frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 22 February 2012
doi: 10.3389/fpsyg.2012.00032

Recognition and memory for briefly presented scenes

Mary C. Potter*

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Edited by:
Gabriel Kreiman, Harvard Medical School, USA

Reviewed by:
Rufin VanRullen, Centre de Recherche Cerveau et Cognition Toulouse, France
Gabriel Kreiman, Harvard Medical School, USA

***Correspondence:**
Mary C. Potter, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 46-4125, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.
e-mail: mpotter@mit.edu

Three times per second, our eyes make a new fixation that generates a new bottom-up analysis in the visual system. How much is extracted from each glimpse? For how long and in what form is that information remembered? To answer these questions, investigators have mimicked the effect of continual shifts of fixation by using rapid serial visual presentation of sequences of unrelated pictures. Experiments in which viewers detect specified target pictures show that detection on the basis of meaning is possible at presentation durations as brief as 13 ms, suggesting that understanding may be based on feedforward processing, without feedback. In contrast, memory for what was just seen is poor unless the viewer has about 500 ms to think about the scene: the scene does not need to remain in view. Initial memory loss after brief presentations occurs over several seconds, suggesting that at least some of the information from the previous few fixations persists long enough to support a coherent representation of the current environment. In contrast to marked memory loss shortly after brief presentations, memory for pictures viewed for 1 s or more is excellent. Although some specific visual information persists, the form and content of the perceptual and memory representations of pictures over time indicate that conceptual information is extracted early and determines most of what remains in longer-term memory.

Keywords: picture perception, rapid serial visual presentation, picture memory, detection, feedforward processing, masking, search

INTRODUCTION

THE PROBLEM

We make three or four eye fixations each second, all day long. That suggests that 250 ms is long enough to identify most objects, but is it enough to recognize a whole scene? How much do we remember about each fixation and for how long? To develop and maintain information about the environment, we need some form of visual memory that spans several fixations. But, carry-over from the preceding fixation lacks detail (e.g., Irwin, 1992; Irwin and Andrews, 1996; Henderson and Hollingworth, 1999). Indeed, we overlook major changes in a scene if the scene is interrupted for as little as 80 ms – the phenomena of change blindness (e.g., Rensink et al., 1997, 2000) and boundary extension (Intraub and Richardson, 1989). We are not blind, however, to changes that affect gist or changes to objects that we are attending or are about to fixate. Thus, the information that we carry over from a fixation seems to be limited and to be meaningful rather than purely visual. Our memory for pictures is poor, however, for unrelated pictures presented in a continuous sequence at rates in the range of eye fixations (Potter and Levy, 1969).

On the other hand, we have good long-term memory for pictures viewed for 1 or 2 s (Nickerson, 1965; Shepard, 1967; Potter and Levy, 1969; Standing, 1973) and we can remember them in considerable detail, whether they represent single objects (Brady et al., 2008) or more complex scenes (Konkle et al., 2010). Some highly distinctive information must be retained from pictures that we have viewed for a few seconds.

HOW LONG DOES IT TAKE TO UNDERSTAND A VISUAL OBJECT OR SCENE?

There is no simple answer to this question: the answer depends on what one's criterion for understanding is. One could measure the time it takes to name the picture, but even for objects with well-known names, the naming time of about 900 ms includes search for the word after one has already recognized what the object is. Another measure of understanding is the time to decide whether the scene or object matches some description, such as "animal." This category detection task turns out to be considerably faster (around 600 ms) than the time to name a picture (Potter and Faulconer, 1975), but still includes the time to generate the yes or no response. A still faster response is the time between the onset of a pair of pictures and the initiation of an eye movement to (for example) the picture of an animal or a face (e.g., Kirchner and Thorpe, 2006; Crouzet et al., 2010): this selective decision can take as little as 100 ms. All the measures just discussed include the time for the information to pass from the retina to the visual cortex as well as decision and response processes that occur after identification (e.g., Potter, 1983). Still shorter times can be obtained by using measures of brain responses such as event related potentials (ERPs) that do not include any overt response.

Single masked stimuli

A different approach is to control the time available for processing a stimulus such as a picture, and to measure the minimum presentation time required for successful identification. However, because of visual persistence (continued activation in the visual system after a stimulus ends), the duration of the physical stimulus is not closely related to the effective duration of the stimulus: for example, a picture presented for only 20 ms followed by a blank screen may be as readily processed as one shown for 100 ms. A common method to solve that problem is to use a backward mask such as a patterned stimulus that follows the picture. Such a mask is thought to interrupt processing of the picture. If that is the case, then by varying the stimulus onset asynchrony (SOA) between the onset of the target and that of the mask, a minimal processing time required for identification can be determined. For example, when a single picture is presented, followed by a visual mask (such as a collage of colored paper cut into small circles and irregular shapes), it is possible to remember as many as half the pictures with a duration as short as 50 ms, and 80% are remembered at a duration of 120 ms (Potter, 1976; see **Figure 3**, discussed below).

A continuing problem with the logic of the masking procedure, however, is that the neural basis for the effect is not well-understood: does the masked stimulus continue to be processed, perhaps unconsciously, after the mask appears, or does processing instantly stop? This question is especially relevant to feedforward models, discussed below. Moreover, a backward mask does not necessarily interrupt all processing – and the amount of interference is a complex function of the visual relation between the target and mask, the semantic (conceptual) relation, and the SOA between target and mask. (For a more complete account of the complexities of backward and forward masking, see Eriksen and Eriksen, 1971, and Breitmeyer and Ogmen, 2006.) With very short SOAs, the visual relation may be a stronger determinant of the effectiveness of the mask than the conceptual relation, but as the SOA increases, the reverse may be the case (e.g., Potter, 1976; Loftus and Ginn, 1984). Indeed, the most important factor may be whether the following mask is itself a stimulus that the viewer needs to attend to and report on: see rapid serial visual presentation (RSVP) below and the discussion of visual versus conceptual masking. I return to the question of masking in Section “Detecting Pictures at Ultra-High Rates: Evidence for Feedforward Processing?”

Perception of objects in settings

A further question is whether knowledge of co-occurrences between objects and settings influences the initial perception of a scene, or whether (as suggested by Hollingworth and Henderson, 1998, 1999) objects and settings in a given picture are first understood independently and only later merged. In one set of studies (Davenport and Potter, 2004), pictured objects such as a football player or a priest were superimposed, either congruently or incongruently, on background settings such as a football field or the interior of a cathedral (**Figure 1**). The pictures were presented for 80 ms, with a backward noise mask, and the participant was instructed to report the foreground object, the background setting, or both. In each case performance was better in the congruent than the incongruent condition, suggesting that objects and background are processed interactively, early in processing. In a further study (Davenport, 2007) one or two objects were presented on a background. The relation between the two objects (whether they would be likely to be present in the same scene or not) had an effect on report that was additive with the effect of congruency with the background: that is, the relationship between the two objects, as well as each of the objects’ relation to the background, influenced report of the objects. Joubert et al. (2007, 2008) carried out similar studies, finding that objects in congruent contexts were responded to faster than in incongruous contexts.

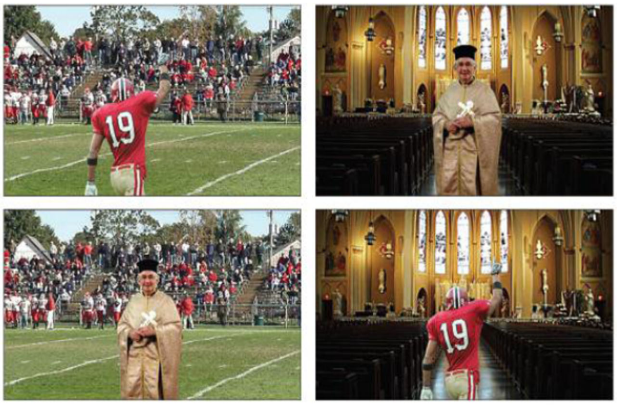

FIGURE 1 | Examples of consistent scenes (top row) and inconsistent scenes (bottom row). In all cases, the central figure or object was photoshopped onto the background picture. Participants viewed only one of the versions of a given picture, and reported either the central object, the background setting, or both. From Davenport and Potter (2004).

RAPID SERIAL VISUAL PRESENTATION

In studies using backward masking of pictures, each trial consists of a single picture and a mask. In normal vision, however, the eyes make a continuous sequence of fixations. What happens when pictures are presented in a continuous stream at durations in the range of eye fixations, and participants try to remember all of them? To investigate this question, Potter and Levy (1969) used RSVP (Forster, 1970) to show participants sequences of 16 unrelated pictures (**Figure 2**). They varied the presentation duration between 125 and 2000 ms. To test recognition memory following the presentation, the pictures were presented one at a time intermixed with 16 new pictures (distractors). Participants responded Yes, Maybe, or No. **Figure 3** shows the proportion of yes responses, corrected for guessing[1]. Clearly, 250 ms was not enough time to assure memory of a picture a few minutes later, as only about half the pictures presented for that duration were correctly recognized. With a presentation of 2 s, more than 90% of the pictures were remembered, consistent with studies showing that long-term memory for pictures viewed for a few seconds is excellent (Nickerson, 1965; Shepard, 1967; Standing, 1973; Brady et al., 2008). Thus, even though a single masked picture may be reported correctly after it is viewed for as little as 50 ms, as shown in the left-hand function in **Figure 3** (Potter, 1976), it takes considerably longer to process pictures to the same level when they are presented in a continuous stream in which all the pictures are to be attended.

[1] A one-high-threshold formula was used to correct for guessing, $P_{\text{corr}} = [P(\text{TY}) - P(\text{FY})]/[1 - P(\text{FY})]$, where TY is a correct yes response and FY is a false yes response. This guessing correction is used in all data figures.

FIGURE 2 | An illustration of an RSVP sequence of pictures.

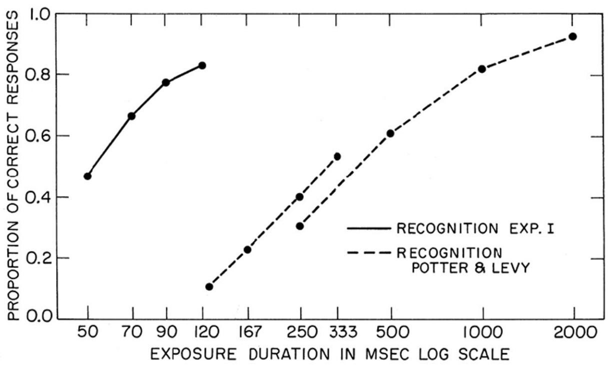

FIGURE 3 | Proportion of pictures recognized following single masked presentations (solid curve, Potter, 1976, Exp. 3) and proportion recognized after RSVP (dashed curves, two groups with different ranges of presentation durations; Potter and Levy, 1969). Data are corrected for guessing (see text footnote 1). Figure from Potter, 1976.

VISUAL VERSUS CONCEPTUAL MASKING

What makes an RSVP sequence hard to remember is not the briefness of the pictures, but the fact that each picture is immediately followed by another. With a single masked picture, the viewer can continue to process the information after the mask appears; evidently that is not possible with a continuous sequence in which all the pictures need to be attended. In a study by Intraub (1980) pictures were presented for 110 ms in an RSVP sequence, and only 20% were remembered later, whereas when a blank interstimulus interval (ISI) was added after each picture, the percent remembered increased steadily as the ISI increased, to 84% with an ISI of 1390 ms: this result shows that a viewer can voluntarily continue to process and code into memory a brief picture after it is no longer in view. Similarly, a study showed that pictures presented for 173 ms in an RSVP sequence were poorly remembered; if a blank of 827 ms was added after each picture, memory was almost as good as if the pictures were each shown for 1000 ms (Potter et al., 2004).

Once the SOA between the picture and the following visual mask is about 100 ms, memory depends little on the actual duration of presentation, but instead on the total uninterrupted time the viewer has to continue to think about the picture. Thus, if a viewer is shown a sequence of pictures that alternate between a short duration of 112 ms and a long duration of 1500 ms, the instruction to attend only to the brief pictures results in memory for about 63% of the brief pictures and only 54% of the long pictures: intention to continue processing the brief pictures actually leads to better memory than for the long-duration pictures (Intraub, 1984). These results show that there is a distinction between visual and conceptual masking: visual masking occurs primarily with short SOAs (under 100 ms), whereas conceptual masking (due to attention to a following stimulus) occurs with SOAs up to 500 ms or more (Potter, 1976; see also Intraub, 1980, 1981; Loftus and Ginn, 1984; Loftus et al., 1988).

DETECTING PICTURES

Given the poor memory for pictures presented at durations in the range of eye fixations, does it take longer than a single fixation to understand a novel scene? Do we even momentarily understand pictures shown for only 250 ms in an RSVP sequence? Intuitively, we may think that if we had understood what a picture was about, we would surely remember it for at least a few minutes. Perhaps viewers fail to remember briefly presented pictures because they did not comprehend them in a single glimpse, whereas normally they could continue to look at something until it is recognized. Yet, when viewing pictures at a rate as high as 10/s, one's impression is that each picture can be seen and understood momentarily: is that an illusion? To address this question, participants were asked to detect a target picture in an RSVP sequence that was named or shown to them before the sequence (Potter, 1975, 1976). Detection was surprisingly good with either kind of cue, even at durations as short as 113 ms/picture (**Figure 4**). It is not surprising that showing the actual picture in advance enabled viewers to detect it; the surprise was that performance was almost as good when viewers had only a name that captured the picture's conceptual gist. Intraub (1979, 1980, 1981) showed that viewers could even detect pictures described by a negative category – for example, they could detect the only picture in a sequence that was not an animal and could report the identity of the non-animal picture, showing that they had understood it.

Further evidence that a picture's identity can be retrieved quickly is shown in a detection study (Potter et al., 2010) in which participants looked for two instances of pictures in a specified category such as "dinner food" or "bird," and had to report the specific identity of each instance (e.g., *swan* and *eagle*). As shown in **Figure 5**, the presentation duration was 107 ms. Participants were able to do this successfully even when the two targets were presented in immediate succession, although they showed an attentional blink for the second target when the SOA between targets was 213 ms, an effect typically observed in search tasks.

Detection and memory when multiple pictures are presented simultaneously

Viewers can process serially presented pictures remarkably rapidly, but can they process two or more pictures presented

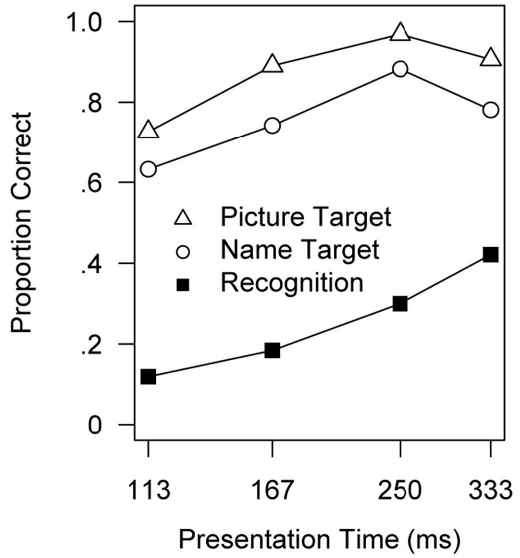

FIGURE 4 | Detection of a target picture in an RSVP sequence of 16 pictures, given a picture of the target or a name for the target, as a function of the presentation time per picture. Also shown is later recognition performance in a group that simply viewed the sequence, and then was tested for recognition. Results are corrected for guessing (see text footnote 1). From Potter (1976).

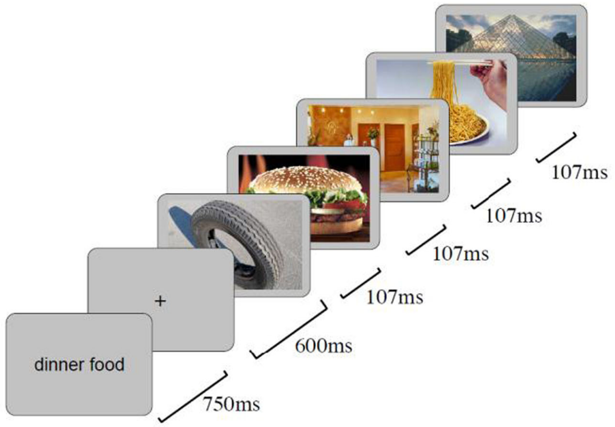

FIGURE 5 | An example of an RSVP sequence in a search experiment in which participants reported the specific names of two exemplars of the search category. Here the exemplars are hamburger and spaghetti. From Potter et al. (2010).

simultaneously? When the task is to detect a specified target, the results suggest that detection is relatively successful with up to four simultaneous pictures, in RSVP streams consisting of eight successive four-item arrays (Potter and Fox, 2009). Each array was a combination of none to four pictures, with texture masks in the non-picture locations. The RSVP sequence was presented at 240, 400, or 720 ms per array. Although accuracy decreased a little as the number of pictures in an array increased, detection was generally good: even at 240 ms per array with four simultaneous pictures, 59% of the targets were detected, with 9% false yeses (cf. Rousselet et al., 2002, 2004a,b). In contrast, when viewers simply tried to remember the pictures, performance was much lower overall, particularly when there was more than one picture in the array.

Rapid memory loss for pictures seen briefly in RSVP: serial position effects in memory testing

People can understand pictures presented briefly, but forget most of them a few minutes later. When the recognition test begins immediately, the first one or two pictures tested are likely to be remembered well, but there is rapid loss over the next several seconds of testing (Potter et al., 2002, 2004; Endress and Potter, in press): that is, there is a strong serial position effect in the memory test. A related question is the effect of serial position in presentation. If information is being lost as more pictures are presented (either because of retroactive interference or because of the passage of time), we would expect a recency effect. (Note, however, that the final picture is not tested in these experiments as it is not masked.) Potter et al. (2002) found no evidence of a recency effect, even with sequences of 10 or 20 pictures. Increasing the memory set size did decrease the extra benefit of early testing somewhat, but not by causing selective forgetting of pictures early in the list.

WHAT IS THE NATURE OF THIS SHORT-LASTING MEMORY FOR PICTURES?

The time course of forgetting after viewing an RSVP sequence of pictures contrasts with that of *change blindness,* the apparently immediate loss of detailed information about a single picture, once it is no longer in view. Change blindness is the inability of viewers to detect a change in one feature of a picture, and it has been observed when a blank interval as short as 80 ms intervenes between the initial and changed versions; at longer intervals, the problem is even more acute (see Rensink et al., 1997; Simons and Levin, 1997). Imposing a short blank between views is necessary to obscure the transient that would mark the location of the change if there were no interval. Change blindness suggests that those details were not perceived in the first place, or that many specifics of a picture are lost immediately, or that the next picture updates the similar preceding picture without leaving a record of the changed details. Change blindness is, however, a very different phenomenon than the forgetting observed after an RSVP sequence. Whereas on a change blindness trial there is no question that the picture remains the same in most respects and is thus seen as the same picture, in the RSVP experiments considered here the question is whether a given test picture is familiar at all. Thus, change blindness studies assess the level of detail in immediate memory for a picture, whereas here we are interested in the persistence of a representation sufficient to make the picture seem familiar when presented again among dissimilar pictures.

Could the short-lasting memory for pictures be *iconic memory* (e.g., Sperling, 1960) or very short-term memory (VSTM) as described by Phillips and his colleagues (Phillips, 1983; Potter and Jiang, 2009)? The answer is no. Iconic memory is a very brief form of relatively literal perceptual memory (although see Coltheart, 1983, for a somewhat different characterization), but it cannot account for the fleeting picture memory found with an immediate recognition test after an RSVP sequence, because iconic memory is eliminated by noise masking and under photopic conditions it lasts no longer than about 300 ms. VSTM is a form of short-lasting visual memory observed in experiments such as those of Phillips and Christie (1977), who presented viewers briefly with a 4 × 4 matrix in which an average of 8 random squares were white, and then tested memory by presenting a second matrix that was either identical to the preceding one or had one white cell added or deleted. VSTM, unlike iconic memory, is capacity-limited, with an estimated capacity of three or four items. In Phillips and Christie's study, the most recent matrix could be maintained for several seconds in VSTM, provided that no other such matrices were presented in the interval and the participant continued to attend to the remembered matrix. In contrast, in RSVP studies multiple pictures are presented and one or more to-be-attended pictures intervene between presentation and testing.

A likely contributor to short-term memory for pictures is conceptual short-term memory (CSTM), a short-lasting memory component proposed by Potter (1993, 1999, 2010) that represents conceptual information about current stimuli, such as the meaning of a picture, or meanings of words and sentences computed as one reads or listens. The reasons for regarding this brief memory representation as conceptual rather than (say) perceptual include its apparent role in rapid selection between two words on the basis of meaning, in relation to context (Potter et al., 1993, 1998), and its putative role in sequential visual search tasks like those considered here in which the targets are defined by meaning or category rather than by physical form. During the brief time that information about stimuli is in CSTM, associative links enable extraction of whatever structure is present (such as sentence structure or the gist of a picture) or allow the stimulus to be compared to a target specification, in a search task. Any momentarily active information that does not become incorporated into such a structure (such as the irrelevant meaning of an ambiguous word, or a non-target picture) will be quickly forgotten.

Conceptual versus visual-perceptual memory

In relation to rapidly presented pictures, the CSTM claim is that some pictures are adequately encoded and consolidated into longer-term memory during even brief viewing, but others are represented only in CSTM and are vulnerable to interference in the first few seconds after viewing. However, we do not know whether the picture representation that persists for several seconds in the studies we have reviewed here is sufficiently abstract to be considered conceptual rather than wholly or partly perceptual. Do viewers remember only the picture's conceptual content or gist, or do they also remember visual features such as color, shape, and layout? Work of Irwin and Andrews (1996), Gordon and Irwin (2000), and Henderson (1997) suggests that this representation of previous fixations may be at least partially conceptual rather than literal, inasmuch as viewers may not notice literal changes that are conceptually consistent with the earlier fixation. Earlier work showed that the gist of a scene is understood quickly even though the scene may then be forgotten (fairly) rapidly (e.g., Potter, 1976; Intraub, 1980, 1981), which is consistent with the assumption that conceptual information is abstracted rapidly. Intraub (1981) showed, however, that viewers can remember some specific pictorial information such as the colors and layout, along with the gist.

The relative roles of such specific pictorial information and more abstract conceptual information were explored in Potter et al. (2004). They contrasted a conceptual and a pictorial recognition test of picture memory. In the pictorial test, participants made yes–no decisions to the five pictures they had just seen (excluding the sixth final picture), mixed with five new pictures. In the conceptual test, they made yes–no decisions to descriptive verbal titles of the pictures, mixed with titles of unseen pictures. The rate of presentation was 173 ms/picture; the 10-item recognition test after each trial took about 8 s. The assumption was that test pictures provide both visual and conceptual information, whereas titles provide only conceptual information. If the benefit of immediate testing is that viewers only briefly preserve purely pictorial information, then the title test should reduce the benefit of early testing, but should be fairly equivalent to the picture test later in testing. That was just what they found, as shown in **Figure 6**. In a more recent study (Endress and Potter, in press) the advantage of testing recognition with pictures rather than titles was maintained throughout the test, suggesting that some more detailed information (perceptual or conceptual) beyond that captured by a title does persist over the 8-s test, even though memory for both forms of information continues to decline.

In a further test of the conceptual basis of memory, Potter et al. (2004) included in the recognition test occasional pictures that matched the title – the gist – of one of the old pictures. They called such a picture a *decoy.* The decoy was visually different from the old picture it replaced: side by side, it was easy to tell that the pictures were not the same. If viewers rely on a conceptual or gist representation of the presented pictures, they

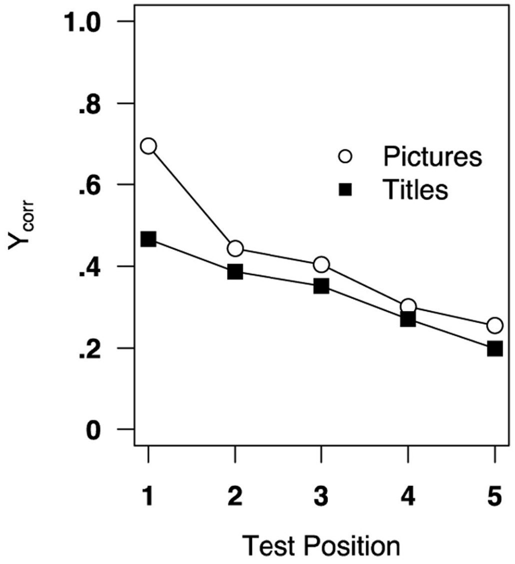

FIGURE 6 | Recognition test of five pictures shown in RSVP for 173 ms/picture; the test used pictures or titles. Guessing-corrected results (see text footnote 1) are shown as a function of relative position in the recognition test, which included five new pictures (distractors). From Potter et al. (2004).

should make false yes responses more often to decoys than to new distractors that are not conceptually similar to any of the pictures they have just seen, and that was what happened. Overall, participants recognized 52% of the old pictures, falsely recognized 30% of the decoys, and falsely recognized 15% of the other new pictures, showing some susceptibility to conceptual decoys. On the other hand, had viewers remembered only the conceptual gist of the picture, equivalent to the information in a title, they would have been as likely to "recognize" the decoys as to recognize the correct picture, and clearly that was not the case. Again, then, there is evidence that viewers can remember some particulars of pictures, beyond the gist, even at high rates of presentation.

SHORT-LASTING MEMORY: SUMMARY

In sum, initial memory for a glimpsed picture (seen for the equivalent of a single fixation) is fairly accurate, but declines markedly over the first few recognition tests (or across an unfilled delay of 5 s). There is some evidence that the initial stronger memory includes specifically visual information, whereas after a delay the memory is primarily conceptual. That is, detailed visual information about a picture is lost more rapidly than conceptual information. Accurate visual information may be important for maintaining and updating scene representations over fixations, but conceptual memory seems to be the basis for longer-term, organized knowledge.

As stated earlier, unlike briefly glimpsed pictures, memory for pictures viewed for a second or more can be highly accurate, at least when viewers are paying attention. Yet, as work reviewed here shows, normal eye fixations are too brief to guarantee good memory. They are, however, long enough to make it highly likely that the viewer will have understood what he or she saw, at least momentarily, allowing the viewer to continue looking or to take appropriate action. The rapid comprehension of the gist of a scene suggests that scenes are initially perceived as wholes – like single objects. Although the gist of pictured scenes can be extracted rapidly, exactly how that is done remains unclear. Work of Oliva and her collaborators has given us some ideas about how visual properties such as layout, texture, color, and the like can enable rapid categorization of natural scenes, street scenes, and interiors (Oliva, 2005).

DETECTING PICTURES AT ULTRA-HIGH RATES: EVIDENCE FOR FEEDFORWARD PROCESSING?

WHAT CONSTITUTES EVIDENCE FOR FEEDFORWARD PROCESSING?

It is widely assumed that under normal viewing conditions perception results from a combination of feedforward and feedback connections (Di Lollo et al., 2000; Enns and Di Lollo, 2000; Lamme and Roelfsema, 2000; Hochstein and Ahissar, 2002). Feedback from higher to lower levels in the visual system takes time, however. At presentation durations of about 50 ms or less with masking, some have proposed that there would not be time for feedback to arrive before the lower-level activity has been interrupted by the mask, so that perception would be restricted to the information in the forward pass of neural activity from the retina through the visual system (Perrett et al., 1992; Thorpe and Fabre-Thorpe, 2001; Hung et al., 2005; Liu et al., 2009). In feedforward models of the visual system (Serre et al., 2007a,b) units which process the stimulus are hierarchically arranged: units representing small regions of space (receptive fields) in the retina converge to represent larger and larger receptive fields and more abstract information along a series of pathways from V1 to inferotemporal cortex (IT) and further on to prefrontal cortex (PFC). Visual experience tunes this hierarchical structure, which acts as a filter that permits recognition of a huge range of objects and scenes in a single forward pass of processing. Yet, there is little direct evidence that the feedforward process is able to identify objects and scenes accurately, without feedback.

Conscious perception

The ability to identify or remember a stimulus is commonly taken to mean that the viewer was conscious of the stimulus, and in the work discussed here I make the assumption that consciousness is shown by the ability to report on the stimulus by responding to a target picture or by recognizing its title or the picture itself, in a memory test. (See, however, evidence for unconscious effects, in Feedforward Processing and Masked Priming.) There is a debate about whether a single forward pass is sufficient for conscious perception. A reentrant process providing feedback may be necessary to achieve understanding and conscious awareness (Lamme and Roelfsema, 2000; Dehaene and Naccache, 2001; Hochstein and Ahissar, 2002). As mentioned earlier, it has been suggested that a threshold duration of about 50 ms must be exceeded if a backward mask is presented or the stimulus will not be consciously perceived: consciousness of a stimulus may require sufficient time "to establish sustained activity in recurrent cortical loops" (Del Cul et al., 2007) or to ignite a network required for conscious perception (Deheane et al., 1998). These authors thus hypothesize that viewers cannot become conscious of a stimulus on the basis of a single feedforward sweep, without time for any feedback. Detection in RSVP at durations of 50 ms/picture or less should be impossible if there is such a threshold, because there is too little time to establish a long-range cortical loop before a picture has been overwritten by subsequent pictures. As reviewed in the next section, however, there is evidence that perception is sometimes possible with very brief, masked stimuli, a result that suggests that feedforward processing may be sufficient for conscious perception under some conditions.

EVIDENCE FOR PROCESSING OF VERY BRIEF STIMULI

RSVP responses: monkey neurons and humans

Recordings of individual neurons in the cortex of the anterior superior temporal sulcus (STSa) of monkeys who viewed a set of pictures of monkey faces and other objects via RSVP at various rates up to 72/s (14 ms) showed that neurons respond to a preferred picture above chance, even at 14 ms (Keysers et al., 2001, 2005). In a detection study with human observers using the same set of pictures, but presenting them in seven-picture RSVP sequences, the participants were shown a target picture before each sequence. They detected the target above chance at 14 ms/picture, although detection improved as the duration per picture was increased. In another condition in the same study, recognition of a target picture was tested immediately after the sequence, instead of being shown before the sequence. Participants were still above chance at 14 ms/picture, but performance was not as good as when they saw the target picture in advance. A possible problem with the human study is that the pictures were repeated across trials and hence became familiar, which might have allowed participants to focus on simple features in order to spot the target.

Further evidence: detection and immediate memory

A study by Potter, Wyble, and McCourt (in preparation) replicated some of the behavioral conditions of Keysers et al. (2001), but crucially, instead of showing the picture target, they gave only a descriptive name for the target (e.g., *smiling couple*), before or immediately after an RSVP sequence of six pictured scenes. Moreover, each picture was presented only once, and none of the pictures were familiar to the participants. Thus, participants had only a conceptual representation of the target they were to detect. The RSVP sequence was presented at durations between 13 and 80 ms. Even at a presentation duration of 13 ms, the targets were detected or recognized above chance: that is, the probability of a correct detection on target-present trials was significantly higher than the probability of a false detection response on target-absent trials. In addition, at the end of each trial participants were shown two pictures, both matching the target name, and asked to indicate which one they had seen. If they had correctly detected the target they were more likely to pick the right picture than if they had failed to detect it. Thus, viewers could detect and retain at least briefly information about named targets they had never seen before, at an RSVP duration as short as 13 ms.

These results are consistent with the claim of the feedforward model that pictures can be understood in a single feedforward sweep even when attention has not been directed to a specific category in advance.

How long does recognition memory last, after a very brief presentation?

Studies of the monkey visual system using single-cell recordings show that cortical neurons that are selective for particular objects can "recognize" multiple objects in parallel at levels as high as the inferior temporal cortex. Something similar in human perception might account for the ability to remember rapidly presented pictures. In monkeys, this initial parallel process is followed within 150 ms by competitive inhibition of all but the one relevant object in a given receptive field, at least when there is a task that defines the relevant stimulus (e.g., Chelazzi et al., 1998; see Rousselet et al., 2004b, for a review). The large and overlapping receptive fields found in the inferior temporal cortex may allow for temporary representation in parallel of several successive pictures presented at a high rate, followed by competitive suppression that favors the most salient picture. That could account for the capacity to detect a target by name immediately after the presentation of six pictures. If high-level representations of several of the pictures in the sequence were activated, however, it is likely that mutual competition would soon decrease their activation. In the experiments with RSVP described above (Potter et al., in preparation), a delay of 5 s in naming the target picture after the sequence did decrease accuracy.

FEEDFORWARD PROCESSING AND MASKED PRIMING

In masked priming studies, a brief presentation of a word becomes invisible when it is followed by a second unmasked word to which the participant must respond (Forster and Davis, 1984; Dehaene and Naccache, 2001). The unreportable prime word still has an effect on the following word, showing that it must have been unconsciously identified. Indeed, the term "priming" implies a process that benefits a later stimulus in the absence of memory for the initial stimulus, the prime. Given that the prime may have been presented for 50 ms or more in typical masked priming experiments (above the threshold for perception with a noise mask), why is the participant not conscious of the prime? In such studies the focus of attention is on the second stimulus, and its longer duration permits it to receive full, recurrent processing that may interfere with retention of the more vulnerable information from the prime that was extracted during the feedforward sweep. When, as in Potter et al. (in preparation), the masking stimulus is the same duration as the preceding target stimulus and is another picture that is to be attended, a duration of 13 ms is clearly sufficient, on a significant proportion of trials, to drive detection, identification, and (at least briefly) recognition memory for the pictures preceding the final picture. Whether the phenomenon of masked priming has the same neural basis as the reportable detection observed with RSVP tasks such as those of Keysers et al. (2001, 2005) and Potter et al. (in preparation) remains to be determined.

DISCUSSION: ULTRA-RAPID PROCESSING AND FEEDFORWARD PROCESSING

Both the results of Keysers et al. (2001, 2005) with monkey neurons and with humans, as well as the results of Potter et al. (in preparation) with humans, show that pictures can be detected and briefly remembered when presented in a short sequence as rapid as 72 or 75 pictures/s. Even when no target is specified in advance, a name presented immediately after the sequence can prompt memory for the corresponding picture. These results support a feedforward model that can extract a picture's conceptual meaning in a single forward sweep of information with an input of only 13 or 14 ms, even when the picture is preceded and followed by other pictures.

But are there other explanations for successful detection when the presentation duration is brief and masked by successive pictures? One possibility is that at high rates of presentation several temporally adjacent pictures are integrated, like double or multiple camera exposures. Certainly the subjective impression in viewing rapid sequences is that the pictures merge into each other visually, as though they were overlaid. Possibly viewers simply recover the target from such a composite representation, rather than detecting it during the feedforward pass.

A related possibility is that following masks do not interrupt processing immediately. As mentioned in Section "Single Masked Stimuli," the neural basis for masking is not well-understood. There is evidence (Keysers et al., 2001, 2005) that neurons activated in higher visual areas may continue to be active for about 60 ms longer than the SOA between stimuli in an RSVP sequence, although the neuron would have maintained activity much longer (e.g., for 350 ms) without a following stimulus. That is, a following stimulus does eventually suppress activity, but only after about

60 ms of overlap. If the time course of activity in STSa neurons in monkeys is representative of the activity in the temporal cortex underlying human performance, then there would be about 73 ms of activity produced by each picture in an RSVP sequence at 13 ms/picture, possibly allowing time for feedback between levels. However, because as many as five different pictures would be active at the same time (at 13 ms/picture), it is not evident that all could be receiving feedback. That, of course, might account for the decrease in the probability of a correct detection as the rate of presentation increases.

Nonetheless, the feedforward hypothesis remains a strong contender as an explanation of picture identification with very brief presentation durations. In the absence of a specific model for how feedback might assist reportable detection of brief targets, the feedforward hypothesis seems the most plausible account.

HOW LONG DOES IT TAKE TO UNDERSTAND A PICTURED SCENE?

Returning to the question considered in the introduction, what can be concluded about the time required to identify a scene? If the question is the minimum exposure duration (prior to a mask) that is required, 13 or 14 ms is sometimes enough, when the mask is another scene. But if the question is the time from arrival at the retina to correct categorization, then the most reliable measures available at present are reaction time measures, the most sensitive of which is an eye movement to the appropriate target in a choice situation. For detection of a face (when a picture with a face is presented together with another picture), that time can be as short as 100 ms, with a mean time of 140 ms (Crouzet et al., 2010); detection of a vehicle takes somewhat longer.

Detection of a pre-specified target is likely to be more rapid than comprehension of a new scene that the viewer is told nothing about. That is one reason that recognition memory for a picture, even immediately after a short RSVP sequence, is less accurate than detection (Keysers et al., 2001), but another reason is that forgetting begins very quickly after presentation (Potter et al., 2002). As reviewed in Section "Rapid Serial Visual Presentation," momentary comprehension is no guarantee of subsequent memory, even seconds later. We comprehend rapidly, but then we forget selectively, on the basis of what is relevant to our current goals and needs.

As reviewed above, the speed of detection has suggested to a number of investigators that accurate comprehension or categorization can occur on the basis of the early feedforward sweep of visual information, without requiring feedback loops from higher to lower levels and back. The behavior of individual neurons in the human inferior temporal cortex and homologous areas in monkey cortex, reviewed elsewhere in this issue, provides another window on the timing of picture categorization that gives some support to the feedforward hypothesis. When a scene is complex or its components are unfamiliar, we undoubtedly require more processing time and probably more than a single fixation to comprehend it. However, a lifetime of knowledge of the world that is built into our visual system appears to allow immediate understanding of most scenes, based on the initial sweep of visual information when the scene is presented.

ACKNOWLEDGMENTS

This work was supported by Grant MH47432 from the National Institute of Mental Health.

REFERENCES

Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. (2008). Remembering thousands of objects with high fidelity. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14325–14329.

Breitmeyer, B. G., and Ogmen, H. (2006). *Visual Masking: Time Slices through Conscious and Unconscious Vision*, 2nd Edn. New York: Oxford University Press.

Chelazzi, L., Duncan, J., Miller, E. K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiol.* 80, 2918–2940.

Coltheart, M. (1983). Iconic memory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 302, 283–294.

Crouzet, S. M., Kirchner, H., and Thorpe, S. J. (2010). Fast saccades toward faces: face detection in just 100 ms. *J. Vis.* 10, 16.1–17.

Davenport, J. L. (2007). Consistency effects between objects in scene processing. *Mem. Cognit.* 35, 393–401.

Davenport, J. L., and Potter, M. C. (2004). Scene consistency in object and background perception. *Psychol. Sci.* 15, 559–564.

Dehaene, S., and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79, 1–37.

Deheane, S., Kergsberg, M., and Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14529–14534.

Del Cul, A., Baillet, S., and Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biol.* 5, 2408–2423. doi:10.1371/journal.pbio.0050260

Di Lollo, V., Enns, J. T., and Rensink, R. A. (2000). Competition for consciousness among visual events: the psychophysics of reentrant visual pathways. *J. Exp. Psychol. Gen.* 129, 481–507.

Endress, A. D., and Potter, M. C. (in press). Early conceptual and linguistic processes operate in independent channels. *Psychol. Sci.* [Epub ahead of print].

Enns, J. T., and Di Lollo, V. (2000). What's new in visual masking? *Trends Cogn. Sci. (Regul. Ed.)* 4, 345–352.

Eriksen, C. W., and Eriksen, B. A. (1971). Visual perceptual processing rates and backward and forward masking. *J. Exp. Psychol.* 89, 306–313.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Percept. Psychophys.* 8, 215–221.

Forster, K. I., and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 680–698.

Gordon, R. D., and Irwin, D. E. (2000). The role of physical and conceptual properties in preserving object continuity. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 136–150.

Henderson, J. M. (1997). Transsaccadic memory and integration during real-world object perception. *Psychol. Sci.* 8, 51–55.

Henderson, J. M., and Hollingworth, A. (1999). High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271.

Hochstein, S., and Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* 36, 791–804.

Hollingworth, A., and Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *J. Exp. Psychol. Gen.* 127, 398–415.

Hollingworth, A., and Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychol. (Amst.)* 102, 319–343.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866.

Intraub, H. (1979). The role of implicit naming in pictorial encoding. *J. Exp. Psychol. Hum. Learn.* 5, 1–12.

Intraub, H. (1980). Presentation rate and the representation of briefly glimpsed pictures in memory. *J. Exp. Psychol. Hum. Learn.* 6, 1–12.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 604–610.

Intraub, H. (1984). Conceptual masking: the effects of subsequent visual events on memory for pictures. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 115–125.

Intraub, H., and Richardson, M. (1989). Wide-angle memories of close-up scenes. *J. Exp. Psychol. Learn. Mem. Cogn.* 15, 1989, 179–187.

Irwin, D. E. (1992). Memory for position and identity across eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 307–317.

Irwin, D. E., and Andrews, R. V. (1996). "Integration and accumulation of information across saccadic eye movements," in *Attention and Performance XVI: Information Integration in Perception and Communication*, eds T. Inui and J. L. McClelland (Cambridge, MA: MIT Press), 125–155.

Joubert, O., Fize, D., Rousselet, G. A., and Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *J. Vis.* 8(13), 11.1–18.

Joubert, O. R., Rousselet, G. A., Fize, D., and Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. *Vision Res.* 47, 3286–3297.

Keysers, C., Xiao, D. K., Földiák, P., and Perrett, D. I. (2001). The speed of sight. *J. Cogn. Neurosci.* 13, 90–101.

Keysers, C., Xiao, D.-K., Földiák, P., and Perrett, D. I. (2005). Out of sight but not out of mind: the neurophysiology of iconic memory in the superior temporal sulcus. *Cogn. Neuropsychol.* 22, 316–332.

Kirchner, H., and Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Res.* 46, 1762–1776.

Konkle, T., Brady, T. F., Alvarez, G. A., and Oliva, A. (2010). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol. Sci.* 21, 1551–1556.

Lamme, V. A. F., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579.

Liu, H., Agam, Y., Madsen, J. R., and Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62, 281–290.

Loftus, G. R., and Ginn, M. (1984). Perceptual and conceptual masking of pictures. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 435–441.

Loftus, G. R., Hanna, A. M., and Lester, L. (1988). Conceptual masking: how one picture captures attention from another picture. *Cogn. Psychol.* 20, 237–282.

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: a demonstration of capacity. *Can. J. Psychol.* 19, 155–160.

Oliva, A. (2005). "Gist of the scene," in *Encyclopedia of Neurobiology of Attention*, eds L. Itti, G. Rees and J. K. Tsotsos (San Diego, CA: Elsevier), 251–256.

Perrett, D., Hietanen, J., Oram, M., and Benson, P. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 335, 23–30.

Phillips, W. A. (1983). Short-term visual memory. *Philos. Trans. R. Soc. Lond. B Biol. Sci* 302, 295–309.

Phillips, W. A., and Christie, D. F. M. (1977). Components of visual memory. *Q. J. Exp. Psychol. (Hove)* 29, 117–133.

Potter, M. C. (1975). Meaning in visual search. *Science* 187, 965–966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *J. Exp. Psychol. Hum. Learn. Mem.* 2, 509–522.

Potter, M. C. (1983). "Representational buffers: the eye-mind hypothesis in picture perception, reading, and visual search," in *Eye Movements in Reading: Perceptual and Language Processes*, ed. K. Rayner (New York: Academic Press), 423–437.

Potter, M. C. (1993). Very short-term conceptual memory. *Mem. Cognit.* 21, 156–161.

Potter, M. C. (1999). "Understanding sentences and scenes: the role of conceptual short term memory," in *Fleeting Memories: Cognition of Brief Visual Stimuli*, ed. V. Coltheart (Cambridge, MA: MIT Press), 13–46.

Potter, M. C. (2010). Conceptual short term memory. *Scholarpedia* 5, 3334

Potter, M. C., and Faulconer, B. A. (1975). Time to understand pictures and words. *Nature* 253, 437–438.

Potter, M. C., and Fox, L. F. (2009). Detecting and remembering simultaneous pictures in a rapid serial visual presentation. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 28–38.

Potter, M. C., and Jiang, Y. V. (2009). "Visual short-term memory," in *Oxford Companion to Consciousness*, eds T. Bayne, A. Cleeremans, and P. Wilken (Oxford: Oxford University Press), 436–438.

Potter, M. C., and Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *J. Exp. Psychol.* 81, 10–15.

Potter, M. C., Moryadas, A., Abrams, I., and Noel, A. (1993). Word perception and misperception in context. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 3–22.

Potter, M. C., Staub, A., and O'Connor, D. H. (2004). Pictorial and conceptual representation of glimpsed pictures. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 478–489.

Potter, M. C., Staub, A., Rado, J., and O'Connor, D. H. (2002). Recognition memory for briefly-presented pictures: the time course of rapid forgetting. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 1163–1175.

Potter, M. C., Stiefbold, D., and Moryadas, A. (1998). Word selection in reading sentences: preceding versus following contexts. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 68–100.

Potter, M. C., Wyble, B., Pandav, R., and Olejarczyk, J. (2010). Picture detection in RSVP: features or identity? *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1486–1494.

Rensink, R. A., O'Regan, J. K., and Clark, J. J. (2000). On the failure to detect changes in scenes across brief interruptions. *Vis. Cogn.* 7, 127–145.

Rensink, R. A., O'Regan, J. R., and Clark, J. J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* 8, 368–373.

Rousselet, G., Fabre-Thorpe, M., and Thorpe, S. J. (2002). Parallel processing in high level categorization of natural images. *Nat. Neurosci.* 5, 629–630.

Rousselet, G. A., Thorpe, S. J., and Fabre-Thorpe, M. (2004a). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Res.* 44, 877–894.

Rousselet, G., Thorpe, S. J., and Fabre-Thorpe, M. (2004b). How parallel is visual processing in the ventral pathway? *Trends Cogn. Sci. (Regul. Ed.)* 8, 363–370.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007a). A quantitative theory of immediate visual recognition. *Prog. Brain Res.* 165, 33–56.

Serre, T., Oliva, A., and Poggio, T. (2007b). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429.

Shepard, R. N. (1967), Recognition memory for words, sentences, and pictures. *J. Mem. Lang.* 6, 156–163.

Simons, D. J., and Levin, D. T. (1997). Change blindness. *Trends Cogn. Sci. (Regul. Ed.)* 1, 261–267.

Sperling, G. (1960). The information available in brief visual presentations. *Psychol. Monogr.* 74, 1–29.

Standing, L. (1973). Learning 10,000 pictures. *Q. J. Exp. Psychol. (Hove)* 25, 207–222.

Thorpe, S., and Fabre-Thorpe, M. (2001). Seeking categories in the brain. *Science* 291, 260–263.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 November 2011; accepted: 28 January 2012; published online: 22 February 2012.

*Citation: Potter MC (2012) Recognition and memory for briefly presented scenes. Front. Psychology **3**:32. doi: 10.3389/fpsyg.2012.00032*

This article was submitted to Frontiers in Perception Science, a specialty of Frontiers in Psychology.