Perception, 2003, volume 32, pages 579 – 592

DOI:10.1068/p3396

The time course of abstract visual representation

Benjamin W Tatler
Sussex Centre for Neuroscience, School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK; e-mail: b.w.tatler@sussex.ac.uk

Iain D Gilchrist
Department of Experimental Psychology, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK

Jenny Rusted
Laboratory of Experimental Psychology, School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK

Received 27 May 2002, in revised form 19 March 2003

Abstract. Studies in change blindness re-enforce the suggestion that veridical, pictorial representations that survive multiple relocations of gaze are unlikely to be generated in the visual system. However, more abstract information may well be extracted and represented by the visual system. In this paper we study the types of information that are retained and the time courses over which these representations are constructed when participants view complex natural scenes. We find that such information is retained and that the resultant abstract representations encode a range of information. Different types of information are extracted and represented over different time courses. After several seconds of viewing natural scenes, our visual system is able to construct a complex information-rich representation.

1 Introduction

Until relatively recently a popular assumption in the literature was that the visual system constructed a richly detailed reconstruction of visual surroundings (eg Marr 1982), integrating information from multiple fixations (eg McConkie and Rayner 1976). However, evidence began to emerge in the 1970s and 1980s that cast doubt upon this intuitively appealing notion, from studies which specifically failed to find evidence for trans-saccadic integration of pictorial information from successive fixations (eg Bridgeman et al 1975; Bridgeman and Mayer 1983; Mack 1970; McConkie and Zola 1979; Whipple and Wallach 1978). Some of the most compelling evidence that the visual system does not form a stable, veridical representation across views comes from studies of change blindness (since Grimes 1996). These experiments demonstrate that participants are unable to detect large changes in a scene, such as the disappearance of a building. Change blindness has often been used to suggest that very little or no pictorial information survives saccades, implying either very sparse or entirely absent veridical pictorial representations (eg Grimes 1996; Rensink et al 1995, 1997, 2000). Others have suggested that the visual system may not require *any* form of visual representation of the world (Dennett 1991; Gibson 1966; MacKay 1973) and that internal representation is largely unnecessary because the world can itself act as its own 'outside memory' (O'Regan 1992; O'Regan and Noë 2001). However, most researchers hold the view that some information is likely to be retained by the visual system, even if retention of veridical pictorial detail is poor, although they are not always in agreement about the nature of retained information or its detail and accuracy.

One possibility to have received much attention over the years is that the information retained by the visual system is abstract rather than veridically pictorial (eg Gibson 1979; Henderson 1994, 1997; Hochberg 1968; Irwin 1991, 1993; O'Regan and Lévy-Schoen 1983; Pollatsek and Rayner 1992). In fact, the possibility of non-pictorial representation has much older roots and can be seen in Hobbes's (1651, 1656) proposition that mental representations might take the form of language-like symbols. If we suppose that representations may be abstract in nature, we can consider the type of information that might be integrated into the representations. Over the years, particular emphasis has been placed upon two main candidates for the type of information that might be retained in any non-pictorial representation: gist and spatial layout. Gist refers to the overall meaning or nature of a scene or image (eg whether a scene is of a kitchen or an office) and is independent of explicit knowledge of the scene's content, layout, or other detail. This type of information can be extracted very rapidly from scenes (Intraub 1980, 1981), even during presentation times as brief as 120 ms (Biederman 1981). Spatial layout refers to the overall arrangement and positioning of items and features within a scene and can be independent of semantics and properties of objects (Hochberg 1968).

Most accounts of scene perception or visual representation implicate one or both of these two forms of abstract information (de Graef 1992; Henderson 1992; Rayner and Pollatsek 1992). Change-detection findings themselves have been used to argue the case for retention of gist and layout information (Aginsky and Tarr 2000; Hollingworth and Henderson 2002; Rensink 2000). Manipulations of semantic consistency of objects in scenes have also been used to suggest that object gist is encoded and retained (eg Friedman 1979; Pezdek et al 1989). Chun and Nakayama (2000) implicated retention of spatial information using a complex search task, and Simons (1996) interpreted observers' failures to detect changes in video sequences in terms of failed veridical representations but preserved abstract spatial information. Spatial priming has also been used to suggest representation of scene layout information (Sanocki and Epstein 1997). Studies of long-term scene memory have suggested that spatial information is extracted, can be accessed immediately after exposure, and can be transferred to longer-term scene schema (eg Mandler and Parker 1976).

While there is agreement that information about gist and spatial layout in scenes is likely to be retained by the visual system, consensus has yet to be reached on whether more detailed (but potentially still abstract) information is also encoded and, if so, what this information might be. Various studies have suggested candidates for extraction. Henderson and Hollingworth (1999) found support for retained object-presence information in a change-detection paradigm. Melcher (2001) found that the number of objects recalled improved steadily as presentation times were increased from 1 to 4 s for computer-generated complex scenes. This result was used to propose a medium-term visual memory for scene content, persisting for a period between several seconds and several minutes. Evidence for the encoding of an inventory for complex scenes can be found within the long-term-memory literature, whereby such information was shown to be available for judgments immediately following visual presentation of a scene and can be retrieved over much longer time scales (eg Mandler and Ritchey 1977). Retention of object identity has been incorporated into several recent accounts of scene perception (eg de Graef 1992; Rayner and Pollatsek 1992; Simons and Levin 1997).

More precise information than merely identity of objects may be encoded and retained. This possibility has been suggested, for example, in the recent work by Henderson and colleagues (eg Henderson 1994; Henderson and Siefert 1999; Hollingworth and Henderson 2002). Using a combination of explicit and implicit measures these authors have suggested that very accurate object information can be encoded and retained, and that perhaps visual representations are much more richly detailed than change-detection researchers have postulated in recent years (Hollingworth and Henderson 2002).

While encoding and retention of a variety of information types has been suggested, the extraction of a specific type of information has most often been studied in isolation within the short-term-memory literature (exceptions to this include Aginsky and

Tarr 2000; and Hollingworth and Henderson 2002). One limitation of this approach is that it reduces the ecological validity of the investigation of scene representation because in every-day vision we very rarely attend to a single type of information. Rather, we are likely to require information that spans a number of types of information; for example, when making a cup of tea (Land et al 1999) we may require information about overall gist, what objects are present in the room, where they are located, and more detailed object information such as colour and shape of the items might also be useful. Under these circumstances, the resource allocation for encoding and storage of each individual information modality might be rather different from the situation where only one type of information need be extracted. This limitation has long been recognised within the study of long-term memory, and Mandler and colleagues advocated the need to use ecologically valid conditions of viewing and encoding, testing multiple types of information concurrently (Mandler and Johnson 1976; Mandler et al 1977). In support of this approach, it has been found that under some circumstances, when participants are required to encode more than one type of information, costs can be seen when compared to isolated encoding situations (Light and Berger 1974; Light et al 1975). In addition, single-modality studies do not allow us to address the question of prioritisation during the construction of any retained representation. It may be that different information types are extracted in a specific order, or that all information types are encoded concurrently.

In the study reported in this paper we investigated the ability of observers to extract and retain a variety of types of information from briefly presented photographic images of complex real-world scenes. Explicit questioning was used to probe retention of gist information, absolute spatial positions of objects in the scene, the content of the scenes (the objects or features present in the images), specific shapes of objects, colours of objects, and the relative distances between objects (spatial relations between items in the scene). After each image, participants were asked about two of these six categories of information, and did not know in advance which they would be questioned about. As a result, it was very hard for observers to predict the questions. Variations in the size, positions, and types of items tested in the questions also reduced predictability (see section 2). In this way, we hoped to avoid the situation in which observers can attend to a single information type and instead promote concurrent processing (if possible) of the different types of information in the scenes, as is more likely to be the case under normal viewing conditions. Studying the representational system under conditions of concurrent processing allows insights into the behaviour of the system in everyday vision and increases the ecological validity of the study. However, we acknowledge that the task in our study was not like all everyday visual experiences: the goal here was to memorise as much of the scene as possible as participants knew that they would be asked questions after viewing each scene. However, we reduced the influence of this by making presentation times brief (only a few seconds) and by using complex scenes in order to reduce the change of memory rehearsal strategies during viewing. Our experiments therefore were perhaps most analogous to the period of initial inspection upon first entering a new scene.

The time course of information extraction was also investigated in this study. Presentation times of scenes were varied such that the processes of information assimilation might be interrupted at different stages in their progress.

One of the methodological problems with studies of this kind is that participants will often have an expectation of the nature of certain scene properties based on previous experience. For example, participants may as a default assume that the walls of a room are magnolia or white. We tested this ‘guess-ability’ in a control experiment in which different participants were asked to make judgments about the scenes without exposure to them.

2 Methods

2.1 *Participants*

Fourteen participants (two male, twelve female) aged 18 to 37 years (mean 21.5 years, SD 5.8 years) took part in the main experiment, viewing natural images presented on a computer monitor. Fifteen different participants (one male, fourteen female) aged 18 to 39 years (mean 24.6 years, SD 6.0 years) took part in the control experiment. These participants did not view the natural images. All had normal or corrected-to-normal vision.

2.2 *Procedure*

Participants viewed 48 photographic images depicting a variety of indoor and outdoor familiar scenes (see figure 1 for two examples). Images were displayed on a 17 inch colour monitor, positioned at a viewing distance of 60 cm. Consequently, the images presented subtended 30 deg horizontally and 22 deg vertically of the participants' visual field.

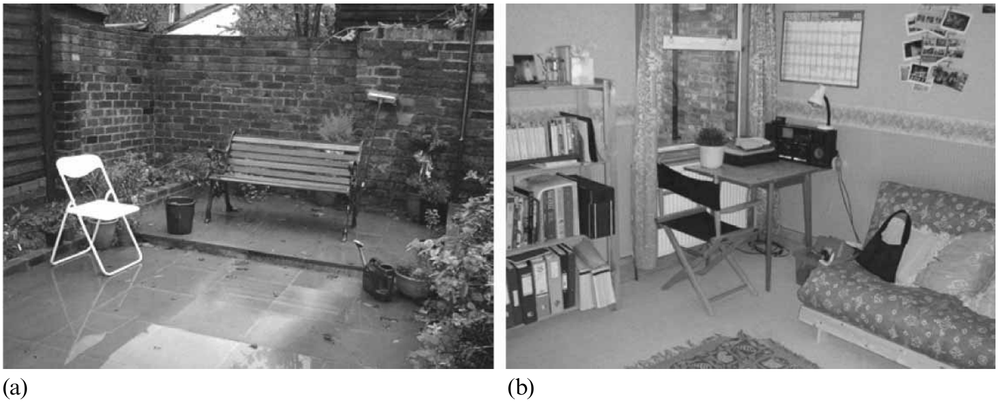

Figure 1. Two typical examples of images viewed by participants in this study. The images covered both indoor and outdoor familiar scenes.

Participants were free to view the images as they wished during each presentation time. Following each image presentation, participants were asked two questions about the image just viewed. These questions covered two of six categories of information about the scene: gist, content, absolute layout, relative layout, object shape, and object colour. Questions were in the form of a four-alternative forced choice in which one of the four options was always correct. The three foils for each question were matched to each other and the target object in terms of size and contextual viability. For presence questions viable foils were items that might have been present in the scene. For shape and colour questions, foils were viable alternatives for the target item. For absolute position, foils were both viable for the type of object tested and were also locations occupied by other items in the scene. Finally, foils for relative distance were other items occurring within the scene. Care was taken in choosing the foils in order to equate discriminability between foils and targets. Results from the control questionnaires (see below) and the performances between the question types validated comparability of the question difficulty.

While we can largely control for target – foil discriminability and contextual viability, it may be that particular objects tested were easier to discriminate perceptually within the scene and could therefore be extracted more easily. It may even be that the information types tested are themselves not perceptually matched—ie colour extraction might not be perceptually comparable to shape extraction. If this were the case, direct quantitative comparisons between question types would be inappropriate. However, qualitative comparisons between question types are still valid, and provide crucial insights into the construction of abstract representations and the differential extraction of information types.

For questions testing gist, content, relative distance, and colour response, options were words. For example for the scene shown in figure 1a, the content question was: "Which one of the following was present: (a) broom, (b) pond, (c) bin, (d) greenhouse?" In the case of relative-distance questions, foils were all items within the scene, but situated further from the question item than the correct response. For example, for the scene depicted in figure 1b, the relative-distance question was: "Which one of the following was nearest to the shelves: (a) settee, (b) lamp, (c) plant, (d) poster?" For questions testing shape, response options were pictorial with foils of viable alternative shape of the target item. In the case of questions testing absolute position, participants were presented with an outline sketch of the scene in which four positions were indicated; foil locations corresponded to positions of items other than the target item within the scene.

Gist questions were always asked as the first of the two questions, if they were to be asked about the presented image, because answering another question about the scene would have provided cues for gist. The two questions asked were counterbalanced within the six categories between images for each participant and between participants for each image. This provided a data set that comprised at least four responses per question per image across the participants. By testing a wide range of aspects of the scene relating to a wide variety of object and feature sizes and positions, we encouraged general viewing strategies by the participants. The inspection period was, therefore, most like the initial inspection that would occur upon first seeing a new scene or visual environment.

The time that the scene was present on the computer monitor was varied randomly between scenes ranging from 1 to 10 s. This made strategic memorising approaches to viewing the images harder for participants to adopt and, crucially, allowed us to interrupt the information accumulation process at different stages. By terminating the available visual input at different times, we could use the performance in subsequent questions to assess information assimilation within a given time frame.

The control experiment was in the form of a questionnaire with 48 sections corresponding to the 48 scenes of the main experiment and was answered by participants who did not view the images. In each section the identity of the 'scene' was stated (ie the gist question was answered) and was followed by the five four-option multiple-choice questions about the unseen 'scene'. The five questions were exactly the same as those used in the main experiment for the corresponding image, covering the same remaining categories of information: presence, shape, colour, absolute position, and relative distance.

There are some limitations to the protocol used here to assess the content and nature of representations. Primarily, the use of explicit questioning might limit our ability to probe any implicit form of represented information. While we acknowledge this limitation, we would point out that the use of four-alternative forced-choice questioning allows for the possibility that responses may reflect implicitly encoded information. Hence responses by participants in our study are likely to comprise a mixture of implicit and explicit information retained from the visual scenes viewed. Unfortunately, the extent to which responses were drawn from implicit knowledge cannot be assessed here.

By varying presentation time, we effectively interrupt the information assimilation process at different points in its progress. Consequently, responses by participants reflect an explicit translation of the state of the representation at the time of interruption to viewing, and can be used to infer the relative extents and progression of information accumulation.

3 Results

3.1 *Control questionnaires*

For each question in the control questionnaires, a χ^2 goodness-of-fit test was performed with the expected frequencies set to 25% (ie chance) for each response option. Twenty-nine out of the two hundred and forty questions showed a non-random distribution of responses by participants, in which one single response was chosen more frequently than any other. These questions demonstrate a general between-participants expectation or bias about the likely answer to the question asked, based on past experience of similar items in similar settings. The 'correct' option (ie that corresponding to the correct answer in the main experiment) was that chosen in eight of these twenty-nine questions. Therefore, in these eight questions in the main experiment (where participants answered questions after viewing the images) it was not possible to determine whether correct answers were because information about that item had been extracted during viewing or because people have general expectations that matched the correct option in these questions. Consequently, the responses to these eight specific questions were excluded from all analyses of the main experimental data.

Trials in which a participant bias toward a particular answer was found (both 'correct' and 'incorrect'), did not show any differences in the relative frequencies between the five types of question tested ($\chi^2 = 2.93$, $p > 0.05$); hence there was no tendency to guess correctly any particular one of the types of information.

3.2 *Information extraction from viewed images*

Figure 2 shows the performance by participants, pooled across all question types for each participant individually ($N = 1306$ responses). A χ^2 test (using expected frequencies calculated from the experimental data) showed that there were significant differences in performance between participants ($\chi^2 = 31.45$, $p < 0.05$). While there were differences in participant performance, most participants performed at about 56% correct (figure 2).

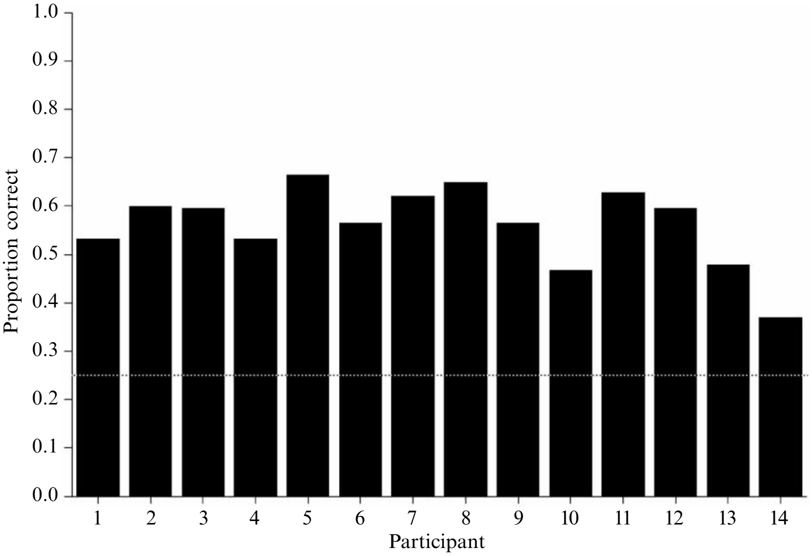

Figure 2. The proportion of questions answered correctly by each of the fourteen participants for all 48 trials of the experiment ($N = 1306$ responses). Chance is 25% (0.25) in this experiment and is indicated by the dotted line in the figure. All participants performed at greater than chance. Performances between participants were very similar.

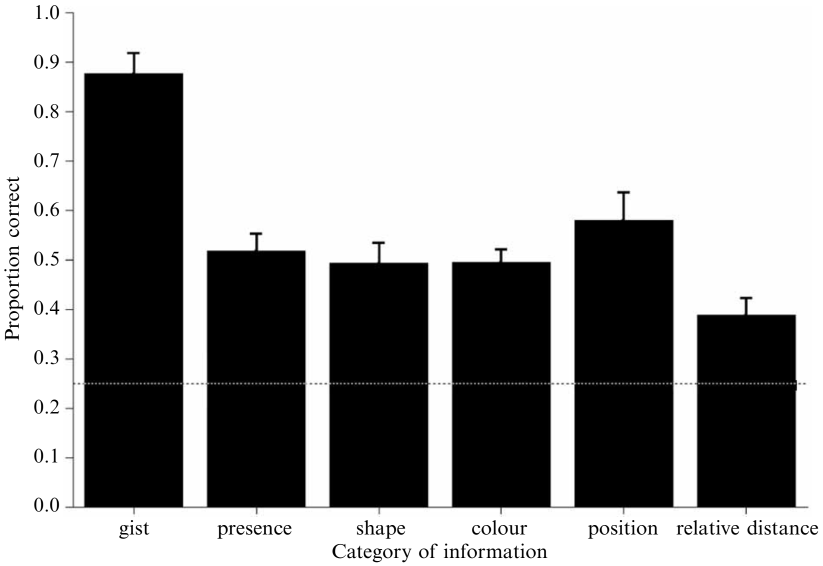

Figure 3. Performance by the fourteen participants in answering questions covering the six categories of information tested in this experiment. Responses were above chance in all questions, with highest performance in response to gist questions and poorest performance in response to questions testing the relative distances of items in the image. Chance is indicated by the dotted line. Error bars indicate standard error between participants.

3.2.1 *Effects associated with the questions.* Figure 3 shows overall performance by all participants in each of the six questions. One-sample t-tests show that performance is above chance in each question (gist: $t_{13} = 15.22$, $p < 0.05$; presence: $t_{13} = 7.75$, $p < 0.05$; shape: $t_{13} = 5.75$, $p < 0.05$; colour: $t_{13} = 9.25$, $p < 0.05$; absolute position: $t_{13} = 5.79$, $p < 0.05$; relative distance: $t_{13} = 3.94$, $p < 0.05$).

After each image presentation, participants were asked two questions. If the representation probed in our study was transient in nature, it would be expected that performance by participants in answering the second question after each image would be much lower than that for the first question. We can test this possibility by comparing performance for each question type according to whether it was asked as the first question after image presentation or the second question. Gist questions are excluded here because they were always asked first if they were to be asked (see section 2). Paired-sample t-tests (with a Bonferroni correction) show that there were no differences in performance when a particular question was asked first or second, for questions testing presence ($t_{13} = 0.66$, $p > 0.05$), shape ($t_{13} = 0.45$, $p > 0.05$), colour ($t_{13} = 0.78$, $p > 0.05$), absolute position ($t_{13} = 0.21$, $p > 0.05$), or relative distance ($t_{13} = 1.39$, $p > 0.05$). Consequently, it appears that there is no significant decay in the representation between the time of the first and second question after presentation. As a result, data from the first and second questions will be pooled in subsequent analyses.

Each participant received only two of the six possible questions about each image. The questions were divided into three sets such that each set contained two of the possible six chosen quasirandomly, but the three sets combined contained each possible question. An analysis of variance (ANOVA) shows the (expected) variation between question types ($F_{5,66} = 19.61$, $p < 0.05$), but that the variation between sets (within questions) was not significant ($F_{2,66} = 3.46$, $p > 0.05$), and that there was no interaction ($F_{10,66} = 1.57$, $p > 0.05$). Data from the three question sets will therefore be grouped in other analyses.

3.2.2 *Effects associated with the type and size of the images viewed.* The images were grouped into three sets, the order of which was varied systematically between participants in each experimental session. There were no differences between the three image sets in any of the questions ($F_{2,234} < 1$).

A Kruskal–Wallis test shows that differences in performances between individual scenes in each of the different question types were not significant (gist: $H_{47} = 24.31$, $p > 0.05$; presence: $H_{44} = 47.49$, $p > 0.05$; shape: $H_{46} = 50.90$, $p > 0.05$; colour: $H_{47} = 50.76$, $p > 0.05$; absolute position: $H_{45} = 40.67$, $p > 0.05$; relative distance: $H_{45} = 45.71$, $p > 0.05$).

Different items were tested in each question for each image. Hence we should consider the possibility that properties of the items tested might influence the ability of participants to extract information about those items. The most obvious characteristic that might be expected to influence performance is the size of the tested object. In spite of a wide variation in sizes of items tested (0.04%–32% of the total screen area), linear regression of the performance data shows that there were no significant trends in performance with item size [presence: slope $(\beta) = 1.4 \times 10^{-2}$, intercept $(\alpha) = 0.44$, $t_{76} = 1.16$, $p(\beta = 0) > 0.05$; shape: $\beta = -1.8 \times 10^{-2}$, $\alpha = 0.57$, $t_{57} = 1.30$, $p(\beta = 0) > 0.05$; colour: $\beta = -1.1 \times 10^{-2}$, $\alpha = 0.55$, $t_{93} = 1.47$, $p(\beta = 0) > 0.05$; absolute position: $\beta = 1.9 \times 10^{-3}$, $\alpha = 0.57$, $t_{49} = 0.10$, $p(\beta = 0) > 0.05$; relative distance: $\beta = 2.0 \times 10^{-4}$, $\alpha = 0.39$, $t_{83} = 0.03$, $p(\beta = 0) > 0.05$]. In these linear regressions, curves are described in terms of increase in proportion of correct responses with unit increase in the percentage of screen area occupied by the item tested.

3.2.3 *The influence of trial number on performance.* Each experimental session lasted for around 45–60 min, and it is possible that performance might have varied over this time as participants became more familiar with the requirements of the experiment. One indication of any such change might come from a comparison of performances in the first, second, and third image set shown to participants during each experiment. However, on grouping by image set, there was no significant effect of trial ($F_{2,234} = 0.06$, $p > 0.05$). If the data are considered on a trial-by-trial basis, linear regression analysis shows that there were no systematic trends in performance over the time course of the experimental session for any of the question types [gist: slope $(\beta) = -1.0 \times 10^{-3}$, intercept $(\alpha) = 0.89$, $t_{228} = 0.62$, $p(\beta = 0) > 0.05$; presence: $\beta = 1.8 \times 10^{-3}$, $\alpha = 0.47$, $t_{206} = 0.69$, $p(\beta = 0) > 0.05$; shape: $\beta = -1.6 \times 10^{-3}$, $\alpha = 0.54$, $t_{216} = 0.65$, $p(\beta = 0) > 0.05$; colour: $\beta = -2.1 \times 10^{-3}$, $\alpha = 0.55$, $t_{220} = 0.88$, $p(\beta = 0) > 0.05$; position: $\beta = 8.3 \times 10^{-4}$, $\alpha = 0.56$, $t_{211} = 0.33$, $p(\beta = 0) > 0.05$; relative distance: $\beta = 4.1 \times 10^{-3}$, $\alpha = 0.29$, $t_{213} = 1.70$, $p(\beta = 0) > 0.05$]. The steepest of these slopes (that of relative distance) is equivalent to an increase in the proportion of correct responses of 0.20 over the 48 trials of the experiment. The lack of systematic change to performance over trials further indicates that performance was not affected by time within the experimental session.

3.2.4 *The influence of display duration on performance.* A further factor that was altered systematically was the presentation duration, which was varied randomly between 1 and 10 s. Figure 4 shows the effect of image presentation time on performance for all question types combined. There is a clear increase in performance with presentation time [slope $(\beta) = 2.3 \times 10^{-2}$, intercept $(\alpha) = 0.44$, $t_{138} = 4.09$, $p(\beta = 0) < 0.05$].

As noted above, a direct quantitative comparison of performance across the different question types is not appropriate. However, a qualitative comparison of how presentation time affects the extent of encoding of information between conditions is possible.

Figure 5 shows that presentation time had differential effects for the six question types. Performance in gist questions was not influenced by viewing time [figure 5a; $\beta = -8.3 \times 10^{-4}$, $\alpha = 0.87$, $t_{114} = 0.09$, $p(\beta = 0) > 0.05$]. However, other question types

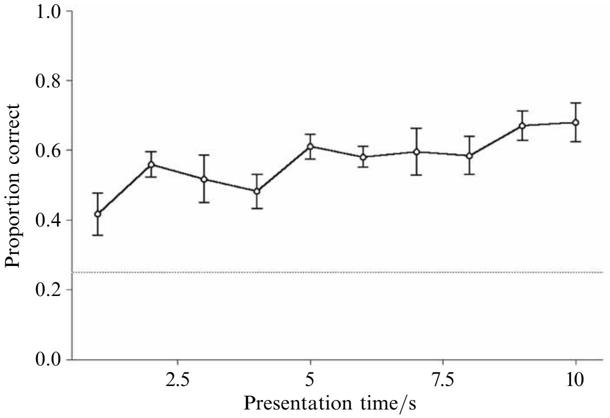

Figure 4. The effect of image presentation time upon the proportion of questions answered correctly by participants. Performance increased over the range of presentation times, which varied from 1 to 10 s. Error bars indicate standard error between participants.

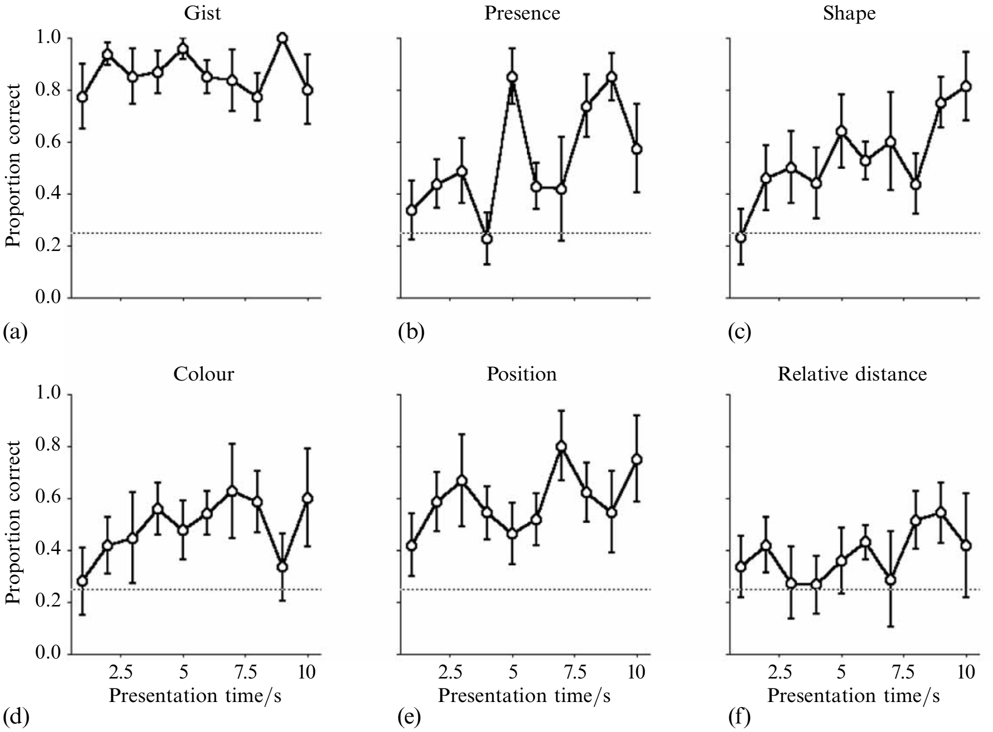

Figure 5. The effect of presentation time upon the proportion of questions answered correctly by participants for each of the six types of question. Performances in response to different types of information were affected by presentation time in different ways. Error bars indicate standard error between participants.

appeared to be influenced in some way by presentation time. Presence and shape questions both appear to show strong effects of presentation time upon performance and this is verified by linear regression of the response data for each of these two question types [presence: $\beta = 4.5 \times 10^{-2}$, $\alpha = 0.29$, $t_{111} = 3.36$, $p(\beta = 0) < 0.05$; shape: $\beta = 4.3 \times 10^{-2}$, $\alpha = 0.29$, $t_{105} = 3.26$, $p(\beta = 0) < 0.05$]. Colour showed an overall insignificant slope [figure 5d; $\beta = 2.1 \times 10^{-2}$, $\alpha = 0.37$, $t_{108} = 1.41$, $p(\beta = 0) > 0.05$]. However, closer examination of the data revealed that initially performance does appear to increase with presentation time, but plateaus after 4–6 s. Trends in performance for absolute position [figure 5e; $\beta = 2.0 \times 10^{-2}$, $\alpha = 0.47$, $t_{114} = 1.40$, $p(\beta = 0) > 0.05$] and relative distance [figure 5f; $\beta = 2.1 \times 10^{-2}$, $\alpha = 0.28$, $t_{104} = 1.56$, $p(\beta = 0) > 0.05$] were both non-significant.

Performances after 1 and 2 s of viewing can be used to consider the early progress of the accumulation of the five types of information tested. Performance for gist questions is above chance after 1 s of viewing ($t_{10} = 4.22$, $p < 0.05$) and remains at this level thereafter. For questions testing absolute position information, performance after 1 s of viewing was no different from chance ($t_{11} = 1.38$, $p > 0.05$), but after 2 s was significantly above chance ($t_{12} = 2.94$, $p < 0.05$). For the other question types, performance was no different from chance after either 1 s (presence: $t_{13} = 0.73$, $p > 0.05$; shape: $t_{12} = 0.18$, $p > 0.05$; colour: $t_{11} = 0.22$, $p > 0.05$; relative distance: $t_{10} = 0.71$, $p > 0.05$) or 2 s of viewing (presence: $t_{12} = 1.99$, $p > 0.05$; shape: $t_{10} = 1.66$, $p > 0.05$; colour: $t_{12} = 1.55$, $p > 0.05$; relative distance: $t_{12} = 1.56$, $p > 0.05$).

4 Discussion

When participants view images of natural scenes and are required to attend to multiple items and aspects of the scene, they are able to extract and represent several different types of information concurrently. Participants perform significantly better than chance in answering questions about all six types of information tested. Previous studies of information extraction from scenes have focused on two main types of information likely to be extracted and integrated into representation: gist (eg Biederman 1981) and absolute spatial layout (eg Hochberg 1968). While our data confirm that both gist and absolute spatial layout can be extracted, they are by no means the only types of information extracted from the images viewed; a wide range of types of information describing the scene is represented. The retention of multiple types of information about scenes is consistent with suggestions by several researchers (Aginsky and Tarr 2000; de Graef 1992; Henderson 1992; Hollingworth and Henderson 2002; Rayner and Pollatsek 1992; Rensink 2000). However, while representations described by these researchers comprise more than one information type, they typically only increase the spectrum of two or three types of abstract information. De Graef (1992) and Rayner and Pollatsek (1992) describe representations in which object identity and spatial layout are the primary information types. Rensink (2000) proposes a tripartite representation comprising gist, layout, and long-term scene schema (an amalgamation of long-term memories of similar scenes and expectations based on previous experience). Aginsky and Tarr (2000) propose that representations include scene layout and object surface properties. Hollingworth and Henderson (2002) suggest that representations contain richly encoded local object details indexed to scene layout. On the basis of data presented in this paper, we can extend and integrate the above ideas by proposing an abstract representation in which all of these types of information are retained.

Data from the control questionnaire can be used to consider the constraints and validity of comparisons between question types in our study. In two hundred and eleven of the two hundred and forty questions, control responses were distributed evenly across the 4 possible response options. This result confirms that target and foils were not discriminable on the basis of contextual viability alone and required exposure and encoding of the scene for discriminations to be made. The remaining twenty-nine questions may have perhaps been less well balanced between all four options, but the distribution of these twenty-nine questions over the five question types were uniform ($\chi^2 = 3.59$, $p > 0.05$). Overall performances in the five object-specific question types show that the questions appeared to be equally difficult because performances were not significantly different from one another (figure 3). It remains, however, that the questions may not have been matched in terms of perceptual discriminability (see section 2). Since we cannot be certain that this was not the case, we feel that it is inappropriate to make direct quantitative comparisons between question types. However, the time courses of information assimilation can be compared qualitatively between the different question types; it is to this issue that our discussion now turns.

The data presented in this paper enable us to comment upon the ways in which representations are built up. After 1 s, performance for gist questions is near maximal and does not change significantly with prolonged exposure, suggesting that assimilation of gist information into representations is accomplished within this short time scale. This time course is consistent with previous studies of the extraction of gist information from scenes, which suggest that gist can be extracted fully within hundreds of milliseconds (eg Biederman 1981). In contrast, performance for all other questions is at chance after 1 s, requiring more prolonged exposure for information accumulation.

After 2 s, performance in response to questions testing absolute position is significantly above chance. However, performance remains no different from chance for presence, shape, colour, or relative distance questions. It appears, therefore, that a sketch of gist and rudimentary spatial layout is constructed soon after the start of exposure to a new scene. Prominence of gist and layout in representations is consistent with the suggestion by Rensink (2000) and with conclusions drawn by Aginsky and Tarr (2000) who suggest that position and presence are more 'salient' in the representational system than surface properties such as colour.

With prolonged exposure to a scene, we are able to add detail to this initial gist-and-layout sketch, by increasing the accuracy of absolute layout information, and by adding information about object presence, shape, colour, and relative distance. For all of these five types of information, assimilation continues over a period of several seconds. Continued assimilation of information over the course of several seconds is consistent with a recent report by Melcher (2001). Melcher found that for viewing complex computer-generated scenes, the number of items recalled by participants increased as viewing time was increased from 1 to 2, and then 4 s. The reported time course of assimilation of object-identity information is much greater than the time courses for object recognition obtained by RSVP techniques, which can be achieved for images presented for times as brief as 32 ms (Delorme et al 2000). Presumably, the discrepancy reflects the time needed to consolidate the object information to form a representation or memory trace, rather than simply to identify it.

While information assimilation continues over the course of several seconds for presence, shape, colour, absolute layout, and relative distance, the precise effect of time appears to vary. For presence and shape information, assimilation has not reached an obvious plateau even after 10 s, suggesting that further viewing would allow increased precision of representation. Conversely, colour reaches a clear plateau after 4 s, with no apparent increase in performance with further exposure. The data for absolute and relative layout information do not reach a maximum during the display duration of this experiment. These data suggest that there are qualitative differences in the ways in which information is assimilated for the various question types.

It is now evident that the overall performances in each question type shown in figure 3 are an underestimation of the potential extent of extraction of each of the categories of information. It appears that the eventual detail of the representation can be quite high, with performances reaching 70%–80% correct, at least in the cases of presence, shape, and absolute position. A comparison of performance on the first and second questions after each image presentation demonstrates that the represented information from which responses are drawn does not become visually masked and survives for at least several seconds after the end of viewing. Hence our data describe a multipartite abstract representation assimilated over a number of seconds of viewing in which eventual detail can be quite rich and that is stable within the time course of several seconds after viewing.

5 Conclusions

Multiple types of information can be extracted concurrently as observers view images of natural scenes. The extent and rate of extraction vary between the different types of information tested, suggesting differing relative priorities for the representation of this information. The proposed abstract representation consists of an early sketch comprising gist information and a crude spatial layout of the scene. With prolonged viewing, layout information is refined and detail is added to the representation in terms of item presence, shape, colour, and relative distances. All types of information apart from gist continue to be assimilated over several seconds of viewing, at differing rates. By the end of 10 s of viewing, representational faithfulness can be quite high, with performances for presence, shape, and absolute position information of around 70% – 80%.

References

Aginsky V, Tarr M J, 2000 "How are different properties of a scene encoded in visual memory?" *Visual Cognition* **7** 147 – 162

Biederman I, 1981 "On the semantics of a glance at a scene", in *Perceptual Organization* Eds M Kubovy, J R Pomerantz (Hillsdale, NJ: Lawrence Erlbaum Associates) pp 213 – 253

Bridgeman B, Hendry D, Stark L, 1975 "Failure to detect displacement of the visual world during saccadic eye movements" *Vision Research* **15** 719 – 722

Bridgeman B, Mayer M, 1983 "Failure to integrate visual information from successive fixations" *Bulletin of the Psychonomic Society* **21** 285 – 286

Chun M M, Nakayama K, 2000 "On the functional role of implicit visual memory for the adaptive deployment of attention across scenes" *Visual Cognition* **7** 65 – 81

Delorme A, Richard G, Fabre-Thorpe M, 2000 "Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans" *Vision Research* **40** 2187 – 2200

Dennett D C, 1991 *Consciousness Explained* (Boston, MA: Little, Brown & Co)

Friedman A, 1979 "Framing pictures: the role of knowledge in automatized encoding and memory for gist" *Journal of Experimental Psychology: General* **108** 316 – 355

Gibson J J, 1966 *The Senses Considered as Perceptual Systems* (New York: Appleton-Century-Crofts)

Gibson J J, 1979 *The Ecological Approach to Visual Perception* (Boston, MA: Houghton Mifflin)

Graef P de, 1992 "Scene-context effects and models of real-world perception", in *Eye Movements and Visual Cognition: Scene Perception and Reading* Ed. K Rayner (New York: Springer) pp 243 – 259

Grimes J, 1996 "On the failure to detect changes in scenes across saccades", in *Perception: Vancouver Studies in Cognitive Science* Ed. K Atkins (New York: Oxford University Press) pp 89 – 110

Henderson J M, 1992 "Object identification in context—the visual processing of natural scenes" *Canadian Journal of Psychology—Revue Canadienne de Psychologie* **46** 319 – 341

Henderson J M, 1994 "Two representational systems in dynamic visual identification" *Journal of Experimental Psychology: General* **123** 410 – 426

Henderson J M, 1997 "Transsaccadic memory and integration during real-world object perception" *Psychological Science* **8** 51 – 55

Henderson J M, Hollingworth A, 1999 "The role of fixation position in detecting scene changes across saccades" *Psychological Science* **10** 438 – 443

Henderson J M, Siefert A B C, 1999 "The influence of enantiomorphic transformation on transsaccadic object integration" *Journal of Experimental Psychology: Human Perception and Performance* **25** 243 – 255

Hobbes T, 1651/1839 "Leviathan, or, The Matter, Form, & Power of a Common-Wealth, Ecclesiastical and Civill", in *The English Works of Thomas Hobbes of Malmesbury* volume 3, Ed. W Molesworth (London: John Bond) entire volume

Hobbes T, 1656/1839 "Elements of philosophy", in *The English Works of Thomas Hobbes of Malmesbury* volume 1, Ed. W Molesworth (London: John Bond) entire volume

Hochberg J, 1968 "In the mind's eye", in *Contemporary Theory and Research in Visual Perception* Ed. R N Haber (New York: Holt) pp 309 – 331

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

Intraub H, 1980 "Presentation rate and the representation of briefly glimpsed pictures in memory" *Journal of Experimental Psychology: Human Learning and Memory* **6** 1 – 12

Intraub H, 1981 "Rapid conceptual identification of sequentially presented pictures" *Journal of Experimental Psychology: Human Perception and Performance* **7** 604 – 610

Irwin D E, 1991 "Information integration across saccadic eye-movements" *Cognitive Psychology* **23** 420 – 456

Irwin D E, 1993 "Perceiving an integrated visual world", in *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence and Cognitive Neuroscience* Eds D E Mayer, S Kornblum (Cambridge, MA: MIT Press) pp 121 – 142

Land M F, Mennie N, Rusted J, 1999 "The roles of vision and eye movements in the control of activities of daily living" *Perception* **28** 1311 – 1328

Light L L, Berger D E, 1974 "Memory for modality: Within-modality discrimination is not automatic" *Journal of Experimental Psychology* **103** 854 – 860

Light L L, Berger D E, Bardales M, 1975 "Trade-off between memory for verbal items and their visual attributes" *Journal of Experimental Psychology: Human Learning and Memory* **1** 188 – 193

McConkie G W, Rayner K, 1976 "Identifying the span of the effective stimulus in reading: literature review and theories of reading", in *Theoretical Models and Processes of Reading* Eds H Singer, R B Ruddell (Newark, NJ: International Reading Association) pp 137 – 162

McConkie G W, Zola D, 1979 "Is visual information integrated across successive fixations in reading?" *Perception & Psychophysics* **25** 221 – 224

Mack A, 1970 "An investigation of the relationship between eye and retinal image movement in the perception of movement" *Perception & Psychophysics* **8** 291 – 298

MacKay D M, 1973 "Visual stability and voluntary eye movements", in *Handbook of Sensory Physiology* Ed. R Jung (Berlin: Springer) pp 307 – 331

Mandler J M, Johnson N S, 1976 "Some of the thousand words a picture is worth" *Journal of Experimental Psychology: Human Learning and Memory* **2** 529 – 540

Mandler J M, Parker R E, 1976 "Memory for descriptive and spatial information in complex pictures" *Journal of Experimental Psychology: Human Learning and Memory* **2** 38 – 48

Mandler J M, Ritchey G H, 1977 "Long-term memory for pictures" *Journal of Experimental Psychology: Human Learning and Memory* **3** 386 – 396

Mandler J M, Seegmiller D, Day J, 1977 "On the coding of spatial information" *Memory & Cognition* **5** 10 – 16

Marr D, 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (San Francisco, CA: W H Freeman)

Melcher D, 2001 "Persistence of visual memory for scenes—A medium-term memory may help us to keep track of objects during visual tasks" *Nature* **412** 401

O'Regan J K, 1982 "Solving the real mysteries of visual perception—the world as an outside memory" *Canadian Journal of Psychology—Revue Canadienne de Psychologie* **46** 461 – 488

O'Regan J K, Lévy-Schoen A, 1983 "Integrating visual information from successive fixations—does trans-saccadic fusion exist?" *Vision Research* **23** 765 – 768

O'Regan J K, Noë A, 2001 "A sensorimotor account of vision and visual consciousness" *Behavioral and Brain Sciences* **24** 939 – 973; discussion 973 – 1031

Pezdek K, Whetstone T, Reynolds K, Askari N, Dougherty T, 1989 "Memory for real-world scenes—the role of consistency with schema expectation" *Journal of Experimental Psychology: Learning Memory and Cognition* **15** 587 – 595

Pollatsek A, Rayner K, 1992 "What is integrated across fixations?", in *Eye Movements and Visual Cognition: Scene Perception and Reading* Ed. K Rayner (New York: Springer) pp 166 – 191

Rayner K, Pollatsek A, 1992 "Eye-movements and scene perception" *Canadian Journal of Psychology—Revue Canadienne de Psychologie* **46** 342 – 376

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

Rensink R A, O'Regan J K, Clark J J, 1995 "Image flicker is as good as saccades in making large scene changes invisible" *Perception* **24** 26 – 27

Rensink R A, O'Regan J K, Clark J J, 1997 "To see or not to see: The need for attention to perceive changes in scenes" *Psychological Science* **8** 368 – 373

Rensink R A, O'Regan J K, Clark J J, 2000 "On the failure to detect changes in scenes across brief interruptions" *Visual Cognition* **7** 127 – 145

Sanocki T, Epstein W, 1997 "Priming spatial layout of scenes" *Psychological Science* **8** 374 – 378

Simons D J, 1996 “In sight, out of mind: When object representations fail” *Psychological Science* **7** 301 – 305

Simons D J, Levin D T, 1997 “Change blindness” *Trends in Cognitive Science* **1** 261 – 267

Whipple W R, Wallach H, 1978 “Direction-specific motion thresholds for abnormal image shifts during saccadic eye movement” *Perception & Psychophysics* **24** 349 – 355