

High-level aspects of oculomotor control during viewing of natural-task images

Roxanne L. Canosa^{*a}, Jeff B. Pelz^a, Neil R. Mennie^b, Joseph Peak^c

^aRochester Institute of Technology, Rochester, NY, USA 14623

^bCenter for Visual Science, University of Rochester, Rochester, NY 14627

^cNaval Research Laboratories, Washington, DC 20375

ABSTRACT

Eye movements are an external manifestation of selective attention and can play an important role in indicating which attributes of a scene carry the most pertinent information. Models that predict gaze distribution often define a local conspicuity value that relies on low-level image features to indicate the perceived saliency of an image region. While such bottom-up models have some success in predicting fixation densities in simple 2D images, success with natural scenes requires an understanding of the goals of the observer, including the perceived usefulness of an object in the context of an explicit or implicit task. In the present study, observers viewed natural images while their eye movements were recorded. Eye movement patterns revealed that subjects preferentially fixated objects relevant for potential actions implied by the gist of the scene, rather than selecting targets based purely on image features. A *proto-object* map is constructed that is based on highly textured regions of the image that predict the location of potential objects. This map is used as a mask to inhibit the unimportant low-level features and enhance the important features to constrain the regions of potential interest. The resulting *importance map* correlates well to subject fixations on natural-task images.

Keywords: Eye movements, Attention, Saliency, Natural images, Natural tasks

1. INTRODUCTION

Visual perception is an inherently selective process. Selective attention is the means by which an individual chooses a subset of the information available from the visual scene 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 a particular scene or image can be represented compactly, thus making optimal use of limited neural resources. An example of the effect of efficient encoding is evidenced from studies on *change-blindness* (Rensink, O'Reagan & Clark¹), which show how observers of complex, natural scenes are mostly unaware of large-scale changes in subsequent viewings of the same scene. Olshausen & Field² provide a biologically plausible account of how a sparse neural code can provide a compact representation of a natural image due to the high level of statistical independence among the principal components that describe the gray-level variance in the image.

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 (Koch & Ullman³). In the saliency map, the value at a coordinate provides a measure of the contribution of the corresponding image pixel to the conspicuity of that image region. There currently exist two approaches to modeling the effects of saliency on viewing behavior – the bottom-up, or stimulus-driven approach, and the top-down, or task-dependent approach.

* Correspondence: rlc8222@cis.rit.edu

The stimulus-driven approach begins with a low-level description of the image in terms of feature vectors, and measures the response of image regions after convolution with filters designed to detect those features. Parkhurst, Law, & Niebur⁴, as well as Itti & Koch⁵ use spatio-chromatic filters at various levels of spatial resolution to detect color, luminance, and oriented-edge features along separate channels. After detection, the features are scaled and linearly summed over all channels to produce a single value of salience for each pixel in the image. In support of the stimulus-driven approach, Theeuwes⁶ found that in a simple search task, search times were dramatically increased when subjects were presented with distracters that differed along a single irrelevant dimension, thus showing that attention can be captured involuntarily and influence the selection process. Bacon & Egeth⁷ refuted the assertion of stimulus-driven attentional capture by showing that goal-directed selection is able to override the salient feature singletons when specific, known, feature groupings are available.

The purpose of the present study is to develop a biologically-plausible model of an attentional mechanism that selects regions of high salience in an image, and which correlates well with the fixation patterns of subjects who viewed images of natural scenes. The model is based on the approach taken by Parkhurst, Law & Niebur⁴, and Itti & Koch⁵ in that it uses oriented spatio-chromatic filters at various resolutions to detect low-level features of high salience. The stimulus-driven model developed by Parkhurst, et al. is sufficient for capturing highly salient regions of simple images, such as fractal patterns. The model, however does not correlate well to fixation patterns when subjects viewed more complex, natural images. This is particularly so when an explicit task has been imposed upon the viewer. In order to take into account viewing behavior in the presence of natural, realistic images, the present study augments the Parkhurst et al. model with a higher-level algorithm to locate proto-objects in the image. The proto-object map is based on highly textured regions in the image that predict the location of potential objects. This map is then used as a mask to both inhibit the unimportant lower-level features and enhance the important features based on high-level and goal-oriented constraints. We call the resulting map an *importance map*.

The motivation for developing an algorithm that takes into consideration the importance of certain image regions is derived from studies showing that visual capabilities in humans are neither general nor insensitive to context, but rather are specific and tailored to the task at hand. Early studies of fixation patterns stressed the importance of low-level features in the image that guide the eye during free-viewing (Hebb⁸, and Kaufman & Richards⁹). The belief was that people tend to fixate first upon edges, lines, and corners as they scan an image, and sequentially build up a perceptual Gestalt of the scene over time. However, these early studies were primarily concerned with spontaneous fixation patterns during free-viewing and ignored the higher-level aspects of eye movement control.

Other early studies (Yarbus¹⁰, Buswell¹¹) showed that high-level cognitive strategies are reflected in the patterns of eye movement traces. Distinctly different patterns of scan paths could be elicited from subjects when they performed context-sensitive tasks. Yarbus¹⁰ found that when subjects viewed I. E. Repin's painting, *An Unexpected Visitor*, depicting a scene of several people greeting a newcomer, a specific question posed to the subject elicited a "signature" pattern of eye movements. Different questions elicited different "signature" patterns.

Noton & Stark¹² conducted similar studies on the task dependencies of visual scan paths and concluded that the visual processing of natural images is essentially serial, with specific features processed in a specific order. Subjects tend to have a characteristic scan path for any particular image, which is repeated for each subsequent viewing of the same image, however there is much variation in scan paths across different subjects viewing the same image and across the same subject viewing different images. More recently, Andrews & Coppola¹³ found that the temporal and spatial patterns of eye movements are highly idiosyncratic, and that the size and frequency of saccades covaried significantly when subjects viewed complex natural scenes as compared with viewing blank images or simple patterns. Further, Deubel¹⁴ suggested a link between object recognition and saccade target selection at the neural level, and Land, Mennie, & Rusted¹⁵ showed that eye movements monitor and guide virtually every action that is necessary to complete an over-learned task such as making tea. For that study, nearly all of the fixations (over 95%) were directed to objects that were relevant for the sub-task currently being executed. This result was

described as a consequence of the goal-driven behavior of the subjects, and is an example of the task-relevancy of visual scan paths documented by Yarbus¹⁰ and Buswell¹¹. 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, as suggested by Pelz & Canosa¹⁶.

2. MODEL DESCRIPTION

2.1 Construction of the model

This section describes in detail the steps that were taken to create the importance map. This map consists of two essential modules – the saliency map, and the proto-object map – as shown in Figure 1. The saliency map is an implementation of the model developed by Parkhurst, Law, and Niebur⁴, and is the area shown in Figure 1 enclosed within the dashed lines. The purpose of the proto-object map is to inhibit regions of the saliency map that do not correspond to expected object locations, and to enhance those regions that do.

2.1.1 The saliency map

The saliency map takes as input the original, RGB formatted image and creates three separate feature channels that are processed in parallel – color, intensity, and orientation. Within the color channel, the image is separated into red, green, blue and yellow color planes, and each plane is sampled at three spatial scales (1:1, 1:2, 1:4) forming a Gaussian pyramid (Burt and Adelson¹⁷). The model originally proposed by Parkhurst, et al. used nine resolution levels in the Gaussian pyramid; the present study uses three for computational efficiency. The second stage of processing simulates the center-surround organization of simple cells and the double-opponent color system in the primate visual system by subtracting a lower resolution color image from a higher resolution opponent color. This results in a representation of “colorfulness” in the image, as expressed in the 2-dimensional R-G and B-Y color opponent space. The two resulting opponent-color pyramids are then scaled to be in the range from 0 – 1, and linearly summed to create the color feature map.

Processing along the parallel intensity and orientation channels follows a similar path. For the intensity channel, the input image is expressed as the average of the RGB digital counts, and again sampled at three spatial resolutions followed by the center-surround computation which forms the intensity feature map. For the orientation channel, the input gray-level image is convolved with oriented Gabor filters at each spatial scale, forming four orientation pyramids – 0°, 45°, 90°, and 135°.

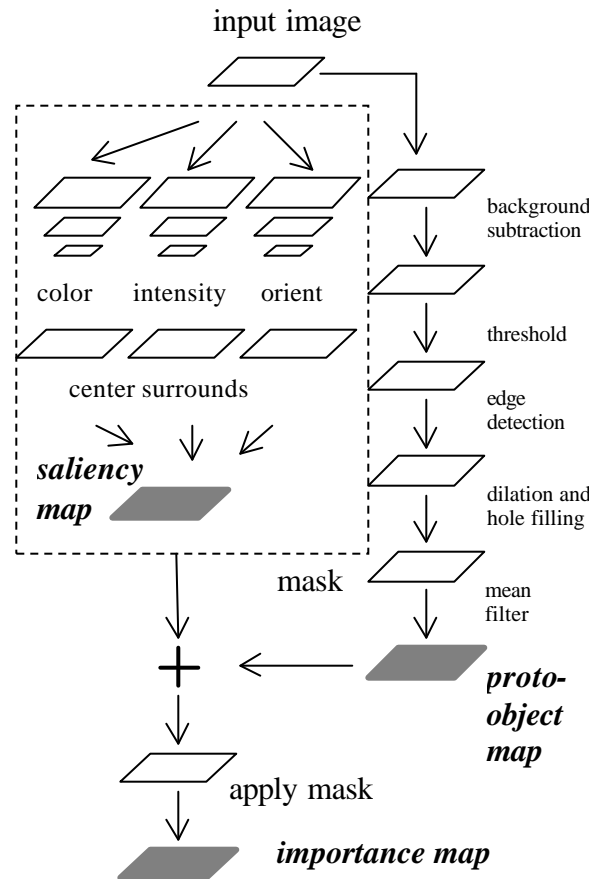


Figure 1 Construction of the *importance map*, area within dotted lines is adapted from Parkhurst, Law & Niebur⁴.

A Gabor filter is derived from a Gaussian modulated sinusoid, and enables a local spatial frequency analysis to detect oriented edges at a particular scale. The justification for using this type of filter is that it simulates the structure of receptive fields in area V1 neurons that have been found to be tuned to particular orientations and spatial frequencies (Hubel & Wiesel¹⁸). Figure 2 is a graphical depiction of the basis functions used to model the spatial response of the Gabor filters.



Figure 2 Basis functions of the Gabor filters used to detect oriented edges. From left, 0°, 45°, 90°, and 135°.

The center-surround computation is performed on each of the oriented pyramids, the pyramids are then scaled and summed to form a single orientation feature map. Once each feature map has been constructed, they are linearly summed together and the result is again scaled to be in the range 0-1.

It should be noted that the Parkhurst, et al. model uses the result of the saliency map to locate a single region of highest saliency as the selection for the next fixation. A winner-take-all mechanism selects the most salient region and moves the fixation to that location; this process is followed by an inhibition-of-return (Posner & Cohen¹⁹) procedure to reduce the saliency at the current fixation point so that the next highest region may be selected for the following fixation. The following adaptation of the model, for which the present study is concerned, does not use an inhibition of return mechanism. This is because the primary purpose of the adapted model is to predict the *likelihood* of a fixation on any particular region in the image, rather than to predict the locations of a sequence of fixations.

2.1.2 The proto-object map and the importance map

The proto-object map is constructed in parallel with the saliency map, and is used to identify regions of potential objects in the image. The algorithm is based on detecting texture from edge densities. The first stage involves subtracting off an estimate of the background from the original image. To do this, the image (1280 x 768) was sectioned into 16x16 blocks (representing ~ 3/4° of visual angle as described in Section 3). The pixels within each block were then set to a single value. For simplicity, the minimum of that block was used:

$$f(x,y) = \min_{(s,t) \in S_{x,y}} \{ g(s,t) \} \quad (1)$$

where $S_{x,y}$ is the area defined by the block. This has the effect of locating regions of relatively uniform intensity in the image, and simulates the figure/ground segmentation of perceptual organization (Rubin²⁰). The background image is then subtracted from the original image to create a foreground image.

The next stage is to threshold the foreground image to create a binary representation of the foreground, which is used for edge detection. A Canny edge operator is used on the binary image to detect both weak and strong edges in the foreground image, including weak edges only if they are connected to strong edges. The Canny operator uses a derivative of Gaussian filter to calculate gradients in the thresholded image, and finds the local maxima of those gradients. The Canny operator has been shown to be optimal for detecting two-dimensional luminance step-edges and closely approximates the operation of simple cells in striate cortex that sum the outputs from lower-level cells (Canny²¹). From the edge image, potential object regions are located. A dilation procedure is used to thicken the edges and fill in holes:

$$A \oplus H = \{ z \mid (H)_z \cap A \neq \emptyset \} \quad (2)$$

H is the structuring element after rotation about the origin and a shift by z pixels. The dilation is the set of all possible displacements z , such that A and H overlap by at least one pixel. Essentially the dilation procedure expands the boundaries of the edges. After the dilation, a majority procedure is used to fill in the

holes between the dilated edges. This results in the creation of a binary object mask. The binary mask is saved for a subsequent procedure. After the creation of the mask, the mask is then smoothed locally with a mean filter:

$$f(x,y) = \frac{1}{(m \times n) \sum_{(s,t) \in S_{xy}} g(s,t)} \quad (3)$$

The mean filter operates on an $m \times n$ subimage window centered at pixel x,y . S_{xy} is the area in the mask defined by the window. The smoothed mask is referred to as the proto-object map, and is used as an added feature channel for the pre-masked saliency map. Finally, the binary object mask is applied to the map obtained after averaging together the color, intensity, orientation, and proto-object maps. This has the effect of further inhibiting regions in the image that are not likely to be objects, and enhancing regions that are. Figure 3 shows the process of creating the mask for one of the input images.

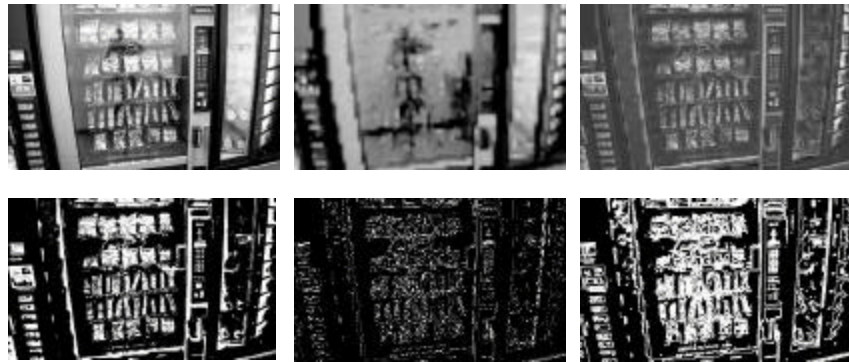


Figure 3 Creation of the object mask. Top row, from left, input image, background estimate, after subtraction of background. Bottom row, from left, thresholded image after background subtraction, after edge detection, and after dilation and majority operations.

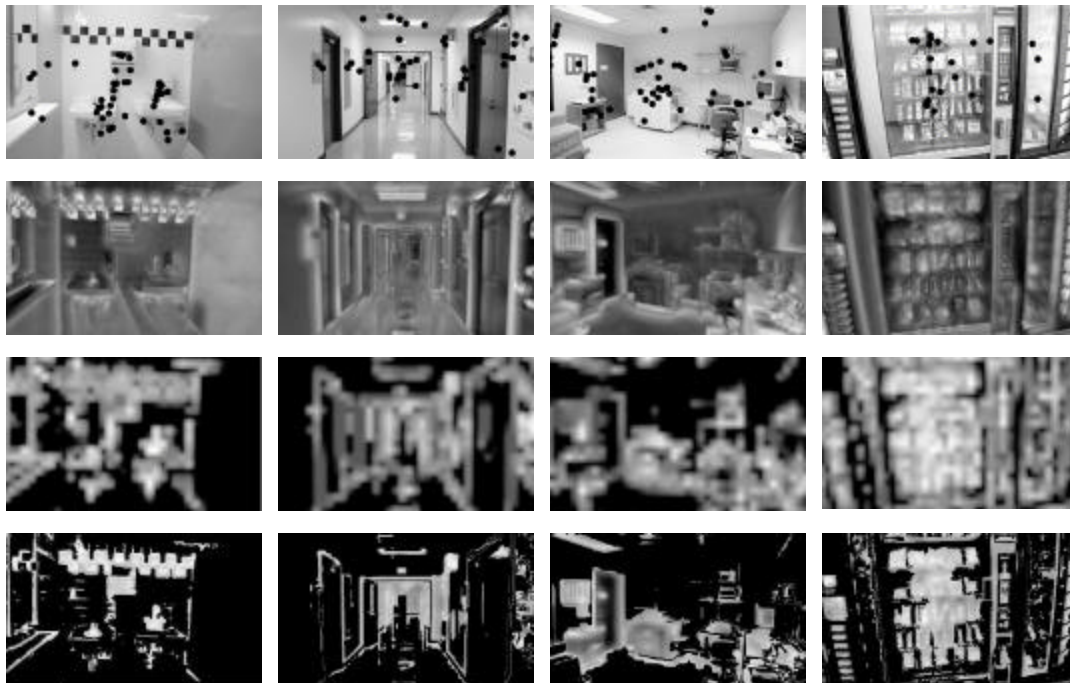


Figure 4 Input images (1st row) – from left, washroom, hallway, office, and vending machine. Saliency map (2nd row), proto-object map (3rd row), and resulting importance map (4th row)

The top row of Figure 4 shows each of the four input images with a single subject's overlaid fixation plot, and the following rows show the resulting maps for the corresponding image. The 2nd row shows the saliency map, the 3rd row shows the proto-object map, and the 4th row shows the final importance map. Regions of high intensity in the maps indicate a high probability of fixation.

3. METHODOLOGY

3.1 Eye movement monitoring

In order to determine the correlation between the computational models of attention previously discussed and the fixation patterns of people viewing natural images, eye-tracking data was collected and analyzed. An Applied Science Laboratories (ASL) model 501 video-based eye-tracking system was used to record eye movement data. The ASL system is head-mounted, and allows the subject to move his/her head freely during the experiment. A near-infrared LED (IRED) illuminates the eye, allowing a retro-reflected bright image of the pupil to be captured by the head-mounted eye camera. Gaze position is calculated at a video field rate of 60 Hz, providing a temporal resolution of 16.7 msec.

Gaze position, in terms of vertical and horizontal eye position coordinates with respect to the display plane, is defined as the integration of eye-in-head and head-in-space positions. The eye-head integration is performed by using the output from the eye-tracker in conjunction with a Polhemus 3-Space Fastrak magnetic head tracker (MHT). The MHT consists of a fixed transmitter placed near the subject, and a receiver that is attached to the headband of the eye-tracker. The display plane position and orientation with respect to the transmitter is measured and recorded. Gaze position is then calculated as the intersection of the line of sight with the display plane, and recorded as (x,y,z) coordinates for position.

A calibration procedure is performed for each subject before the experiment begins, and checked at the end of the experiment. To perform the calibration, a grid of 9 points is projected onto the display plane. The subject holds his/her head steady while fixating each of the 9 points sequentially. To ensure that the head does not move during calibration, a semiconductor laser that is mounted to the headband of the eye-tracker projects to one of the 9 points in the field. The subject is asked to maintain the projection of the laser on the point during the calibration. After calibration, the head is free to move. The accuracy of the eye-tracker after calibration was measured to be within approximately 1° of visual angle.

A 50 inch Pioneer Plasma monitor was used to display the images. The screen size was 1280 x 768 pixels, with a screen resolution of 30 pixels/inch. The display area subtended a visual field of 60° horizontally and 35° vertically at a viewing distance of 38 inches. At this distance, approximately 22 pixels covers 1° of visual angle.

3.2 Data Collection

Eleven paid subjects (7 male, 4 female) from the RIT community participated in the experiment, all with normal or corrected to normal vision, and all were naïve with respect to the purpose of the experiment. The experiment lasted approximately 1 hour.

Each subject viewed a total of 180 images divided over 2 sets of 90 images each. Each set was labeled either 'free-view', where the subject was instructed to freely view each image as long as desired, and 'multi-view,' where the subject was given explicit instructions before viewing a group of images – again, with no time limit. For the purpose of this study, only the 4 images shown in Figure 3 were used for the data analysis, however eye movement data was collected for all 180 images and will be used as part of a subsequent study. During free-view each of the 4 images was viewed freely. During multi-view, the same 4 images were viewed 3 separate times. Before each multi-view viewing, the subject was given a verbal instruction to imagine, where the instruction was specific for the context of the image. For example, for the hallway image (2nd from the left in Figure 3) the instruction was either "Put something in the garbage," "The fire alarm has just gone off," or "Find a bathroom". The different tasks are labeled M1, M2, and M3.

Table 1 shows the instructions for the various multi-view tasks:

Image	Task M1	Task M2	Task M3
Washroom	Wash your hands	Fill a cup with water	Comb your hair
Hallway	Put something in the garbage	Find a bathroom	The fire alarm just went off
Office	Get supplies from the closet	Work at the computer	Make a photocopy
Vending	Check for Skittles	Buy a Snickers bar	Check for change

Table 1 Instructions for the multi-view set

All of the subjects viewed the multi-view images in the same order each of the 3 times they were displayed, however, the order of the task for a given image was changed for some of subjects. Six subjects were labeled group “A” and were given the tasks in one order, and 5 subjects were labeled “B” and given the tasks in a different order.

Software was developed to analyze the raw data to remove blinks and track losses, find fixations, and eliminate raw data samples that coincided with a saccadic eye movement. The fixation finder uses a velocity threshold of approximately 45° per second to determine the beginning and end of a fixation, and a fixed window size of approximately 1° to collect all data points within a particular fixation.

4. RESULTS

4.1 Comparison of the maps to fixation densities

A metric was developed to measure the correlation between the density of subjects’ fixations on a particular image and the corresponding saliency, proto-object, and importance maps. First, the mean conspicuity value for each map was calculated. A conspicuity value is defined to be the value at an x,y coordinate in a map after all of the computations for that map have been completed. The mean value of conspicuity is the expected value at any random location in the map, and is a measure of the overall conspicuity of a particular map. Figure 5 shows the mean conspicuity value for each of the 3 maps. The number of samples for each of the maps is the size of the input image – $1280 \times 768 = 983,040$.

Next, the mean conspicuity of fixations is calculated. For each fixation, the x,y coordinate in the image at fixation is found, and the conspicuity value from a particular map at that location is extracted. That value is defined to be the conspicuity of the fixation. If the fixation conspicuity is not significantly different from the mean conspicuity of a map, then that map is not a good predictor of the fixation because any random location would do just as well. If, on the other hand, the fixation conspicuity is significantly greater than the mean of a map,

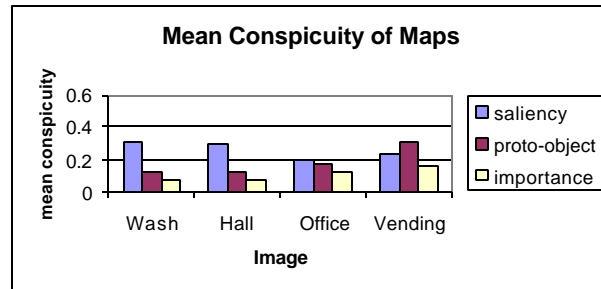


Figure 5 Mean conspicuity of maps

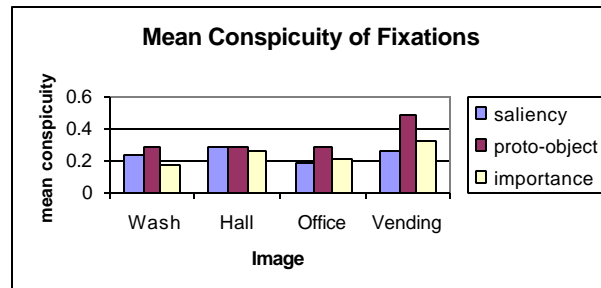


Figure 6 Mean conspicuity of fixations – all subjects, all tasks

then the map is a good predictor because the fixation is on a highly conspicuous region of the image. Figure 6 shows the mean conspicuity of fixations, over all subjects and all tasks – free-view, M1, M2, and M3. The number of samples for Wash is 790, for Hall is 904, for Office is 813, and for Vending is 1416. The sample size varied primarily because the self-paced viewing patterns were different for the four images. A statistical t-test was performed to compare the means of the fixation conspicuities to the means of the corresponding map conspicuities. In each case, the means were found to be significantly different ($p < 0.005$, $\alpha = 0.05$). However for the saliency map, the mean values for fixation were significantly *less* than the mean values for the map for the Wash, Hall, and Office images. This indicates that subjects had a tendency to look away from the regions that were indicated as highly salient in that map. For both the object map and the importance map, the fixation means were significantly higher than the map means for all images.

A comparison of the mean conspicuity of fixations is not sufficient to make any meaningful conclusions about the relative effectiveness of each map in predicting fixation locations. This is because, as evidenced in Figure 5, the mean conspicuity varies greatly between maps. For example, the mean map conspicuity for the Wash image is 0.32 for the saliency map, whereas it is 0.07 for the importance map. This is an indication that either every pixel in the importance map is much lower in conspicuity than the saliency map, or the conspicuous regions are more spatially constrained in the importance map, or both. A review of Figure 4 shows that the second case is the situation. The implication is that fixations on regions of the importance map that carry no conspicuity value will significantly lower the correlation, as defined in Section 4.1, whereas fixations on regions that carry a high conspicuity value will significantly raise the correlation. For the saliency map, virtually any fixation will give some value of conspicuity, and some fixations will give a high value.

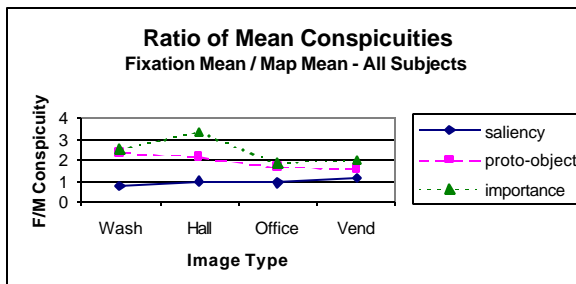


Figure 7 Ratio of the map means to the fixation means for all subjects over all tasks

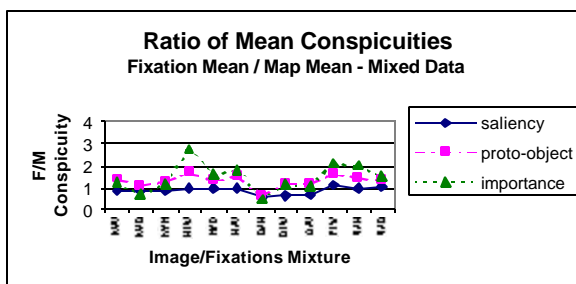


Figure 8 Ratio of the map means to the fixation means – mixed maps/fixations. Labels along the x-axis indicate map/fixation pattern for free-view.

A more meaningful comparison is to compare the ratios of fixation means to map means, as shown in Figure 7. This will have the effect of normalizing the fixation conspicuity for each map. Figure 7 shows that for every image the proto-object map gives a higher value for conspicuity than the saliency map, and the importance map gives a higher value than either the proto-object map or the saliency map alone. This is an indication that the importance map makes use of the low-level feature information from the saliency map (color, intensity, and edge orientation) as well as the location of potential objects in the scene to determine conspicuity.

An important consideration is how well each map correlates to random fixation locations. If a particular map correlates very highly with any fixation sequence, regardless of the actual image viewed, then it is evidence that the map is not a true indicator of either conspicuous low-level features or high-level goals. Figure 8 shows the ratio of fixation means to map means, where the maps and the fixations do not correspond. In other words, the fixations are for a particular image, and the maps are for a *different* image.

Figure 8 indicates that both the proto-object map and the importance map give values of mean conspicuity that are greater than random in the mixed case, although not as high as for the non-mixed case. One would expect that the conspicuities would be closer to random because the fixation patterns are irrelevant to the map. It is possible that the fixation patterns, while they are not relevant for the particular map under consideration, contain a spatial bias that is also a reflection of the same spatial bias inherent in the maps. For example, the highest ratios of mean conspicuities for the mixed data occur with the Hallway and Vending maps (from the left on Figure 8 – data points 4,5,6, and 10,11,12 respectively). Referring to Figure 4, it appears that both the proto-object map and the importance map have values of high conspicuity towards the center of the image for two of the four input images – Hallway and Vending. It is possible that those maps reflect the bias of the photographer – that is, a bias toward centering the “important” areas of the scene in the viewfinder – and that bias may also be present in the viewing patterns of the subjects, regardless of the image.

4.2 Examination of central bias

Figure 9 shows histograms of fixation distances from the center of each image, over all subjects and for all tasks for each image.

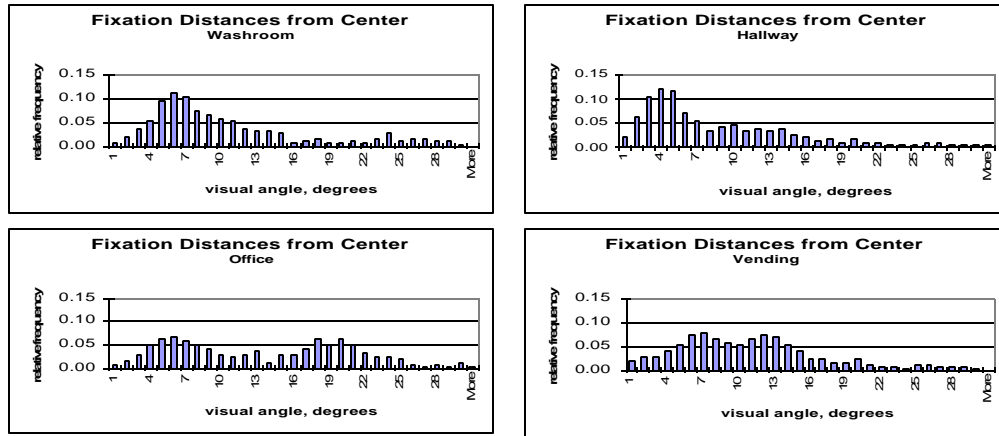


Figure 9 Histograms of fixation distance from the center of each image.

From the histograms, it can be seen that most of the fixations are within $\pm 10^\circ$ of the center of three of the four images, even though that area comprises less than 20% of the total image area. At a distance of $\pm 10^\circ$, 63% of the Washroom fixations, 66% of the Hallway fixations, 44% of the Office fixations, and 52% of the Vending fixations are 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 images.

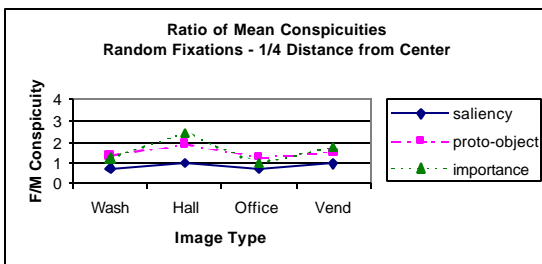


Figure 10 Ratio of mean conspicuities for random fixations 1/4 distance from center of images

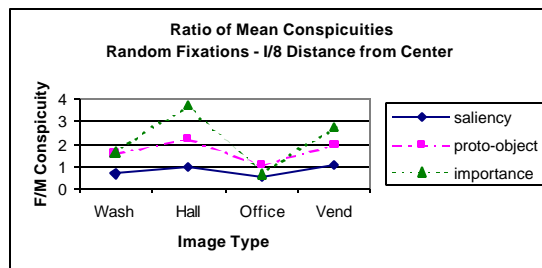


Figure 11 Ratio of mean conspicuities for random fixations 1/8 distance from center of images

If a uniformly spatially distributed random sequence of fixations is generated, it is expected that the ratios of mean conspicuities will approach unity as the sample size increases. However, as indicated by Figure 9, the fixation locations for three of the four images are not uniformly distributed across the image region, but are spatially biased toward the center of the image. Therefore, a random fixation sequence was generated that constricted all fixations to a $\frac{1}{4}$ image size window ($29.1^\circ \times 17.5^\circ$) located at the center of the image, and another sequence that constricted all fixations to a $\frac{1}{2}$ image size window ($14.5^\circ \times 8.7^\circ$). Figures 10 and 11 show the ratios of fixation means to map means for the constricted fixations. From Figures 10 and 11 it is clear that conspicuity is high for both the proto-object and importance maps when fixations are constricted to the center of either the Hallway or Vending image, and somewhat so for the Washroom image. This is in close agreement with the mixed map/fixations data of Figure 8, and reflects the finding that the mixed data is not truly random, but rather exemplifies the general tendency to look towards the center of an image. While some proposed models include an imposed, explicit bias toward the center to improve performance, both the proto-object and importance maps preserve the central tendency without artificial means. The Hallway and the Washroom scenes elicited the strongest central tendency bias on subjects' fixation patterns, and this is reflected in the location of regions of highest saliency toward the center of the corresponding high-level proto-object and importance maps for those images. The high-level maps of the Office image, for which subjects did not preferentially view the center, maintain a close to random saliency for the randomly generated fixations, as expected.

4.3 Comparison across tasks

As part of the experimental procedure, subjects viewed each image four times – once for free-view, and once for each of the three multi-view tasks as given in Table 1. It would be interesting to know if the multi-view instructions elicited a stronger correlation between map conspicuity and fixation conspicuity than the free-view condition. Figure 12 shows a comparison of the conspicuity ratios for each of the conditions for the saliency map. There appears to be no difference between any of the viewing conditions for this map. Figure 13 shows the same comparison for the importance map. This figure shows that there is some difference in the conspicuity ratios for each of the viewing conditions. The free-view condition elicits a lower correlation between map conspicuity and fixation conspicuity for all but the Vending M2 and M3, and Wash M3 conditions. This preliminary analysis suggests that a specific task posed to the viewer will tend to be more “object-oriented” than if there is no particular task. This is in agreement with the conclusions of Land et al.¹⁵ and Pelz and Canosa¹⁶ who showed that people have a strong tendency to look at objects in the environment that are important for current and future actions.

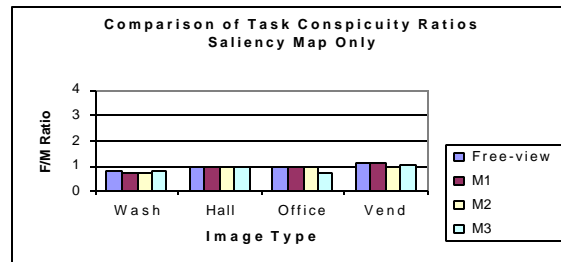


Figure 12 Saliency map task comparisons

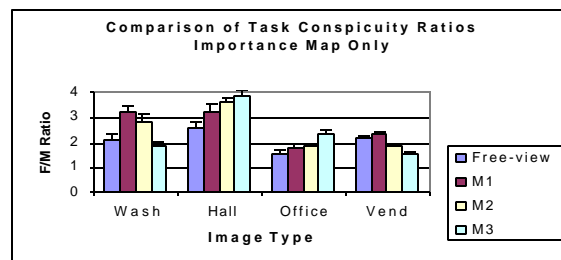


Figure 13 Importance map task comparisons

5. DISCUSSION AND CONCLUSION

The purpose of the present 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 patterns. A high-level proto-object map was constructed to identify regions that contain potential

objects in the image, and was used in conjunction with a low-level saliency map that locates features corresponding to colorfulness, luminosity, and oriented edges. The proto-object map simulates the top-down constraints relating to task-relevancy, expectation of location (central bias), and goal-oriented behavior. The saliency map simulates the bottom-up responses of neurons in terms of the center-surround organization of receptive fields, lateral inhibition, and color-opponent cells. We call the resulting fusion of the two maps an importance map because it has the effect of enhancing regions of the image that people consider important and inhibiting regions that people do not consider important. The result is a biologically inspired model that is in agreement with empirical evidence as well as with our *intuition* of where people will look in natural, realistic images. In support of the concept of the importance map, Moran & Desimone²² provide physiological evidence showing that failure to attend to effective stimuli (effective in terms of a high cell response for that stimulus) will not cause a cell to fire. In other words, a region of the scene that exhibits high salience according to the predicted response of low-level detectors will not be activated in the absence of focused attention to that region. Focused attention on relevant objects may be the “glue” that binds together properties of the scene and our perception of a continuous, coherent reality.

The model of saliency developed by Parkhurst, et al. is a biologically plausible, realistic account of how low-level image features may predict cell responses. It has been shown to be particularly useful for predicting fixation locations in images where color information predominates and orientation and luminosity have a lesser role, such as in pseudo-colored fractal images. The importance map, on the other hand, does not include any information from the scene about color or luminance differences, and must rely on the low-level metrics for this information. This is perhaps the reason for an increase in the conspicuity ratio when the saliency map is used in conjunction with the object-oriented proto-object map.

The importance map is a better predictor of actual behavior than the saliency map alone because the importance map does not rely solely on local image properties to predict fixation locations. Parkhurst et al. noted the limitations of a non-object approach: “subjects often examined objects on table tops independent of their salience,” suggesting that global, scene-dependent strategies may play an important role in determining fixation locations.

The importance map also includes an inherent global bias towards the center of the image when that bias is warranted, at least for the four images that were used as part of this study. In order to simulate the empirically-observed central bias, the saliency map when used alone must include a spatial frequency reduction toward the periphery of the image, and is modeled by Parkhurst et al. using a 2-D Gaussian blur centered on the image. An obvious extension to this study will be to examine the central-bias issue with a broader range of images, including images containing people, images of outdoor scenes, and spatially incoherent images such as fractal and other abstract images.

Another situation to be considered is the effect of block size during 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 potential objects in a wider variety of scenes. Also, a better measure of the conspicuity ratio may be obtained by measuring the conspicuity values over a broader region, and integrating the result, instead of measuring it at each pixel as is done in the current study. A lower granularity for the conspicuity ratio may prove more useful for determining conspicuity over a greater spatial range in the image.

In 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. This approach suggests that 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 density 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.

ACKNOWLEDGEMENTS

Thanks to Roger Gaboriski and Carl Reynolds for help with the saliency map code generation, and Jason Babcock for help with eye-tracker data collection software. This work was supported in part by the Naval Research Laboratories; the New York State Office of Science, Technology, and Academic Research; and the Xerox Corporation.

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), pp. 368-373, 1997.
2. B.A. Olshausen and D.J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, **381**, pp. 607-609, 1996.
3. C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, **4**, pp. 219-227, 1985.
4. D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, **42**, pp. 107-123, 2002.
5. L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, **40**, pp. 1489-1506, 2000.
6. J. Theeuwes, "Perceptual selectivity for color and form," *Perception and Psychophysics*, **51**, pp. 599-606, 1992.
7. W.F. Bacon and H.E. Egeth, "Overriding stimulus-driven attentional capture," *Perception and Psychophysics*, **55**(5), pp. 485-496, 1994.
8. D.O. Hebb, *The Organization of Behavior*, John Wiley & Sons, New York, 1949.
9. L. Kaufman and W. Richards, "Spontaneous fixation tendencies for visual forms," *Perception and Psychophysics*, **5**(2), pp. 85-88, 1969.
10. A. Yarbus, *Eye Movements and Vision*, Plenum Press, New York, 1967.
11. G.T. Buswell, *How People Look at Pictures: A Study of the Psychology of Perception in Art*, The University of Chicago Press, Chicago, 1935.
12. D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognizing patterns," *Vision Research*, **11**, pp. 929-942, 1971.
13. T.J. Andrews and D.M. Coppola, "Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments," *Vision Research*, **39**, pp. 2947-2953, 1999.
14. H. Deubel, and W.X. Schneider, "Saccade target selection and object recognition: Evidence for a common attentional mechanism," *Vision Research*, **36**(12), pp. 1827-1837, 1996.
15. M.F. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, **28**, pp. 1311-1328, 1999.
16. Pelz, J.B. and Canosa, R., "Oculomotor behavior and perceptual strategies in complex tasks," *Vision Research*, **41**, pp. 3587-3596.
17. P.J. Burt and E.H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, **31**(4), pp. 532-540, 1983.
18. D. H. Hubel and T.N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *Journal of Physiology*, **195**, pp.215-243, 1968.
19. M.I. Posner and Y. Cohen, "Components of visual orienting," in H. Bouma and D.G. Bouwhuis (eds), *Attention and Performance*, **X**, pp. 531-556, Erlbaum, Hillsdale, New Jersey, 1984.
20. E. Rubin, *Visuell Wahrgenommene Figuren*, Glydenalske boghandel, Kobenhaven, 1921.
21. J.F. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**, pp. 769-798, 1986.
22. J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science*, **229**, pp. 782-784.