Integration of form across saccadic eye movements

Mary Hayhoe, Joel Lachter
Center for Visual Science, University of Rochester, 274 Meliora Hall, Rochester, NY 14627, USA
Jerome Feldman
International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA
Received 26 April 1990

Abstract. To perceive a stable world, one must somehow be able to relate visual information from successive fixations. Little is known, however, about the nature of the integrative process. By using a task which requires the integration of spatial position information from different fixations, it is demonstrated that visual information from previous fixations is preserved in a world-centered representation which is precise enough to support judgements of geometric shape. It is also shown that successive views are aligned with respect to common visual features, indicating that visual stability may be normally accomplished by a visual matching strategy in combination with cancellation by an eye-position signal.

1 Introduction
Movements of the eyes, head, and body lead to continual motion of the image on the retina. High velocity saccadic eye movements, which occur several times a second, move the high-acuity foveal region around the scene and impose a frame-like structure on the visual input. Nonetheless, the visual world appears to be stable and of uniform clarity. To explain these phenomena it has often been assumed that the perceptual system constructs a representation of a scene that is integrated across successive fixations, and that successive views are aligned by an eye-position signal (the corollary discharge) which cancels the retinal displacement (von Holst and Mittelstaedt 1950/1971; Sperry 1950). However, this ‘cancellation theory’, as it has been referred to (Matin 1986), indicates nothing about the nature of the visual integration process. One explicit suggestion, made originally by McConkie and Rayner (1967), is that retinotopic images are added together in an extended image whose coordinate system is fixed relative to positions in the world (see also Breitmeyer et al 1982; Jonides et al 1982). Despite its plausibility this idea has little empirical support. Numerous attempts to show some ‘fusion’ or addition of images from different fixations have been unsuccessful, and this has cast doubt on the existence of an integrative process and an extended global image (Pylyshyn 1989; Pollatsek and Rayner 1991).

Previous attempts to demonstrate integration across saccadic eye movements fall into two classes. Some have tried to demonstrate the addition of images from successive fixations (Irwin et al 1983; O’Regan and Levy-Schoen 1983; Rayner and Pollatsek 1983; Irwin et al 1988). For example, Irwin and colleagues used a task requiring subjects to find the missing dot in an array of twenty-five dots when half the dots were presented before a saccade, and the rest after the saccade. This task requires some sort of point-by-point integration of the two images, and is found to be impossible over a range of conditions. Several reports of positive results in this and in similar tasks were shown to be artifacts of display persistence (Bridgeman and Mayer 1983; Rayner and Pollatsek 1983; Irwin et al 1988; Irwin et al 1990). Others have looked for some sort of facilitation of perception (eg of a word or picture) by presenting a stimulus in a peripheral location before the saccade. While priming
effects are found, they do not appear to be linked tightly to spatial location nor to the particular visual features of the peripheral preview stimulus (McConkie and Zola 1979; Rayner et al 1980; McConkie et al 1982; Pollatsek et al 1984; Pollatsek et al 1990).

For example, changing the case of the letters in a word has no effect on naming times or on reading. These studies clearly demonstrate that there is no low-level, photograph-like addition of images across saccades. This raises the question: what is the nature of the visual information which is retained from prior views? The current conception of short-term visual memory is that there is a rapidly decaying visual analogue representation followed by a nonvisual, postcategorical representation that contains only poorly specified and rapidly decaying information about spatial location (Mewhort et al 1981; Dixon 1986; Irwin and Yeomans 1986). The failure to find low-level integration of images has been taken to indicate that information derived from separate fixations is combined only in this later, postcategorical memory (O’Regan and Levy-Schoen 1983; Irwin et al 1988; Pollatsek and Rayner 1991).

Indeed, O’Regan and Levy-Schoen (1983) have suggested that visual information from previous views, including spatial position information, is retained only to the extent that it is semantically encoded (e.g., “in front of the red table”).

However, it is not necessarily for integration to occur in a purely symbolic or purely pictorial way, and something between these extremes seems more likely. Many of the previous investigations, which grew out of studies of reading, have looked for an enhancement of visual processing of the current foveal image by the peripheral stimulus that preceded a saccade to that spatial location. However, the aspects of our perceptual experience which led to the assumption that an integrative process must exist do not require this sort of enhancement, but rather the building of a representation that combines information from prior foveal views to produce a more complete scene description than is available from a single view. For example, actions like orienting towards an object not currently in view, crossing a busy street, or judging the shape of a large building require integration in this sense. It seems unlikely that these tasks would be subserved by integration which occurred only in semantic memory. However, we know very little about the precision with which we can perform such tasks. Perception of form, for example, is generally thought to be based on the retinal image information available in a single fixation. In this experiment we therefore asked how well a simple shape judgement can be made when only part of the shape information is available in one fixation. If such a task can be performed relatively easily, then the spatial information retained from previous views must be quite precise. In addition, good performance would point towards an integrative process that operates on precategorical information, since in this task the form emerges only in the integrated representation. Some ability to do this task would be expected on the basis of the fact that subjects are able to make judgements about the visual direction of a point of light in the dark and can detect changes in its position which occur during a saccadic eye movement. However, it is not clear from these experiments how well subjects would manage to extract form information, since the magnitude of displacement required for detection of a change is somewhat variable, and thresholds as large as 20%–30% of the saccade magnitude are often reported (see e.g., Bridgeman et al 1975; Bridgeman and Stark 1979; Li and Matin 1990).

2 Methods
The particular task we chose was one of angle judgement. Subjects were presented with three points in succession, which defined a triangle (see figure 1). Subjects tried to judge whether the top angle in the triangle defined by the time-integrated view was acute or obtuse. We measured the precision with which subjects could make these judgements both when the eyes were stationary and when they moved between the
dot presentations. In the latter condition the location of the dots must be retained in a frame independent of eye position, if the task is to be performed accurately. Ability to perform this task therefore indicates how precisely subjects can retain spatial position information and whether they can integrate this information into a representation which can support angle judgements. There were three conditions in all. In the first condition subjects fixated the fixed reference point while the three points were flashed in sequence. This task simply requires integration over time and provides a baseline by which to judge performance in the other conditions. In the other two conditions subjects were asked to move their eyes to the approximate anticipated location of each dot, so that all the dots fell in roughly the same retinal region, close to the fovea. In the second condition the fixation point remained present, allowing alignment of successive frames relative to this point. In the third there was no fixation point, so eye position information was required to compute the location of the dots. This allowed us to compare the roles of visual cues and of eye position cues in relating successive views.

Each dot was presented for 100 ms, and the interval between the dots was varied from simultaneous presentation to 800 ms. Because of the brief duration, no eye movements were possible for simultaneous presentation. The largest side of the triangle subtended approximately 15 deg of visual angle. Data for each of the three

![Diagram](image)

**Figure 1.** (a) The stimulus, integrated over time. The points (filled circles) were presented in order from left to right, and the top angle of the triangle was judged acute or obtuse. When the fixed reference point was present it was located as shown by the asterisk. A typical eye movement trace during the eye movement conditions is shown. The squares show the location of the eye when the spot was presented. (b) Sequence of events in a trial.
conditions were collected in blocks of 160 trials, with the interstimulus interval held constant. Within the block, subjects made acute/obtuse judgements for angles ranging from 79° to 101° in steps of 2°. The different angles were generated by displacement of all three dots. The triangles were randomly rotated by ±15° and translated by ±1 deg from trial to trial to decorrelate the positions of the individual dots from the angle to be judged. (For one subject, MH, the rotations were ±6°, with no translations.) Psychometric functions were generated by using the method of constant stimuli, and thresholds for discriminating the angle from 90° were estimated by using a criterion of 75% correct. No systematic biases were observed for acute/obtuse responses. Data were collected in three sessions, with a total of 48 or more trials at each angle.

No bite bar or chinrest was used in these experiments so subjects' heads were not fixed except when eye position was recorded, as described below. Point stimuli were displayed on an oscilloscope with P31 phosphor, and were viewed binocularly, at a distance of 0.8 m, through a red cutoff filter. This roughly equated rod and cone sensitivity and reduced the visibility of phosphor persistence. Stimuli were presented at approximately 1 log unit above threshold. In a separate experiment screen persistence was measured by opening a shutter at varying time intervals following the offset of the stimulus. Detectability measured by d' fell to zero by 7 ms. All experiments were run in the dark. Feedback was given after each trial.

In a separate set of trials, eye position was measured with a Dual Purkinje Image Eyetracker (Stanford Research Institute). Eye position was calibrated at the beginning of each block of 40 trials, with the use of a twenty-five point grid spanning 10 deg of visual angle and a linear estimate of x and y gains. Calibrations were accurate within 15 min arc.

3 Results
The results are shown in figure 2 which plots thresholds for discriminating the angle from 90° as a function of interstimulus interval. Subjects performed the task with little difficulty. With simultaneous presentation of the three dots, thresholds of about 2°–4° were obtained. Overall, there was surprisingly little effect of interposing a dark interval in between presentation of the dots. As the interstimulus interval increases to 800 ms, thresholds increase by only a factor of about 2 over all conditions. For all the conditions, spatial memory therefore decays quite slowly. Performance was better when there was a reference point. These conditions are indicated by the filled circles and by the open triangles. It is of interest here that there appears to be little systematic difference between these two conditions. There is some indication of a systematic elevation of thresholds in the eye movement condition for subjects KK and JL. A small elevation of threshold might be expected from the fact that subjects may occasionally fail to reach the anticipated location of the dot during the interstimulus interval, so that the eye will be in motion during the presentation of a dot. Data collected on both JL and AI for smaller rotations (±6° cf MH), where the location of the points is more predictable, showed no difference between the two conditions. On the whole, when a reference point is visible it makes little difference whether or not saccades are made. Thus there does not appear to be an added cost to integrating over saccades beyond a slow decay with time. Subjects are able to make fairly fine spatial form judgements over quite long time intervals, independent of eye position, which indicates that they are using an underlying spatial representation that allows them to integrate precise spatial information in a world-based store.

(1) No improvement with practice was observed in data collected over several months. Two of the observers were experienced in psychophysical judgements. The other two were inexperienced.
In the third condition, where the reference point is removed, subjects must use eye-position information to compute the location of the dots. This condition is shown by the open squares in figure 2. In this condition accurate judgements were still possible, although performance declined for all subjects. There was considerable variability between subjects in this condition. Subject MH performed almost as well as with a reference point, whereas subject JL could not perform the task to 75% criterion level even at 11° deviations from a right angle, which were the largest ones we explored. Although some subjects can, if necessary, use eye position information to integrate in a world-based frame, the system does distinctly better when given a fixed reference point.

It is of interest to know how these angle thresholds translate into the spatial precision with which each point is known. It is not possible to do this precisely, but we can get some idea of how they are related. If all the errors were in the estimated position of only one of the points (that is, the other two were known exactly), and if the direction of the error was in the direction from the true location such that it maximized the change in angle, then a threshold of 2° in this experiment would correspond to a positional error of only 14 min arc. This value increases in proportion to the threshold, so that a 6° threshold, for example, would correspond to a 42 min arc displacement in one of the points. In reality, of course, this error will be dependent on the direction of error in each of the points on a particular trial, so this value corresponds to a lower limit for the total error in the three points. Also, the error will in fact be distributed roughly equally between the three points. Given the retinal eccentricities and the magnitude of the eye movements involved, this performance is remarkably good.

![Figure 2. Thresholds for discriminating an angle from 90° as a function of interstimulus interval. Data for four subjects are shown for each of the three conditions. The filled circles show thresholds for the condition where subjects fixated a continuously present point. The triangles indicate the condition where the fixed reference point remained visible, but subjects moved their eyes before each point was presented, to its anticipated location. The open squares show performance when similar eye movements were made, but the reference point was removed. (No eye movements were attempted during simultaneous presentations in this condition.) The arrows for subject JL indicate failure to reach 75% criterion at the largest deviations from a right angle that were used (±11 deg; sim indicates simultaneous presentation of the dots. Standard errors are ±1 SEM between sessions.](image-url)
It is clearly important to know whether in fact subjects were moving their eyes as instructed. Eye position was monitored in a separate set of trials for each of the subjects. Subjects performed as expected with little difficulty. Figure 1 shows a typical eye-movement trace during a trial. The squares show the location of the eye when the spot was presented. For the conditions requiring eye movements, the average change in direction of gaze between the presentations of the first two dots was 3.2°. Between the last two dots it was 9.2°. (This average is computed from the four subjects, with 40 trials per condition.) Thus the judgement could not be made in a retinal frame. During fixation trials the average change in eye position between presentations of the dots was 7 min arc. Although we cannot guarantee that this reflects performance in the trials in which the data in figure 2 were collected, major deviations would be surprising, as the appearance of the dots provided subjects with feedback as to whether they were moving their eyes approximately to the correct locations.

A second important consideration is whether subjects were basing their judgements on all three dots. Despite the random rotations of the triangles there is inevitably some correlation between the angle to be judged and the positions of individual dots. To investigate whether the judgements were based on all three dots, control trials were run where only the first two dots were presented. The slope of the line defined by these two dots is partly correlated with the size of the angle to be judged, so above-chance performance would be expected if judgements are based on this information alone. This still requires some integration across eye position, but of a simpler kind. The 'threshold' for 75% correct performance on control trials was calculated, and these data are presented in table 1, along with thresholds for experimental trials collected in the same sessions. (In each block of trials, half were experimental and half were control. Each threshold is calculated on the basis of 32 trials per point at each of ten angles.) Although subjects' performance with two dots was above chance, as expected, it was substantially better with three dots. For AI and KK, performance with two dots was so poor that a threshold could not be computed in many instances. Performance on control trials was somewhat better for MH, for whom the rotations were only ±6°, but it was still comparatively poor. (Control trials were also run for subject JL, but only at the smaller rotations. Like MH, the performance of JL on experimental trials was substantially better.)

![Table 1](image-url) The 'threshold' for 75% correct performance for experimental (3 dots) and control (2 dots) trials.

<table>
<thead>
<tr>
<th>Trial conditions</th>
<th>MH 3-dots</th>
<th>MH 2-dots</th>
<th>AI 3-dots</th>
<th>AI 2-dots</th>
<th>KK 3-dots</th>
<th>KK 2-dots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixation point present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>simultaneous</td>
<td>1.3</td>
<td>3.1</td>
<td>4.2</td>
<td>4.3</td>
<td>1.5</td>
<td>8.8</td>
</tr>
<tr>
<td>400 ms ISI* with fixation</td>
<td>3.3</td>
<td>5.3</td>
<td>8.6 **</td>
<td>4.6 **</td>
<td></td>
<td></td>
</tr>
<tr>
<td>400 ms ISI moving eyes</td>
<td>2.4</td>
<td>4.8</td>
<td>4.7</td>
<td>9.0</td>
<td>5.4</td>
<td>**</td>
</tr>
<tr>
<td>No fixation point simultaneous</td>
<td>2.5</td>
<td>6.8</td>
<td>4.5</td>
<td>7.0</td>
<td>2.1 **</td>
<td></td>
</tr>
<tr>
<td>400 ms ISI moving eyes</td>
<td>4.3</td>
<td>7.4</td>
<td>7.4 **</td>
<td>4.6 **</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* ISI = interstimulus interval.
** Was not possible to obtain threshold.
4 Discussion
This experiment demonstrates that it is possible to make accurate judgements of form from information garnered from separate fixations. Both the precision of the spatial position information and the relatively slow decay of this information are unlike the properties of the theories of memory representation suggested by a number of previous investigators (Mewhort et al 1981; Dixon 1986; Irwin and Yeomans 1986). Instead, the results point towards a map-like, or spatially ordered representation precise enough to support geometrical judgements. This conclusion is supported by the findings of Matin et al (1981) who demonstrated good vernier discrimination for flashes separated in time, and also by the recent experiments of Palmer and Ames (1989) who found that precise memory of line length and of simple geometrical shapes was maintained across eye movements. The slow decay of the memory revealed in the present experiment makes sense given that integration may often be required over a sequence of saccades. While it is hard to draw definitive conclusions from a single experiment of this sort, the perceptual memory(2) revealed here is not well described as semantic or propositional as suggested most explicitly by O'Regan and Levy-Schoen. Nor is it consistent with the suggestion of Pollatsek and Rayner, who propose that the integration occurs at a postcategorical level where spatial location is not directly represented (Pollatsek and Rayner 1991). On the basis of these data it seems more likely that the integration is performed at some intermediate level of visual representation. One explicit statement of the idea of a spatially ordered intermediate visual buffer, which seems more appropriate to describe the present results, has been made by Feldman (1985), who proposed that features are assembled in a head-centered frame called a stable feature frame. This differs from the image addition idea in that the stable feature frame contains intermediate level visual features as a set of parameters at a given spatial location. The formation of the stable feature frame precedes object recognition. Since the form information in this experiment does not emerge until after integration, our results are consistent with integration at a precategorical level, like the stable feature frame.

The sort of visual memory revealed here, which integrates information from different locations in an extended spatial representation, is consistent with classic experiments which show that objects viewed through an aperture can be identified from a sequence of partial views. When subjects track an aperture moving across a figure, the form can be perceived (Rock and Halper 1969; Rock and Sigman 1973). This demonstrates integration in a spatial frame for the pursuit system, similar to our result with saccades. This suggests that fundamentally similar processes operate both during pursuit and during saccadic eye movements to construct a world-centered representation, despite the fact that different neural mechanisms control the two types of movement and that they generate different perceptual effects (Matin 1986). It is hard to compare the precision of the form judgements in the two types of experiment, although substantial perceptual distortions commonly occur with pursuit movements. These distortions are correlated with tracking errors (Mack 1986). In another type of aperture-viewing experiment the aperture is fixed and the object moves behind it (Parks 1965; Hochberg 1968). Perception of the form in this situation is due partly to the painting of the image on the retina by pursuit eye-movements (eg Anstis and Atkinson 1967; Haber and Nathanson 1968; Morgan et al 1982). When the eye is stationary, form judgements are more difficult but are clearly still possible (Hochberg 1968; Morgan et al 1982; Park and Kosslyn 1989).

(2) It is hard to evaluate whether this is a perceptual or a memory representation. The three points are clearly not visible simultaneously, although one has a clear 'sense' of a triangle.
Our results also have implications for the mechanisms of visual stability. Most treatments of how we make the transformation from retinotopic to world-centered coordinates implicitly assume the existence of some integrative process and deal with the issue of what information is used to align successive views appropriately. Much attention has focussed on the relative importance of inflow versus outflow eye-position signals, and it is clear that eye-position information is used to assign a location to a point of light in the dark (Matin 1986). The role of eye-position cues as the visual environment becomes more complex is less clear. A number of authors have argued that an actual cancellation by an eye-position signal is unnecessary in normal scenes because visual relationships are preserved during saccades (Gibson 1966; MacKay 1973). The importance of visual reference frames is also supported by recent experiments which indicate that in normal visual scenes, visual information dominates over eye-position signals in judgements of visual direction, when visual and efference copy signals are in conflict (Matin et al 1982; Stark and Bridgeman 1983; Bridgeman and Fishman 1985). Visual stability requires that the appropriate relation between frames is achieved as well as preserving the relation between the scene and the observer, which is reflected in visual direction judgements. When the relation between successive frames is not consistent with the corollary discharge signal, as when the eyes are partially paralyzed, transient instabilities result (Stevens et al 1976; Stark and Bridgeman 1983). This indicates that the eye-movement signal is actively operating even in a normal lighted environment to relate successive views. However, the results of the present experiment clearly indicate a visual matching strategy in relating successive views, since subjects do better when successive frames can be aligned with respect to the fixed reference point. The variability between subjects in the absence of a reference point indicates that there is a wide variability in the fidelity of the eye-position signal or in the ability to use it. This subject variability essentially disappears in the reference-point condition. This, together with the superior performance in this condition, suggests that the visual reference information is used to refine eye-position information, and that the spatial locations of objects in successive frames are assigned with respect to objects common to the frames.

Acknowledgements. This work was supported by NIH grants EY05729 to Mary Hayhoe, and EY01319 to the Center for Visual Science, and an NSF predoctoral fellowship to Joel Lachter.

References
Bridgeman B, Fishman R, 1985 “Dissociation of corollary discharge from gaze direction does not induce a straight-ahead shift” Perception & Psychophysics 37 523 – 528
Dixon P, 1986 “Attention and interference in the perception of brief visual displays” Journal of Experimental Psychology 12 133 – 148
Feldman J A, 1985 “Four frames suffice: A provisional model of vision and space” Behavioral and Brain Sciences 8 265 – 289
Haber R N, Nathanson L S, 1968 “Post-retinal storage? Some further observations on Parks’ camel as seen through the eye of a needle” Perception & Psychophysics 3 349 – 355

Irwin D E, Yeomans J M, 1986 “Sensory registration and informational persistence” Journal of Experimental Psychology 12 343 – 360


Jonides J, Irwin D E, Yantis S, 1982 “Integrating visual information from successive fixations” Science 215 192 – 194


Parks T E, 1965 “Post-retinal visual storage” American Journal of Psychology 78 145 – 147


Rock I, Sigman E, 1973 "Intelligence factors in perception of form through a moving slit" Perception 2 357–369

