Atten Percept Psychophys (2017) 79:1674–1681
DOI 10.3758/s13414-017-1349-y

Natural scenes can be identified as rapidly as individual features

Piers D. L. Howe[1]

Published online: 5 June 2017

Abstract Can observers determine the gist of a natural scene in a purely feedforward manner, or does this process require deliberation and feedback? Observers can recognise images that are presented for very brief periods of time before being masked. It is unclear whether this recognition process occurs in a purely feedforward manner or whether feedback from higher cortical areas to lower cortical areas is necessary. The current study revealed that the minimum presentation time required to identify or to determine the gist of a natural scene was no different from that required to determine the orientation or colour of an isolated line. Conversely, a visual task that would be expected to necessitate feedback (determining whether an image contained exactly six lines) required a significantly greater minimum presentation time. Assuming that the orientation or colour of an isolated line can be determined in a purely feedforward manner, these results indicate that the identification and the determination of the gist of a natural scene can also be performed in a purely feedforward manner. These results challenge a number of theories of visual recognition that require feedback.

Keywords Object recognition · Scene perception · Model selection

It is generally agreed that visual perception involves both feedforward and feedback processes. Initial activation in the lower visual cortical areas propagates up to higher cortical areas during the feedforward pass. This activity then propagates back to the lower cortical areas to refine and modulate their activity (Di Lollo, 2012; Hochstein & Ahissar, 2002; Lamme & Roelfsema, 2000). There is less agreement, however, as to which aspects of an image can be analysed by the initial feedforward pass and which aspects become consciously accessible only when feedback from higher cortical areas reaches the lower cortical areas (Evans & Treisman, 2005; Lamme, 2003, 2006; Pascual-Leone & Walsh, 2001; Potter, Wyble, Hagmann, & McCourt, 2014). Potter et al. (2014) claimed that the gist of a natural scene can be extracted by the initial feedforward pass. In their experiment, they presented observers with a sequence of six images, one at a time, in the same location of space. They showed that for this Rapid Serial Visual Presentation (RSVP) paradigm, observers were able to determine whether any of the images in the RSVP sequence belonged to a particular conceptual category, such as *picnic* or *harbor with boats* at above chances levels of performance, even when each image in the RSVP sequence was presented for only 13 ms. They claimed that this RSVP presentation time was too short to allow for feedback to occur. In particular, they argued that the time necessary for information to propagate from the primary visual cortex (V1) to higher visual cortical areas and back to primary visual cortex would be at least 50 ms, which follows from the assumption that the information would need to transverse five synapses during the round trip, and it takes a minimum of 10 ms to transverse each synapse (Tovée, 1994). Their finding that the gist of the images can be extracted even when each image is presented for just 13 ms thus suggests that the gist can be extracted by the initial feedforward pass, without the need for feedback (Potter et al., 2014).

Potter et al. (2014) assumed that each image in the RSVP sequence was processed in V1 only for the duration for which it was presented. In their experiments, they ensured that the image queried at the end of the trial (i.e., the target image) was

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

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

never the first or last image in the RSVP sequence. Thus, the processing of the target image in V1 would be interfered with (i.e., masked) by the processing of the images presented both before and after it (Intraub, 1984; Loftus, Hanna, & Lester, 1988; Loschky, Hansen, Sethi, & Pydimarri, 2010). However, natural scenes are unlikely to be perfect masks. Such images will often contain large regions where there are no edges. These regions are unlikely to mask the corresponding regions in the target image (Maguire & Howe, 2016). When Maguire and Howe (2016) used an RSVP paradigm comprised entirely of natural scenes, they replicated the finding of Potter et al. and found that observers could determine the gist of a target image even when that image was presented for only 13 ms. Maguire and Howe then repeated their experiment but replaced all the nontarget images with images that comprised a large number of randomly oriented overlapping lines (see their Experiment 4). Such images contained high-contrast edges across their entire extent. They found that when the target image was preceded and followed by these masks, observers could not extract the gist of the natural scene target image, even when it was presented for 27 ms. The findings of Howe and Maguire therefore contradicted those of Potter et al., but they did not allow Maguire and Howe to determine whether or not the target image in the RSVP stream was processed in a purely feedforward manner. Specifically, they had no way of determining whether the minimum RSVP presentation time they observed was longer than would be expected if the processing was purely feedforward.

The purpose of the current study is to address this concern. Assuming that it takes a minimum of 10 ms to transverse a synapse (Tovée, 1994), establishing a feedback loop from higher cortical areas to lower cortical areas would be expected to take, at a minimum, an additional 20 to 30 ms compared to processing an image in a purely feedforward manner, since the information would need to transverse through an additional two to three synapse as it propagates from higher cortical areas back to lower cortical areas. It follows that if the minimum RSVP presentation time require to process an image is equal to the minimum RSVP presentation time required for feedforward processing, we can conclude that the image is likely to be processed in a purely feedforward manner. In this study, I measured the minimum RSVP presentation time needed to determine the colour and orientation of a line presented in isolation, as it is likely that these can be processed in a purely feedforward manner (Evans & Treisman, 2005; Hochstein & Ahissar, 2002; Treisman & Gelade, 1980). By comparing this time to the minimum RSVP presentation required to discriminate one natural scene image from another I could investigate whether natural scenes can also be processed in a purely feedforward manner. This data showed that the gist of a natural scene could be determined as rapidly as an individual features could be extracted, suggesting that the former occurs in a feedforward manner.

It is possible that natural scenes can be discriminated based solely on their disjunctive (i.e., unbound) features (Evans & Treisman, 2005). However, sometimes one needs to determine how the features are associated together, an issue known as the binding problem (Treisman, 1996; Treisman & Schmidt, 1982; Wolfe & Cave, 1999). For example, if a visual scene contains two lines, one blue and the other red, you need to solve the binding problem before you can determine which line is blue and which is red. Because it is thought that the binding problem is solved by attentional processes (Treisman & Gelade, 1980), it is possible that it requires feedback. If true, the minimum RSVP presentation time required to solve the binding problem should be longer than that needed to identify an isolated feature. The second, third, and fourth experiments addressed this issue. The second and third experiments measured the minimum RSVP presentation time required to determine the orientation or colour, respectively, of an isolated line, while the fourth experiment measured the minimum RSVP presentation time required to bind colour to orientation. For all three experiments, it was found that the minimum RSVP presentation time was approximately the same.

I also investigated the depth of processing of a natural scene. In the first experiment, I presented the RSVP sequence, then a test image, and then asked the observer whether the test image matched the target image in the RSVP sequence. While this task tests whether the observer has some awareness of the target image, it does not require the observer to necessarily extract the gist of the target image. To investigate this issue, I arranged for the fifth experiment to be identical to the first experiment, except that I replaced the test image with a test phrase, such as *airport* or *swimming pool,* and asked whether the target image matched this category. This ensured that the observer extracted the gist of the target image. Given that identification and classification are simultaneous (Grill-Spector & Kanwisher, 2005) and classification is automatic and obligatory (Greene & Fei-Fei, 2014), I was not surprised that I obtained a similar result in the fifth experiment as in the first experiment.

Finally, I investigated the minimum RSVP presentation time required for observers to perform a task that requires feedback. Specifically, in Experiment 6, observers were shown a scene containing a number of line segments and asked to indicate if there were exactly six line segments. Since there were too many line segments for observers to subitize (Kaufman, Lord, Reese, & Volkmann, 1949), observers would not have been able to enumerate them all at once. Since there were too many for them all to be simultaneously stored in visual short-term memory (Luck & Vogel, 1997), there would likely be at least one feedback loop to V1 during the processing. Consistent with this expectation, the minimum RSVP presentation time required to perform this

task was longer than that to perform the other tasks, thereby proving that not all tasks have the same minimum RSVP presentation time.

Method

The stimuli were presented on a 21-inch CRT monitor using MATLAB® running PsychToobox (Brainard, 1997; Pelli, 1997). The resolution was 1280 × 1024, and the refresh rate was 85 Hz. All stimuli were viewed at a distance of 60 cm in a dark room. All participants gave informed consent, and these experiments were approved by the Human Ethics Advisory Group in the School of Psychological Sciences at the University of Melbourne.

In total, six experiments were run. These experiments were similar in design to Experiment 4 of Maguire and Howe (2016). A power analysis revealed that to replicate the significant finding of that experiment at a presentation time of 53 ms with a $\beta = 0.95$ and an $\alpha = 0.05$ would have required a sample size of just four participants. Adopting a more conservative approach, I opted instead to use a sample size of 15 participants in each of my experiments.

Experiment 1—Natural scenes

This experiment was similar to Experiment 4 of Maguire and Howe (2016), except that more finely grained presentation times were used for the RSVP sequence, and the test image was presented only after the sequence. At the start of the trial, the participant was told how many trials were remaining and was invited to click the mouse to start the RSVP sequence. Then a six-item RSVP sequence was shown (see Fig. 1). All images in the RSVP sequence were presented for the same duration as each other. I will refer to this duration as the RSVP presentation time. Five of the images in the RSVP sequence were line masks, comprising a large number of overlapping randomly orientated and randomly coloured lines. These masks were specifically designed to have a large number of high-contrast edges, so as to be effective at disrupting processing in the early visual cortical areas, such as the primary visual cortex (Maguire & Howe, 2016). Each mask was individually constructed and different from all the other masks and subtended approximately 7.3 × 7.3 degrees of visual angle (°). Previous work has shown that such images make especially effective masks for the scene images considered in the current study (Maguire & Howe, 2016). One of the images in the RSVP sequence was the target image. The target image was a natural scene, taken from a publicly available database (Konkle, Brady, Alvarez, & Oliva, 2010), as detailed in the appendix. This image was randomly selected from one of 37 categories

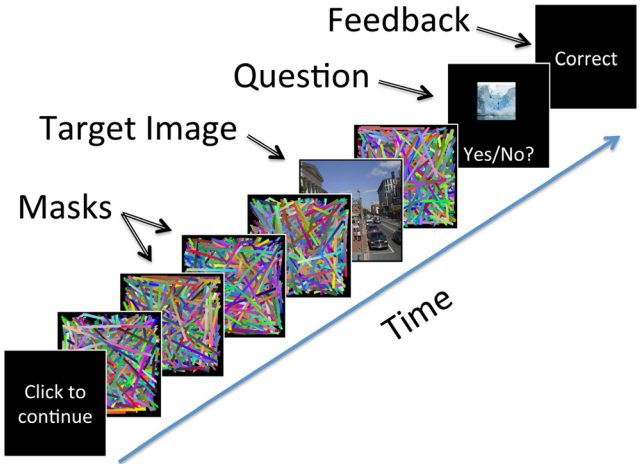

Fig. 1 An example trial from the first experiment. The trial starts with the observer being informed of the number of trials remaining in the experiment and being invited to click the mouse to start the RSVP sequence. The six-image RSVP sequence is then presented, with five of the images being masks and one being a natural scene (the target image). The observer is then shown an image and asked whether this test image matches the target image. Feedback is immediately given. The other experiments used a very similar trial structure. (Colour figure online)

independently for each trial and for each observer. Each trial used a different target image. The target image was assigned to a random point in the RSVP sequence, except that it could never be the first or last image to ensure that it was always both forward and backward masked. After the RSVP sequence finished, a test image was shown and remained visible until the observer indicated with the mouse whether this image matched the target image shown during the RSVP sequence. On 75% of the trials it did. On the remaining trials, the test image was drawn from one of the other 36 categories. Thus, this experiment measured the minimum RSVP presentation time required for the observer to determine whether or not the test image matched the target image. At the end of the trial, the observer was given immediate feedback on whether he or she had responded correctly.

The experiment started with 20 practice trials. In these practice trials, each image in the RSVP sequence was presented for 150 ms. In the main experiment, there were 12 blocks of trials, with each block containing 20 trials. Each block used a different RSVP presentation time, but all trials within a block used the same RSVP presentation time. For the first six blocks, the presentation times were 82.1 ms, 70.6 ms, 58.8 ms, 47.0 ms, 35.3 ms, and 23.5 ms, respectively. These presentations times were then repeated over the next six blocks.

The average age of the participants was 24.5 years, and 12 were female. All had normal or corrected-to-normal visual acuity, achieving a minimum of 20/25 visual acuity as tested with a Good-Lite® Near Vision Chart. They also had normal colour vision as confirmed by the Ishihara Test for Colour Blindness. The experiment took approximately 45 minutes to run.

Experiment 2—Orientation

Experiment 2 was identical to Experiment 1 except that the target image was a line that subtended 1.25° × 0.25° and was presented on a black background (CIE xyY: .329 .495 .490). These lines had the same width as the lines in the masks. The orientation of the line was random, and its colour was randomly selected from one of the following seven values in CIE xyY colour space: red (.496 .314 34.2), green (.273 .521 91.9), blue (.160 .100 24.4), orange (.449 .374 50.3), purple (.263 .154 46.9), cyan (.203 .279 99.4), and yellow (.375 .472 110). The position of the line was also random, except that the line was constrained to appear in a position that ensured that it was entirely overlapped by the masks. The test image comprised a line that was located at the centre of the screen. This line was either identical to the line in the target image or had the same colour but was rotated by 90 degrees (see Fig. 2). Thus, this experiment measured the ability of the observers to detect the orientation of the line in the target image. The average age of the participants was 24.5 years, and 12 were female.

Experiment 3—Colour

Experiment 3 was identical to Experiment 2 except that on those trials where the test image differed from the target image, it did so only in colour. Thus, the line in the test image always had the same orientation as the line in the target image but sometimes had a different colour, selected from one of the other six possible colour values that the line could take. The experiment therefore measured the ability of the observers to detect the colour of the line in the target image. The average age of the participants was 23.6 years, and 12 were female.

Experiment 4—Parallel lines

Experiment 4 was identical to Experiment 2 except that the target image comprised two parallel lines, each with a different colour, with the colours chosen from the seven colours used in Experiment 2. Each line had the same dimensions as the line in the target image in Experiment 2. The two lines were separated by half a line length (i.e., 0.63°). In the test image, the pair of lines was either identical to the pair presented in the target image, or the colours of the two lines were reversed. Thus, the colours and the orientations of the lines in the test image were always the same as those in the target image, but on 25% of the trials the colours were switched between the two lines. This experiment therefore tested the ability of the observers to bind the correct colour with the correct line. In other words, it tested the observer's ability to solve a simple form of the binding problem. Observers could solve this binding problem by focusing on just one of the two lines. However, to do this would still require feedback to earlier cortical areas (e.g., V1) from higher cortical areas as the receptive field sizes of neurons in the higher cortical areas are too large to allow the two lines to be distinguished (Hochstein & Ahissar, 2002). Detailed perception such as this could only be achieved by earlier cortical areas, as only these areas have receptive field sizes small enough to distinguish the lines. The average age of the participants was 25.1 years, and eight were female.

Experiment 5—Verbal test

Experiment 5 was identical to Experiment 1 except that that at the end of the RSVP sequence the observer was not presented with a test image but was instead presented with a short phrase that described a category of natural scenes. Please see the appendix for a listing of the 37 categories. The observer was required to determine whether the target image corresponded to this category. This experiment therefore measured the minimum RSVP presentation time the observer needed to determine the gist of the target image. The average age of the participants was 22.2 years, and 13 were female.

Experiment 6—Counting to six

Experiment 6 was identical to Experiment 1 except that the test image comprised a black background and five, six, or

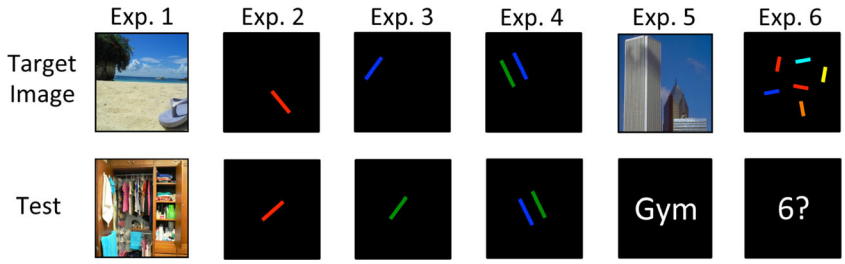

Fig. 2 In Experiments 1–4, observers were shown a target image and then a test image. On 75% of the trials, the test image matched the target image, but in the remaining 25% of the trials, the test image was a decoy, examples of which are shown above. In Experiment 5, no test image was shown. Instead, a word or phrase describing a category of images was shown, and the observer was asked to indicate whether the target image was an example of that category. In Experiment 6, observers were shown either five, six, or seven lines and asked if there were exactly six lines. (Colour figure online)

seven lines. Each line had a random orientation and a random colour, chosen from the seven possible colours used in Experiment 2. The observer's task was to determine if exactly six lines had been shown. This occurred on 75% of the trials. The average age of the observers was 21.9 years, and 13 were female.

Results

For each trial, I recorded whether any timing errors occurred for the presentation of any of the images in the RSVP sequence. Any trial that contained a timing error were excluded from the subsequent analysis. Timing errors occurred on only 0.1% of the trials. For each RSVP presentation time and for each participant I calculated *d'* using the log-linear method (Hautus, 1995), as was done in our previous study (Maguire & Howe, 2016). This was done to obtain a pure measure of sensitivity, independent of any bias (Green & Swets, 1966). Figure 3 shows the *d'* values averaged across participants for each experiment. I fitted the *d'* values using a piecewise linear function comprising two linear components. The function was specified by two parameters, t_o and *m*. The first component was a line with a *d'* of zero that ran from an RSVP presentation time of zero to a RSVP presentation time of t_0. At this point, the second component started. This was a straight line with slope *m* that was joined to the end of the first component. Using the MATLAB® function *fminsearch*, I fitted this piecewise linear model to the first four data points for each experiment, corresponding to RSVP presentation times of 23.5 ms to 58 ms. These fits are represented by the dotted lines in each subplot in Fig. 3. As can be seen from these figures, the model provides good fits for the data points for Experiments 1–5, in the vicinity of t_0. For these experiments, I am able to obtain reliable estimate of t_0. For Experiment 6, the observers were not able to do the task, even at the longest RSVP presentation time. It would therefore have been inappropriate to fit the model to the data set for that experiment. The results are shown in Table 1. Using bootstrapping with replacement (Efron & Tibshirani, 1998), I tested whether t_0 was significantly different in different experiments. In brief, for each experiment, I would select, with replacement, 15 participants. Using their data I would calculate t_0. By repeating this process 10,000 times, I could measure the probability that a bootstrapped sample for one experiment would generate a t_0 that would exceed the t_0 generated by a bootstrapped sample in another experiment. If this probability was larger than 0.975 or smaller than 0.025, I would be able to conclude that the t_0 values in the two experiments were significantly different. As there were no significant differences between the t_0 values for Experiments 1–5 (Table 2), I averaged across these experiments to obtain a mean t_0 value of 34.7 ms (±2.0 ms *SD*).

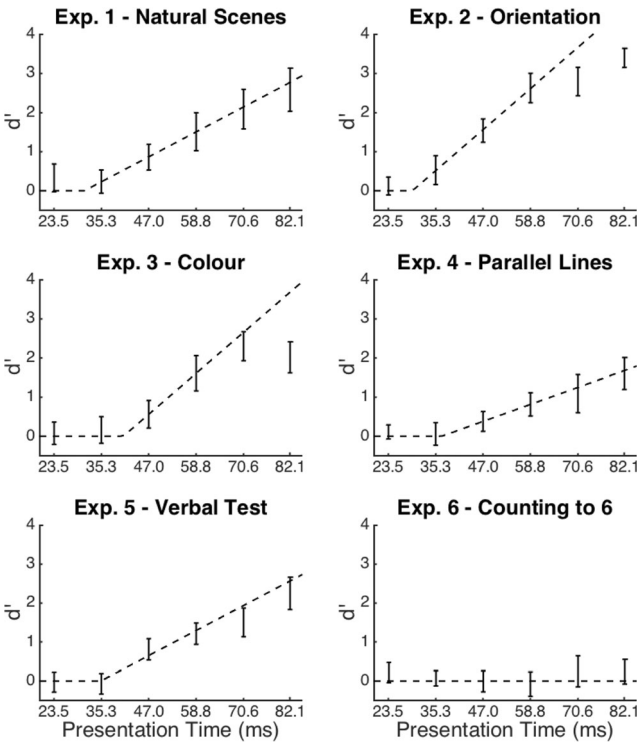

Fig. 3 The results from the six experiments showing the average data. *Error bars* denote the data and represent 95% confidence intervals. The *dotted lines* represent the piecewise linear model fit. From this fit I was able to determine t_0, which is the minimum RSVP presentation time for nonzero *d'*

General discussion

The aim of my experiments was to determine to what extent images can be processed in a purely feedforward manner and whether there is evidence that some images can only be processed when feedback occurs from higher cortical areas to lower cortical areas. To make this determination, I assumed that the colour and orientation of an isolated line could be processed in a purely feedforward manner. I found that the

Table 1 The results for Experiments 1–5

Experiment	t_0 (ms)	r^2	Adjusted r^2
1	31.1 (± 5.8)	0.976	0.960
2	29.5 (± 2.3)	0.999	0.999
3	40.8 (± 3.9)	0.997	0.996
4	36.8 (± 6.2)	0.996	0.993
5	35.2 (± 1.7)	0.983	0.972

The results for Experiment 6 are omitted as the model could not fit that data set; t_0 is the minimum RSVP presentation time required for *d'* to be greater than zero. Standard deviation is in brackets; r^2 represents the model fit. It is the proportion of the total variance of the fitted data points that the model can account for. Adjusted r^2 corrects for the number of predictors in the model, taking into account the sample size

Table 2 Results from the bootstrap significance testing for Experiments 1–5 (Efron & Tibshirani, 1998)

	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Exp. 1	.63	.16	.46	.47
Exp. 2		.06	.28	.08
Exp. 3			.56	.22
Exp. 4				.69

Each cell represents the proportion of the bootstrapped trials where t_0 for the experiment corresponding to the row is less than t_0 for the experiment corresponding to the column. In no cells was $p < .025$ or $> .975$, so there is no evidence that t_0 is systematically different in any pair of experiments, even when Bonferroni corrections are not performed

minimum RSVP presentation time required to identify the orientation and colour of an isolated line was 29.5 ms and 40.8 ms, respectively. Bootstrapping with resampling revealed that these estimates were not significantly different. Combining them, I concluded that a minimum RSVP presentation time of 35.1 ms is required for feedforward processing to occur. If feedback processing from higher cortical areas to lower cortical areas were to occur in addition to this, then it would be expected to take at least an additional 20 to 30 ms, as discussed earlier. I found that the minimum RSVP presentation time required for an observer to be able to determine whether a natural scene test image presented at the end of an RSVP sequence matched the natural scene image presented during the RSVP sequence was 31.1 ms. From this I concluded that natural scenes can be processed in a purely feedforward manner.

It has been suggested that, to some degree, observers may be able to identify natural scenes by detecting the unbound (i.e., disjunctive) features present within the scene (Evans & Treisman, 2005). For example, if an observer detects that an image contains a large amount of blue, he can conclude that the image is more likely to depict the sea than a desert. However, in other situations, the observer needs to determine which features are associated together. In this case, the observer would need to solve the binding problem (Roskies, 1999; Treisman, 1996; Wolfe & Cave, 1999). In the fourth experiment, I tested the minimum RSVP presentation time that observers would need to solve a simple form of the binding problem. Surprisingly, I found that this time was only 36.8 ms, indicating that, at least for simple conjunctions, even the binding problem can be solved without feedback from higher cortical areas to the primary visual cortex. This result directly contradicts Reverse Hierarchy Theory, which assumes that although basic features such as colour and orientation can be discriminated on the initial feedforward pass, feedback from higher cortical areas to lower cortical areas is needed to discriminate conjunctions of features (Hochstein & Ahissar, 2002). These results are consistent with the suggestion that the binding problem is solved by independently registering features to their relevant location (Vul & Rich, 2010). Features that are perceived to belong to the same location are consciously perceived (i.e., bound) together, thereby solving the binding problem. It follows from this theory that binding features together should take no more time than registering the individuals features, which is what we found.

The fifth experiment investigated the depth of processing of natural scenes. In the first experiment, observers were required to determine whether a test image presented at the end of the trial matched the natural scene image presented in the RSVP stream. They may have been able to do this without determining the gist of the test image. Experiment 5 repeated Experiment 1 but replaced the test image with a verbal word or phrase, such as *airport*, so as to ensure that observers would need to extract the gist of the target image. The minimum RSVP presentation time required to do this task was only 35.2 ms, which was not significantly longer than that in Experiment 1. This indicates that even the gist of a natural scene can be extracted in a feedforward manner.

The sixth experiment tested a fundamental assumption of this study. It verified that a process that is assumed to require feedback to V1 had a longer RSVP presentation time than the task described above. For this experiment I chose an enumeration task where observers had to determine whether exactly six line elements had been shown in the test image. As this number is outside the subitizing range (Kaufman et al., 1949), it is unlikely that observers would be able to process the stimulus all at once. Since they could not store the entire stimulus in short-term memory (Luck & Vogel, 1997), it is likely that the processing of the stimulus would involve activity feedback to the primary visual cortex. Consistent with this expectation, it was found that observers required a greater RSVP presentation time to perform this task than was required in the experiments described above. Indeed, *d*' was not significantly greater than zero even for the maximum presentation time of 82.1 ms, $t(14) = 1.62$, $p = .13$, $r^2 = .16$.

In my analysis, each model fit produced two numbers, t_0 and m, where t_0 represents the minimum RSVP presentation time required for d' to exceed zero and m represents the rate at which d' then increases as a function of RSVP presentation time. For the purposes of this study, I have been ignoring m as only t_0 allowed me to distinguish between feedforward and feedback processing. In future experiments, it would be interesting to investigate what factors influence m, but these investigations were beyond the scope of the current study.

Limitations

In this study I made a number of assumptions that need to be acknowledged. Alluded to earlier, the main assumption I made was that the colour and orientation of an isolated line can be processed in a purely feedforward manner. It could be that this

assumption is false and that the conscious perception of an image always requires reentrant feedback to be established from higher visual cortical areas to lower ones (Lamme, 2003, 2006; Lamme & Roelfsema, 2000; Pascual-Leone & Walsh, 2001). If so, this would invalidate both the conclusions of this study and many theories of visual perception (Evans & Treisman, 2005; Serre, Kreiman, et al., 2007a; Serre, Olivia, & Poggio, 2007b).

Another assumption that I made is that masking can disrupt visual processing, at least in the earlier cortical areas (Potter et al., 2014). This is not an unreasonable assumption as the mask used in the current experiment was specifically designed to disrupt processing in the primary visual cortex (Maguire & Howe, 2016). However, this does not mean that processing in the primary visual cortex is instantly halted by the presentation of the mask. Thus, I do not claim that the presentation time is identical to the time the image was processed for in V1. This is why it was important to measure the minimum presentation time required to process a stimulus that I can assume is processed in a purely feedforward manner. If I find that a second stimulus requires the same minimum processing time, this would then be convincing evidence that the second stimulus can also be processed without feedback to the primary visual cortex from higher cortical areas.

A final assumption that I made was that a process that requires both a feedforward pass and a feedback pass would necessarily take longer than a process that requires just a feedforward pass. As discussed above, assuming that feedback requires the information to transverse two to three synapses as it propagates from higher cortical areas to lower ones and that it takes a minimum of 10 ms to transverse each synapse (Tovée, 1994), a process that requires a feedback pass should take a minimum of 20 to 30 ms more than one that requires only a feedforward pass. Given the accuracy by which I was able to measure the minimum RSVP presentation times, this difference would have been trivial to detect. Indeed, in Experiment 6 I found that the minimum RSVP presentation time required for a process that one can reasonably assume requires feedback is much greater than the minimum RSVP presentation times reported in Experiments 1–5.

Conclusion

I found that the minimum RSVP presentation time required to extract the gist of a natural scene was approximately the same as that required to determine the orientation and colour of an isolated line. Under the assumption that the latter can be processed in a feedforward manner, it follows that extracting the gist of a natural scene can also be processed in a feedforward manner. These finding challenges theories that posit that feedback from higher cortical areas to lower cortical areas is necessary to recognise images (Di Lollo, 2012; Lamme & Roelfsema, 2000). I also found that the minimum RSVP presentation time required to solve a simple form of the binding problem was the same as that required to process a single feature. This contradicts theories that assume that feedback from higher cortical areas to lower cortical areas is required to solve the binding problem (Hochstein & Ahissar, 2002) but is consistent with those that instead assume that features are processed and registered independently of each other (Vul & Rich, 2010). Finally, I found that the minimum RSVP presentation time required to process a stimulus that would be expected to require feedback was indeed significantly longer than that required to register an individual feature. This is further evidence that, had extracting the gist of a natural scene required feedback, this paradigm would have been able to detect this.

Acknowledgements The author would like to thank Molly Potter for very helpful discussions.

Appendix

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

From that data set, I selected only those image categories that contained 68 examples so as to ensure that I had enough examples in each category. This left 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

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences, 16*(6), 317–321.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap* (Vol. 57). New York, NY: Chapman & Hall/CRC.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1476–1492.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley & Sons.

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obigatory: Evidence from Stroop-like paradigm. *Journal of Vision, 14*(1), 14.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16*, 152–160.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d'*. *Behavior Research Methods, Instruments, & Computers, 27*(1), 46–51.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791–804. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12467584

Intraub, H. (1984). Conceptual masking: The effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory & Cognition, 10*, 115–125.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *American Journal of Psychology, 62*(4), 498–525.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The fole of categories in visual long-term memory. *Psychological Science, 21*(11), 1551–1556.

Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences, 7*, 12–18.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*, 494–501.

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Cognitive Sciences, 10*, 494–501.

Loftus, G. R., Hanna, A. M., & Lester, L. (1988). Conceptual masking: How one picture captures attention from another picture. *Cognitive Psychology, 20*, 237–282.

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.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281. doi:10.1038/36846

Maguire, J. F., & Howe, P. D. L. (2016). Failure to detect meaning in RSVP at 27 ms per picture. *Attention Perception and Psychophysics, 78*(5), 1405–1413.

Pascual-Leone, A., & Walsh, V. (2001). Fast backprojections from the motion to the primary visual are necessary for visual awareness. *Science, 292*, 510–512.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9176953

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*(2), 270–279. doi:10.3758/s13414-013-0605-z

Roskies, A. L. (1999). The binding problem. *Neuron, 24*(1), 7–9, 111–125. doi:10.1016/S0896-6273(00)80817-X showArticle Info

Serre, T., Kreiman, G., Kouch, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007a). A quantitative theory of immediate visual recognition. *Progress in Brain Research, 165*, 33–56.

Serre, T., Olivia, A., & Poggio, T. (2007b). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 6424–6429.

Tovée, M. J. (1994). How fast is the speed of thought. *Current Biology, 4*(12), 1125–1127.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology, 6*(2), 171–178.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7351125

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*(1), 107–141.

Vul, E., & Rich, A. N. (2010). Independent sampling of features enables conscious perception of bound objects. *Psychological Science, 21*(8), 1168–1175.

Wolfe, J. M., & Cave, K. R. (1999). The psychophysical evidence for a binding problem in human vision. *Neuron, 24*(1), 11–17, 111–125. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10677023