Atten Percept Psychophys
DOI 10.3758/s13414-016-1096-5

Failure to detect meaning in RSVP at 27 ms per picture

John F. Maguire[1] · **Piers D. L. Howe**[1]

Abstract The human visual system has the remarkable ability to rapidly detect meaning from visual stimuli. Potter, Wyble, Hagmann, and McCourt (*Attention, Perception, & Psychophysics*, *76*, 270–279, 2014) tested the minimum viewing time required to obtain meaning from a stream of pictures shown in a rapid serial visual presentation (RSVP) sequence containing either six or 12 pictures. They reported that observers could detect the presence of a target picture specified by name (e.g., *smiling couple*) even when the pictures in the sequence were presented for just 13 ms each. Potter et al. claimed that this was insufficient time for feedback processing to occur, so feedforward processing alone must be able to generate conscious awareness of the target pictures. A potential confound in their study is that the pictures in the RSVP sequence sometime contained areas with no high-contrast edges, and so may not have adequately masked each other. Consequently, iconic memories of portions of the target pictures may have persisted in the visual system, thereby increasing the effective presentation time. Our study addressed this issue by redoing the Potter et al. study, but using four different types of masks. We found that when adequate masking was used, no evidence emerged that observers could detect the presence of a specific target picture, even when each picture in the RSVP sequence was presented for 27 ms. On the basis of these findings, we cannot rule out the possibility that feedback processing is necessary for individual pictures to be recognized.

Keywords RSVP · Scene perception · Visual perception · Recognition · Visual masking

The human visual system has the remarkable ability to detect meaning from a visual stimulus and to bring it to conscious awareness. Our normal mode of vision is aided by rapid eye movements, allowing three or four briefly presented different views of the external world to be observed every second. The human visual system naturally processes images at this presentation rate; however, it can detect meaning from images presented at far shorter durations (Keysers, Xiao, Foldiak, & Perrett, 2001; Potter, 1975, 1976).

Potter, Wyble, Hagmann, and McCourt (2014) showed observers a series of single pictures in a rapid serial visual presentation (RSVP) sequence. They found that observers could determine the presence or absence of a specific picture even when the pictures in the sequence were presented for just 13 ms each. The implication that observers can process a picture that is presented for just 13 ms challenges established feedback theories of visual perception that postulate that neural activity needs to propagate from the primary visual cortex up to higher cortical areas and back to the primary visual cortex before recognition can occur at the level of detail required for an individual picture to be detected (Bar et al., 2006; Del Cul, Baillet, & Dehaene, 2007; Di Lollo, 2012; Koivisto, 2012; Lamme, 2006; Lamme & Roelfsema, 2000; Tononi, 2004). It is highly unlikely that this feedback process can occur within 13 ms. Indeed, Potter et al. (2014) argued that it would take a minimum of 50 ms for feedback to occur. This estimate is consistent with the findings of Lamme and Roelfsema (2000), who reported that the response latencies at

✉ John F. Maguire
j.maguire@cfa.vic.gov.au

Piers D. L. Howe
pdhowe@unimelb.edu.au

[1] School of Psychological Sciences, University of Melbourne, 12th Floor Redmond Barry Building, Melbourne, Victoria 3010, Australia

Published online: 05 April 2016

any hierarchical level of the visual system are about 10 ms after those at the previous level. Assuming that a minimum of five levels would need to be traversed as the activity propagates from V1 to higher cortical areas and back again, this would imply that this feedback process is unlikely to occur in less than 50 ms. Thus, the Potter et al. finding that recognition can occur within 13 ms suggests that recognition at the level of an individual picture can occur in a purely feedforward manner.

However, the findings of Potter et al. (2014) may have been confounded due to inadequate masking. Masking occurs when the visual perception of a stimulus is impaired by the presentation of a temporally adjacent and (usually) spatially overlapping stimulus; forward masking occurs when the target picture is preceded by the mask; backward masking occurs when the target picture is followed by a mask (Breitmeyer & Öğmen, 2006; Keysers & Perrett, 2002). In Potter et al. (2014), the target picture was never the first or last picture in the RSVP sequence. Consequently, it was both forward and backward masked by the other pictures in the sequence. However, unless the target picture is adequately masked, portions of it may persist as an iconic memory within the visual system for 200–300 ms after its presentation (Atkinson & Shiffrin, 1968; Kovacs, Vogels, & Orban, 1995; Sperling, 1960), and it is not clear that the target picture was effectively masked by the other pictures in the sequence, as these were pictures of natural scenes, and so had not been specifically designed to act as masks.

Potter et al.'s (2014) use of natural scenes as masks was not without justification. Natural scenes had previously been used as masks in visual detection studies (Intraub, 1984; Potter, 1976) and been found to be the most effective masks in a study that compared four different mask types (natural scenes, scene textures, phase-randomized scenes, and white noise; Loschky, Hansen, Sethi, & Pydimarri, 2010). However, many of the natural scenes used by Potter et al. (2014) contained extended areas where there were no high contrast edges (e.g., expanses of sky). Presumably, these areas would not have masked the corresponding areas of the target picture, thereby allowing these portions of the target picture to be processed for longer than the specified presentation duration, possibly allowing for feedback connections to be established.

This same masking confound may have occurred in other visual detection studies. For example, Evans, Horowitz, and Wolfe (2011) performed an RSVP study in which observers viewed a sequence of six images presented in rapid succession to determine whether a particular precued target image was present. The target image was a natural scene and, if present, was always the second image in the sequence. The other images were colored texture synthetic masks created using Portilla and Simoncelli's (2000) algorithm. It was found that observers could perform this task at 83 % accuracy even when each picture in the sequence was presented for just 20 ms. As with the Potter et al. (2014) study, it is possible that in the Evans et al. study the target images were not adequately masked by the preceding and following images, so the effective presentation time of the target images may have been longer than 20 ms. Indeed the masks used in the Evans et al. study have characteristics similar to those of the 1/*f* noise masks that we used in our second experiment and that we found to be the least effective of the four types of mask that we investigated.

With the present study, we sought to address this potential confound by first replicating the findings of Potter et al. (2014) using natural scene masks. Once this was done, we then extended their study by developing more effective masks so as to better control the effective presentation time of the target pictures. Various stimuli can be used to mask a target picture. A 1/*f* noise mask has visual properties similar to those of natural scenes (Field, 1987) and has been commonly used in visual detection studies of natural scenes (Greene & Oliva, 2009; Loftus & Ginn, 1984; Serre, Oliva, & Poggio, 2007). Another common form of masking used in visual detection studies is geometric masks (Davenport & Potter, 2004; Lähteenmäki, Hyönä, Koivisto, & Nummenmaa, 2015; Naccache, Blandin, & Dehaene, 2002). Lähteenmäki et al. suggested that high-contrast edges and complex color patterns are more effective masks than the commonly used scrambled rectangular picture masks. So our study tested the effectiveness of geometric masks containing overlapping colored squares. Lastly, in our study we used a mask comprising densely packed lines. This was done because edges strongly drive V1 activity (Hubel & Wiesel, 1962), so an image comprising many lines is likely to be an effective mask for activity in this cortical area. Thus, in total, our study used four mask types: natural scenes, 1/*f* noise, colored squares, and masks solely comprising different colored lines (Fig. 1)

Method

Stimuli and apparatus

The stimuli were presented using MATLAB running Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) on a 21-in. CRT monitor with 1,280 × 1,024 resolution and a 75-Hz refresh rate. All stimuli were centrally located on the monitor, viewed at a distance of 60 cm in a dark room, and subtended approximately 7.3 × 7.3 deg of visual angle (°). The target stimuli were photographs from natural scenes sourced from a publically available collection (see the Appendix). Every picture was presented only once, and all were novel to the observers. Four separate experiments were run. The experiments differed only in the type of mask used (Fig. 1). These experiments used natural-scene masks, 1/*f* noise masks, geometric masks, and colored line masks.

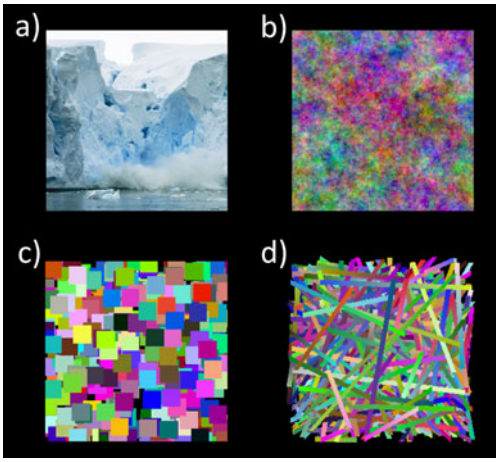

Fig. 1 The four masks used in the experiment: (**a**) natural-scene mask, (**b**) 1/*f* noise mask, (**c**) geometric mask, and (**d**) colored lines mask

Procedure

Other than using different masks, all four experiments were identical and used the principal design features of Potter et al.'s (2014) study, except that, to maximize the chances of participants identifying the targets, in our study the targets were both precued and postcued, whereas in the Potter et al. study they were either precued or postcued (but not both). Whereas Potter et al. used both 12-item and six-item RSVP sequences, in our study we utilized only a six-item RSVP sequence, again so as to make it easier for the participants to detect the target images.

Each trial started by advising the participants of how many trials remained in that block and precueing the participant to the target picture for that trial using a one or two word description of the target (e.g., *golf course*, *kitchen*, *beach*, etc.). This written message remained on the screen until the participant clicked the computer mouse to start the trial. A centrally located fixation cross was then presented on a blank screen for 200 ms. The six-item RSVP sequence immediately followed, after which the participants were immediately reminded of the target identity and asked to indicate by clicking on the words "Yes" or "No" with the computer mouse whether they had seen the target image in the RSVP sequence. The written reminder of the target identity remained on the screen until the participant had responded. In all, 37 possible image categories were presented, as listed in the Appendix. At most one example of each image category was presented in the RSVP sequence, thereby ensuring that our cue (i.e., the target description) was unambiguous, even when natural-scene masks were used. Figure 2 shows a typical trial for the experiment that used natural-scene masks. The other experiments used an identical procedure, except that the natural-scene masks were replaced with masks that depended on the experiment. For example, in the final experiment, line masks were used. For that experiment, five of the six images in the RSVP sequence would comprise line masks, and one of the images would be a natural scene.

Design

Each experiment was run separately using a different set of participants. For a given experiment, each participant began with a practice block of 20 trials in which all the RSVP pictures were presented for 136 ms each. Eight blocks of the main experiment then followed. As with Potter et al. (2014), within a block, all of the RSVP images were presented for the same duration. However, different blocks utilized different image durations. These durations were presented in descending order (80, 53, 27, and 13 ms), and then repeated. Thus, in the first block each image in each RSVP sequence was presented for 80 ms. Each block contained 20 trials: 15 target-present and five target-absent. The target picture was presented equally often in Positions 2 to 5 of the RSVP sequence in a random order. The target was never presented in Position 1 or 6, to ensure that it was always both forward and backward masked by the other pictures in the RSVP sequence. For those experiments that did not use natural-scene masks, every trial contained exactly one natural-scene image. Thus, regardless of whether or not the target image was present, a natural-scene image was always presented in the RSVP sequence. This meant that just detecting the presence of a natural scene among the masks was not sufficient to perform the task. Participants needed to determine the identity of the natural scene to determine whether it corresponded to the target category.

Participants

A power analysis revealed that to replicate the significant finding of Experiment 1 of Potter et al. (2014), at $t = 13$ ms with $\beta = .95$ and $\alpha = .05$, would require a sample size of 11 participants. Adopting a more conservative approach, we opted instead to use a sample size of 16 participants in each of our experiments, since this was the number used by Potter et al. in their experiment. For our experiments, our participants were all first-year psychology students, and all gave informed consent before participating in the experiment. All had normal or corrected-to-normal vision, achieving a minimum of 20/25 visual acuity, as tested with the Good-Lite Near Vision Chart, and normal color vision, as tested by the Ishihara Test for Colour Blindness. In the experiment that used 1/*f* masks, one participant had to be replaced for responding "Yes" to more than 50 % of all target-absent trials, since this suggested that she was randomly guessing (Potter et al., 2014).

Results and discussion

Experiment 1: Natural-scene masks

Stimulus presentation times were checked, and trials were discarded if computer timing errors occurred. Across participants, an average of 0.3 % of trials were discarded for this

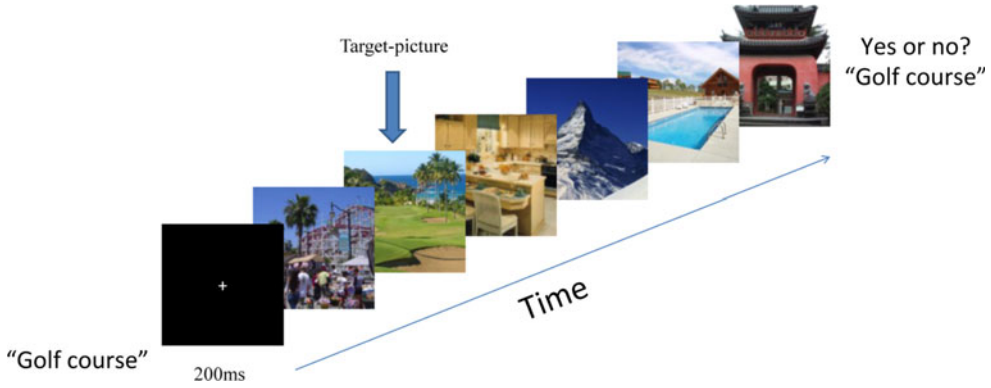

Fig. 2 Representation of a typical trial in the experiment that used natural-scene masks. The precue (e.g., "Golf course") was presented prior to the appearance of a fixation cross. The RSVP sequence then occurred, after which the participant was immediately reminded of the target's identity and asked whether or not it had been present

reason. To compensate for a possible response bias, detection sensitivity was calculated using d' according to the log–linear method (Hautus, 1995). A one-way repeated measures analysis of variance (ANOVA) revealed a significant main effect of duration, $F(3, 45) = 6.86$, $p = .001$, $\eta_p^2 = .314$: Detection accuracy increased with increasing presentation durations. Figure 3 displays the mean d's for all four durations in all four experiments.

We successfully replicated the findings of Potter et al. (2014). In particular, one-tailed single-sample t tests showed that d' was significantly above chance ($d' = 0$) at all durations, including the crucial 13-ms duration (see Table 1). Comparing our results to the precueing condition of Experiment 1 of Potter et al., we find that our d' values are similar to what they reported at 13 ms, but are considerably less than they reported at 80 ms. The target names used in our study were more generic than those used by Potter et al., and this might have contributed to this difference (Potter & Hagmann, 2015). Additionally, unlike the Potter et al. study, in our study observers did not receive feedback on their accuracy, and this may have further reduced d' at the longer durations.

Experiment 2: 1/f noise masks

To test the assertion that natural scenes may not be effective masks, in our second experiment 1/*f* noise masks were used. These masks are commonly used and provide a more consistent masking pattern across the entire extent of the target picture than do natural-scene masks.

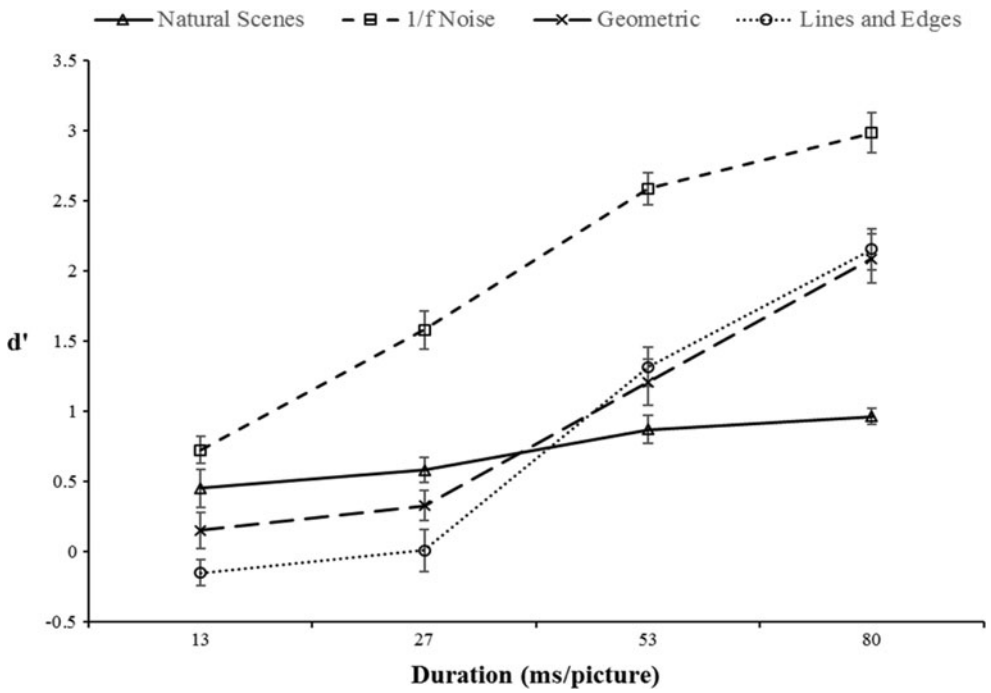

Fig. 3 Results for the four experiments: d' as a function of picture duration and mask type. Error bars depict the standard errors of the means

Table 1 Experiment 1: Descriptive statistics and *t* test results for *d′* at each presentation duration

Duration	Mean	Standard Deviation	*t*	*p*	Cohen's *d*
13	0.45	0.53	3.38	.004	0.85
27	0.58	0.36	6.46	<.001	1.62
53	0.87	0.41	8.59	<.001	2.15
80	0.96	0.23	16.47	<.001	4.12

df = 15

An average of 0.08 % of trials were discarded due to computer timing errors. A one-way repeated measures ANOVA showed a significant main effect of duration, $F(3, 45) = 67.27$, $p < .001$, $\eta_p^2 = .818$. Detection accuracy decreased with decreasing presentation durations. Figure 3 displays the mean *d′*s for all four durations.

One-tailed single-sample *t* tests showed that *d′* was significantly above chance at all durations (see Table 2). A two-way mixed-effects ANOVA comparing the two experiments showed a significant main effect of duration, $F(3, 90) = 66.29$, $p < .001$, $\eta_p^2 = .688$; a main effect of mask type, $F(1, 30) = 215.41$, $p < .001$, $\eta_p^2 = .878$; and significant interaction between duration and mask, $F(3, 90) = 25.52$, $p < .001$, $\eta_p^2 = .46$. This showed that the natural-scene masks generated lower mean *d′* values at all durations; however, the difference in mean *d′* values between mask types decreased with decreasing durations. The differences in mean *d′* values between the two types of masks were significant at 80, 53, and 27 ms (*p*s < .001, *M*s = 2.02, 1.71, and 1.0, respectively) and nonsignificant at 13 ms ($p = .10$, $M = 0.27$).

Experiment 3: Geometric masks

The findings of Experiment 2 prompted the use of a different mask type. In Experiment 3, geometric masks comprising overlapping colored squares were used. These masks contain more lines and edges than 1/*f* noise masks, so they should more strongly drive V1 activity (Hubel & Wiesel, 1962) and be more effective at impairing the perception of the target pictures.

Across participants, on average 0.02 % of trials were discarded due to computer timing errors. A one-way repeated measures ANOVA revealed a significant main effect of duration, $F(3, 45) = 52.92$, $p < .001$, $\eta_p^2 = .779$; detection accuracy increased with increasing presentation durations. One-tailed single-sample *t* tests showed that *d′* was significantly above chance at the 80-, 53-, and 27-ms durations, but was not significantly above chance when the duration was 13 ms (see Table 3). A Bayesian analysis of *d′* at 13 ms with a Cuachy prior width of 1, conducted using the JASP statistics package, revealed an odds ratio of 1.6:1 in favor of the null hypothesis ($d' = 0$), implying "anecdotal" support for the null hypothesis (Love et al., 2015).

A two-way mixed effects ANOVA comparing Experiments 1 and 3 revealed a significant main effect of duration, $F(3, 90) = 53.87$, $p < .001$, $\eta_p^2 = .642$, but not of mask type. A significant interaction between duration and mask, $F(3, 90) = 18.80$, $p < .001$, $\eta_p^2 = .385$, showed that at longer durations (80 and 53 ms), the natural-scene masks resulted in lower mean *d′* values, whereas at shorter durations (27 and 13 ms), the geometric masks resulted in lower mean *d′* values. The difference in mean *d′* values between mask types was significant ($p < .001$, $M = 1.12$) at 80 ms, and nonsignificant at 53 ms ($p = .089$, $M = 0.34$), 27 ms ($p = .078$, $M = –0.25$), and 13 ms ($p = .119$, $M = –0.30$).

Experiment 4: Colored line masks

In Experiment 3 we used geometric masks, which contain more lines and edges than 1/*f* noise and natural-scene masks. Since these geometric masks were found to be more effective than the previous masks, in Experiment 4 we went one step further, and constructed masks solely comprising lines.

Across participants, on average 0.08 % of trials were discarded due to timing errors. A one-way repeated measures ANOVA showed a significant main effect of duration, $F(3, 45) = 67.74$, $p < .001$, $\eta_p^2 = .821$. As before, detection accuracy increased with increasing presentation durations. One-tailed single-sample *t* tests showed that *d′* was significantly above chance at the 80- and 53-ms durations; however, it was not significantly different from chance at the 27- and 13-ms durations (see Table 4). A Bayesian analysis of *d′* with a Cauchy prior width of 1, conducted using the JASP statistics package, revealed odds ratios of 5.1:1 and 12.3:1 in favor of

Table 2 Experiment 2: Descriptive statistics and *t* test results for *d′* at each presentation duration

Duration	Mean	Standard Deviation	*t*	*p*	Cohen's *d*
13	0.73	0.38	7.65	<.001	1.91
27	1.58	0.54	11.64	<.001	2.91
53	2.59	0.45	23.04	<.001	5.76
80	2.98	0.58	20.58	<.001	5.14

df = 15

Table 3 Experiment 3: Descriptive statistics and *t* test results for sample *d′* at each presentation duration

Duration	Mean	Standard Deviation	*t*	*p*	Cohen's *d*
13	0.15	0.52	1.19	.127	0.30
27	0.33	0.42	3.11	.004	0.78
53	1.21	0.65	7.41	<.001	1.85
80	2.09	0.70	11.93	<.001	2.98

df = 15

Table 4 Experiment 4: Descriptive statistics and *t* test results for sample *d′* at each presentation duration

Duration	Mean	Standard Deviation	*t*	*p*	Cohen's *d*
13	–0.15	0.37	–1.62	.937	–0.41
27	0.01	0.60	0.06	.477	0.02
53	1.32	0.56	9.34	<.001	2.33
80	2.16	0.59	14.55	<.001	3.64

df = 15

the null hypothesis (*d′* = 0) for the 27-ms and 13-ms duration conditions, implying "moderate" and "strong" support for the null hypothesis, respectively (Love et al., 2015).

A two-way mixed effects ANOVA was used to compare Experiments 1 and 4. This revealed a significant main effect of duration, $F(3, 90) = 68.50$, $p < .001$, $\eta_p^2 = .695$, but not of mask type. A significant interaction between duration and mask type, $F(3, 90) = 28.88$, $p < .001$, $\eta_p^2 = .49$, showed that at the 80- and 53-ms durations, natural-scene masks resulted in lower mean *d′* values; however, at the 27- and 13-ms durations, the lines/edges masks resulted in lower mean *d′* values. The differences in the mean *d′* values between mask types were significant at all durations: 80 ms ($p < .001$, $M = 1.19$); 53 ms ($p = .015$, $M = 0.45$); 27 ms ($p = .003$, $M = –0.57$); and 13 ms ($p = .001$, $M = –0.60$).

General discussion

The main finding of this study was that natural-scene masks are significantly less effective than masks comprising lines/edges at durations of 13 and 27 ms. Presumably, this is because natural-scene masks often contain extended areas where there are no high contrast edges; therefore, corresponding areas of the target picture are likely not to be adequately masked. Our lines/edges masks were specifically designed to have high-contrast edges covering the entire image. When these masks were used, we were unable to find any evidence that meaning can be detected in an RSVP stream at 13 ms, or even at 27 ms, per image. Indeed, a Bayesian analysis showed "strong" evidence that observers could not detect meaning at 13 ms per image.

Potter et al. (2014) previously showed that participants could detect the presence or absence of a target image in a natural-scene RSVP stream at 13 ms per image, and on this basis claimed that it is possible to determine the meanings of pictures in a purely feedforward manner. Even though our finding clearly does not support this claim, since we found that, when adequate masking was used, no meaning can be detected at 13 or even 27 ms, we cannot conclude on that basis that visual processing must necessarily involve feedback. For example, it is plausible that processing might require each image in the RSVP sequence to be presented for 53 ms, even if no feedback is involved. This would be compatible with our finding that meaning can be reliably detected in RSVP at 53 ms per picture, regardless of the mask used.

From Fig. 3, it appears that it may be possible for observers to detect meaning even when each image is presented for just 40 ms. For the sake of argument, suppose that later studies were to show that this were indeed the case. Would this then rule out feedback models of visual processing? There are at least two reasons why it would be difficult to make such a claim. First, without knowing the minimum time it takes to establish feedback, it would be hard to interpret this hypothetical result. Although it is clear that feedback cannot be established within 13 ms, it is conceivable that 40 ms might be long enough to establish feedback. Without a reliable estimate for the minimum time it takes to establish feedback, this finding could not be interpreted. Second, even if we were to ignore the previous difficulty, we would still need to be able to demonstrate that the mask we used to limit the processing time of the target really did prevent further processing when it is presented, or else the effective target presentation duration might be longer than expected. Although we have demonstrated that our lines/edges mask is more effective than the natural-scene masks, we have not demonstrated that it instantly disrupts all V1 activity when it is presented. Thus, we could not be completely certain of the effective target presentation time when using this mask. For example, if this mask were to take 10 ms to completely disrupt V1 activity, then a stimulus presented for 40 ms and then masked might have an effective presentation time of 50 ms.

For the sake of argument, let us now instead suppose the converse, and that later studies will show that it is not possible to detect meaning when an image is presented for 40 ms, providing it is adequately masked, and instead that observers need at least 53 ms to process an image. Let us further suppose that it is shown that feedback can be established in 40 ms. Would these two findings rule out feedforward models of visual processing? There are at least two reasons why it would be difficult to make this claim. First, the fact that feedback does occur within the required timespan does not prove that feedback is necessary for visual recognition. It could be that feedforward processing allows for visual recognition, and feedback processing is needed only for refinement. Alternatively, it could be that feedback is not required for refinement, but rather plays a different role in perception. For example, the main role of feedback might be instead to direct attention to a particular location (Macknik & Martinez-Conde, 2007). Either way, even if feedback were to be shown to operate in the correct timeframe, this would not prove that it was necessary for visual recognition. The second reason is that the masks might interfere with and slow down feedforward processing. It might be the case that if feedforward processing were not slowed down by masking,

visual recognition could occur when images are presented for just 13 ms. If that were true, this would argue strongly against feedback being essential for recognition (Potter et al., 2014). In summary, regardless of whether or not later studies demonstrate that visual recognition can occur for images presented for just 40 ms, it is unlikely that we could use this fact to determine whether or not visual recognition can occur in a purely feedforward manner.

We found that detection accuracy declined with shorter presentation durations across all mask types. However, which masks were the most effective depended on the picture duration. At shorter durations, colored squares and lines/edges masks were more effective than natural-scene masks. However, at longer durations natural-scene masks proved to be the most effective. These findings echo those of Potter (1976), who reported that perceptual masks similar to our geometric masks become ineffective for durations of 100 ms or more, but that natural-scene masks continue being effective for durations up to 300 ms. However, in Potter's study the task was later memory for the pictures, not immediate detection of a particular picture.

In the past, findings such as this have been explained in terms of perceptual and conceptual masking (Intraub, 1984; Loftus & Ginn, 1984; Potter, 1976). Masks such as our geometric masks and the lines/edges masks were thought to act primarily by disrupting the perceptual processing of the target image. Specifically, they were thought to disrupt the initial visual processing of the raw stimulus input in the early cortical areas. Conversely, natural-scene masks were thought to act not only by disrupting visual processing, but also by disrupting the conceptual encoding of the identity of the target image in higher cortical areas. Consequently, these masks would continue to be effective even when presented after the initial processing of the target image had occurred, providing that the target image had not yet been conceptually encoded. This would naturally explain why our natural-scene masks would be more effective than both our geometric masks and our lines/edges masks at longer stimulus durations.

However, more recent studies have suggested that conceptual masking may be less important than was previously thought (Loschky et al., 2010). Loschky et al. found that when the interstimulus interval between the target image and the mask was 82 ms, a mask comprising natural scene textures was almost as effective at masking the identity of a target image of a natural scene as a mask comprising a second natural scene (i.e., a natural-scene mask). Since the natural-scene texture mask had higher-order statistics similar to those of the natural scenes, but lacked any conceptual meaning (i.e., the mask resembled noise and did not depict any recognizable image), Loschky et al. concluded that the masking of the target image was due primarily to spatial masking, and not to conceptual masking.

Since it appears that conceptual masking may not be significant, we instead propose a different explanation for the interaction between mask type and picture presentation duration found in our experiments. Specifically, we postulate that different masks have varying levels of effectiveness at driving activity in different parts of the visual processing stream. Presumably, the more a mask can drive neural activity in a given cortical area, the more effective that mask will be at masking activity in that cortical area. The geometric and lines/edges masks both contain a lot of edges, so they would be effective at driving and masking activity in lower/earlier cortical areas such as the primary visual cortex (V1), since activity in these areas is primarily driven by edges (Hubel & Wiesel, 1962). However, higher/later visual cortical areas are more strongly driven by different stimuli. For example, the inferior temporal cortex (IT) contains neurons that respond strongly to complex objects such as faces and hands, but only weakly to simple stimuli such as lines and isolated edges (Desimone, Albright, Gross, & Bruce, 1984). As such, we would not expect the geometrical and lines/edges masks to be the most effective at driving and masking the activity of these neurons. Indeed, we would expect natural images to be more effective at driving and masking these higher/later cortical areas, because these natural-scene masks contained complex objects and higher-order statistics. However, because these natural-scene masks contained fewer edges than the geometric or lines/edges masks, we would expect the natural-scene masks to be less effective at driving and masking activity in lower/earlier visual cortical areas such as the primary visual cortex. At short picture presentation durations, the mask will be presented while processing is occurring primarily in the lower/early visual cortical areas. Since the geometric and lines/edges masks will likely drive these areas more strongly than the natural-scene masks, it follows that for short picture presentation durations, the geometric and lines/edges masks will be the most effective. Conversely, with longer picture presentation durations the mask will be presented after visual processing has moved to higher/later visual cortical areas. Since natural scenes are likely to be the most effective at driving these cortical areas, we would expect the natural-scene masks to be the most effective ones at longer picture durations. In summary, at short picture durations the most effective masks should be the ones that most strongly drive activity in earlier cortical areas (i.e., geometric and lines/edges masks), whereas for longer picture durations the most effective masks should be the ones that most strongly drive activity in later cortical areas (i.e., the natural-scene masks).

Turning our attention now to the $1/f$ noise masks, we note that since these masks contain neither lines nor complex objects, they should be less effective at driving both lower/early and higher/later visual cortical areas, which would explain why they are the least effective masks, regardless of the picture presentation duration.

In our study, observers were required to perform quite a complex recognition task; for example, they would sometimes

need to determine whether a picture depicted the inside of church, as opposed to a kitchen, a cavern, a cave, or any of the other 33 possible scene categories listed in the Appendix. Other studies have addressed a similar question, but using a simpler task. For example, Haynes and Rees (2005) asked observers to determine the orientation of a briefly presented line. Although observers could perform this task with a high degree of accuracy when the line was presented for 50 ms, performance fell to chance when the stimulus was backward masked and presented for 33 ms. Thus, like us, they found that observers could not recognize an object when it was presented for less than 50 ms and adequately masked. This suggests that the temporal limit observed in our study may apply regardless of the exact nature of the categorization task, provided that the task requires detailed perception, such as determining the exact orientation of a line or the precise category of a natural scene.

Our results seem to be particularly compatible with reverse hierarchy theory (Hochstein & Ahissar, 2002). This theory suggests that conscious awareness may exist on a continuum between *vision at a glance* (general gist of a scene) using larger visual fields in higher cortical areas, and *vision with scrutiny*, using smaller visual fields in lower cortical areas. Commencing after the feedforward process, conscious awareness develops from higher cortical areas in a top-down manner, using feedback processes to gradually include detailed information from the lower cortical areas. If this higher-level *gist* finds matching picture details in the lower cortical areas, then detailed conscious awareness is attained; if not, the picture is not perceived in sufficient detail to be recognized (Bar et al., 2006). So, in this study, adequately masked target pictures presented for 13 and 27 ms may have provided a set of possible percepts at higher cortical areas, primed by the feedforward process. However, given that there was presumably insufficient time for multiple feedback connections to be established, these possible percepts may have failed to find matching neuronal activity at lower cortical areas. Consequently, the set of possible percepts may have been discarded, and the target pictures may not have been perceived in sufficient detail to be recognized by the observers.

In conclusion, the take-home message of our study is that before it can be claimed that a picture can processed after being presented for a certain duration, it must first be established that the masking is adequate to prevent processing from occurring beyond the stated duration. Proving that a mask is adequate is very difficult, because doing so requires proving that it immediately stops all processing once it has been presented. In our study we developed more effective masks than those used in previous work. We found that reliable recognition occurred when each image in the RSVP sequence was presented for 53 ms, but not when each image was presented for 27 ms. On the basis of this data, we cannot rule out the possibility that visual processing requires feedback.

Appendix

The pictures used in our study depicted a wide range of everyday natural scenes and were sourced from a publically available collection by Konkle, Brady, Alvarez, and Oliva (2010), which can be accessed here: http://konklab.fas.harvard.edu/#.

From that data set, we selected only those image categories that contained 68 examples, so as to ensure that we had enough examples in each category. This left us with 37 categories, as follows: *airport*, *amusement park*, *bar*, *barn*, *bathroom*, *beach*, *bedroom*, *bridge*, *campsite*, *canyon*, *castle*, *cave*, *cavern*, *cemetery*, *church*, *classroom*, *closet*, *conference room*, *construction site*, *desert*, *foyer*, *golf course*, *greenhouse*, *gym*, *hair salon*, *iceberg*, *kitchen*, *library*, *lobby*, *mountain*, *playground*, *sea port*, *skyscraper*, *street*, *swimming pool*, *temple*, and *underwater*.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York, NY: Academic Press.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., . . . Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*, 449–454. doi:10.1073/pnas.0507062103

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Breitmeyer, B. G., & Öğmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision* (2nd ed.). Oxford, UK: Oxford University Press.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15,* 559–564. doi:10.1111/j.0956-7976.2004.00719.x

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology, 5*, e260. doi:10.1371/journal.pbio.0050260

Desimone, R., Albright, T. D., Gross, C. S., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience, 4,* 2051–2062.

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences, 16,* 317–321. doi:10.1016/j.tics.2012.04.007

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological Science, 22,* 739–746. doi:10.1177/0956797611407930

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, 4,* 2379–2394.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58,* 137–176. doi:10.1016/j.cogpsych.2008.06.001

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d'*. *Behavior Research Methods, Instruments, & Computers, 27,* 46–51. doi:10.3758/BF03203619

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8,* 686–691. doi:10.1038/nn1445

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

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, 160,* 106.

Intraub, H. (1984). Conceptual masking: The effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 115–125. doi:10.1037/0278-7393.10.1.115

Keysers, C., & Perrett, D. I. (2002). Visual masking and RSVP reveal neural competition. *Trends in Cognitive Sciences, 6,* 120–125. doi:10.1016/S1364-6613(00)01852-0

Keysers, C., Xiao, D. K., Foldiak, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience, 13,* 90–101.

Koivisto, M. (2012). Is reentry critical for visual awareness of object presence? *Vision Research, 63,* 43–49. doi:10.1016/j.visres.2012.05.001

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science, 21,* 1551–1556. doi:10.1177/0956797610385359

Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Sciences, 92,* 5587–5591.

Lähteenmäki, M., Hyönä, J., Koivisto, M., & Nummenmaa, L. (2015). Affective processing requires awareness. *Journal of Experimental Psychology: General, 144,* 339–365. doi:10.1037/xge0000040

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10,* 494–501. doi:10.1016/j.tics.2006.09.001

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23,* 571–579. doi:10.1016/S0166-2236(00)01657-X

Loftus, G. R., & Ginn, M. (1984). Perceptual and conceptual masking of pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 435–441. doi:10.1037/0278-7393.10.3.435

Loschky, L. C., Hansen, B. C., Sethi, A., & Pydimarri, T. N. (2010). The role of higher order image statistics in masking scene gist recognition. *Attention, Perception, & Psychophysics, 72,* 427–444. doi:10.3758/APP.72.2.427

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., . . . Wagenmakers, E.-J. (2015). JASP (Version 0.7) (Computer Software: https://jasp-stats.org).

Macknik, S. L., & Martinez-Conde, S. (2007). The role of feedback in visual masking and visual processing. *Advances in Cognitive Psychology, 3,* 125–152.

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science, 13,* 416–424. doi:10.1111/1467-9280.00474

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. doi:10.1163/156856897X00366

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40,* 49–70.

Potter, M. C. (1975). Meaning in visual search. *Science, 187,* 965–966. doi:10.1126/science.1145183

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509–522. doi:10.1037/0278-7393.2.5.509

Potter, M. C., & Hagmann, C. E. (2015). Banana or fruit? Detection and recognition across categorical levels in RSVP. *Psychonomic Bulletin & Review, 22,* 578–585. doi:10.3758/s13423-014-0692-4

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics, 76,* 270–279. doi:10.3758/s13414-013-0605-z

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104,* 6424–6429. doi:10.1073/pnas.0700622104

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11, Whole No. 498), 1–30.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5,* 42. doi:10.1186/1471-2202-5-42