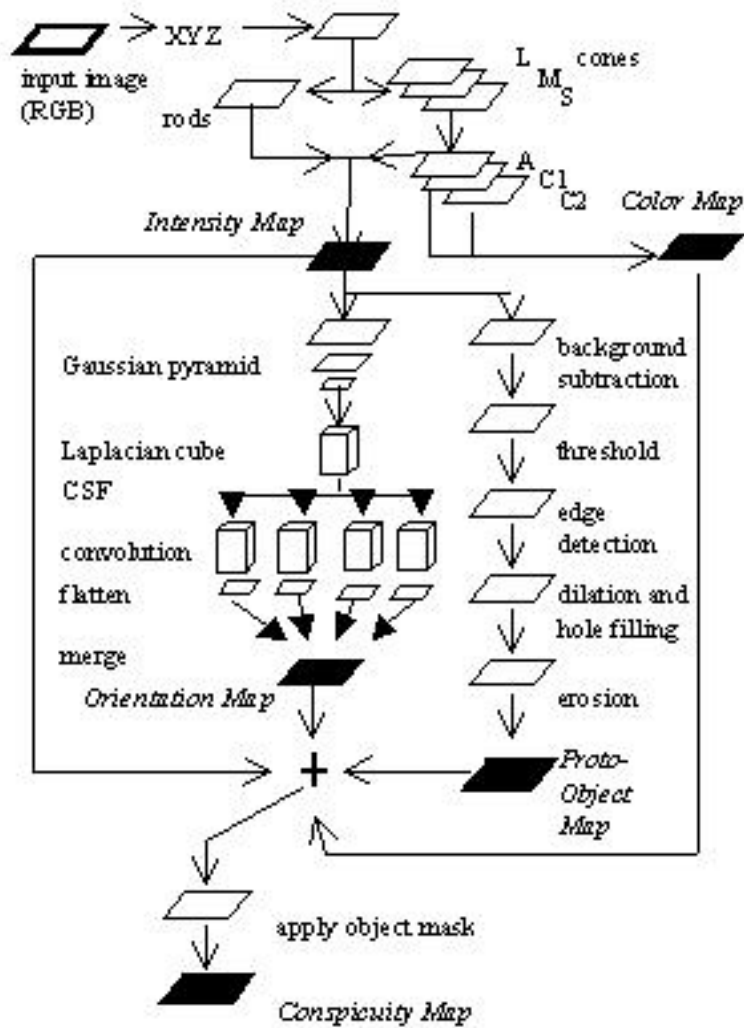


Figure 1: Construction of the conspicuity map.

using the transformation matrices as shown in Eqs. (1) and (2)<sup>12</sup>

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

$$rod = -0.702X + 1.039Y + 0.433Z. \quad (2)$$

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,  $V(\lambda)$ .<sup>13</sup> The cone responses are from the Hunt-Pointer-Estevéz responsivities.<sup>14</sup> 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 represent chromaticity differences in the input image, and simulate the subjective appearance of color resulting from the 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 Eq.(3), is the same transformation that is used in the CIE color appearance model of CIECAM97.<sup>12,14,15</sup> In Eq. (3),  $A$  refers to the achromatic channel,  $C1$  refers to the R/G color opponent channel, and  $C2$  refers to the B/Y color opponent channel. After the calculation of the rod and cone signals, the low-level feature maps are computed

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

## 2.2. 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 has been implemented in earlier studies,<sup>3,4</sup> however the computational steps that derive the maps implemented here 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 defined as the vector distance from the origin (neutral point) to a point corresponding to the  $C1$  and  $C2$  signal, expressed in the two-dimensional R/G and B/Y color opponent space. The result is given in Eq. (4). This value is calculated for every pixel, resulting in the color map

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

Combining the output from the rod signal with the output from the achromatic color opponent channels 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. The differential weightings of the rod and cone signals that will result in an achromatic output which is monotonic with luminance is given in Eq. (5)<sup>12</sup>

$$A_t = A + rod/7. \quad (5)$$

$A$  refers to the achromatic luminance information originating from the cone signal, and  $rod$  refers to the luminance information originating only from the rod signal, as given in Eq. (2). The oriented edge module takes as input the intensity signal. The purpose for using the intensity signal for the computation of edge information is to maintain a single representation of luminance throughout the model. Earlier models<sup>3,4</sup> used the average value of the RGB digital counts from the original image to create a luminance image for input into the oriented edge module. Here, the intensity signal represents the luminance output from the retinal receptors, and will be used to simulate the center-surround organization of receptive fields from retinal ganglion cells.

To create the oriented edge map, the first stage of processing is the computation of a multi-resolution Gaussian pyramid of intensity images from the original intensity signal.<sup>16</sup> To create the Gaussian pyramid, the intensity signal is sampled at seven spatial scales ( $1:1$ ,  $1:2$ ,  $1:4$ ,  $1:8$ ,  $1:16$ ,  $1:32$ ,  $1:64$ ) relative to the size of the original input image, 1280 by 768 pixels. Each level is then 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 (CSF). Contrast sensitivity is defined as the sensitivity of an observer to sinusoidal gratings of varying frequencies, and is a measure of an average retinal ganglion cell's receptive field size. A typical primate CSF is depicted in Figure 2, which shows that frequencies in the range of 6-7 cycles per degree are most readily detected. The effects

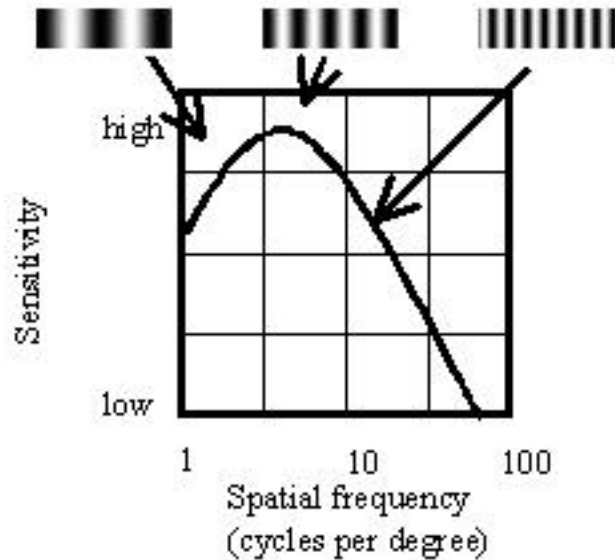


Figure 2: Contrast sensitivity function with example spatial frequencies.

of 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 visibility of each edge according to how sensitive the human visual system is to the frequency 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, each kernel is transformed into the frequency domain using the Fast Fourier Transform algorithm.

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. For each bandpass filter, the maximum frequency at which a response is found is given by the Nyquist (folding) frequency, as given in Eq. (6)

$$f_{max} = \frac{1}{\Delta x}. \quad (6)$$

where  $\Delta x$  refers to the sampling distance in the spatial domain, and is given in units of degrees per pixel.  $\Delta x$  is found by dividing the width of the viewing screen in degrees ( $60^\circ$  for a viewer seated 38 inches away from a 50 inch display) by the width of the viewing screen in pixels (1280 pixels). Thus,  $f_{max}$  is given in cycles per degree, which corresponds to the x-axis of the contrast sensitivity function of Figure 2.

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 each edge image, either enhancing the edge if the human visual system is particularly sensitive to that frequency, or inhibiting that edge if it is not. Eq. (7) gives the contrast sensitivity function used to model the weights<sup>17</sup>

$$CSF = 2.6 \cdot (0.0192 + 0.114f) \cdot e^{-(0.114f)^{1.1}}. \quad (7)$$

Each bandpass filter uses  $f_{max}$  of Eq. (6) for  $f$  in Eq. (7). The CSF value that results is the weight that is applied to the corresponding edge image that has the same frequency response as the particular bandpass filter.

The final step in the creation of the oriented edge map is to find the amount of edge information in the image at varying spatial orientations. This is accomplished by convolving the edge image with Gabor filters at four orientations  $-0^\circ$ ,  $45^\circ$ ,  $90^\circ$ , and  $135^\circ$  relative to the displayed orientation of the image. A Gabor filter<sup>18</sup> is derived from a Gaussian modulated sinusoid, and enables a local spatial frequency analysis to detect oriented edges at a particular scale. The biological justification for using this type of filter is that it simulates the structure of receptive fields in area V1 neurons. These neurons have been found to be tuned to particular orientations, as well as to specific spatial frequencies.<sup>19</sup> The Gabor function used to generate the

filters is given in Eq. (8)

$$f(x, y) = \cos(x\cos\theta - y\sin\theta) \cdot \left( \frac{1}{\sqrt{2\pi\sigma}} \right) \cdot e^{-\frac{1}{2} \left( \left( \frac{x}{\sigma} \right)^2 - \left( \frac{y}{\sigma} \right)^2 \right)}. \quad (8)$$

In Eq. (8),  $\sigma = 12$  (for an image resolution of 1280 x 768 pixels) and  $\theta = 0^\circ, 45^\circ, 90^\circ,$  and  $135^\circ$ . Four convolution kernels were created from the Gabor function (one for each value of  $\theta$ ) using a kernel size of 15 x 15, a grating with a step size of 0.1 ( 1/number of samples, where number of samples = 10), and two cycles of the sine wave. Figure 3 is a graphical depiction of the Gabor basis functions created from Eq. (8). These are used to model the receptive fields of area V1 neurons, and determine the response to visual stimuli of various orientations. The output after convolution with



Figure 3: Basis functions of the Gabor filters used to model the tuning of receptive fields in area V1 of striate cortex. From left,  $0^\circ, 45^\circ, 90^\circ,$  and  $135^\circ$ .

the Gabor filters is four oriented Laplacian cubes – one for each of the orientations. Each of the four cubes is then flattened into a single plane by merging together (summing) the six levels of the oriented edge cube, after each level has been normalized to be in the range from zero to one. This creates four oriented edge signals. The four signals are then normalized once again, and linearly summed together to create the oriented edge map. A future implementation of the edge module may apply a non-linear weighting function to each level and/or each orientation signal before merging, to improve the correspondence between edge representation and human visual response.

Once the color map, the intensity map, and the oriented edge map have been generated, they are each linearly scaled from zero to one before a final merge to create a single low-level saliency map. This is accomplished by summing the values from each map together on a pixel-by-pixel basis.

### 2.3. *The high-level proto-object map*

The proto-object map is constructed in parallel with the edge orientation 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 calculated from the pre-processing stage as input. Background estimation is accomplished by sectioning the image into 16x16 pixel blocks, and setting all of the pixels within a block to a single value. For simplicity, the minimum pixel gray level of each block was used; however future implementations may use a weighted combination of gray level values for each block. A block corresponds to approximately  $\frac{3}{4}^\circ$  of visual angle, assuming a viewing distance of 38 inches. The effect of this step is that regions of relatively uniform intensity in the image are localized, *i.e.*, broad areas of low texture, which correspond to potentially low information, are considered an estimate of the background. This definition of background assumes certain properties of the image, and has been found empirically to be most valid for outdoor scenes, yet the algorithm also performed well for many indoor scenes from the test database. A foreground image is created by subtracting the background from the original gray level image to segment areas of the image corresponding to the foreground. The result after subtraction simulates the effect of figure/ground segmentation of perceptual organization.<sup>20</sup>

The next stage applies a threshold to the foreground image to create a binary foreground image, which is subsequently used for edge detection. A Canny edge operator is used to detect both weak and strong edges in the binarized foreground image, including the weak edges only if they are connected to strong edges. This produces a more robust representation of foreground information, in that shadows and spurious edge noise will not be included. Essentially, the Canny operator uses a derivative-of-Gaussian filter to calculate gradients in the binary image, and then finds the local maxima of those gradients. The Canny operator has been shown to be optimal for detecting two-dimensional step edges, and closely approximates the operation of simple cells in striate cortex that sum the outputs from lower-level cells.<sup>21</sup>

Regions corresponding to potential objects in the image are grown into larger regions by using morphological operators on the foreground image. The rationale for this step is that highly textured areas are likely to be perceptually interesting, and may predict the location of potentially useful objects in the scene. Therefore, the texture from edges should be merged



Figure 4: Sample input images with overlaid fixation plots ( $1^{st}$  row) – from left, washroom, hallway, office, vending. Low-level saliency maps ( $2^{nd}$  row), proto-object maps ( $3^{rd}$  row), and final conspicuity maps ( $4^{th}$  row).

into a larger region, corresponding to the potential object. The result is called the *proto-object* map; *proto-* because object detection and recognition is neither required nor useful at the early stages of visual perception. The goal at the end of this stage is to eliminate as much information as possible from further processing, and allow a *gateway* through which potentially useful information may pass. The proto-object map is used along with the color map, the intensity map, and the oriented edge map as an additional channel to detect conspicuous areas. For the proto-object map, conspicuous areas are defined as those areas which may have task-relevant information.

Once the four maps have been merged into a single map (again, a summation after normalization), 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 4 shows four example input images and the corresponding low-level and proto-object maps for each image. The bottom row of Figure 4 shows the result after merging the maps and masking. The bright areas correspond to highly conspicuous regions in the example images. These areas are where the model predicts a viewer’s visual attention is most likely to be captured. [6pt]

### 3. Verification of Model – Experimental Method

Eye-tracking data was collected and analyzed for the purpose of determining the correlation between the model and the fixation locations of people

viewing natural scenes. An Applied Science Laboratory model 501 head-mounted eye-tracker was used to record gaze positions, which were sampled 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 per inch was used to display the images. The display area subtended a visual field of  $60^\circ$  horizontally and  $35^\circ$  vertically at a viewing distance of 38 inches. At this distance, approximately 21 pixels cover  $1^\circ$  of visual angle.

### *3.1. Data collection*

Eleven subjects participated in the eye-tracking experiment, which lasted approximately 45 minutes. All subjects had 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 to ensure that the recorded fixation location corresponded to the gaze position on the image. Custom software was developed that uses the calibration data to correct for any detected slippage (drift) of the headgear during the run of the experiment, and to correct for any detected inaccuracies (offsets) in the reported fixation location. After drift and offset correction the average angular deviation from the (nine) calibration points was  $0.73^\circ \pm 0.06^\circ$  at the start of the experiment, and  $0.56^\circ \pm 0.04^\circ$  at the end of the experiment.

Each subject viewed a total of 152 color images divided into two sets of 76 images each. The two sets were labeled *A* and *B*, and were counter-balanced between observers. The image database includes a wide range 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 for as long as desired, and *multi-view*, where the subject was given an explicit instruction before viewing each image. The current study uses only a small subset of the multi-view data, therefore the multi-view instructions that are important for this study are discussed in the following section.

Free-view always preceded multi-view. During free-view, the presentation order of the images was randomized, whereas during multi-view, the images were presented in a fixed order. Six subjects viewed image set *A* as free-view and image set *B* as multi-view. The remaining five subjects viewed image set *B* as free-view and image set *A* as multi-view. Each subject viewed all 152 images in the database; several images were viewed more than once.

#### 4. Results

A metric was developed to measure the correlation between the density of subjects' fixation locations on a particular image and the prediction of fixation locations from the model. The metric compares the conspicuity value from the map of the fixated regions to the expected conspicuity value from any random location in the map. The metric is referred to as the  $F/M$  ratio, as it is the ratio of the **F**ixation mean to the **M**ap mean.

The mean conspicuity of fixations,  $F$ , is defined as the average conspicuity value extracted from the map at the x,y-coordinates of the fixation locations, for all fixations on a particular image (all subjects). This value is found by first generating a map for a particular input image using the model depicted in Figure 1. Next, the x,y-coordinates of the fixation locations for that image are determined from the eye-tracking data, using a custom fixation-finding algorithm. The fixation-finder uses an eye movement velocity threshold of  $110^\circ$  per second to detect the beginning and end of a fixation, and assigns the centroid of all data points falling within the temporal fixation window to be the x,y-coordinate of the fixation location.

For each fixation, the x,y-coordinate is used as an index into the map to extract the conspicuity value at that location. Since the x,y-coordinate is the location of a single pixel in the input image, information from the map must be integrated over an area corresponding to foveal coverage. The area of the visual field covered by the fovea at a distance of 38 inches covers approximately  $1^\circ$ , therefore, a 21x21 pixel window (corresponding to  $1^\circ$  visual angle) is centered on the map at the x,y-coordinate, and all conspicuity values falling within the window are averaged together to find the conspicuity of a particular fixation.

The conspicuity of all fixations on an image is found in a like manner, and those values are averaged together to determine the mean conspicuity of fixation,  $F$ . The mean conspicuity of a map,  $M$ , is simply the average value of the map generated from the model. The  $F/M$  ratio is the ratio between the mean conspicuity of fixation 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 fixation is nearly the same as the mean value of the map. Any random distribution would do just as well. If the  $F/M$  ratio is less than one, then the map is also not a good predictor because the fixations tend to be on regions that the map has predicted viewers will look *away*

from – the fixations have relatively low conspicuity values as compared to the mean value of the map. However, if the  $F/M$  ratio is greater than one, then the map is a good predictor because the fixations tend to be on regions of the image that the model has assigned a high conspicuity value relative to other regions.

To compare the predictive power of the model using several different feature parameters, four maps were generated for each of the 152 images in the database. These maps were computed as given in Eqs. (9) through (12)

$$CIEmap = (C + I + E)/3. \quad (9)$$

$$Pmap = P. \quad (10)$$

$$CIEPmap = \left( (C + I + E + P)/4 \right) \cdot P. \quad (11)$$

$$C\_map = (C \cdot w_1 + I \cdot w_2 + E \cdot w_3 + P \cdot w_4) \cdot P \cdot w_5. \quad (12)$$

In Eqs. (9) through (12),  $C$  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,  $CIEP$  refers to the final conspicuity map with equal weighting for each of the feature maps, and  $C\_map$  refers to the final, weighted conspicuity map. In Eq. (12),  $w_1$  through  $w_5$  refer to the weights applied to the individual feature maps before merging for the final conspicuity map. Section 4.2 discusses the procedure used to determine the feature weights.[6pt]

#### ***4.1. Free-view Fixation/Map correlation***

A comparison of the  $F/M$  ratio for each of the map computations given in Eqs. (9) through (12) is shown in Figure 5, for the 152 images under the free-view condition. The mean  $F/M$  ratio for the CIE map, which uses only low-level information from the image, is close to one, implying that the correlation between fixation locations and map conspicuity values is close to random. Subjects tend to look at regions rated as highly conspicuous by the CIE map as frequently as they look at any other region in the image. The  $P$  map shows a significantly higher correlation to fixation locations than does the CIE map, indicating that an object, or rather the potential location of an object, plays an important role in predicting where people will look in a complex, natural scene. The CIEP map (unweighted features), which uses the  $P$  map as an added feature along with the low-level information, has a higher correlation to fixation locations than either the CIE map or the  $P$  map alone does. This shows that when low-level information is combined with

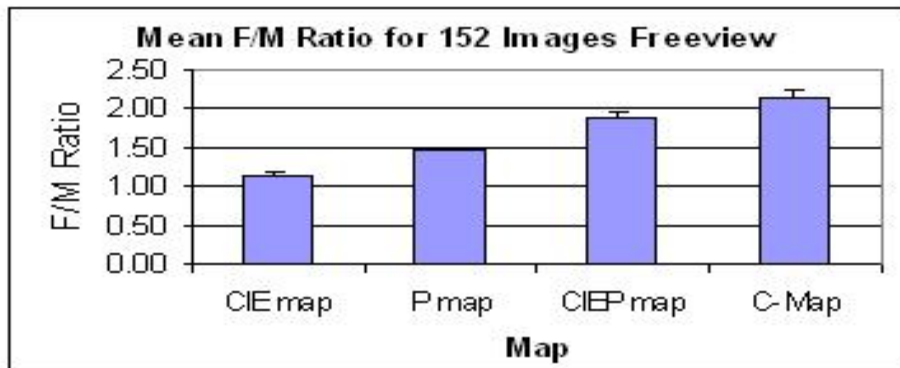


Figure 5: Mean F/M ratios for four maps generated from 152 images.

potential object location, fixation locations can be predicted with higher accuracy. The C\_map (weighted features) shows the highest correlation to fixation locations, implying that a differential weighting of the feature maps can produce a higher correlation.

The improvement in F/M ratio as the maps include more information about objects or potential objects shows that attention is likely to be directed to those objects in a scene, rather than to highly salient, non-object areas. This is an indication that both perceptual relevancy as well as feature salience guides patterns of oculomotor behavior in humans. It has been suggested that salience is an important indicator for predicting fixations in the absence of an explicit or implied task,<sup>3,4</sup> however this study shows that even in a free-view condition, subjects are more likely to look at locations in the field that offer potential relevancy.[6pt]

#### 4.2. Determination of map weights

There is no reason to believe that each of the low-level feature maps and the high-level proto-object map as constructed in Eqs. (9) through (11) should contribute equally to the perceptual conspicuity of a particular image region, or that an equal weighting is computationally equivalent to bottom-up human visual processing. Therefore, an attempt to derive the optimal weight for each feature map was conducted on a per-image basis. It must be emphasized that the feature weights were determined using the results from the eye-tracking data, *i.e.*, the weights were assigned such that the assignment gave an optimal, or near-optimal F/M ratio. The purpose of using this technique is not to show that any particular weighting scheme is superior for predicting fixation locations, but rather to show that a uniform

weighting of the feature maps is not optimal. Further study is required to determine a weighting scheme independent of eye tracking data that is useful over a broad range of image types.

Two methods were used to find an optimal or near-optimal feature map weighting – random weight generation and a genetic algorithm. An optimal weighting is defined as one that yields the highest F/M ratio possible for any particular image, *i.e.*, the best correlation between the map conspicuity and the subjects’ fixation locations. Thus, an optimal weighting must take into account human viewing behavior. It must be emphasized that an approximation to optimal is used as the criterion for weighting the maps, because an exhaustive search of all possible combinations of feature weights is computationally prohibitive.

The random weight generation method assigns a random number between -1 and 1 to each of the five weights in Eq. (12), and then normalizes  $w_1$  through  $w_4$  to sum to one. The weights are then applied to the five feature maps before computing the final conspicuity map, and the F/M ratio for the resulting map is calculated. Ten thousand trials were run on three example images using the random weight generation method and the maximum F/M value from the trials was found. Table 1 shows the weights that were applied to the feature maps of the three example images. Figure 6 shows those three example images with an overlaid fixation plot from a single subject’s eye movement trace; beneath each image is the C-map for that image with the weights as given in Table 1 applied.

Table 1: Maximum F/M ratios and associated weights for the three example images using the random weight generation method.

Image	F/M max	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$
A1	2.81	1.17	-0.8	-0.09	0.72	0.33
A28	2.69	0.27	0.8	-0.05	-0.01	0.85
B17	1.62	0.57	0.44	0.08	-0.08	0.09

An optimal map-weighting scheme is heavily image dependent, at least for the three example images given here. It is interesting to note from Table 1 that for some images, a negative feature weight produces the highest F/M ratio. For example, the map for image A28 (middle image of Figure 6) correlates highly to fixation locations when the map slightly de-emphasizes oriented edges ( $w_3$ ) and strongly emphasizes intensity ( $w_2$ ). In other words, subjects tend to look away from edges and towards bright regions in the image, perhaps because the more interesting region in the image is the central, bright object and not the boundary between the lake and sky. For image

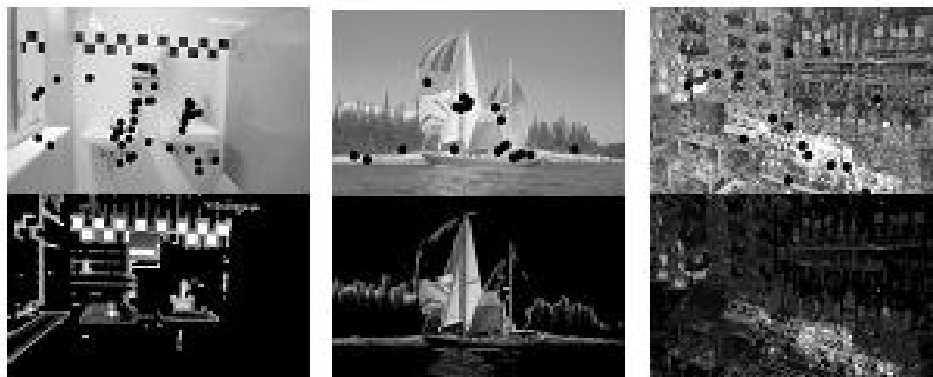


Figure 6: Example images with overlaid fixation plots (top row) and corresponding weighted conspicuity maps (bottom row) From left, images are A1, A28, and B17.

B17, color and intensity share nearly equal weighting; where fixations are not as heavily influenced by object locations. This is probably due to the relatively abstract composition of image B17 as compared to the other two images.

The second method used to find an optimal feature weighting uses a genetic algorithm.<sup>22</sup> The motivation for using a genetic algorithm for this task is to reduce the amount of computation required for finding a near-optimal solution. A total of 2410 generations were run for each of the 152 images in the database, however a solution frequently converged before all generations were completed. The parameters used for the weight finding genetic algorithm are as follows:

fitness criteria: highest F/M ratio  
max number of generations: 300  
size of population in each generation: 10  
number of genes (weights) in a chromosome: 5  
probability of crossover: 0.5  
probability of mutation: 0.05  
scale of mutations: 0.1

For every generation of ten chromosomes, two were selected as parents and the remaining eight were eliminated. The parents mated and produced eight new children with crossovers and mutations according to the given parameters. Table 2 shows the final F/M ratios and associated weights for the three example images of Figure 6, and the number of generations until a solution converged.

Table 2: Maximum F/M ratios and associated weights for the three example images using the genetic algorithm weight generation method.

Image	F/M max	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	#gens
A1	2.81	1.07	-0.64	0.05	0.52	0.21	246
A28	2.69	-0.05	1.14	-0.12	0.03	0.55	185
B17	1.60	0.43	0.53	0.13	-0.09	0.62	58

The primary advantage of using a genetic algorithm over the random weight generation method for this problem is a reduction of computation, assuming constant overhead for both methods. Since the maximum F/M ratio is nearly the same for both methods, the genetic algorithm should be chosen because near optimal weights are found in fewer steps.

Regardless of the method used to find the weights, some *a priori* knowledge about the image must be considered before assigning weights to features. If images are classified according to a visual similarity metric (color, texture, semantic category) prior to processing, weights may be assigned according to both image class and results from knowledge of prior fixation locations on exemplars from each class. For example, the three example images of Figure 6 depict three distinct image categories – indoor uncluttered scene, outdoor uncluttered scene, and indoor cluttered scene.[6pt]

### 4.3. Examination of central bias - top down effect?

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, sometimes images exhibit a spatial bias that locates the most interesting part of the image in the center. The bias may reflect that of the photographer; that is, photographers frequently will center the most important area of the scene in the view-finder when capturing an image. In this way, the photographer directs viewing behavior. Can the conspicuity map capture the idea of “interesting“ areas without an artificial bias toward the center?

An analysis of fixation distances from the center of each of the four example images from Figure 4 found that approximately half of the fixations were within  $\pm 10^\circ$  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  $\pm 6^\circ$  from center, 30% of washroom fixations, 37% of hallway fixations, 17% of office fixations, and 23% of vending fixations are found. Inscribing a circle with a radius of  $6^\circ$  at the center of the image covers 49,867 pixels, or 5% of the total image area. If fixations were evenly distributed across an image, then one would expect 5%

of all fixations to fall in the central area. Accounting for the expected 5%, the washroom image exhibits a central bias of 25%, the hallway image shows 32% bias, the office image shows 12% bias, and the vending image shows 18% bias. From this data it can be concluded subjects preferentially fixated the center of the hallway image most frequently, followed by the washroom image, the vending image, and least frequently the office image. The office image shows the least central bias most likely because the objects in that image are not centrally located, as they are in the other three images. Thus, fixations cannot be considered randomly distributed across the image space for at least three of the four images.

To test the conspicuity map for a preserved central bias, a random sequence of fixations was generated that restricted all fixations to a 1/16 image size window ( $14.5^\circ \times 8.7^\circ$ ). Figure 7 shows how the spatial bias affects the F/M ratio of the four images of Figure 4, using both the low-level feature saliency map (CIE map) and the weighted high-level conspicuity map (C\_Map). The F/M ratio is close to one for all four images when the

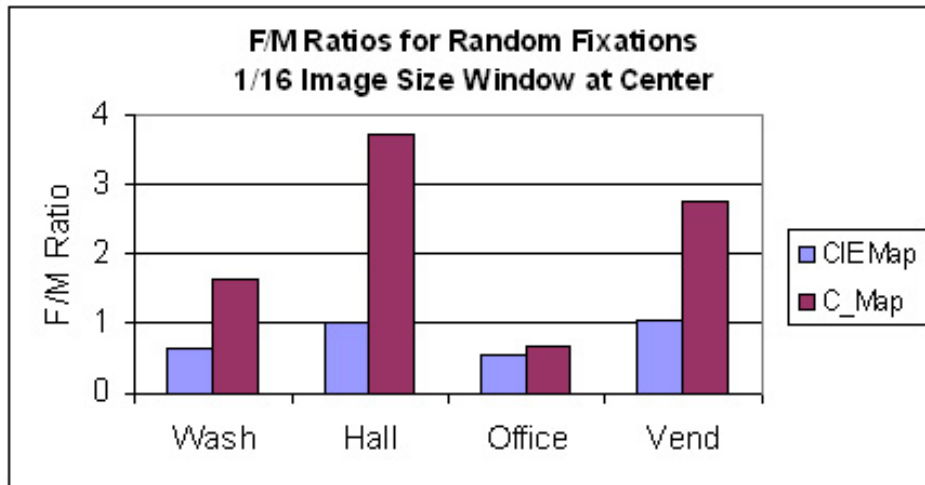


Figure 7: F/M ratios for random fixations restricted to 1/16 image size window centered in image.

low-level maps are used alone. This indicates that low-level features are not sufficient for predicting fixation locations when the fixations are biased to fall in the central area of the image. The F/M ratio is high when the weighted conspicuity map is used with three of the four images (washroom, hallway, and vending), and low for the office image. This is in agreement with the examination of subject central bias described earlier, where the

hallway image was found to bias fixations most closely toward the center, followed by the washroom image, the vending image, and finally the office image.

The analysis of random fixations shows that the high-level maps are able to incorporate object information, and can accurately simulate the general tendency to look towards the center, when that tendency is warranted. Other models include an imposed, explicit bias toward the center to improve performance, however, the object-oriented model presented here preserves the central bias without artificial means, and more importantly, does so when only that bias is suggested by image content. This can be considered a top-down mechanism for predicting fixation location in the absence of externally provided semantic content.

#### ***4.4. Multi-view Fixation/Map correlation***

The perceptual conspicuity model presented here correlates well with the fixation locations of subjects who view natural images under the free-view condition. Does the model also correlate well to fixation locations when a specific task has been imposed upon the viewer?

During the multi-view part of the eye-tracking experiment, the subject was given an explicit instruction before viewing a set of images. One such set of images was the four interior scenes as shown in Figure 4. During free-view, these four images were mixed in with 72 other images in the database, and all were viewed in a random order for as long as desired by the subject. During multi-view, these four images were blocked together and viewed three separate times, each time in a fixed order.

Before each multi-view presentation, the subject was given a verbal instruction that consisted of a task or situation to be imagined while viewing the image. Each image had three instructions associated with it, one instruction for each of the three separate viewings. For example, for the hallway image, the instruction was either “put something in the garbage,” “the fire alarm has just gone off,” or “find a bathroom.” The purpose of varying the instruction for each viewing of the same image is to determine the extent to which fixation location is dependent upon the imposed task. If fixation locations are highly dependent upon the imposed task, then support is offered for the hypothesis that vision and visual perception is an inherently active and selective process, rather than a passive process whereby information is merely collected, processed, and perhaps stored for later retrieval. Active vision, rather than passive viewing, is the means by which an individual is able to integrate specific, local aspects of the scene with goal-oriented behavior. Consequently, an active vision hypothesis presumes that the purpose

of vision is to serve the needs of the individual as those needs arise during the course of daily living. The implication for a machine vision system is that active vision, and knowledge about human visual perception during an active task, may reduce computational load and make better use of limited resources.

Table 3 shows the instructions that were given for each of the three viewings of the four images. The three sets of instructions are labeled  $T1$ ,  $T2$ , and  $T3$ . Figure 8 shows a comparison of the F/M ratios for the low-level

Table 3: Instructions for the multi-view part of the eye-tracking experiment. Each row indicates image viewed – from top, washroom, hallway, office, and vending.

Task T1	Task T2	Task T3
“wash your hands”	“fill a cup with water”	“comb your hair”
“put something in the garbage”	“find a bathroom”	“the fire alarm just went off”
“get supplies from the closet”	“work at the computer”	“make a photocopy”
“check for Skittles”	“buy a Snickers bar”	“check for change”

saliency maps and the conspicuity maps computed under the three multi-view tasks, and the free-view condition, for each of the four images. From

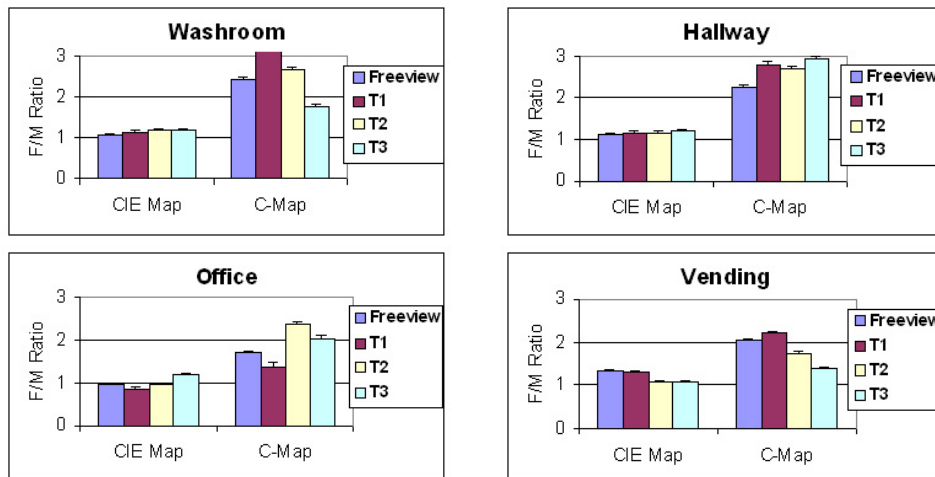


Figure 8: Comparison of the F/M ratios for the three multi-view tasks and the free-view condition for the four images of Figure 4. Top row from left: washroom, hallway. Bottom row from left: office, and vending.

Figure 8, it may be observed that the low-level saliency maps do not correlate well to subjects’ fixation locations for either the free-view or the multi-view conditions. However, adding information about potential object locations in the environment not only improves the correlation, but does so in a task-

dependent way. The F/M ratios for the conspicuity maps vary according to the nature of the task, with some tasks being more “object-oriented” than others. For example, for the washroom image, task T1 (“wash your hands”) gives the highest F/M ratio and task T3 (“comb your hair”) gives the lowest. This observation can be interpreted as indicating that washing one’s hands requires more fixations on objects in the environment than does combing one’s hair. Since the conspicuity map is designed to place more emphasis on potential objects, the correlation will be higher when a task requires object manipulations.

## 5. Discussion

The purpose of this study was to develop a biologically plausible representation of the highly conspicuous regions of an image, and to determine the correlation between this representation and people’s viewing behavior. A high-level proto-object map was constructed to identify regions in the image that contain potentially useful objects. This was used in conjunction with a low-level saliency map to locate features in the scene corresponding to colorfulness, luminosity, and oriented edges. The saliency map simulates the bottom-up response of neurons in terms of the center-surround organization of receptive fields, lateral inhibition, color-opponent processes, and contrast sensitivity. The proto-object map simulates high-level perception in the form of figure/ground segmentation, and a top-down constraint simulates a central location bias when that bias is warranted.

The result is a topographic map of perceptual conspicuity, where high values in the map correspond to perceptually conspicuous regions in the image, and low values correspond to regions that are not likely to be fixated. The map is in agreement with empirical studies of visual behavior as confirmed by eye-tracking experiments, and also confirms our intuition of where people are likely to look when viewing natural images.

Other models of visual salience<sup>3,4</sup> have used a biological approach with limited success. The earlier models were able to successfully predict fixation locations in images where color information predominates, and luminance and edges have a lesser role, such as in passive viewing of pseudo-colored fractal images. The perceptual conspicuity map developed here is a better predictor of *active* visual behavior because it uses an object-oriented approach. A preference for objects has been noted in the earlier studies that use only low-level feature information:

... subjects often examined objects on table tops independent of their salience.<sup>3</sup>

This suggests that global, scene-dependent strategies play an important role in determining fixation locations.

The perceptual conspicuity map includes an inherent global bias toward the center of the image when potentially important objects are located there. In order to simulate the empirically observed central bias, the conspicuity map should, in a future implementation, include a spatial frequency reduction from the most conspicuous region toward the periphery of the image. This would account for the fall-off in visual acuity in the non-foveal regions of the retina, and would be an additional constraint for biological plausibility since important objects tend to be foveally located during an active task. The fall-off could be implemented as a Gaussian blur centered on the most perceptually conspicuous region of the image, centrally located or not. Another valuable extension to this study would be to examine the central bias issue with a broader range of images that includes more non-centrally located objects of interest. Also, one should consider the effect of varying the block size during the construction of the proto-object map. The current study uses a fixed block size of 16x16 pixels, however it is possible that an adaptive block size according to the expected image type, may prove most useful for locating objects in a wider variety of scenes.

In addition, a higher correlation between the map and fixation locations might be obtained by extracting the conspicuity values from the map using an adaptive window size, according to top-down constraints such as the expected size of potentially useful objects, and expected location. A Bayesian network could be employed to take into account evidence from the scene and incorporate prior knowledge about the imposed task to reason about a fixation strategy. The size of the fixation location window should also be allowed to vary over a greater spatial range in the image to promote flexibility across a wider range of image content.

## **6. Conclusion**

Fixation locations are not completely deterministic, yet they are also not completely random. This fact is obvious even without eye-tracking studies, however, little work has been done to discover how exogenous and endogenous factors interact to determine the target of the next saccade. This is particularly true when considering high-level factors such as motivation, prior learning, and experience, where the visual system is used primarily as a tool to monitor, assist, and assess the immediate environment during ongoing activity. For this reason, an accurate description of the high-level parameters that determine how attentional resources should be allocated

cannot be considered in isolation of the environment, or of the current task under execution.

A description of visual behavior during the over-learned task of making tea<sup>9</sup> found that the eyes monitor and guide virtually every action that is necessary to complete this activity. The visual salience (*i.e.*, color, luminance contrast, texture) of objects was not an important indicator of fixation locations. Rather, it was the object's relevance to the task that accurately predicted the saccadic landing position. This is a consequence of the goal-driven behavior of people, as noted earlier.<sup>7</sup> The visually salient properties of an image may be important for determining which region is fixated next for free-viewing non-representational images such as fractal patterns, but not for active vision during an ongoing task in a complex environment, where strategic behavior and the formulation of a plan of action is required.

This study provides evidence for the hypothesis that people preferentially fixate objects relevant for potential actions implied by the semantics of the scene, rather than selecting targets based purely on image features. Success with predicting fixation densities in natural images requires not only knowledge about the salient properties of low-level image features, but also an understanding of the observer's goals, including the perceived usefulness of an object in the context of an implicit or explicit task.

The analysis provided here found that there is virtually no correlation between the low-level salient properties of natural images and fixation locations. People simply do not look at something unless there is a need to do so. In terms of activation theory, the salient properties of any particular region are promoted to the level of awareness only after they have been coupled to a relevant object. Even though feature salience is an inherent property of the image or scene, it is the location of objects that determines if and how much of the salient properties are perceived. Thus, perceptual conspicuity can be described as the modulation of feature salience due to task preference for certain objects.

In summary, locating highly conspicuous regions of an image must ultimately take into consideration the implicit semantics of the scene – that is, the meaningfulness of the contents of the scene for the viewer. Objects as well as their locations 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 con-

straints in a way that is task-relevant, goal-oriented, and ultimately most meaningful for the viewer. A machine vision system that successfully emulates the capabilities of the human visual system (and not all necessarily need to do this) must eschew an explicit representation in favor of a limited representation that takes into account information mostly from objects that have current relevance. The ability to suggest the next camera position based on task constraints, coupled with a limited amount of processing at that position, will provide a distinct advantage for systems that seek to emulate the superior perceptual capabilities of the human visual system. Overall, the advantages are compelling enough to warrant further study.

## References

- [1] R.A. Rensink, J.K. O'Reagan, and J.J. Clark, To see or not to see: The need for attention to perceive changes in scenes, *Psychological Science*, **8**(5)(1997), pp. 368-373.
- [2] C. Koch, and S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, *Human Neurobiology*, **4**(1985), pp. 219-227.
- [3] D. Parkhurst, K. Law, and E. Niebur, Modeling the role of salience in the allocation of overt visual attention, *Vision Research*, **42**(2002), pp. 107-123.
- [4] L. Itti, and C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, *Vision Research*, **40**(2000), pp. 1489-1506.
- [5] D.O. Hebb, *The Organization of Behavior*, (John Wiley & Sons, New York, 1949).
- [6] L. Kaufman, and W. Richards, Spontaneous fixation tendencies for visual forms. *Perception and Psychophysics*, **5**(2)(1969), pp. 85-88.
- [7] A. Yarbus, *Eye Movements and Vision*, (Plenum Press, New York, 1967).
- [8] G.T. Buswell, *How People Look at Pictures: A Study of the Psychology of Perception in Art*, (The University of Chicago Press, Chicago, 1935).

- [9] M.L. Land, N. Mennie, and J. Rusted, The roles of vision and eye movements in the control of activities of daily living, *Perception*, **28**(1999), pp. 1311-1328.
- [10] K.A. Turano, D.R. Gerguschat, and F.H. Baker, Oculomotor strategies for the direction of gaze tested with a real-world activity, *Vision Research*, **43**(2003), pp. 333-346.
- [11] J.B. Pelz, and R. Canosa, Oculomotor behavior and perceptual strategies in complex tasks, *Vision Research*, **41**(2002), pp. 3587-3596.
- [12] S.N. Pattanaik, J.A. Ferwerda, M.D. Fairchild, and D.P. Greenberg, A multi-scale model of adaptation and spatial vision for realistic image display, *Proceedings of the SIGGRAPH*, **98**(1998), pp. 287-298.
- [13] G. Wyszecki, and W.S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd edn. (Wiley, New York, 1982).
- [14] M.D. Fairchild, *Color Appearance Models*, (Addison-Wesley, Reading, MA, 1998).
- [15] R.W.G. Hunt, *The Reproduction of Color* 5th edn. (Fountain Press, Kingston-upon-Thames, England, 1995).
- [16] P.J. Burt, and E.H. Adelson, The Laplacian pyramid as a compact image code, *IEEE Transactions on Communications*, **31**(4)(1983), pp. 532-540.
- [17] J.L. Manno, and D.J. Sakrison, The effects of a visual fidelity criterion on the encoding of images, *IEEE Transactions of Information Theory*, **20**(4)(1974), pp. 525-535.
- [18] D. Gabor, Theory of Communication, *IEEE Proceedings*, **93**(1946), pp. 429-441.
- [19] D.H. Hubel, and T. N. Wiesel, Receptive fields and functional architecture of monkey striate cortex, *Journal of Physiology*, **195**(1968), pp. 215-243.
- [20] E. Rubin, *Visuell Wahrgenommene Figuren*, (Kobenhaven, Glydenalske boghandel, 1921) from S. E. Palmer, *Vision Science: Photons to Phenomenology*, (MIT Press, Cambridge, MA, 1999).

- [21] F. F. Canny, A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**(1986), pp. 769-798.
- [22] J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (University of Michigan Press, Ann Arbor, MI, 1975).