

## *Chapter 3*

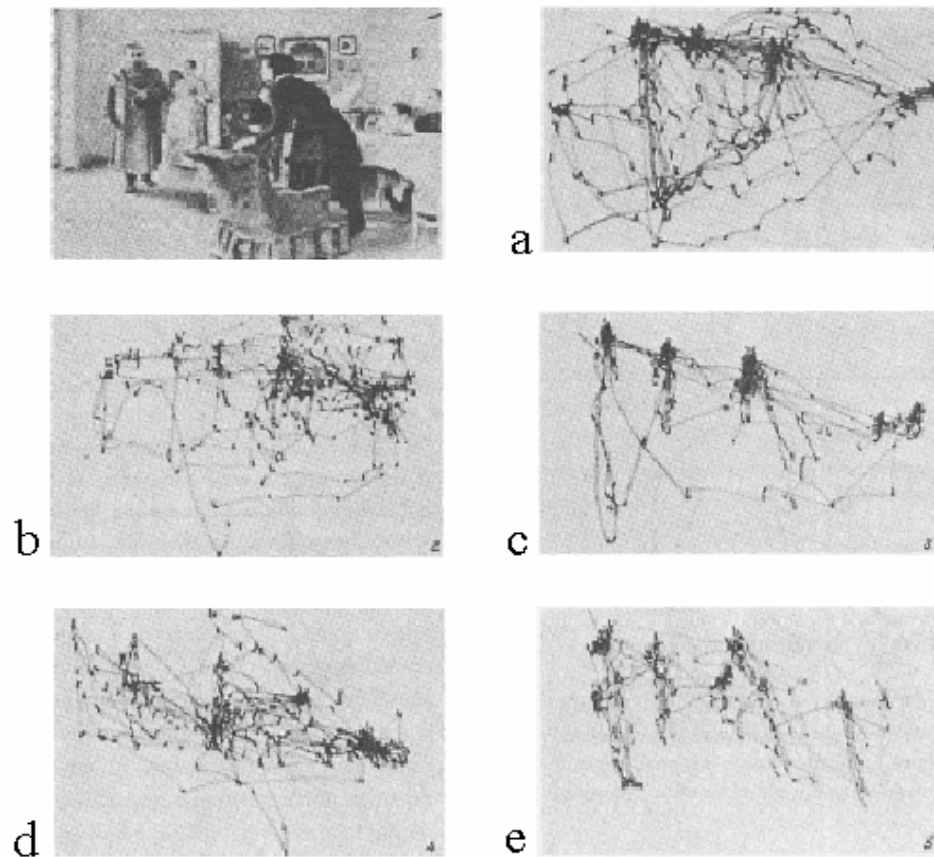
# Building Object Representations

### **3.1 Introduction**

#### *3.1.1. Background*

The retina is structured in such a way that central vision (in the fovea) receives the highest resolution processing. This resolution degrades toward the periphery. This structure is also reflected in the brain: more cortical area in primary visual cortex is devoted to processing central vision than peripheral vision. We move our eyes around as we view a scene, or as we read, so that high resolution processing can be dedicated to different locations.

Some of the very first eye movement recordings were made by Buswell (1935). He noted that eye movements were not random, but very idiosyncratic. Yarbus (1967) made the very important observation that cognitive, or “top-down” factors play a role in where observers look. For example, the question the observer needs to answer will influence where he looks in the image (Figure 3.1).



**Figure 3.1.** Subjects were asked to view the painting “Unexpected Visitor” under different conditions. (a) Free viewing. (b) Estimate the economic level of the people. (c) Judge their ages. (d) Guess what they had been doing before the visitor’s arrival. (e) remember the clothes worn by the people.

Although eye movements are considered idiosyncratic, the locations at which observers fixate have been shown to correlate with what is perceived as “interesting” in the image (Mackworth and Morandi, 1967). Observers also look toward high curvature in the image, but do not fixate the contour; they fixate just inside of it, at the interior of “angles” (Kaufman & Richards, 1969). Melcher & Kowler (1999) showed that first fixations to an object land on the center of gravity. The same phenomenon has been demonstrated for rendered three dimensional objects (Vishwanath & Kowler, 2003).

This result is not surprising when the task is to localize the object. Little work has been done to characterize the fixations that follow localization.

It is not a new idea that observers might learn objects by isolating features with fixation and build a description based on the relation of parts over subsequent fixations (Hebb, 1968). Norton and Stark (1971) formally proposed that observers use “scanpaths” – the spatiotemporal pattern of eye movements – to encode and retrieve memories. This would suggest that observers employ a certain top-down inspection strategy when attempting to recognize objects and scenes. But how do we know which strategy they employ in a given situation? If the class of the object and/or scene is already known, we might be able to predict the most probable fixation locations (Oliva, Torralba, Castelano & Henderson, 2003).

What happens when all observers have is bottom-up information, for example, when children are learning new objects? There is obscure evidence that children explicitly trace contours with their eyes during perception (Zinchenko, Chzhi-Tsin & Tarakanov, 1962), but this finding needs to be replicated. Several attempts have been made to predict “salient” areas of an image based on bottom-up information (Itti & Koch, 2000; Kadir & Brady, 2001; Nothdurft, 2002). Success of the algorithms has been judged through inspection by the authors, rather than by a rigorous comparison with human data.

### *3.1.2. Efficient Information Processing*

It is widely accepted that observers look at “interesting” points in an image. We would like to propose that, in fact, observers look at “informative” points in an image.

Consider the case of encoding a new object. The information might be the shape, color and texture of the object, how many parts it has and in what context it is found. Shape is one of the strongest cues for recognition, so it is not unreasonable to believe that children trace contours when learning objects for the first time because they may be encoding shape. Given unlimited time, the most accurate shape information might be attained in this manner. Evolutionarily, however, it makes more sense that observers would try to gather less accurate, but still useful information in as few fixations as possible. Such a strategy would force fixations more toward angles, just inside contours with high curvature.

### 3.1.3. *Summary of Approach*

In this chapter we outline two eye movement experiments. In the first, we probe behaviors when the object class is known. Subjects are asked to identify large words without context. We find that, based on the knowledge of object class, subjects employ a sensible and robust information gathering strategy for their first fixation. Subsequent fixations are modified to focus on areas where the information is uncertain.

In the second experiment, we probe behaviors in the absence of top-down information. Subjects are asked to learn and then recognize new objects with no known category. We compare their eye movements with those of a model that places its next fixation where it will maximize its knowledge about the orientations of the object contour. Using *sequential information maximization*, the model observer constructs the most informative path which can be compared with human behavior.

## 3.2 Recognizing Known Objects

In reading tasks, eye movements have been studied extensively. They are a bit easier to deal with because reading is a two dimensional problem, the structures of the “objects” are well defined, and the task is clear. When reading a sentence, we also have contextual information from the surrounding text (the “scene”). In this situation, people tend to skip small words such as “the” and “or”, and they apply a single fixation to the words they do not skip, near 40% of its total length (for a review, see Rayner, 1992).

In these studies, we’ve removed the context and enlarged the word to 4 times the normal reading size. We wondered how observers would modify their normal reading behavior to correctly identify the words.

### 3.2.1. *Experimental Methods*

This work was a collaborative effort between the Malik and Schor research groups at UC Berkeley. Collaborators included Greg Mori and Shrikant Bharadwaj. The experiments were conducted in the Schor eye movement lab.

#### *Subjects*

I and my collaborators, along with one naïve observer, were subjects in these experiments. One collaborator was also naïve about the purpose of the experiment.

#### *Stimuli*

The stimulus was a five or six letter word presented in one of four quadrants surrounding a central fixation marker. Each letter subtended one degree in width, about

four times larger than the normal reading size. We made the words large so that subjects would not be able to foveate it with a single glance. In three conditions, the words were masked by covering two letters of the word with an additional letter. Different letter masks were used so that subjects could not learn to ignore the mask. The absolute position of the word was jittered to prevent subjects from anticipating its location in a quadrant. The words were white on a black background.

### Design

The design consisted of four conditions, presented in random order. There were 25 trials for each condition, for a total of 100 trials. The conditions were: unmasked (control), front, middle and end masked.

### Procedure

We fixed subjects' head position with a bite bar, and they viewed the stimulus monocularly. Corrective lenses were placed in the viewing apparatus, if needed to view the stimulus clearly. This was done using the calibration stimulus as a target. Subjects fixated the central marker and pressed a button to begin the trial. Subjects were instructed to look at the word, identify it and return to fixation when finished. After every 5 trials, the subject came off the bite bar and reported the last 5 words he had seen.

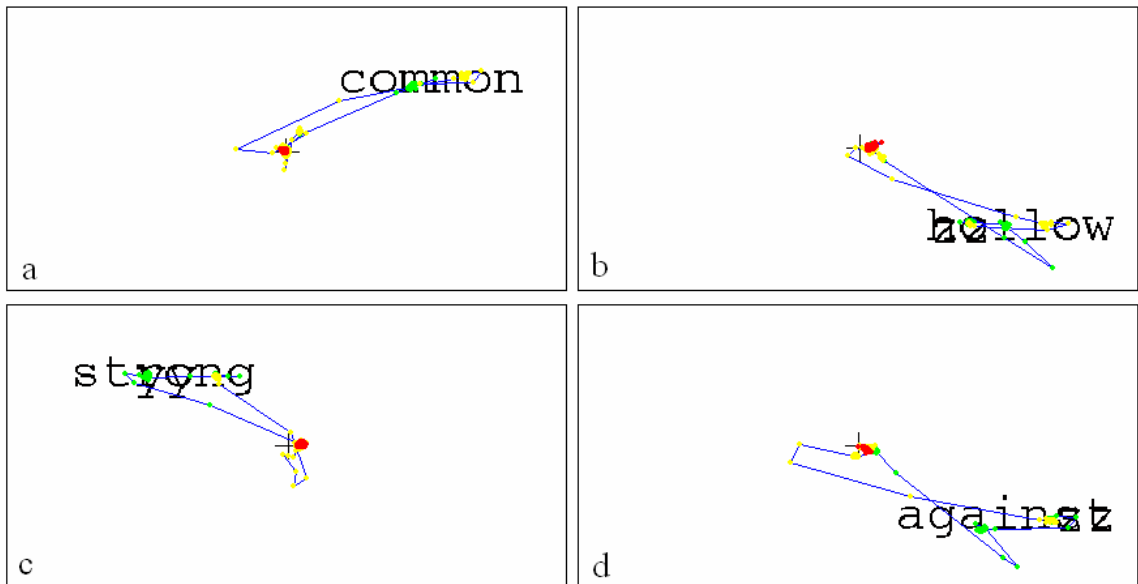
### Apparatus

Horizontal and vertical eye positions were measured with an SRI Dual Purkinje Image Eye Tracker. Eye position was calibrated using a 3 by 5 grid of fixation crosses.

Subjects were asked to fixate each marker in turn, cued to move to the next with a beep. The calibration was repeated twice, with the sequence of fixations reversed to remove any biases in the calibration due to overshooting of the target. Eye position was mapped onto the stimulus using a projective transformation across each square of the grid.

### 3.2.2. Results and Conclusion

In the unmasked condition, subjects were able to identify the word correctly in about one second. In all of the conditions, subjects fixated near 35% of the total word length, consistent with normal reading behavior (Rayner, 1992, p333-354). Subjects may have adopted this strategy because they know there would be a word presented. This is the location of the root of the word, arguably the most informative. Masking only affected later saccades – subjects looked toward the masked area before the word could be reported correctly (Figure 3.2).



**Figure 3.2.** Example of fixation patterns for the four mask locations a) none b) front c) middle and d) end. Green, yellow and red dots represent early, intermediate and late eye positions, respectively.

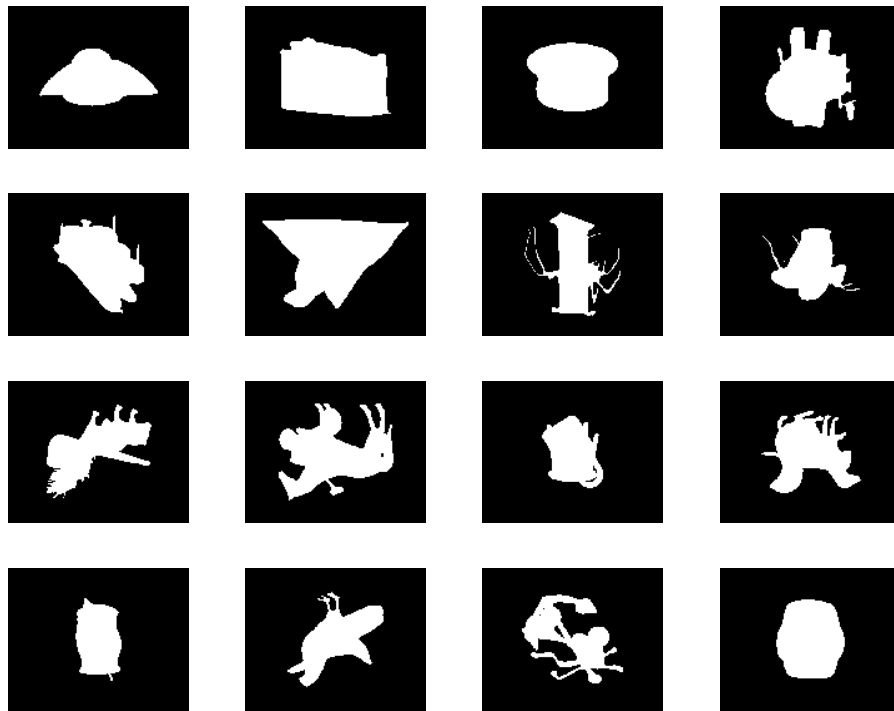
### 3.3 Learning New Objects

We saw in the previous experiment that observers executed a seemingly programmed routine initially, when the object class was known. Next, we investigate how observers study an object they have never seen before, that has no known category. Presumably they must rely heavily on bottom-up information. Because eye movements will be affected by the task, we constructed a learn-recognize paradigm using novel objects. Observers had to create new memory representations for each object, and later match them to make a recognition judgment.

#### 3.3.1. *Experimental Methods*

##### *Stimuli*

To obtain a dataset that was novel, yet as “object-like” as possible, we chose to start with the Snodgrass and Vanderwart (1980) line drawing dataset of 260 objects. Silhouettes were created for a given object, which was then inverted and overlapped with a second inverted object silhouette (Figure 3.3). The result was a series of 50 novel objects, each with unique “parts”. For the recognition phase of the study, we included 200 additional distracter objects, constructed in the same manner as the 50 target objects. The set of distracter objects contained some of the same “parts”, but the conjunction of all parts was never the same. This served to make the discrimination between objects and distracters more challenging so that observers had to encode the object carefully.



**Figure 3.3** Novel objects are constructed by superimposing the inverted silhouettes of two objects from the Snodgrass and Vanderwart dataset (1980).

The novel object was scaled to subtend roughly 30 degrees at a viewing distance of 24 inches. The object appeared to the right or left of a fixation marker. The objects were made this large to ensure that the object could not be foveated with a single fixation. A one degree object could easily be encoded in a single fixation, for example. The absolute position of the object was jittered up to 2 degrees in both the x and y direction. The objects were presented as white silhouettes on a mid-gray background.

### Subjects

Three subjects between the ages of 20 and 30 participated in the experiment. All had experience as psychophysical observers, but none had ever participated in an eye tracking experiment before. The subjects had little or no previous visual experience with the Snodgrass and Vanderwart dataset. None of the subjects had any experience with the silhouettes of the original objects or with the novel objects we constructed. Refer to CPHS protocol 2003-8-50.

### Design

The experiment consisted of 4 learning phases and 4 recognition phases, conducted over 5 consecutive days. A learning phase occurred on days 1 through 4, and a recognition phase occurred on days 2 through 5. Subjects slept between each learning and recognition phase, to allow for memory consolidation. A learning phase consisted of the same 50 target objects, presented in random order. The objects could appear left or right of the fixation marker with equal probability. During recognition phases, 10 target objects were mixed in series with 40 new distracters. Subjects never saw the same distracter object twice.

### Procedure

All phases of the experiment were run in a quiet, dimly lit room to reduce distractions. On the first day, subjects were tested for eye dominance. We tracked the position of the dominant eye as subjects viewed the stimulus monocularly. We performed an initial calibration and saved the settings to speed up the calibration step

before each phase of the experiment. Once the subject was comfortable and calibrated within the setup, he was shown a set of 50 trials. On each trial, the subject first fixated the marker, and then pressed a key to cue the object when ready. After a 1 second delay, an object appeared for 5 seconds. On learning trials, subjects were told to “look at the object and remember it for later”. On recognition trials, subjects were asked “is this an object that appeared in the learning set?” and they responded with a confidence rating of 1 (definitely no) to 5 (absolutely yes). If they completed their investigation within the allotted 5 seconds, they were instructed to return to fixation. On average, a session of 50 trials took about 15 minutes.

### Apparatus

We used an Arrington ViewPoint Eye Tracker to measure eye position during the trials. A small infrared LED illuminates the subject’s eye, and an IR sensitive camera captures video at a rate of 30 frames per second. We used the pupil only method, which segments the pupil from the image, fits an ellipse to its border and marks the center of the pupil as the current eye position. Subjects’ head position was fixed with a bitebar and forehead rest.

To calibrate subjects, we first adjusted the settings to get good pupil segmentation for the full range of eye movements with respect to the stimulus. Eye position is mapped onto the stimulus by presenting a grid of 25 points across the stimulus space. The points were measured one at a time, preceded by a square that decreased in size to a point, creating a smooth pursuit to the desired location. The result of a good

calibration is a clear, rectilinear mapping, which can be visualized with the ViewPoint software.

The stimulus was also presented using the ViewPoint software “picture list”, and displayed on a PC installed with the video capture card. The experiment was controlled using Matlab, but the results were written to a ViewPoint data file. Subject’s confidence ratings were recorded as “markers” for each trial.

### *3.3.2. Computing Fixations*

The data file contains raw eye position measurements which must be converted into fixation locations and dwell time. We count a fixation as a cluster of 5 or more measurements that lie in circular area with a radius of one degree. The fixation location is taken as the centroid of the cluster of points. Dwell time is proportional to the number of points in the fixation cluster.

### *3.3.3. Results and Discussion*

Subjects reported that they were unsure of their task during the first learning phase, but that it became clear what they were supposed to do after the first recognition phase. As the days progressed, subjects reported that they were becoming familiar with the objects and that the recognition task was getting easier.

On average, dwell times rise, peak, and then decline toward the end of any given trial. There is no apparent difference in the location, amplitude or shape of the peak across phases for learning or recognition trials. We conclude that subjects do not change their processing strategy (or become more efficient) as they become familiar with the

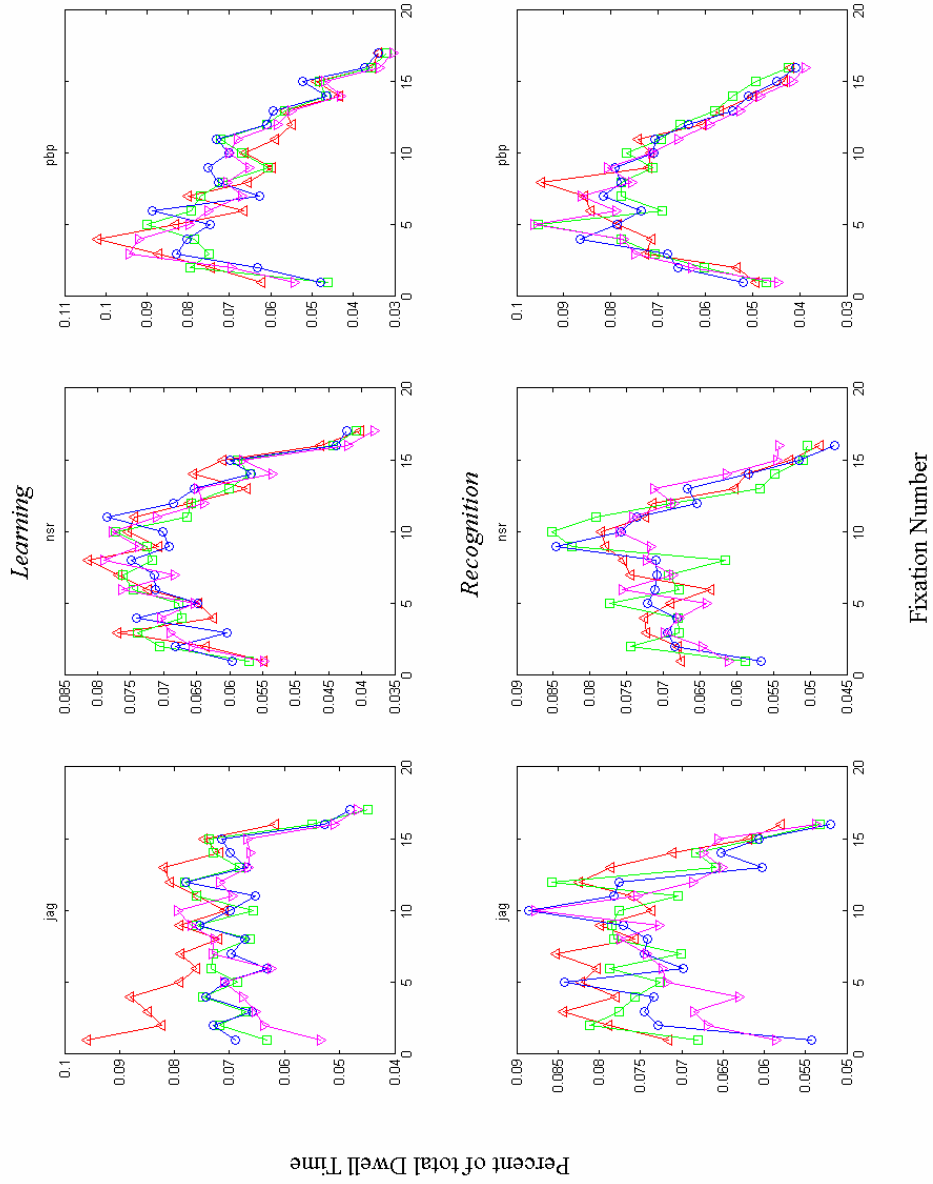
experiment. For subjects nsr and pbp we see a rightward shift in the peak for recognition trials as compared to learning trials. This may suggest that the more intense processing load occurs a little later for recognition trials. It must be noted that this data is quite noisy, and no significant changes were actually detected (Figure 3.4).

### **3.4 Model**

In this section we describe a model of human eye movements that employs a strategy of *sequential information maximization*. Legge, Hooven, Klitz, Mansfield and Tjan (2001) took a similar approach in constructing an ideal observer model for reading. Our situation is simpler in the sense that we are ignoring top-down effects, but it is more complicated in that we are dealing with two dimensional stimuli and must define “information” in terms of image properties. We are working with silhouettes, so the information clearly resides in the contour. We will assume that the exact location of each contour point is computed during the first fixation. The remaining information is in the orientation of these edges.

#### *3.4.1. Representing Information*

We represent orientation information as a probability distribution over the possible orientations. When we have no information, the probability distribution is flat. When we have complete information, the distribution will equal one at the correct orientation, and zero every else. We can represent the nature of this distribution by computing its entropy.



**Figure 3.4.** Percent of total dwell time as a function of fixation number. Learning phases are in the top row and recognition phases are in the bottom row. The red (^), green (□), blue (o) and magenta (v) curves represent phases 1 through 4 respectively. See the text for a discussion.

$$H = -\sum_{\theta} p_{\theta} \log_2 p_{\theta}$$

Entropy will start off at its maximum (no information) and decrease towards zero as we gain information.

### 3.4.2. Cortical Magnification Factor

If we want our model observer to gather orientation information in a way similar to human observers, we must take into account the reason we move our eyes about: people only have high resolution computing power in the fovea. To mimic this biological architecture, we construct a polar grid around fixation  $f_i$  and compute the average entropy of orientations for each bin. The spacing of the grid is based on the human cortical magnification factor (Rovamo & Virsu, 1979). Because our stimuli appear to the right and left of a fixation marker, we chose to compute the factor  $M$  by combining estimates for nasal and temporal retina.

$$M = (1 + 0.31E + 0.000095E^3)^{-1} M_0$$

where  $M_0 = 7.99$  mm/deg and  $E$  is eccentricity measured in degrees. By integrating this function we can map the radius in degrees along the retina to the distance in millimeters along the cortex. The 8 angular bins are equal, and the radial bins increase to maintain equal computing area in the cortex. Specifically, the first bin edge is at ~7mm

in cortex (1 degree) and continues in equal steps of  $\sim 7$  mm until a retinal radius of 26.5 degrees is reached (Figure 3.5).

### 3.4.3. Updating Information Over Fixations

The orientation uncertainty, or entropy, of an edge point  $e$  after fixation  $f_i$  depends on the history of fixations and can be written as  $H(e, f_0 \dots f_i)$ . Before the first fixation, we have no information about edge orientations so  $H(e, f_0) = 3$ , corresponding to a uniform distribution over eight possible orientations. For a candidate fixation location,  $f_i$ , we compute the entropy  $H(B, f_i)$  of bin  $B$  from the groundtruth orientations of each edge point that lie within the bin. We then update the information at every edge point according to the equation

$$H(e, f_0 \dots f_i) \xleftarrow{\text{update}} \min\left(H(e, f_0 \dots f_{i-1}), H_{e \in B}(B, f_i)\right).$$

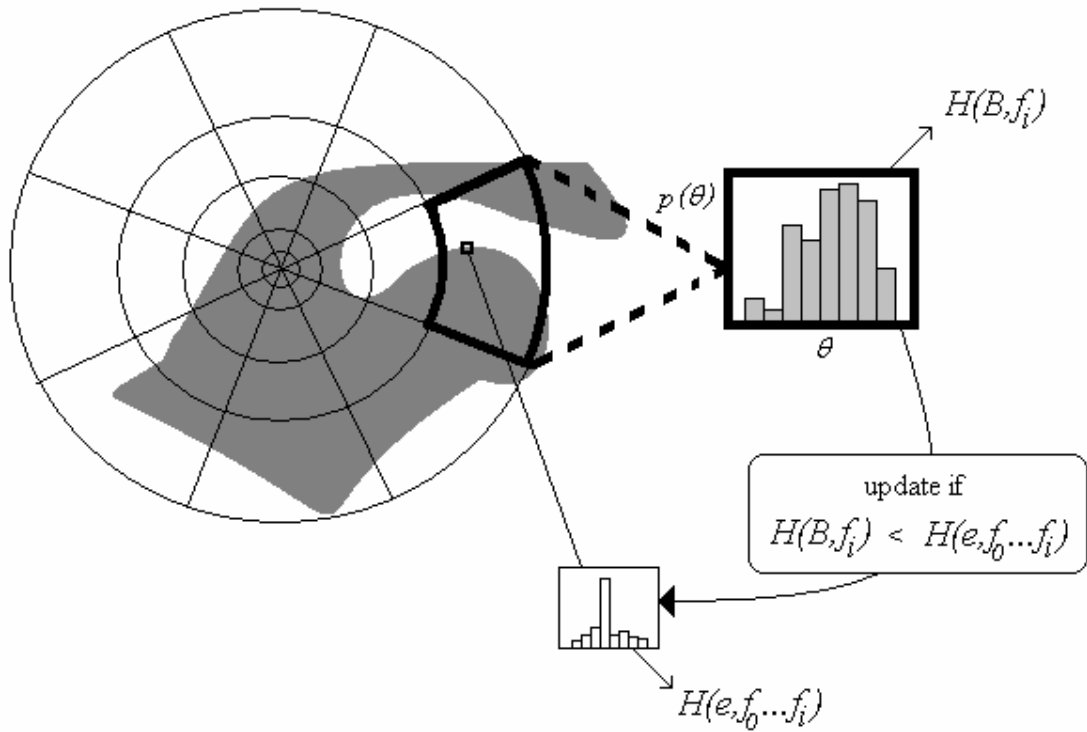
The mean entropy across edge points after fixation  $f_i$  is

$$\bar{H}(f_0 \dots f_i) = \frac{\sum_e H(e, f_0 \dots f_i)}{\# \text{ edges}}$$

and the *information gain* is

$$\text{Gain}(f_i) = \bar{H}(f_0 \dots f_{i-1}) - \bar{H}(f_0 \dots f_i).$$

This process is shown in Figure 3.5. The model observer chooses the fixation that maximize the information gain.



**Figure 3.5.** Schematic of the model observer for sequential information maximization. A polar grid based on the human cortical magnification factor is centered at a possible fixation  $f_i$ . The entropy in bin  $B$  is computed based on the all of the groundtruth edge orientations in the bin. The probability distribution of orientations at an edge point is updated to the bin distribution when the bin entropy is less than the current edge point entropy. The fixation that maximizes information gain is chosen as the next fixation.

#### 3.4.4. Groundtruth Orientation Edge Map

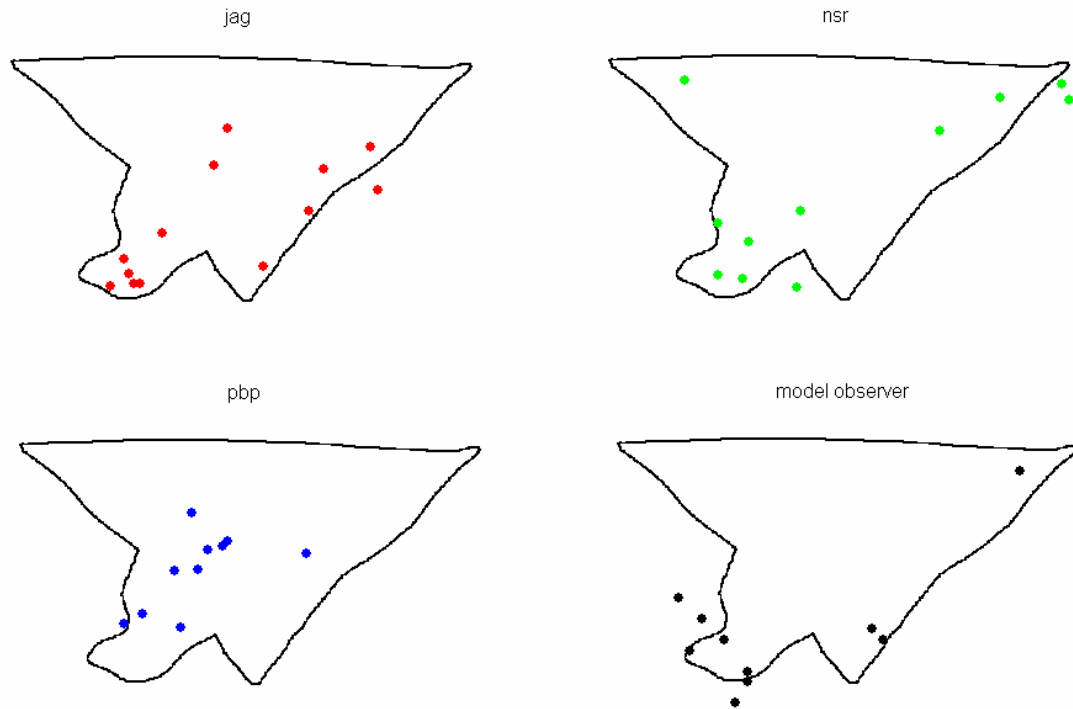
In order to compute the distribution of orientations in a bin, we need groundtruth information about the orientation of each edge point. We get a binary location map by applying a standard zero-cross edge process (available in Matlab) to the object silhouette.

We compute an orientation energy map by filtering the silhouette with Gabors having spatial frequencies 0.5c/deg to 16 c/deg in octave steps, and 8 orientations from 0 to  $2\pi$  in equal steps. The responses in each of the resulting 48 channels are normalized by their maximum response and then summed over scales. We then label each pixel as having the orientation corresponding to the orientation channel with maximum response. A final discrete orientation edge map is obtained by multiplying the location map and orientation energy map.

### 3.5 Results and Discussion

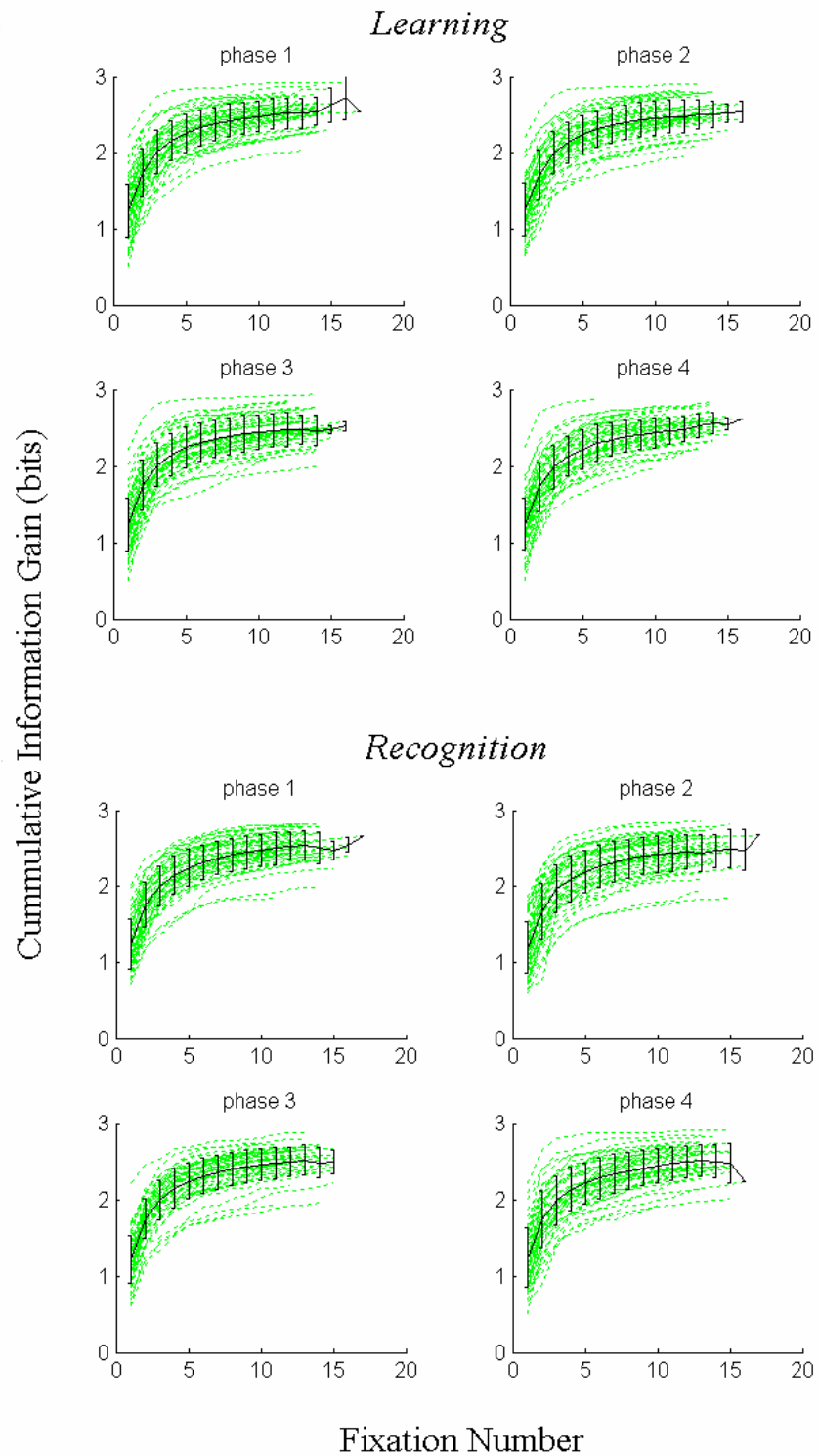
The model observer finds the most informative sequence of fixations  $\{f_0 \dots f_i\}$  by adopting a strategy of *sequential information maximization*. The resulting fixations occur near curved protrusions of the object and away from predictable straight contours, similar to human behavior (Figure 3.6). While the fixations do not align precisely with those of humans, they are much closer than the predictions of current saliency models (e.g., Kadir & Brady, 2001). These models would predict fixations on contour points of maximum curvature (corners, sharp bends, etc). Of course, the idiosyncratic nature of human behavior makes quantitative comparison of fixation locations difficult. We can, however, make a quantitative comparison by comparing the *cumulative information gain* for different fixation sequences. Figure 3.7 shows the cumulative information gain for subject nsr for the different trials within each experimental condition (dotted). There are some trials for which this subject accumulates information very quickly, and others for which he is slower. The subject can make as many as 18 fixations during the 5 seconds presentation, but he collects most of his information in the first 5 fixations. The average

behavior is also plotted for each condition (Figure 3.7, solid curve). The cumulative information reaches a plateau just below 2.5 bits on average for all phases of the experiment. From this observation we conclude that the observer is executing a consistent strategy in terms of information gain.



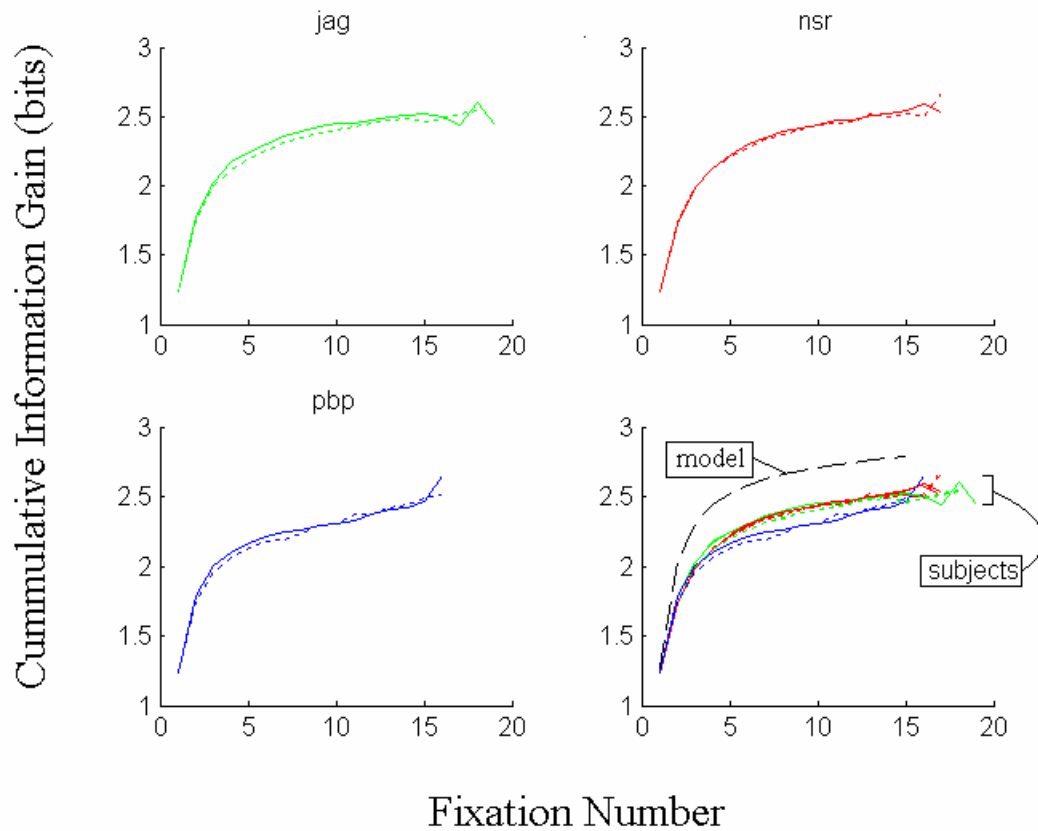
**Figure 3.6.** Example of fixation locations for the three human observers. The path based on sequential information maximization is on the bottom right.

How does observers' performance change from learning to recognition phases? Figure 3.8 plots the average learning (solid) and recognition (dotted) cumulative information curves for the three subjects. There is little difference in the shape of these curves, again confirming that behavior is consistent for a given subject in a given experimental phase. When the curves for all three observers are superimposed (lower



**Figure 3.7.** Summary of information acquisition for subject nsr. The curves for individual trials (50) during each testing phase are traced with a dotted line. The solid line represents the average behavior for that experimental condition. Error bars represent the standard deviation of the mean.

right panel of Figure 3.8) we see that subject pbp adopts a slightly less efficient strategy compared the other two subjects. We've also plotted the performance of the model observer, which selected the fixation that maximized information gain. There is a clear difference between the model and observers, on average.



**Figure 3.8.** The solid curves represent the average amount of information gathered during the learning phases for each subject. The dotted curves are for the recognition phases. Subjects gather information at nearly the same rate in learning and recognition trials. In the lower left panel, we can compare the curves between subjects and the model observer.

The model observer we have described is quite simple, and can be easily modified to better align with the human data. First, we have assumed that edge locations are

known as soon as the first fixation is made. A simple extension of the model can include uncertainty about edge locations as well as uncertainty about their orientations. Second, the size of the bins in the log polar grid can be modified. The smaller a given bin, the more certain we are about the information it contains. We chose our bin size to align with known biology, using the cortical magnification factor as a guide. A better strategy would be to choose bin sizes that will mimic our behavioral orientation sensitivity at different eccentricities. It is difficult to guess at what this sensitivity will be from orientation discrimination thresholds alone because they do not account for crowding and other possible phenomena, but known thresholds could be used initially. Another approach would be simply to tune bin size until we achieve performance similar to that of the human observers. In this way, the model could be used to probe the limits of an observer's ability to localize and judge the orientation of edges during object learning and recognition. Thirdly, the model does not account for partial information. For example, a first fixation might suggest that a given edge point is tilted leftward with  $H=1.5$ . A second fixation might now suggest that the same edge point is tilted rightward with  $H=1.4$ . By combining partial information, we now know that the edge is vertical, but the current model will update the edge to be tilted rightward because that distribution has the lower entropy. Lastly, our model assumes perfect memory. Once the orientation of an edge point has been resolved, it is not forgotten. It is common to see subjects make "confirmatory" eye movements to previous fixation locations towards the end of a trial. It might be that our certainty about the current information degrades with time. Such a property could also be added to the model.

In conclusion, our simple model observer based on *sequential information maximization* from edge orientations provides a first attempt at modeling the dynamic behavior of human eye movements when learning and recognizing complex objects. Our results are promising, and suggest that eye movements based on low-level stimulus features might be predicted in the future.