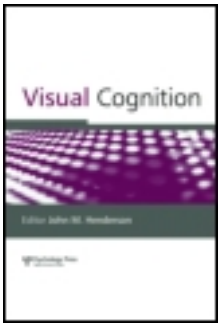

Visual Cognition
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pvis20

Minimal use of working memory in a scene comparison task

Daniel Gajewski [a] & John Henderson [a]
[a] Michigan State University, East Lansing, MI, USA
Version of record first published: 01 Oct 2010.

To cite this article: Daniel Gajewski & John Henderson (2005): Minimal use of working memory in a scene comparison task, Visual Cognition, 12:6, 979-1002

To link to this article: http://dx.doi.org/10.1080/13506280444000616

VISUAL COGNITION, 2005, 12 (6), 979–1002

Psychology Press
Taylor & Francis Group

Minimal use of working memory in a scene comparison task

Daniel A. Gajewski and John M. Henderson
Michigan State University, East Lansing, MI, USA

Eye movement behaviour in hand-eye tasks suggests a preference for a ''just in time'' processing strategy that minimizes the use of working memory. In the present study, a scene comparison task was introduced to determine whether the preference holds when the task is primarily visual and when more complex naturalistic scenes are used as stimuli. In two experiments, participants made same or different judgements in response to simultaneously presented pairs of scenes that were identical or differed by one object. The number of fixations per scene glance and the number of fixations intervening between glances to corresponding objects suggest that frequently one object at a time is encoded and maintained in visual working memory. The same pattern of results was observed in a third experiment using word and object arrays. Overall, the results suggest a strong general bias toward minimal use of visual working memory in complex visual tasks.

Because the highest level of acuity is limited to a rather small portion of the visual field, a region roughly the size of one's thumb nail when viewed at arm's length, saccadic eye movements are needed to direct the eyes toward points of interest in a scene. The fact that our visual experience seems to include more detail than can be provided by a single fixation suggests a role for memory in scene perception; however, the exact role played by memory in scene perception is a topic of debate. Some theorists have proposed that the online visual representation of a scene includes a highly detailed representation of the object

Please address all correspondence to: Daniel A. Gajewski, Department of Psychology, Michigan State University, East Lansing, MI 48824-1117, USA. Email: dan@eyelab.msu.edu

This research was supported by a fellowship from the MSU graduate school on behalf of the NSF IGERT program in Cognitive Science to Daniel A. Gajewski. This work was also supported by the National Science Foundation (BCS-0094433 and ECS-9873531) and the Army Research Office (DAAD19-00-1-0519). Aspects of this data were presented at the 43rd annual meeting of the Psychonomic Society, Kansas City, MO, and the third annual meeting of the Vision Sciences Society, Sarasota, FL. The opinions expressed in this paper are those of the authors and do not necessarily represent the views of the Department of the Army or any other governmental organization. Reference to or citations of trade or corporate names does not constitute explicit or implied endorsement of those entities or their products by the author or the Department of the Army.

http://www.tandf.co.uk/journals/pp/13506285.html DOI:10.1080/13506280444000616

of the current fixation, abstract visual memory representations of the three–four most recently attended items, as well as a representation of the structure or layout of the scene (e.g., Hollingworth & Henderson, 2002; Irwin & Andrews, 1996; Irwin & Zelinsky, 2002; but see Rensink, 2000). Others have suggested that the online visual representation is limited to information that is relevant to the ongoing task (Ballard, Hayhoe, Pook, & Rao, 1997; Hayhoe, Bensinger, & Ballard, 1998). According to a framework called *active vision* (Aloimonos, Weiss, & Bandopadhay, 1987), the purpose of vision is not to build a general representation of the environment. Rather, visual information is believed to be extracted from the environment in the service of specific behavioural goals. Importantly, this framework favours ''just in time'' processing strategies where the acquisition of visual information is delayed until the point in time when it is needed.

A body of work done by proponents of this framework supports a preference to minimize the use of visual working memory. For example, Ballard, Hayhoe, and colleagues (Ballard, Hayhoe, & Pelz, 1995; Hayhoe et al., 1998) have explored this issue using a block-copying task. Participants were given a display that was divided into three regions—a model region, a source region, and a workspace region. The model region contained a pattern of coloured blocks to be copied and the source region provided a set of coloured blocks to be moved into the third area, the workspace region. Inferences about the information retained at any moment during the task were made on the basis of the eye movement patterns. The most frequently observed eye movement pattern began with a saccade to the model, then to the source area, and then to the model again before directing gaze to the workspace in order to guide the placement of the selected block into the appropriate space in the emerging replica. The fact that the same block was fixated twice at different points during the operation suggested that different information was acquired during each glance. That is, the first glance, prior to fixating the source area, appears to have been used to determine the colour of the block that was needed, and the second glance, prior to positioning the selected block into the workspace, appears to have been used to determine the spatial position of the selected block. The authors considered this pattern memoryless because it suggested a preference to delay the acquisition of information until the point in time when it was needed, thereby minimizing the use of visual working memory. Importantly, this preference could not be attributed to capacity limitations because when the distance between the model and the workspace was increased, a manipulation designed to increase the cost associated with acquiring information, the frequency of the memoryless pattern decreased. The preferred strategy, therefore, appears to reflect issues of cognitive economy.

A question that follows from this work is the extent to which these findings will generalize to a task that is primarily visual. The work of Ballard, Hayhoe, and colleagues (Ballard et al., 1995; Hayhoe et al., 1998) arose out of a

computational framework that is at least partially concerned with the functional use of visual information to guide the programming of motoric interactions with the world. As a result, the focus has been on tasks that involve the coordinated use of the hand and the eyes. The suggestion is that the algorithmic complexity of hand–eye tasks can be reduced by allowing the point of fixation to serve as the centre of an external reference frame and by representing in memory only the information needed at each step in the operation (Ballard et al., 1997). The crux of this theoretical perspective is that eye movements play a role in variable binding because the fixation can be used as a deictic device that points to objects in the world that participate in the execution of behavioural programs. Thus, any task that requires the organization of sequential operations, even tasks that do not require the coordination of their visual and motor components, should similarly reveal preferences to minimize the use of memory. However, the possibility remains that the preferred strategy will differ when the visual task does not demand this kind of interaction with the environment. The bias towards minimal short-term representation in the block-copying task might be driven by its motor component.

In addition to testing the generality of the minimal memory preference in a task that is primarily visual, our interest is in the use of visual working memory in the perception of scenes. Therefore, we tested the minimal memory hypotheses using stimuli that approach the complexity of natural viewing situations. Participants were asked to make same-or-different judgements about side-by-side copies of rendered scenes. The pairs of scenes were either identical or differed by one object. Two primary measures were used to make inferences about the use of memory in this task. The first measure was intended to provide an indication of how many objects are looked at in one scene prior to directing the eyes to the other scene. If the preferred strategy is to minimize the use of visual working memory, gaze shifts from one scene to the other should be frequent, with the number of objects fixated during each glance of a scene close to one. The second measure was intended to provide an indication of the number of objects fixated between the first glance to an object in one scene and the first glance to the corresponding object in the other scene. If the minimal memory strategy is preferred, participants should most often direct the eyes immediately from one object in one scene to the corresponding object in the other scene, holding just the one object in memory in order to make the comparison.

The appropriate use of the number of objects fixated as an index of memory use, however, depends upon two assumptions: (1) Objects are the level of analysis in scene perception; and (2) objects are the unit of capacity of visual working memory. The first assumption could arguably be satisfied by the demands of the task given that it is at the level of the object that two scenes may differ; however, an object level of analysis in scene viewing is also supported by empirical work. Nelson and Loftus (1980), for example, showed that the likelihood that an object will be encoded sufficiently to support discrimination in a

recognition task declines sharply with increased distance from the centre of the nearest fixation. In addition, changes made to objects in change detection tasks are reliably detected only if the object is fixated before the change (Hollingworth & Henderson, 2002). The second assumption is satisfied by several studies suggesting that integrated objects are the unit of capacity and that between three and four objects can be held in visual working memory (Irwin, 1992; Irwin & Andrews, 1996; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001; but see Wheeler & Treisman, 2002). For example, Luck and Vogel (1997; Vogel et al., 2001) tested memory for briefly presented arrays of bars that varied in colour and orientation. Performance declined with increasing array size but was unaffected by the need to remember only the colour, only the orientation, or the conjunction of both features. This result was also found when the items in the array were conjunctions of four features. Thus, it was the number of objects that limited performance, not the number of features. These studies suggest that if memory is used maximally in the scene comparison task, at least three–four objects will be fixated in one scene prior to directing gaze to the other scene.

The interest in the number of objects fixated, however, does not come without complications when naturalistic scenes are used as stimuli. Ideally, one would like to take the fixation data for each participant and count the number of objects fixated during a given scene glance. In order to count the number of objects fixated, one would have to assign each fixation to an object. We were attracted to using scenes because of the complexity offered by them, but the things that make scenes complex, such as partial occlusions and the clustering of objects, also make it difficult to assign fixations to objects. In addition, a fair amount of subjectivity would be introduced by this approach. The issue would be more easily resolved if object gazes were comprised of only one fixation. One could then simply count the number of fixations per scene glance. However, scene-viewing studies suggest that this is not the case. Experiments run in this laboratory typically find between 1.5 and 2 fixations per object on average. Henderson, Weeks, and Hollingworth (1999), for example, report an average of 1.7 fixations per gaze on target objects. Our solution was to take the number of fixations during the first glance to strategically placed critical objects as an index of the number of fixations that generally contribute to an object glance in this task. Using this estimate, predictions based on the number of objects fixated per scene glance can be expressed in terms of numbers of fixations per scene glance.

In this study there were two measures of primary interest. The first dependent measure was the mean number of fixations made in each scene before gaze is directed to the other scene. In order to test competing hypotheses about the use of memory in this task, models were generated for both minimal and maximal use of memory. Figure 1A illustrates the type of viewing pattern one would expect if the use of memory were minimized. The two squares on the left represent a pair of scenes comprised of objects, and the grid on the right

represents the number of objects fixated during successive glances to the scenes. If memory is minimized, participants should most frequently begin by fixating one object in one of the scenes and then immediately fixate the corresponding object in the other scene to make the comparison. Then, if the number of shifts between scenes is also minimized, participants should select the next item from the same scene. If this pattern continues, the expected number of objects per scene glance is simply the number of objects fixated divided by the number of scene glances. In this case its 1.75. Figure 1B illustrates the expected viewing pattern given a preference to make full use of memory. Given the evidence for a capacity of three–four items, predictions for maximal use of memory were made conservatively on the basis of three objects. In this scenario, participants should fixate three objects in the first scene before directing their eyes to the corresponding objects in the other scene, and three new objects should be fixated before returning to the original scene. The expected number of objects per scene glance here is 5.25.

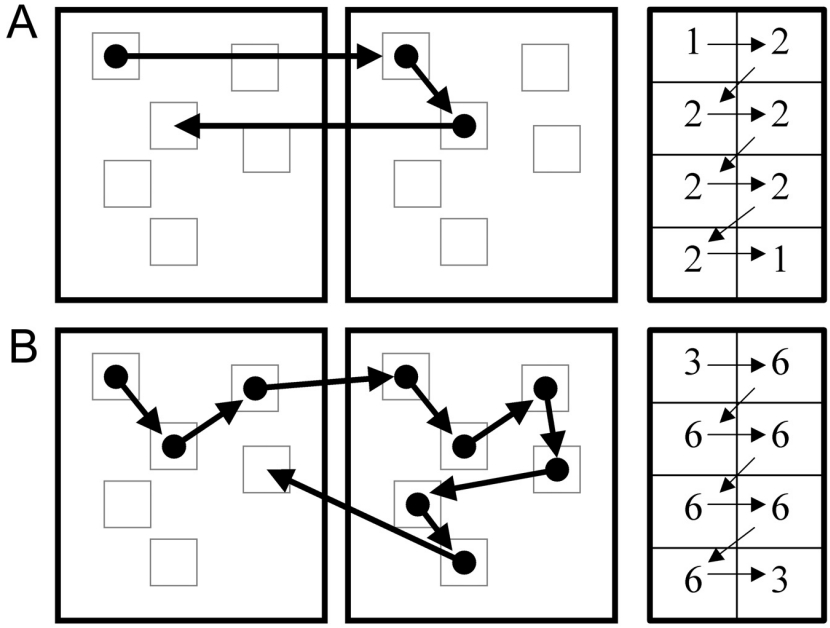

Figure 1. (A) The two large squares on the left represent a pair of scenes comprised of objects. The expected viewing pattern given minimal use of memory is illustrated by the flow of the arrows. The right side of the figure represents a model of the number of objects fixated during each successive scene glance. In this eight-glance model, each scene is entered four times and the expected number of objects fixated per scene is 1.75. (B) An eight-glance model was similarly generated based on the full use of memory. Here, the expected number of objects fixated is 5.25.

Note that the expected values generated by these models depend on the number of times each scene is viewed. In these eight-glance models, each scene is entered four times. As the number of glances to each scene increases, the expected number of objects fixated per scene approaches two and six for minimal and maximal memory, respectively. Therefore, the average number of scene glances was determined from the data in order to appropriately formulate predictions. Finally, in order to compare the predicted values to those experimentally derived, the expected number of objects fixated per scene glance was multiplied by the mean number of fixations per object.

The second measure of interest is the number of fixations that intervened between the first fixation on the first critical object fixated and the first fixation on the corresponding object in the other scene. This analysis provides an indication of how frequently participants went from one object directly to the corresponding object in the other scene. If the use of working memory is minimized, subjects should most frequently look directly from one object to the corresponding object in the other scene. If the number of fixations per object is between one and two, the most frequent number of intervening fixations should be between zero and one. If the use of memory is maximized, however, subjects should most frequently fixate at least two objects prior to fixating the corresponding object in the other scene. The most frequent number of intervening fixations, therefore, should be between two and three.

EXPERIMENT 1

In Experiment 1, participants made same-or-different judgements in response to 32 pairs of scenes. In three conditions the pairs of scenes differed in terms of one object in one of the scenes: An object was removed (deletion), replaced with an item from a different basic-level conceptual category (different-type), or replaced with a different exemplar from the same category (different-token). In the fourth condition, both scenes were identical (same). All four conditions occurred with equal frequency. Judgements of same versus different were indicated with a button press. Each image remained visible for as long as was needed to make the decision; however, participants were instructed to press the appropriate button as soon as the decision was made.

Method

Participants. Twelve Michigan State University undergraduate students participated in the experiment for course credit. All participants had normal vision and were naive with respect to the hypotheses under investigation.

Stimuli. Thirty-two scene images were computer-rendered from three-dimensional (3-D) wire-frame models using a commercial rendering program. Wire-frame models were acquired commercially, donated by 3-D graphic artists,

or developed inhouse. Each model depicted a typical, human-scaled environment. Base scenes were rendered from these models. To create the different-type, different-token, and deletion conditions, the target objects were replaced or removed in the models, and the scenes were rerendered. All scene images subtended 7.8° × 5.9° visual angle at a viewing distance of 1.13 m. Each image was comprised of two versions of scenes side-by-side along the middle of the display with a small gap in between. The remainder of the display was a neutral grey. Figure 2 shows a sample stimulus image. Target objects subtended 1.14° on average along the longest axis and 0.71° along the shortest axis. The objects used for the different-type and different-token conditions were chosen to be similar in size to the no-change target object in the corresponding scene. Target objects were placed so that they were offset from the centres of the scenes. They were placed in uncluttered regions of the scenes so that fixations on them could be easily identified. Nothing about the target objects themselves identified them as different from the other objects in the scenes.

Apparatus. Eye movements were monitored using a Generation 5.5 Stanford Research Institute Dual Purkinje Image Eyetracker (Crane, 1994; Crane & Steele, 1985). The eyetracker has a resolution of 1' of arc and a linear

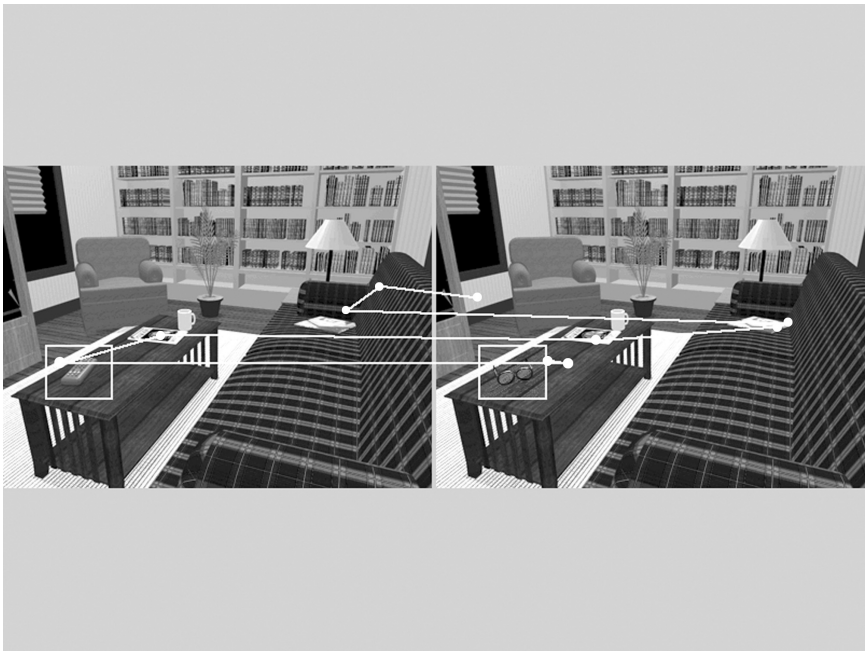

Figure 2. Sample scene illustrating regions used for critical objects and a scan pattern used by one of the participants on one of the trials. Full-colour versions were used in the experiment.

output over the range of the visual display used. A bite-bar and a forehead rest were used to maintain the participant's viewing position and distance. The position of the right eye was tracked, though viewing was binocular. Signals were sampled from the eyetracker using the polling mode of the Data Translations DT2802 analogue-to-digital converter, producing a sampling rate slightly faster than 1000 Hz.

Stimuli were displayed at a resolution of 800 × 600 pixels × 256 colours on an NEC Multisync P750 monitor driven by a Hercules Dynamite 128/Video graphics card. The screen refresh rate was 143 Hz. The room was dimly illuminated by an indirect, low-intensity light source.

Button-presses were collected with a button panel connected to a dedicated input–output (I/O) card. The eyetracker, display monitor, and I/O card were interfaced with a 90 MHz, Pentium-based microcomputer. The computer controlled the experiment and maintained a complete record of eye position and time values, and button press events and times, over the course of each trial.

Procedure. Upon arriving for the experimental session, participants were given a written description of the experiment along with a set of instructions. The description informed participants that their eye movements would be monitored while they viewed images of naturalistic scenes on a computer monitor. Participants were instructed to view pairs of scenes and to judge as quickly and accurately as possible whether they were the same or different. They were explicitly told that differences between scenes would be due to differences between one object in one scene and the corresponding object in the other scene and they were given examples of the types of differences that could be found. Participants were instructed to press the right button as soon as such a difference was detected or the left button as soon as they were satisfied that the two scenes were the same. Following review of the instructions, the experimenter calibrated the eyetracker. Calibration was considered accurate if the computer's estimate of the current fixation position was within +/— 5 min arc of each marker. The participant then completed the experimental session. Calibration was checked every three–four trials, and the eyetracker was recalibrated when necessary.

A trial consisted of the following events. First, a fixation screen was shown. When the participant fixated a central box in this screen (as indicated by a computer-generated display of its estimated fixation position), the experimenter initiated the trial. Each image remained visible for as long as was needed to make the decision; however, subjects were instructed to perform the task both quickly and accurately. Each participant viewed 32 scenes. There were four conditions to the experiment: Three different conditions (different-type, different-token, and deletion) and one same condition. Scenes were assigned to conditions via a Latin-square design so that each scene appeared in each condition an equal number of times across participants. The order of scene presentation (and hence the order of condition presentation) was determined

randomly for each participant within each session. Each session lasted approximately 35 min.

Results and discussion

Single factor within-subjects analyses of variance (ANOVAs) were performed to evaluate the general performance measures of accuracy and reaction time prior to the analyses of primary interest. Accuracy varied marginally across the four conditions, $F(3, 33) = 2.54$, $MSE = 0.02$, $p = .07$, and was highest in the same condition (mean = 0.95). Overall performance was fairly high: The mean proportion correct was .88.

The analysis of reaction times was based on correct responses only. Reaction times varied across the four conditions, $F(3, 33) = 12.16$, $MSE = 20.57$, $p < .001$; however, there were no differences in reaction time when only the three different conditions were considered, $F < 1$. Viewing times were longer in the same condition (mean = 13.9 s) than in the three different conditions (mean = 4.8 s), $F(1, 11) = 13.20$, $MSE = 37.68$, $p < .01$. This difference is expected given that subjects would have to perform an exhaustive search in order to make a decision in the same condition but not in the others.

To generate and test predictions about the use of memory in this task, a series of analyses were performed using the eye-tracking data. First, the number of scene glances was determined to generate minimal and maximal memory models for the expected numbers of objects fixated per scene. Second, an estimate of the number of fixations that generally contribute to an object gaze was determined on the basis of the target objects and this estimate was multiplied by the expected number of objects fixated per scene to express the predictions in terms of the dependent measure—the number of fixations per scene. The observed numbers of fixations per scene were then tested against each of the predicted values. To compare these values, object- and scene-sized regions were defined for each display. Fixation data was recorded in a pixel-based coordinate system and tabulated as a function of occurrence in these regions. The analyses were based on correct trials only. Eye-tracking data from all 12 subjects went into the analyses; however, 21 trials were eliminated (approximately 5%) because of track loss.

Since the predicted number of objects fixated per scene depended on the total number of scene glances, the number of times each scene was fixated was tabulated for each of the two scene regions for every trial. The values for each of the pairs of regions were summed and the averages were entered into the analysis for each subject and condition. The total number of scene glances varied by condition, $F(3, 33) = 12.46$, $MSE = 24.93$, $p < .001$; however, there was no effect when only the three different conditions were considered, $F < 1$. There were more scene glances made in the same condition (16.5) than in the three different conditions (mean = 6.4), $F(1, 11) = 13.69$, $MSE = 44.96$, $p < .01$. Again, this

TABLE 1
Summary of the observed and predicted values in Experiment 1

	Same	*Different*
Expected objects per scene		
Minimum	1.9	1.7
Maximum	5.6	5.0
Expected fixations per scene		
Minimum	2.3	2.0
Maximum	6.8	6.0
Observed fixations (S.E.) per scene	2.73 (.11)	2.46 (.10)

would be expected given exhaustive searching in the same condition. The expected numbers of objects fixated per scene were therefore formulated for minimal and maximal memory based on 17-glance models for the same condition and 6-glance models for the mean of the different conditions (see Figure 1). The resulting values are listed in the top section of Table 1.

An estimate of the number of fixations that are generally made on an object was generated by counting the number of fixations that occurred during the first glance within regions defined for each of the target objects. The object-sized regions were 0.25 degrees larger than the smallest rectangle that would enclose each object (see Figure 2). The regions were slightly larger than the objects to reduce the number of fixations that would be inappropriately excluded from the analysis due to minor errors in calibration and saccade targeting. The mean fixation count for each of the two regions was averaged and entered into the analysis for each subject and for each condition. The number of fixations per object did not vary by condition, $F(3, 33) = 0.79$, $MSE = 0.02$, $p = .51$. On average, 1.2 fixations contributed to the first glances within the target object regions. This value was used to compute the expected number of fixations per scene for minimal and maximal use of memory: The expected numbers of objects fixated per scene derived from each of the models were multiplied by the number of fixations per object to evaluate the observed numbers of fixations per scene. Table 1 presents the values resulting from these computations.

The first measure of interest, the number of fixations per scene, was computed by dividing the total number of fixations in each of the two scene regions by the total number of scene glances for every participant and every trial. The number of fixations per scene varied by condition, $F(3, 33) = 4.08$, $MSE = 0.09$, $p < .05$. However, there was not a reliable difference in the number of fixations per scene when only the three different conditions were considered, $F < 1$. The number of fixations per scene was greater in the same condition (2.73) than in the three different conditions (mean = 2.46), $F(1, 11) = 6.99$, $MSE = 0.06$, $p < .05$. The direction of this difference is consistent with the predictions generated by the models (see Table 1).

The observed values for both the same and different conditions were compared to the respective expected numbers given minimal and maximal use of memory using one-sample *t*-tests. The observed values were greater than values predicted by minimal memory in the same condition, $t(11) = 3.77$, $p < .01$, as well as in the different conditions, $t(11) = 4.85$, $p < .01$. However, the predicted values for maximal memory were greater than the observed values in the same condition, $t(11) = 35.73$, $p < .001$, as well as in the different conditions, $t(11) = 36.84$, $p < .001$. Importantly, the means were considerably closer to those predicted by the minimal memory models. The average difference between the observed and predicted values was 0.3 fixations per scene for minimal memory versus 3.8 fixations per scene for maximal memory. Thus, although the observed values fall outside of the range of predictions made by either strategy, they clearly favour a minimal memory interpretation.

The second analysis of interest is the number of fixations that intervened between the first fixation on the first critical object fixated and the first fixation on the corresponding object in the other scene. However, since by definition this analysis includes only trials where both critical items were fixated, it is important to know how frequently this occurs. Thus, the proportion of trials where both items were fixated was examined prior to the analysis of primary interest. Both analyses were accomplished using the object-sized regions described above. The proportion of trials with both objects fixated differed reliably across the four conditions, $F(3, 33) = 6.17$, $MSE = 0.034$, $p < .01$. This effect, however, is completely driven by the deletion condition. The proportion of trials with both objects fixated did not differ when the deletion condition was excluded from the analysis, $F < 1$. The proportion of trials with both objects fixated was less in the deletion condition (.43) than the average of the other three conditions (.67), $F(1, 11) = 20.37$, $MSE = 0.018$, $p < .01$. This result is not surprising given that object presence can be detected from eccentricities of 4 degrees (Henderson, Williams, Castelhano, & Falk, 2003). Thus, excluding the deletion condition, both objects were fixated on the majority of trials.

The number of fixations that intervened between the first fixation on the first critical object fixated and the first fixation on the corresponding object in the other scene was determined for all trials where both critical objects were fixated. If the use of memory is minimized in this task, participants should most frequently direct the eyes from one object immediately to the corresponding object in the other scene. If memory use is maximized, participants should frequently fixate two objects in between glances to each of the two critical objects. The mean number of fixations per object can be used to determine the number of intervening fixations that would be expected in each of these cases. Given an average of 1.2 fixations per object, if memory is minimized in this task, the most frequent number of intervening fixations should be somewhere between zero and one. However, a mode of between two and three intervening fixations would be expected if the use of memory were maximized. The most frequent number of

intervening fixations was either zero or one in all four conditions of this experiment. Additionally, in all conditions zero and one intervening fixations were more frequent than two and three intervening fixations. Figure 3 shows the distribution of intervening fixations collapsed across conditions and pooled across subjects. This analysis suggests that participants frequently directed their eyes from one item immediately to the other and provides rather strong evidence in favour of the minimal memory hypothesis.

To convey more intuitively what is going on in this task, a scan pattern used by one of the participants on one of the trials is included in Figure 2. The hypothesized one-to-one viewing strategy is evident in the frequent shifts made between scenes. Support for the minimal memory hypothesis was provided by two analyses designed to make inferences about the use of memory in this task: the number of fixations per scene and the number of intervening fixations. The numbers of fixations per scene were somewhat greater than predicted by minimal memory; however, the observed values were considerably smaller than predicted had participants been making full use of memory. The intervening fixations analysis suggests that participants frequently held a single item in memory.

Taken together, the analyses suggest a preference to minimize the use of memory in a task that is primarily visual. However, because different levels of encoding were needed to detect the differences between scenes in this experi-

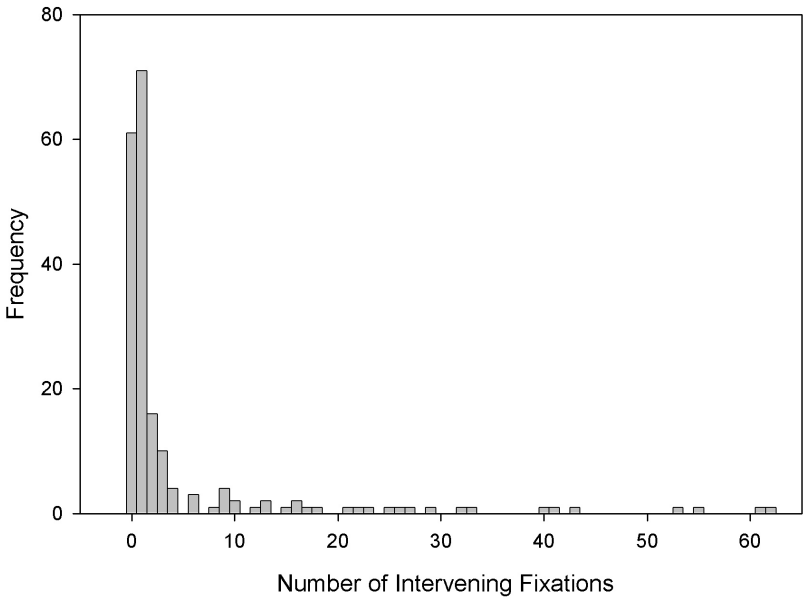

Figure 3. Histogram of the number of intervening fixations in Experiment 1.

ment, the observed viewing strategy might depend on the nature of the difference between scenes. Deletions and different-types required only the encoding of the presence and the identity of the objects. Different-tokens could only be detected on the basis of the visual details of the objects. Although we assume the items can be encoded as integrated objects, one could argue that the need to compare tokens produced a bias to encode the objects in terms of their features. If this were the case, having to detect this kind of difference could have elevated the demand on capacity and the one-to-one viewing strategy would not have reflected minimal use of memory. Given that the nature of the differences between scenes varied randomly from trial to trial, the overall viewing pattern we observed might therefore have been driven by a bias in the encoding strategy. To address this issue, in Experiment 2 we ran only the different-type and different-token conditions, and we ran them in separate blocks.

EXPERIMENT 2

The purpose of Experiment 2 was to examine the possibility that the pattern of viewing depends on the type of encoding needed to make the discrimination. To test this hypothesis, the nature of the difference between scenes was administered in blocks. By isolating the kind of difference that needed to be detected, it was assumed that the encoding strategy would be stable within each block and the generality of the minimal memory preference could be put to a more stringent test. In this experiment, only two kinds of differences between scenes were used: Different-types and different-tokens. In all other respects, Experiment 2 was identical to Experiment 1.

Method

Participants. Sixteen Michigan State University undergraduate students participated in the experiment in exchange for course credit or were paid. All participants had normal vision, were naive with respect to the hypotheses under investigation, and had not participated in Experiment 1.

Stimuli and apparatus. The stimuli and apparatus were the same as in Experiment 1, except that the deletion condition was eliminated.

Procedure. The procedure was the same as in Experiment 1, except the session was divided into two blocks. At the beginning of each block, participants were explicitly told the nature of the difference that could occur between scenes and examples were provided. As a result of the blocking procedure, the experiment was treated as a 2×2 factorial design with two levels of the kind of difference between scenes (different-type and different-token) and the two levels of same versus different. Scenes were assigned to conditions via a Latin-square design so that each scene appeared in each condition an equal number of times

across participants. The order of blocks was counterbalanced and the order of scene presentation was determined randomly for each participant within each session. In addition, the initial point of fixation was moved from the centre to the top of the screen. Because the initial fixation was not under viewer control, this change reduced the possibility that fixations would be inappropriately assigned to one of the critical regions.

Results and discussion

In this experiment, the kind of difference that was possible between scenes varied between blocks. Although the order of blocks was counterbalanced across participants, it is possible that the strategy used by participants in the second block was systematically influenced by the strategy adopted in the first block. To insure that this was not the case, all analyses included tests for order effects. Two within-subjects factors of the kind of difference and same versus different were entered into ANOVAs along with order as a between-subjects factor. Because there were no order interactions, these tests will not be discussed below.

Performance measures of accuracy and reaction time were evaluated prior to the analyses of primary interest. Participants were again more accurate on the same conditions relative to the different conditions, $F(1, 14) = 6.71$, $MSE = 0.01$, $p < .05$; however, there was no effect of the kind of difference between scenes and no interaction, Fs < 1. Accuracy was quite high in this experiment. The mean proportions correct were .97 and .90 in the same and different conditions, respectively.

Reaction times were included for correct responses only. Participants took longer in the same conditions relative to the different conditions, $F(1, 14) = 76.51$, $MSE = 7.92$, $p < .001$; however, there was no effect of the kind of difference between scenes and no interaction, Fs < 1. The mean reaction times were 10.5 s and 4.4 s in the same and different conditions, respectively.

Data from all 16 subjects went into the scene- and object-level analyses; however, 38 trials were eliminated (approximately 7%) because of track loss. The mean number of scene glances was determined in order to generate predictive models. Scene glances were more frequent in the same relative to the different conditions, $F(1, 14) = 64.68$, $MSE = 11.13$, $p < .001$. However, there was no effect of the kind of difference between scenes and no interaction, Fs < 1. The mean number of scene glances was 12.7 and 6.0 in the same and different conditions, respectively. Therefore, the expected number of objects fixated per scene was formulated for minimal and maximal memory based on 13-glance models for the same conditions and 6-glance models for the different conditions. The resulting values are listed in the top section of Table 2.

The number of fixations per object was determined by counting the number of fixations that occurred during the first glance within each of the object-sized regions as discussed in Experiment 1. There were fewer fixations per object in the same conditions relative to the different conditions, $F(1, 14) = 7.13$, $MSE =$

TABLE 2
Summary of the observed and predicted values in Experiment 2

	Same	*Different*
Expected objects per scene		
Minimum	1.8	1.7
Maximum	5.5	5.0
Expected fixations per scene		
Minimum	2.2	2.2
Maximum	6.6	6.5
Observed fixations (S.E.) per scene	2.81 (.10)	2.30 (.08)

0.027, $p < .05$. However, there was no effect of the kind of difference between scenes and no interaction, Fs < 1. There was an average of 1.2 and 1.3 fixations per object in the same and different conditions, respectively. The expected numbers of objects fixated per scene derived from each of the models was multiplied by these values in order to evaluate the observed numbers of fixations per scene. Table 2 presents the values resulting from these computations.

The number of fixations per scene was computed by dividing the total number of fixations in each of the two scene regions by the total number of scene glances for every trial. The number of fixations per scene was greater in the same relative to the different condition, $F(1, 14) = 147.29$, $MSE = 0.029$, $p < .001$. However, there was no effect of type of difference between scenes and no interaction, Fs < 1. The mean number of fixations per entry was 2.81 and 2.30 in the same and different conditions, respectively.

The analyses thus far demonstrate that the pattern of results was not affected by the kind of difference participants expected to find between the scenes. Because there were no effects of the kind of difference on any of the preceding measures, the observed values for both the same and different conditions were collapsed across the kind of difference and these were compared to the predicted values using one-sample *t*-tests. The observed values were greater than the predicted values for minimal memory in the same conditions, $t(15) = 5.88$, $p < .001$, but not in the different conditions, $t(15) = 1.23$, $p = .24$. The predicted values based on maximal memory were greater than the observed values in the same conditions, $t(15) = 36.34$, $p < .001$, as well as in the different conditions, $t(15) = 52.95$, $p < .001$. The average difference between the observed and predicted values was 0.1 fixations per scene for minimal memory versus 4.0 fixations per scene for maximal memory. Thus, the results of this analysis closely mirror the pattern found in Experiment 1.

The second analysis of interest is the number of fixations that intervened between the first fixation on the first critical object fixated and the first fixation on the corresponding object in the other scene given that both critical objects were fixated. In addition, the proportion of trials that both objects were fixated was determined. Both objects were fixated more frequently in the different

conditions (mean = 0.89) relative to the same conditions (mean = 0.71), $F(1, 14) = 15.13$, $MSE = 0.03$, $p < .01$. This difference is not surprising given that correct detections require the inspection of both critical items. If participants were biased toward reporting that the scenes were the same, correct responses could have been made in the same conditions whether both critical items were inspected or not. There was, however, no effect of the kind of difference between scenes and no interaction, Fs < 1. In both cases both critical objects were fixated on the majority of trials. As in Experiment 1, the most frequent number of intervening fixations was either zero or one in all four conditions. Figures 4A and 4B shows the distribution of intervening fixations collapsed across the same and different conditions for the different-type and different-token conditions, respectively. In both cases the mode was zero intervening fixations and zero and one intervening fixations were better than twice as frequent as two and three intervening fixations.

In sum, the results of Experiment 2 are nearly identical to those of Experiment 1. The number of fixations per scene was slightly more than would be predicted by minimal memory in the same conditions but not in the different conditions. In addition, there were many fewer fixations per scene than predicted by maximal memory. The number of intervening fixations was even more aligned with the minimal memory view in Experiment 2. The mode of zero intervening fixations strongly suggests a preference for the one-to-one viewing strategy associated with minimized use of working memory. Importantly, however, there were no differences in the viewing patterns when participants were looking for different-types versus different-tokens. If the preferred viewing strategy observed in Experiment 1 was driven by the level of encoding needed to detect differences in the visual details of the objects, the viewing patterns should have diverged here. This was clearly not the case. The results of Experiment 2 suggest that the minimal memory preference is general—visual working memory is minimized whether it is the identity of the object or the visual details that need to be remembered. An alternative possibility, however, is that the visual details of the objects were encoded in both cases. That is, visual differences existed between the different-tokens as well as the different-types. If participants opted to use the same level of encoding in both blocks, the one-to-one viewing strategy might still be limited to the case when visual details are encoded. In order to test this possibility, a version of the task was designed that eliminated judgements based on the visual details of the objects.

EXPERIMENT 3

To ensure that the viewing strategy observed in Experiments 1 and 2 is not limited to cases when the visual details of the objects are encoded, an object array version of the task was introduced. Participants compared an array of pictures of objects to an array of the names of those objects. The arrays were the

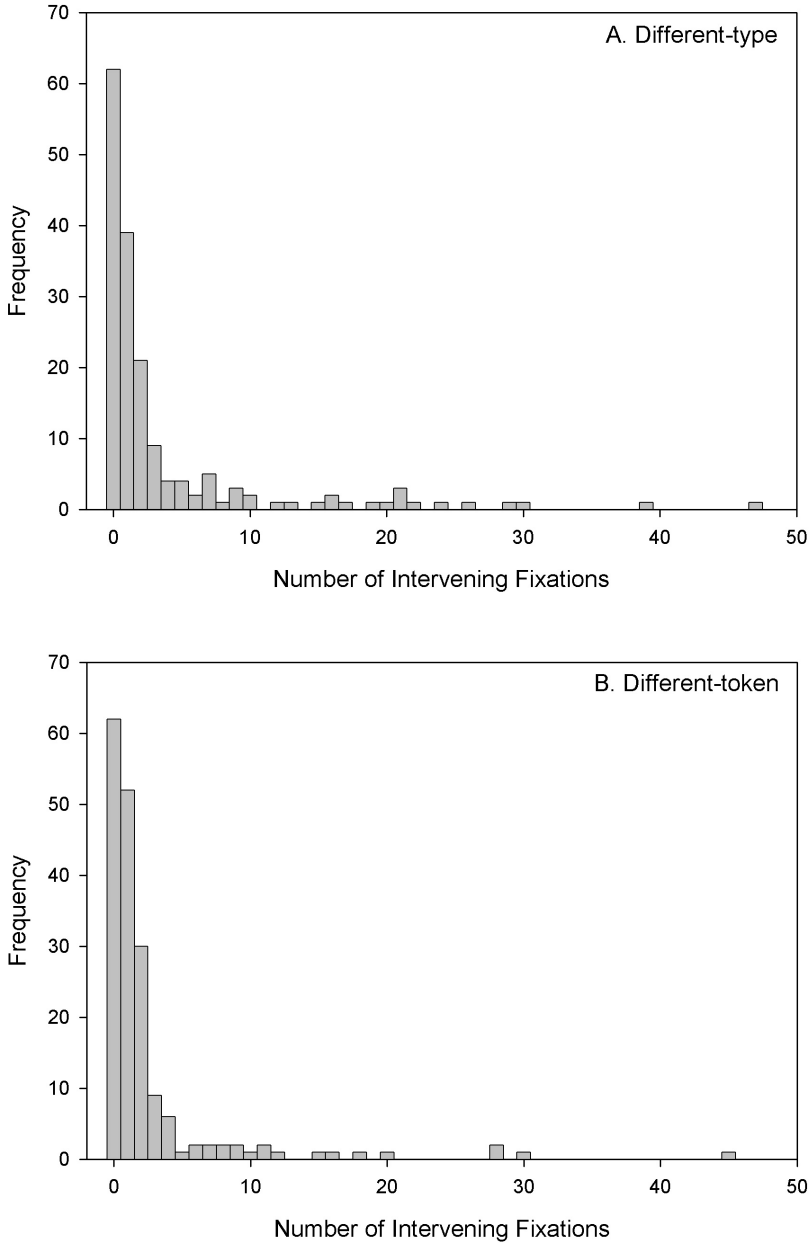

Figure 4. Histogram of the number of intervening fixations in the different-type (A) and different-token (B) conditions of Experiment 2.

same if all of the pictures and names matched, and different if one pair of pictures and names mismatched. Importantly, encoding the visual details of the objects did not assist in the execution of this task. The pictures needed to be encoded at the level of the object's identity in order to detect differences between arrays. If the minimal memory preference is limited to the case when visual details are encoded, the viewing pattern should diverge here. An additional benefit of using object arrays is that fixations can be more readily assigned to objects. As a result, the number of items fixated per array and the number of intervening items can be determined directly.

Method

Participants. Eight Michigan State University undergraduate students participated in the experiment for course credit. All participants had normal vision and were naive with respect to the hypotheses under investigation.

Stimuli. Sixteen images were constructed containing two arrays with a neutral area on top so that the initial fixation would be outside either of the arrays. Figure 5 shows a sample of an image used in this experiment. The two arrays subtended $7.7^\circ \times 9.9^\circ$ visual angle; the neutral area subtended $1.9^\circ \times 15.7^\circ$ visual angle. The three areas were demarcated by light grey borders against a dark grey background. The left-hand arrays were filled with pictures of objects and the right-hand arrays contained the words that named the objects. The object pictures were taken from the Hemera Photo-objects 50,000 Premium Image Collection and were sized so that the largest dimension, vertical or horizontal, subtended 1.9° visual angle. The font size was selected so that the largest words would fit within the same square regions. The objects and words were positioned so that there was a spatial correspondence between matching items in the two arrays. Pictured objects, such as the moose, the fan, and the bug in the sample image (Figure 5), were located in the same position in the picture array as the words that named them in the word array. A total of 184 objects were selected so that they could be named by one-syllable words. Each of the 16 images contained 11 object/word pairs. To create arrays for the different condition, one item in one of the arrays was randomly selected and replaced with one of the eight remaining items. There were no item repetitions.

Apparatus. The apparatus was the same as in Experiments 1 and 2.

Procedure. The procedure was the same as in Experiment 1, except that there were only two conditions in this experiment—same and different. Participants were provided with a visual example that illustrated the nature of

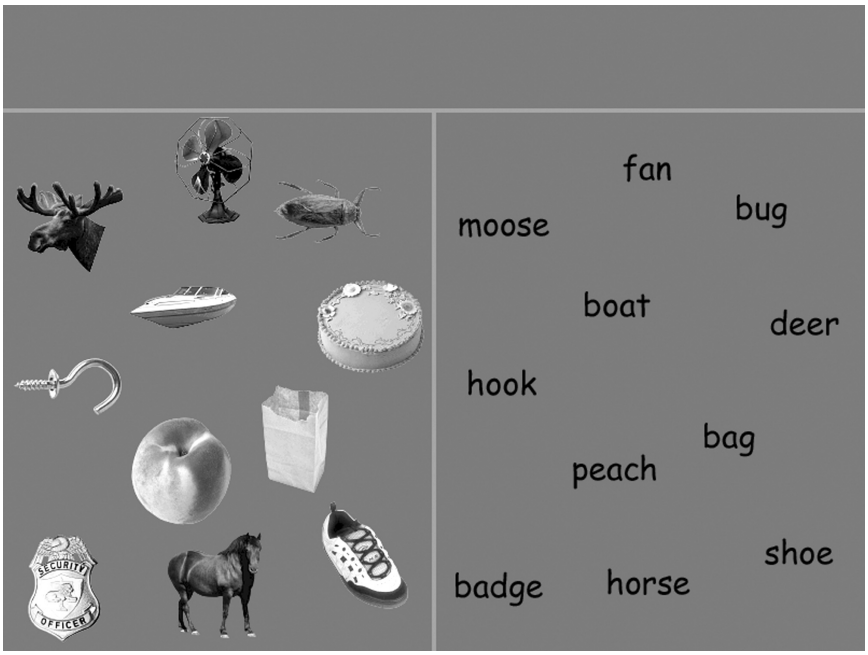

Figure 5. Sample image from the different condition of Experiment 3. Note that the location in the picture array that corresponds to the location of the word ''deer'' contains a picture of a cake. Full-colour displays were used in the experiment.

differences that could be found between arrays. Each session lasted approximately 20 min.

Results and discussion

Performance measures of accuracy and reaction time were evaluated prior to the analyses of primary interest. Accuracy was moderately high (mean = 0.85) and did not reliably differ between the same and the different conditions, $F(1, 7) = 0.50$, $MSE = 0.07$, $p = .50$. Reaction times were entered for correct responses only. Participants viewed the arrays for longer in the same condition (8.5 s) relative to the different condition (6.7 s), $F(1, 7) = 27.12$, $MSE = 0.71$, $p < .01$.

Data from all 16 subjects went into the scene- and item-level analyses; however, 12 trials were eliminated (approximately 9%) because of track loss. The mean number of array glances was determined in order to generate predictive models. Array glances were more frequent in the same relative to the different condition, $F(1, 7) = 18.49$, $MSE = 2.36$, $p < .01$. The mean number of array glances was 14.0 and 10.7 in the same and different conditions, respectively. Therefore, the expected number of items fixated per array was formulated

for minimal and maximal memory based on 14- and 11-glance models, respectively. The resulting values are presented in Table 3.

In this experiment, due to the clean separation between items, we could directly determine the number of items fixated per array. This was accomplished by tabulating the number of item regions that receive fixations during each glance to each of the arrays. Because fixations that fall outside of these regions do not contribute to the analysis, it is important to determine the mean proportion of fixations that fall within the item-sized regions. This was accomplished by tallying the total number of fixations within the item regions and dividing by the total number of fixations falling within regions defined for each of the arrays for every trial. The mean proportion of fixations falling within the item regions was .89 and did not vary between the same and the different conditions, $F(1, 7) = 0.45$, $MSE = 0.00$, $p = .52$. Thus, a large majority of the fixations were assigned to items in the arrays.

The mean number of items fixated per array glance was computed by dividing the total number of items fixated by the total number of array glances for every trial. The number of items fixated per array glance did not reliably differ between the same (1.76) and different (1.67) conditions, $F(1, 7) = 1.04$, $MSE = 0.03$, $p = .34$. However, both means were tested because of the reliable differences found in the number of array glances. The observed values were numerically less than the predicted values based on minimal memory, but they did not reliably differ in the same condition, $t(7) = -1.77$, $p = .12$, nor in the different condition, $t(7) = -1.968$, $p = .09$. The predicted values based on maximal memory, however, were greater than the observed values in the same condition, $t(7) = 49.23$, $p < .001$, as well as in the different condition, $t(7) = 59.54$, $p < .001$. Thus, the results of this analysis are in agreement with the predictions based on the minimal memory hypothesis.

In this experiment we could also determine the actual number of intervening items fixated. Because there was no need to make assumptions about the number of fixations per item, predictions for this measure were more clear: minimal memory predicts a mode of zero intervening items, and maximal memory predicts a mode of two intervening items. Figure 6 shows the distribution of

TABLE 3
Summary of observed and predicted values in Experiment 3

	Same	*Different*
Expected items per array		
Minimum	1.9	1.8
Maximum	5.6	5.5
Observed items (S.E.) per array	1.76 (.08)	1.67 (.06)

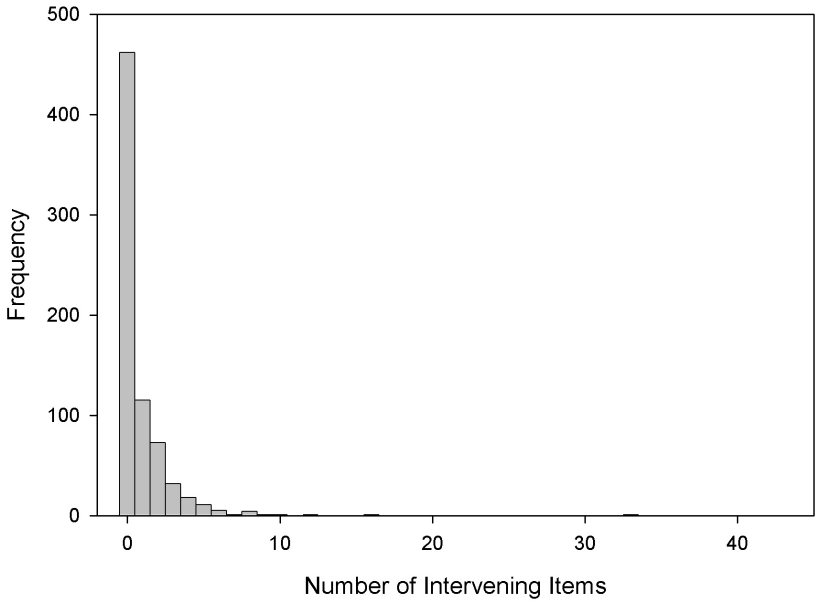

Figure 6. Histogram of the number of intervening items in Experiment 3.

intervening items collapsed across conditions. The most frequent number of intervening items was zero in both conditions.

The pattern of results found in Experiment 3 was consistent with those found in Experiments 1 and 2. In fact, these analyses most unambiguously support the preference for a one-to-one viewing strategy in this task. The number of items fixated per array was fewer than predicted by maximal memory but did not differ reliably from the values predicted by minimal memory. In addition, the intervening items analysis suggests that participants most frequently hold a single item in memory. Importantly, the persistence of this viewing pattern despite the irrelevance of the visual details suggests a minimal memory preference that is general.

GENERAL DISCUSSION

The purpose of this study was to test the generality of the minimal memory preference found by Ballard et al. (1995; Hayhoe et al., 1998). Is the preference to minimize the use of visual working memory limited to the case when visual information must be coordinated with basic motoric interactions with the environment? In order to test the minimal memory hypothesis, a scene com-

parison task was introduced. Participants compared side-by-side simultaneously presented pairs of scenes that were either identical or differed in terms of one object. Eye movements were examined to make inferences about the use of memory in this task. If the minimal memory strategy extends to this primarily visual task, participants would be expected to move the eyes frequently back and forth between scenes holding just one object in memory. There were two dependent measures of primary interest: The number of fixations per scene and the number of fixations between glances to corresponding objects. The observed number of fixations per scene was compared to the number of fixations per scene predicted by models generated with one and with three items occupying working memory. The observed values were reliably greater than predicted by minimal memory in the same and different conditions of Experiment 1 and in the same but not the different conditions of Experiment 2; however, these differences were always very small relative to the differences between the observed values and those predicted by maximal memory. In addition, the number of items fixated per array glance in Experiment 3 did not differ from the predictions based on minimal memory. Strong evidence in favour of the minimal memory view was obtained from the intervening fixation analyses of Experiments 1 and 2 and the intervening item analysis of Experiment 3. These data suggest that corresponding objects are most frequently fixated one after the other.

The pattern of results found in Experiment 1 suggested that the preference to minimize the use of visual working memory generalizes to tasks that are primarily visual. However, the presence of multiple kinds of differences between scenes left open the possibility that the observed viewing pattern depended on the level of encoding needed to detect the differences between scenes. Only the different-token condition required the encoding of the visual details of the objects. Although we assume the objects can be encoded as integrated units, having to detect this kind of difference might have biased participants towards a feature-based encoding strategy. If this were the case, capacity could no longer be expressed in terms of objects and the one-to-one viewing strategy would no longer reflect a minimal memory preference. To explore this possibility, different-types and different-tokens were run in separate blocks in Experiment 2. By isolating the kind of differences that were possible between scenes, it was assumed that a single encoding strategy would be adopted within each block. If the one-to-one viewing strategy observed in Experiment 1 was driven by the demands of the different-token condition, this pattern should only have been observed in the different-token condition of Experiment 2. The failure to find a difference between blocks in this experiment suggests a preference to minimize the use of memory even when only the identity of the objects need be encoded. However, the possibility remained that participants opted to encode the visual details in the different-type condition despite the fact that the identity would have been sufficient. Experiment 3 tested this possibility by introducing an

object array version of the task. Participants compared arrays of object pictures to arrays of words that named the objects. In this task, items needed to be processed to the level of the object's identity to make the comparisons, but participants would have no need to encode the visual details of the objects. The persistent observation of the one-to-one viewing strategy in Experiments 2 and 3 suggests that the minimal memory preference generalizes to the case when only the identity of the object need be encoded.

Because one of the primary goals of this study was to gain a better understanding of the use of visual working memory in scene perception, we chose to begin by using naturalistic scenes as stimuli. Although hypotheses need to be tested using stimuli that approach the complexity of normal viewing situations, using scenes as stimuli introduces challenges to experimentation. In order to overcome these challenges, inferences were made on the basis of one pair of objects in each of the pairs of scenes. The number of fixations on critical objects was used as an index of the number of fixations that generally contribute to a gaze on an object, and this number was used to express the predicted number of objects fixated per scene in terms of the number of fixations per scene. The consistency of the results found in the scene and array versions of the task confirm the validity of this approach. Although the purpose of Experiment 3 was to test for differences that might have resulted from the level of encoding, an additional benefit of using object arrays is that fixations on all the objects could be tabulated. As a result, the actual number of items fixated per array glance and the actual number of intervening items could be determined. Interestingly, the observed number of items fixated per scene and intervening items fit even more closely to the predictions based on minimal memory. This suggests that the reduced level of support coming from the fixations per scene analyses of Experiments 1 and 2 might be the result of saccade targeting errors. However, even given a preference to minimize the use of memory, some fixations would be expected to reflect processes other than the explicit coding of one object to be compared to the other. These fixations could be associated with other aspects of processing such as searching for items to check as well as keeping track of which items have already been inspected.

What does minimal memory in a scene comparison task tell us about memory in scene perception? An issue that has received considerable attention by those studying scene perception is the nature of the information that accumulates in memory as a result of visually exploring a scene. Based on capacity estimates for visual working memory of three–four objects (Irwin, 1992; Irwin & Andrews, 1996; Luck & Vogel, 1997; Vogel et al., 2001), theorists have proposed that the active representation includes the three–four most recently attended items (e.g., Hollingworth & Henderson, 2002; Irwin & Andrews, 1996; Irwin & Zelinsky, 2002). The present study suggests a functional capacity that is more limited. Consistent with the ''just in time'' processing strategy favoured by proponents of the active vision framework, the visual system appears to have

a bias towards minimal online representation. The viewing behaviour observed in this study suggests that it is easier to sample the environment with eye movements than to keep representations active in working memory, but it says nothing about the storage or the retrievability of these representations with the passage of time. In this sense, eye movements are cheaper than memory.

REFERENCES

Aloimonos, J., Weiss, I., & Bandopadhay, A. (1987). Active vision. *International Journal of Computer Vision*, *1*, 333–356.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*, 66–80.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*, 723–767.

Crane, H. D. (1994). The Purkinje image eyetracker, image stabilization, and related forms of stimulus manipulation. In D. H. Kelley (Ed.), *Visual science and engineering: Models and applications* (pp. 15–89). New York: Marcel Dekker.

Crane, H. D., & Steele, C. M. (1985). Generation-V dual-Purkinje-image eyetracker. *Applied Optics*, *24*, 527–537.

Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research*, *38*, 125–137.

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 210–228.

Henderson, J. M., Williams, C. C., Castelhano, M. S., & Falk, R. J. (2003). Eye movements and picture processing during recognition. *Perception and Psychophysics*, *65*, 725–734.

Hollingworth, A., & Henderson, J. M. (2002) Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 113–136.

Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 307–317.

Irwin, D. E., & Andrews, R. (1996). Integration and accumulation of information across saccadic eye movements. In T. Inui & J. L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 125–155). Cambridge, MA: MIT Press.

Irwin, D. E., & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Perception and Psychophysics*, *64*, 882–895.

Luck, L. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 391–399.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*, 17–42.

Vogel, E. K., Woodman, G. F., & Luck, L. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 92–114.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*, 48–64.