The Limits of Feedforward Vision: Recurrent Processing Promotes Robust Object Recognition when Objects Are Degraded

Dean Wyatte, Tim Curran, and Randall O'Reilly

Abstract

■ Everyday vision requires robustness to a myriad of environmental factors that degrade stimuli. Foreground clutter can occlude objects of interest, and complex lighting and shadows can decrease the contrast of items. How does the brain recognize visual objects despite these low-quality inputs? On the basis of predictions from a model of object recognition that contains excitatory feedback, we hypothesized that recurrent processing would promote robust recognition when objects were degraded by strengthening bottom–up signals that were weakened because of occlusion and contrast reduction. To test this hypothesis, we used backward masking to interrupt the processing of partially occluded and contrast reduced images during a categorization experiment. As predicted by the model, we found significant interactions between the mask and occlusion and the mask and contrast, such that the recognition of heavily degraded stimuli was differentially impaired by masking. The model provided a close fit of these results in an isomorphic version of the experiment with identical stimuli. The model also provided an intuitive explanation of the interactions between the mask and degradations, indicating that masking interfered specifically with the extensive recurrent processing necessary to amplify and resolve highly degraded inputs, whereas less degraded inputs did not require much amplification and could be rapidly resolved, making them less susceptible to masking. Together, the results of the experiment and the accompanying model simulations illustrate the limits of feedforward vision and suggest that object recognition is better characterized as a highly interactive, dynamic process that depends on the coordination of multiple brain areas. ■

INTRODUCTION

The human visual system is perhaps the fastest, most accurate object recognition system in the world. Research has indicated that the brain can reliably differentiate between complex visual categories in as little as 100–150 msec (Liu, Agam, Madsen, & Kreiman, 2009; VanRullen & Thorpe, 2001; Thorpe, Fize, & Marlot, 1996; see also Johnson & Olshausen, 2003, for a more conservative estimate). Much of our understanding of how the brain is capable of such robust recognition comes from a "standard" model of object recognition, which posits that visual features are extracted rapidly over a feedforward hierarchy of processing stages corresponding to brain areas along the ventral visual stream (VanRullen, 2007; Riesenhuber & Poggio, 1999; see Serre, Kreiman, et al., 2007, for a review). Although this feedforward model has accounted for a wide range of visual findings—from detailed neural tunings (Cadieu et al., 2007; Freedman, Riesenhuber, Poggio, & Miller, 2003) to overt psychophysical measures (Serre, Oliva, & Poggio, 2007)—it remains to be fully reconciled with anatomical data that indicate nearly equal densities of forward and backward projecting neurons throughout the visual pathways (Sporns & Zwi, 2004; Felleman & Van Essen, 1991). Under a purely feedforward view of object recognition, feedback might play a role in secondary, after-the-fact processes like feature-based attention (Hochstein & Ahissar, 2002) or in the initial learning and development of the visual system, but it is not necessary for core object recognition operations (DiCarlo, Zoccolan, & Rust, 2012).

Single-cell recordings from neurons in the dorsal visual stream, however, suggest a more primary role for recurrent feedback during vision. Neurons in early visual areas (e.g., V1, V2, V3) receive top–down excitation from higher-level areas (e.g., V5/MT), which has been shown to strengthen the responses in the lower-level areas and improve discriminability in figure-ground segregation tasks (Hupe et al., 1998; Lamme, Super, & Spekreijse, 1998). This strengthening dynamic is quite rapid, with latencies as short as 10 msec after the first stimulus-driven responses (Hupe, James, Girard, & Bullier, 2001), suggesting that if a similar dynamic takes place within the ventral visual stream, it might fit within the strict temporal constraints of object recognition. Recently, a cortical model of object recognition characterized by recurrent connections between hierarchically adjacent visual areas was shown to exhibit increased robustness to a visual occlusion degradation on stimuli when compared with a class of purely feedforward models (O'Reilly, Wyatte, Herd, Mingus, & Jilk, under review). Similarly, primate studies have indicated that visual occlusion attenuates the responsiveness of neurons in the inferotemporal cortex (IT) area, yet robust recognition

University of Colorado Boulder

performance remains intact (Nielsen, Logothetis, & Rainer, 2006; Kovacs, Vogels, & Orban, 1995). Motivated by these findings, the present research asks whether the brain's robustness to visual degradations such as occlusion stems from the recurrent connectivity of the ventral visual stream.

To demonstrate our hypothesis, we refer to the model from O'Reilly et al. (under review; Figure 1) to illustrate how recurrent connectivity can give rise to robust object recognition. When recognizing a stimulus, feedback from high-level visual areas like IT cortex can reinforce and strengthen neural responses in lower-level extrastriate areas. This strengthening dynamic is especially important when a stimulus is degraded (by partial occlusion, contrast reduction, etc.), as the first visual responses will only reflect the noisy partial information encoded by the fovea. The now-strengthened responses in early visual areas provide further bottom–up support for the "hypothesis" conveyed by high-level areas, increasing the information available to further downstream neurons involved in the recognition decision. This latter point has been demonstrated in single-cell recordings of IT neurons, which continue to convey new information not captured in their initial spikes over the full time course of their responding (Heller, Hertz, Kjaer, & Richmond, 1995; Rolls & Tovee, 1995; see also Perrett, Oram, & Ashbridge, 1998). On the basis of these findings, we predicted that the mutual reinforcement between the bottom–up, stimulus-driven signals and top–down, conceptual signals would create stable, reliable percept, preserving successful recognition in the face of degradations like visual occlusion and reduced contrast. We refer to this type of processing as "recurrent processing" throughout this article.

To test the aforementioned predictions psychophysically, we needed a means to control the amount of influence from recurrent feedback during object recognition. Backward masking has been suggested to selectively disrupt feedback (see Di Lollo, Enns, & Rensink, 2000; Lamme & Roelfsema, 2000, for reviews) and has been used in a variety of experiments to dissociate the effects of feedforward and recurrent processing (e.g., Boehler, Schoenfeld, Heinze, & Hopf, 2008; Fahrenfort, Scholte, & Lamme, 2007, 2008; Serre, Oliva, & Poggio, 2007; Lamme, Zipser, & Spekreijse, 2002). In our experiment, volunteers categorized visual object stimuli that were degraded by either an occlusion manipulation or contrast reduction. If recurrent processing does indeed promote robust recognition when stimuli are degraded, categorization performance should be substantially impaired when a mask follows a stimulus that is heavily degraded (high occlusion or low contrast) because the mask will disrupt the dynamics depicted in Figure 1 from manifesting across brain areas. In contrast, performance should remain intact when a mask follows a relatively clear stimulus because the well-specified stimulus will result in a stable representation that is already sufficient for categorization. Consistent with these ideas, we found a significant interaction between the mask and occlusion and the mask and contrast, such that the mask impaired the categorization of heavily degraded stimuli more than clear stimuli.

We expand upon our experimental results by using the model from O'Reilly et al. (under review) to simulate an

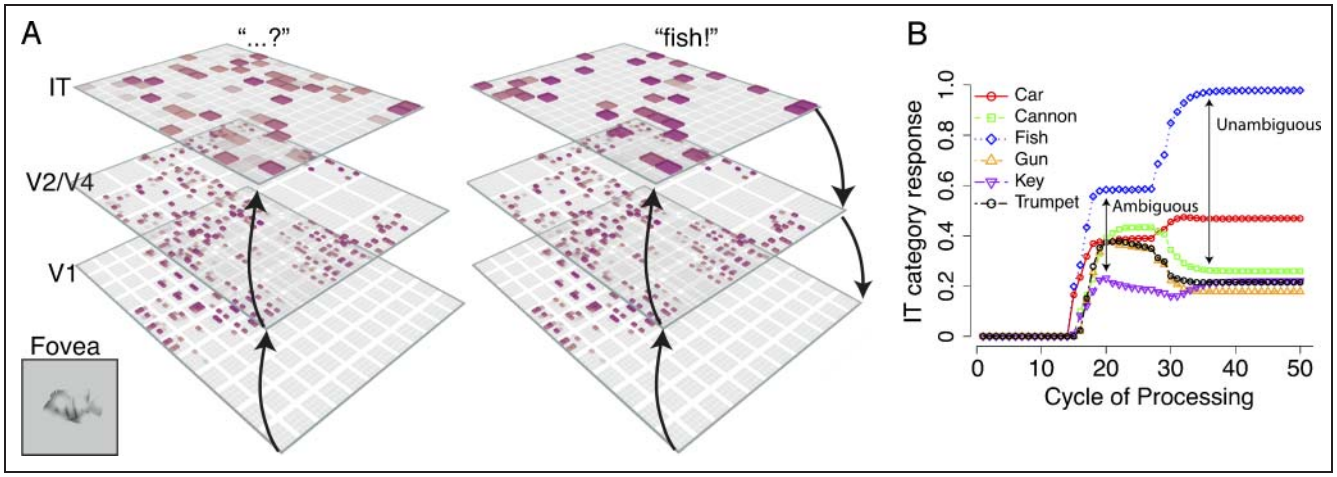

Figure 1. Demonstration of recurrent processing during degraded object recognition. (A) The ventral visual stream consists of primary visual cortex (V1), extrastriate cortex (V2, V4), and inferior temporal cortex (IT). Object recognition in the ventral stream is traditionally conceptualized as a series of feedforward computations across these brain areas. When stimuli are degraded, the resulting bottom–up signals are weak and often noisy, producing an ambiguous pattern of activation across the highest levels of the hierarchy (left). When signals between adjacent layers are propagated recurrently, they reinforce each other despite the degraded bottom–up stimulus, ultimately resulting in a strong, stable, pattern of activation at the highest level (right). Note that many of the IT units that were weakly active in the feedforward model have been correctly suppressed in the recurrent model producing a representation that consists of a small number of strongly active units. (B) The initial feedforward responses in IT cortex contain some information that can be used in the recognition decision, but they reflect the simultaneous activation of neurons shared across a large number of categories, resulting in an overall ambiguous representation. This ambiguity is exacerbated when stimuli are underspecified, which happens with degradations such as occlusion or contrast reduction. Recurrent processing, however, dynamically strengthens and suppresses ambiguous responses, thus creating additional contrast between competing categories and reducing the overall ambiguity in the recognition decision.

isomorphic version of the behavioral experiment with identical stimuli. The model produced a close fit of the experimental results as well as provided an intuitive explanation of interactions between masking and degradations. Specifically, backward masking interacted selectively with the recurrent processing that was necessary to resolve the identity of heavily degraded stimuli, whereas the identity of relatively clear stimuli could be resolved relatively quickly and thus were less susceptible to the effects of the mask.

Together, the results of the experiment and modeling illustrate the limits of feedforward processing during object recognition. Under idealized viewing conditions, visual processing proceeds rapidly in a relatively straightforward manner through the visual pathways without much influence from recurrent feedback. However, when stimuli are less well specified—as is often the case with real-world inputs that contain various environmental factors like occlusion, diffuse lighting, and complex shadows—object recognition depends heavily on the extensive recurrent connectivity of the visual pathways to strengthen object representations and preserve recognition. Overall, our results are consistent with the burgeoning view that vision and object recognition are highly interactive processes governed by moment-to-moment neural dynamics between recurrently connected brain areas (e.g., Roland, 2010; Spivey, 2007; Bar, Kassam, Ghuman, Boshyan, & Schmidt, 2006).

METHODS

A total of 19 volunteers from the University of Colorado at Boulder participated in the experiment as part of their introductory psychology course credit (11 men, 8 women; mean age = 18.9 years). All participants reported normal or corrected-to-normal vision and gave informed consent before the experiment in accordance with the human subjects policy at the University of Colorado.

Experimental Stimuli

During the experiment, participants were required to categorize images from six real-world occurring object categories: *cannon*, *car*, *fish*, *gun*, *key*, and *trumpet* (Figure 2). The specific categories used in the experiment were chosen because of their sharing a horizontal axis of canonical orientation, preventing participants from using coarse orientation information as a cue for category membership. The images themselves were taken from the CU3D-100 data set (http://cu3d.colorado.edu), which consists of three-dimensional object model exemplars that are rendered to 320 × 320 bitmap images with variations in view and lighting (±20° in-depth rotations including a random 180° left–right flip along the horizontal axis and overhead lighting positioned uniformly randomly along an 80° overhead arc). The images were processed with the SHINE toolbox (Willenbockel et al., 2010) to convert their colorspace to grayscale and to normalize luminance across categories.

Seven different exemplars from each category were used during the full experimental session. Before beginning the actual experiment, participants were shown two images of two exemplars from each category (24 total images) to become familiar with the basic visual structure of the experimental images. Twenty images of the remaining five exemplars from each category were used during the experiment itself (600 total images). During the familiarization phase, subjects were informed that the specific images they were viewing would not be used in the experiment itself. The familiarization phase was self-paced, but participants never took longer than 1 min in practice to examine the 24 images.

Occlusion was manipulated by constructing a filter that constituted a circle with a radius of 5% of the image size whose edges were softened with a Gaussian. The size of this filter was 96 × 96 pixels. The filter was applied to the image at random locations by taking a weighted average between the background gray intensity of the image and the pixel intensities at the location of application. Two levels of occlusion were used in the experiment. Control trials were characterized by a small amount of occlusion, during which the filter was applied 29 times to the image. During high occlusion trials, the filter was applied 73 times. In both cases, application of the filter was an iterative process such that the filter could be applied to the same location more than once.

Contrast was independently manipulated to determine whether it benefited from the putative benefits of recurrent processing in the same manner as occlusion. Control trials held the image at its original, full contrast, whereas during low contrast trials, the image's contrast was scaled 25% of its original range. Contrast reduction occurred before the occlusion manipulation was applied, and the background gray level was held constant during the entire image manipulation process.

Pattern masks were constructed by sampling patches of the original object images and assembling them into a new 320 × 320 image. The size of the sampled patch varied between 16 × 16 and 64 × 64 pixels and was randomly sampled from a region surrounding the bounding box of the object in each image. The resulting image patches were placed at random into a new image with the same background gray level as the original object images. Like the occlusion algorithm, the patches were placed into the new image in an iterative manner and were allowed to overlap. A total of 416 samples were taken across the 600 source images. A total of 600 masks were pregenerated for use in the experiment.

Examples of the experimental stimuli are depicted in Figure 2A–C.

Experimental Procedure

During the experiment, volunteers were seated approximately 45 cm from a Gamma-corrected CRT monitor running at a resolution of 1024 × 768 at 120 Hz. Under this

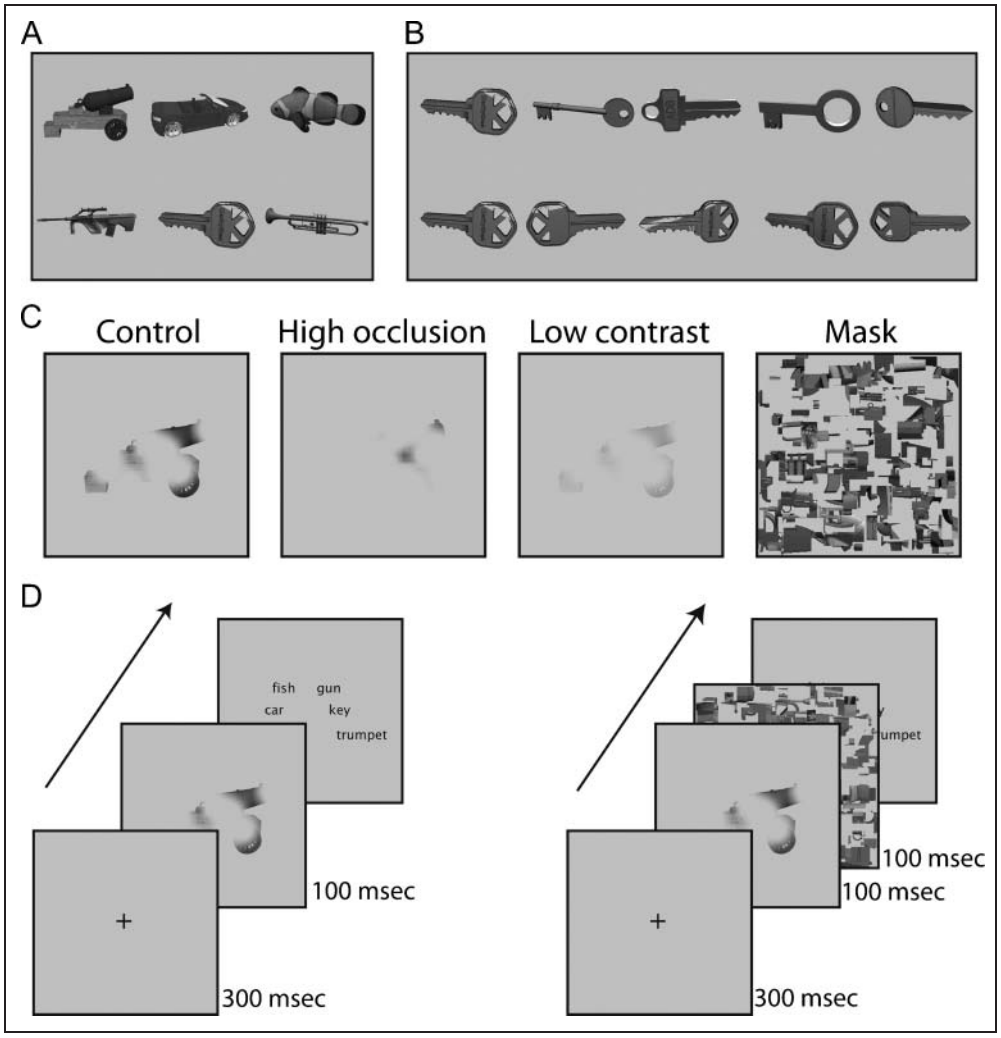

Figure 2. Experimental stimuli and trial schematic. (A) One exemplar from each of the six categories used in the experiment. From left to right, top to bottom: *cannon, car, fish, gun, key,* and *trumpet*. (B) The five exemplars from the *key* category used in the experiment (top) and five instances of a single exemplar with variations in view and lighting (bottom). (C) The effects of the degradation manipulations used in the experiment. "Control" trials contained a low amount of occlusion and were presented at full contrast. "High occlusion" trials contained a high amount of occlusion and were presented at full contrast, whereas "Low contrast" trials contained a low amount of occlusion and were presented at 25% contrast. Pattern masks were constructed from patches of the original images. (D) Trials consisted of a 300-msec fixation cross, followed by a 100-msec stimulus. On trials that contained a mask, the mask was presented directly after the 100-msec stimulus and remained onscreen for an additional 100 msec. All trials were followed by a response screen that contained the names of the six categories and remained onscreen until the subject responded (or for a maximum of 5000 msec).

configuration, stimuli subtended approximately 16° of visual angle on the display. The Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997) was used to synchronize the display of stimuli with the monitor's refresh interval.

The experiment contained eight trial types, reflecting the full factorial crossing of the variables: 2 levels of Occlusion (low, high) × 2 levels of Masking (masked, unmasked) × 2 levels Contrast (low, high). On each trial, the participant was presented with a fixation cross for 300 msec, followed by the degraded object stimulus. On unmasked trials, the object stimulus remained visible for 100 msec. On masked trials, the object stimulus was replaced after 100 msec by a randomly selected mask, which remained visible for an additional 100 msec. Participants were then presented with a response screen that contained the six category names. Subjects' responses were collected via a QWERTY keyboard using the S, D, F, J, K, and L keys. The arrangement of the category names on the response screen was isomorphic with the placement of participants' fingers on the keyboard to facilitate their responding without having to explicitly recall the key associated with their response. Subjects were required to respond within 5000 msec. The response screen remained visible until the subject responded. The ordering of events within a single trial is depicted in Figure 2D.

All trial types were intermixed and presented randomly in blocks of 50 trials. The experiment consisted of 1000 total trials. Of the 600 images that were selected from the CU3D-100 data set for the experiment, 400 were repeated exactly once (random per participant). However, the occlusion manipulation was applied in an on-line fashion before each trial was presented, ensuring that no two degraded stimuli were ever likely to be the same. Participants were given feedback after each trial regarding whether their response was correct or incorrect.

RESULTS

Statistical Analysis

A total of three volunteers were excluded from statistical analysis—two for accuracy levels that were far below 1.5 times the interquartile range of the data across

conditions and one for failing to complete the full experiment. The remaining 16 volunteers were included in the final analysis.

We were interested in the effect of the mask across three key trial types: low occlusion under high contrast, high occlusion under high contrast, and low occlusion under low contrast (denoted as "Control," "Occlusion," and "Contrast" in Figure 3A). This focus excludes the high occlusion under low contrast trial type from the full factorial design of the experiment, which was subject to the floor effects because of multiple, compound sources of degradation. We also included in the analysis a repetition factor to determine whether there was an effect of repeating a subset of the stimuli (see Methods).

A repeated-measures ANOVA with p values corrected for violations in sphericity using the method described in Geisser and Greenhouse (1958) indicated that there was a significant effect of the Mask, $F(1, 15) = 112.99$, $p < .001$, such that recognition accuracy was lower when the mask was present compared with when it was absent. There were also significant effects of the Trial Type, $F(2, 30) = 1512.73$, $p < .001$, and Stimulus Repetition, $F(1, 15) = 10.07$, $p = .006$. The latter result indicates that participants exhibited improved performance for repetitions of stimuli (which always occurred in the last 400 trials of the experiment), consistent with a training effect. We explore these results in more detail next to determine whether they interact with the trial type.

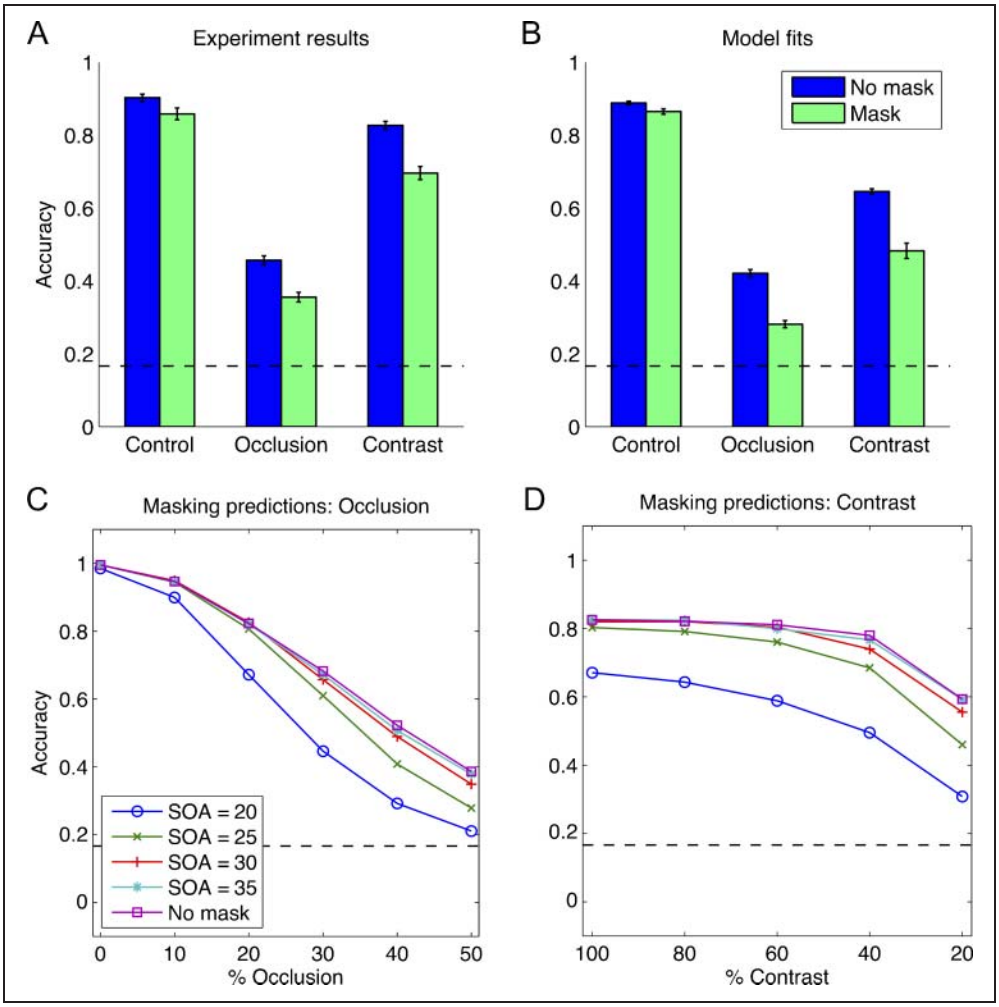

Figure 3. Results of the experiment and model simulations. (A) The effect of masking was significant across three key conditions: low occlusion under high contrast ("Control"), high occlusion under high contrast ("Contrast"), and low occlusion under low contrast ("Contrast"). Furthermore, there were significant interactions between the mask and occlusion and mask and contrast such that recognition was differentially impaired when degraded stimuli were masked ("Occlusion," "Contrast") compared with when relatively clear stimuli were masked ("Control"). Blue bars correspond to unmasked trials, green bars to masked trials. (B) The model from O'Reilly et al. (under review) produced a close fit to the experimental results in an isomorphic version of the experiment with identical stimuli. (C) The model also demonstrates the more general interaction between the mask and occlusion across a broad range of mask latencies and occlusion levels (with contrast fixed at the level in the experiment's control condition, 100%). SOA is stimulus onset asynchrony, that is, mask latency (in model cycles), where smaller values correspond to shorter latencies. (D) The model predicts a similar general interaction between the mask and contrast (with occlusion fixed at the level in the experiment's control condition, 20%). In all plots, the dashed line indicates chance performance for the six-way categorization task (1/6 or 0.1667).

The interaction between the Mask and Trial Type was found to be significant, $F(2, 30) = 10.17$, $p < .001$, indicating a differential impairment of the mask depending on the specific trial type. Paired t tests (with p values corrected for false discovery rate using the method described in Benjamini & Hochberg, 1995) indicated that the effect of the Mask was significantly different between the "Control" and "Occlusion" conditions, $t(15) = -2.32$, $p = .03$, and the "Control" and "Contrast" conditions, $t(15) = -5.73$, $p < .001$. The significant interactions between the Mask and Occlusion and between the Mask and Contrast indicate that participants' performance on the categorization task was differentially impaired when heavily degraded stimuli were masked compared with a relatively clear viewing conditions. This pattern of results suggests that recurrent processing helps to resolve the identity of both heavily occluded and contrast reduced stimuli and that backward masking interferes with this resolution process.

The interaction between the Mask and Stimulus Repetition failed to reach significance, $F(1, 15) = 2.79$, $p = .16$, as did the interaction between the Trial Type and Repetition, $F(2, 30) = 3.31$, $p = .07$. Thus, the training effect because of stimulus repetition was not significantly for masked/unmasked repetitions or across trial types. The three-way interaction between the Mask, Trial Type, and Stimulus Repetition also failed to reach significance, $F(2, 30) = 2.15$, $p = .14$.

To further explore the nature of the interactions between the mask and occlusion and the mask and contrast, we simulated an isomorphic version of the behavioral experiment with model from O'Reilly et al. (under review). The model produced a close fit of the experimental results and provided an intuitive demonstration of how masking interferes with the recurrent processing necessary to resolve the identity of degraded stimuli.

Model Simulations

The model from O'Reilly et al. (under review) was used to model participants' experimental data, as well as to provide insight into how occlusion and contrast interact with mask latencies not tested in the human experiment (see Appendix for detailed simulation methods). To accomplish this, the model was first trained to categorize the images from the same six categories used in the experiment and was then tested with the same occlusion and contrast manipulations as used in the experiment as well as an equivalent masking manipulation. Masking was implemented by clamping an input image into the model's inputs and iterating the model for a variable number of processing cycles before replacing the image with a random pattern mask. In fitting subjects' data, occlusion and contrast were fixed at the same levels used in the experiment whereas the onset of the mask was varied as a free parameter.

The model produced a close fit of data with a simulated mask latency of 25 cycles, with substantial interactions between the mask and occlusion and the mask and contrast (Figure 3B). Occlusion produced a larger overall impairment in recognition than contrast, providing a better absolute fit of subjects' data. Absolute accuracy levels in the Contrast condition of the modeling results did not reach the same level as participants' accuracies, but the magnitude of the mask effect was comparable. Better absolute fits are possible by varying contrast as an additional free parameter (e.g.,, 40% contrast; see Figure 3D). The absolute accuracy difference between the model and subjects in the Contrast condition simply indicates that the model's contrast threshold is limited compared with human observers. The general interactions between the mask and contrast and the mask and occlusion, however, remains at all levels across a broad range of mask latencies and degradation levels, which we also tested using the model.

To explore these general interactions, all three variables were varied in a continuous manner. The onset of the mask was varied from 20 to 35 model cycles and crossed with occlusion, which was varied from 0% to 50% (with contrast fixed at the level in the experiment's control condition, 100%) as well as contrast, which was varied from 100% to 20% (with occlusion fixed at the level in the experiment's control condition, 20%). In both cases, the model produced an interaction between the mask and degradation (Figure 3C–D). Generally speaking, as the amount of degradation increased, recognition became increasingly susceptible to the effect of masking, even at relatively long latencies. For short latency masks (e.g., 20 cycles of processing in the model), it is even possible to impair recognition at low levels of degradation and the multiplicative effects of these short latency masks extend to higher levels of degradation.

The general predictions of the model in Figure 3C–D suggest that the amount of recurrent processing necessary to preserve the recognition of an occluded image varies monotonically with the amount of signal in the stimulus, consistent with the idea that processing within recurrent neural circuits strengthens responses that may have been weakened by visual degradation. The dynamics of the model illustrate precisely how this strengthening through recurrent processing unfolds over time as well as how it is disrupted by masking. We demonstrate these dynamics by comparing the similarity of the model's responses to a degraded stimulus to the responses to a clear stimulus during each cycle of processing (see Appendix for details). The results of this analysis are shown in Figure 4 for occlusion, but the same general principles hold true for contrast reduction and presumably other forms of stimulus degradation.

Under relatively clear viewing conditions, there is sufficient signal in the stimulus to drive the responses in early visual areas (denoted V2/V4 in the model), which in turn drive the responses in successively higher-level areas (IT and category-tuned neurons in the model), rapidly recovering the correct stored representation (Figure 4A).

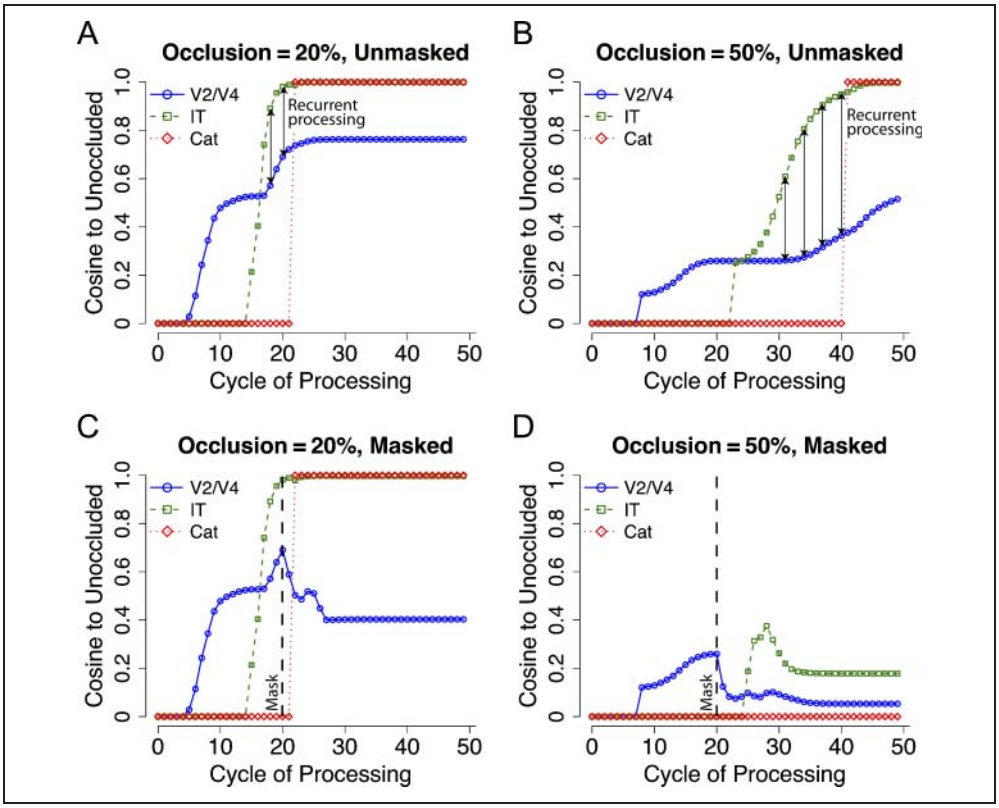

Figure 4. Recurrent processing dynamics. These plots were created by comparing the similarity of the model's responses to an occluded stimulus to the responses to an unoccluded stimulus during each cycle of processing on a single trial. Similarity was measured as the cosine of the angle between the activation vectors of each of the model's layers (V2/V4 = extrastriate cortex, IT = inferotemporal cortex, Cat = category-tuned units). (A) Under relatively clear viewing conditions (Occlusion = 20%, the level used in the experiment's control condition), there is sufficient signal in the stimulus to drive extrastriate units, which in turn quickly produces an IT activation pattern that is a close match to a stored pattern. These factors translate to the activation of the correct category unit. (B) At higher levels of occlusion (e.g., 50%, the level used in the experiment's high occlusion condition), significant recurrent processing is necessary for correct categorization. There is little signal in the stimulus, and extrastriate responses are weak and asymptote early. As the IT activation pattern approaches a stored pattern, it reinforces extrastriate responses, which in turn provide additional support to IT units, ultimately activating the correct category unit. (C, D) Masking the image after 20 cycles of processing impairs recognition of the heavily occluded image but leaves recognition of less occluded stimulus intact. The mask specifically interferes with the recurrent processing between the IT and extrastriate cortex, which begins shortly after 20 model cycles in the heavily occluded case. These same dynamics are present under less occluded case but are less critical and have already completed by the time the mask appears.

Recognition under heavy degradation is still possible but proceeds more slowly. Early visual areas exhibit weak, ambiguous responses because of an underspecified stimulus, but reinforcement from higher-level areas can strengthen and rectify them, ultimately pulling the system as a whole into a stable attractor that is a close, if not exact, match with the stored representations across brain areas (Figure 4B).

Recurrent dynamics between visual areas play a more critical role in driving the activity of category-tuned neurons when stimuli are heavily degraded. Excitatory reinforcement from higher-level visual areas provides additional input to neurons that were attenuated or even prevented from firing because of the weakened bottom–up stimulus, which in turn provides additional bottom–up signals in the absence of the visual information itself. These additional signals are propagated downstream to neurons that are involved in the recognition decision, where they accumulate over time until there is sufficient "evidence" to activate a population of category-tuned neurons. These same dynamics are present under relatively clear viewing conditions but are quicker to manifest and do not last quite as long as when stimuli are heavily degraded because the system converges at an overall faster rate (Figure 4A, compared with Figure 4B). When these well-specified stimuli are masked, the mask influences the responses of early visual areas, but the strong, stable representations that have coalesced throughout higher-level areas are unaffected by the mask (Figure 4C). If we were to take a snapshot of the brain at the same point during the processing of a heavily occluded stimulus, high-level visual areas would be engaged in recurrent processing with earlier areas, and responses would just be beginning to stabilize (Roland, 2010; Lamme & Roelfsema, 2000). Interjecting a mask during this time interrupts the recurrent communication between areas and prevents the full recovery of the stored representation, impairing categorization (Figure 4D).

It is specifically the recurrent connectivity between visual areas that promotes the strengthening and representation completion dynamics described here. Purely feedforward models do not demonstrate completion effects because bottom–up signals are limited by the information present at the fovea. Without top–down reinforcement from higher-level areas, early visual responses remain attenuated from degradation, limiting new information from becoming available to downstream neurons (Figure 5). This effect can be viewed in accordance with research suggesting that backward masking selectively disrupts feedback (see Di Lollo

et al., 2000; Lamme & Roelfsema, 2000, for reviews). If the brain is engaged in recurrent processing of a degraded stimulus and is forced to process a new incoming stimulus (e.g., a pattern mask), categorization processes only have access to whatever information was recovered before the mask was encoded. If a stimulus is heavily degraded, far less information will be available in the purely bottom–up signal, and similarly, far less information will be recoverable before the encoding of a mask.

Finally, we address the possibility that the observed interactions between masking and occlusion and masking in contrast reduction in participants' data were because of a ceiling effect. Strictly speaking, participants' performance was not at ceiling because the levels of occlusion and contrast used in the experiment produced a significant masking effect in the control condition, $t(15) = 3.01$, $p = .01$. Nevertheless, our model can actually be interpreted as predicting a genuine ceiling effect, in the sense that the recognition process completes to near-asymptotic levels before the onset of the mask in the low levels of degradation, whereas the mask interferes with the ongoing processing in higher levels of degradation. This is not a general statistical confound, but instead represents the exact mechanistic prediction that a recurrent system like the brain makes. Furthermore, this prediction is otherwise difficult to motivate for a purely feedforward model, which strictly should not benefit from additional processing time, because no additional information or constraints become available to the system with increased processing time (Figure 5).

DISCUSSION

The goal of the present research was to better understand the feedforward- and feedback-based excitatory dynamics that give rise to robust object recognition processes. Visual object recognition in cortex has traditionally been conceptualized as the net computational result of feedforward processing performed in the ventral visual stream (DiCarlo et al., 2012; Serre, Kreiman, et al., 2007; Serre, Oliva, & Poggio, 2007; VanRullen, 2007; Riesenhuber & Poggio, 1999). Despite the prominence of this view, anatomical data indicate massive levels of recurrent feedback connectivity throughout the visual streams (Sporns & Zwi, 2004; Felleman & Van Essen, 1991). Although feedback is commonly held to influence perception at a later time point than feedforward processing (Hochstein & Ahissar, 2002), single-cell recordings from the dorsal visual stream suggest that feedback plays a primary and immediate role in the computations that give rise to perception by strengthening neural responses in early visual areas (Hupe et al., 1998, 2001; Lamme et al., 1998). Consistent with these dorsal stream data, our results suggest that feedback plays a similar role in the ventral visual stream during object recognition.

By using backward masking to selectively disrupt influence from recurrent processing mechanisms (Boehler et al., 2008; Fahrenfort et al., 2007, 2008; Lamme et al., 2002), we were able to infer its function during the recognition of degraded visual stimuli. We found that masking differentially impaired the recognition of stimuli that were heavily occluded or had restricted contrast compared with relatively clear control stimuli, suggesting that recurrent processing was crucial for recognition of degraded stimuli, but not as important for clear stimuli. This effect was characterized in our results by significant interactions between the mask and occlusion and the mask and contrast.

We were able to closely model the effect of masking on degraded object recognition using a cortical model of object recognition from the literature (O'Reilly et al., under review) without any modifications. The model provided a close fit to subjects' data using the same levels of degradation used in the experiment and also produced a more general interaction effect of masking across a broad range of levels for occlusion and contrast and mask latencies. Although the behavioral experiment did not vary mask latency, the general effect reported in the literature is an impairment in accuracy for short mask latencies with

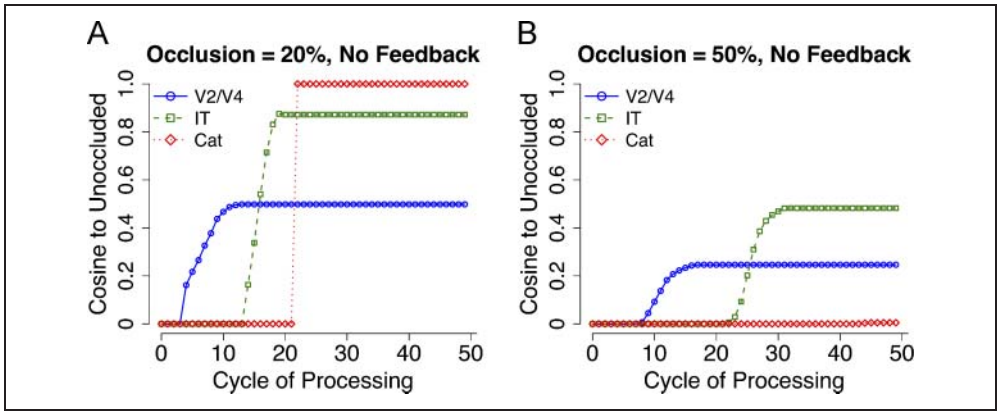

Figure 5. The role of feedback in recurrent processing dynamics. Removing feedback connectivity (e.g., from IT neurons back to extrastriate areas V2 and V4) causes activations across areas to asymptote quickly after the first responses. Without top–down reinforcement from feedback, the similarity of an area's activation pattern to the corresponding unoccluded activation pattern is simply a function of the amount of occlusion. This holds true for IT activation patterns, which often fully completed to the unoccluded activation pattern in the recurrently connected model. (A) With 20% occlusion, the purely bottom–up signals are often strong enough to drive correct categorization without significant recurrent processing. (B) With 50% occlusion, the purely bottom–up responses are weakened compared with the unoccluded responses. Without significant recurrent processing between areas to strengthen these signals, categorization can become impaired.

asymptotic gains in accuracy for longer latencies (Bacon-Mace, Mace, Fabre-Thorpe, & Thorpe, 2005; Rieger, Braun, Bulthoff, & Gegenfurtner, 2005). We expanded on these results by demonstrating that degradation interacted with mask latency in a multiplicative manner because of the increasing amount of recurrent processing that was needed to amplify increasingly degraded stimulus signals.

Our findings are also generally consistent with previous work that has used neuroimaging to study the effect of backward masking on object recognition processes. For example, Grill-Spector, Kushnir, Hendler, and Malach (2000) varied the presentation duration of object stimuli followed by a pattern mask and measured fMRI activation in the lateral occipital complex (LOC), an object-selective region that is thought to contain response complexity and specificity similar to primate IT cortex. Their results indicated that LOC activation decreased with shorter presentation duration, and this activation decrease caused a near-identical impairment in behavioral recognition performance. Similar findings have been demonstrated using magneto-encephalography recording, where the effect of backward masking is a decrease in the amplitude in magneto-encephalography response components related to discrimination (Noguchi & Kakigi, 2005). The exact same effect has been reported for the figure-ground segregation tasks, which have been shown to crucially depend on recurrent processing for successful segmentation (Fahrenfort et al., 2007, 2008). Altogether, these results suggest that object-related signals grow stronger over time up to the point of response saturation or until a new stimulus, such as a mask, enters the processing stream. Our modeling simulations suggest that it is specifically recurrent excitation between brain areas that underlie the amplification of signals over time.

In light of our findings and those of others, it is worth considering more specifically how exactly masking interferes with and ultimately disrupts the signal amplification properties of recurrent processing. Lamme and Roelfsema (2000) suggest that masking creates a mismatch between feedforward- and feedback-based neural responding. Following the presentation of an initial stimulus, responses propagate through the ventral visual stream in a feedforward manner, ultimately reaching processing sites that contain backprojections to earlier areas. If a mask is encoded at a short latency following the initial stimulus, mask-specific responses will propagate forward through the ventral visual stream at the same time that responses encoding the initial stimulus are propagating backward through recurrent connections, creating a mismatch between the feedforward- and feedback-based neural responses and impairing the perception of the original stimulus. Consistent with this view, masking has been shown to cause a decoupling in the functional connectivity (i.e., coactivation) between low-level and high-level visual areas, which has the psychological effect of greatly reduced perceptual visibility (Haynes, Driver, & Rees, 2005; Dehaene et al., 2001).

Our work expands on Lamme and Roelfsema's (2000) theory of recurrent processing and masking by demonstrating the function that the coupled processing between recurrently connected visual areas serves in the context of degraded object recognition. The model dynamics demonstrated in our simulations indicate that recurrent connections between hierarchically adjacent areas provide excitatory reinforcement to each other. This reinforcement is crucial when inputs are underspecified, such as when stimuli are occluded or have reduced contrast, because bottom–up signals are weak and ambiguous. Feedback from higher-level areas can evoke responses in early areas, providing additional bottom–up reinforcement in the absence of the stimulus-driven input. In our simulations, this extra excitation caused the layer that corresponded to IT cortex to form a close (often exact) match to a stored pattern, regardless of the level of degradation. This effect has been described as "object completion," during which missing object features are filled in and has been observed in human LOC imaging studies (Lerner, Harel, & Malach, 2004; Lerner, Hendler, & Malach, 2002; Kourtzi & Kanwisher, 2001). Our model predicts that it is specifically recurrent processing that underlies object completion effects, because purely feedforward models do not exhibit these completion effects. This simple prediction could be explicitly tested by extending our experiment with fMRI methods.

Cumulatively, our results provide insight into exactly why robust recognition is possible in the face of occlusion, although the responses of IT neurons are attenuated (Nielsen et al., 2006; Kovacs et al., 1995). Although the initial responses from IT neurons are weak and ambiguous, recurrent processing between IT and other areas amplifies the responses and rectifies them over time progressively approximating the "object-complete" representation associated with unoccluded stimuli. Recurrent processing and object completion effects take substantial time to manifest due to the multiple bottom–up and top–down constraints that must be integrated into the responses. Accordingly, Kovacs et al. (1995) found that shape-selective IT responses took longer to manifest for occluded images compared with unoccluded images, suggesting that object-complete responses were driving shape selectivity.

Recurrent processing also explains why information analyses of IT responses have indicated that the amount of information embedded in their responses increases over time (Heller et al., 1995; Rolls & Tovee, 1995). Recurrent processing ensures that novel information not explicitly present in the ambiguous stimulus-driven signals becomes integrated into the cumulative response. Furthermore, backward masking has been shown to greatly reduces the amount of information that can be read out of IT responses (Rolls, Tovee, & Panzeri, 1999), presumably because irrelevant information not correlated with the original stimulus-driven response becomes integrated into the total response decreasing the effective signal-to-noise ratio. Given that occluded stimuli require

substantially more recurrent processing than clear stimuli to evoke selective responses in IT neurons, it makes sense that masking would cause a more severe impairment in the recognition of degraded stimuli.

The model of recurrent processing advocated here that is centered around the amplification of signals can be contrasted with "predictive coding" models of recurrent feedback (e.g., Rao, 1999; Rao & Ballard, 1997; see also Friston, 2009, 2010). These models propose that the fundamental role of feedback in the brain is to generate internal predictions of incoming information. These predictions are compared with the incoming information to compute a residual error signal, which is propagated forward to the next processing area where another prediction is generated, repeating the process as the signal ascends the hierarchy of processing areas. The primary difference between predictive coding models and the model of recurrent feedback advocated here is that the former requires feedback to be inhibitory to compute the residual error signal. Recurrent feedback in the brain is exclusively excitatory, and although there are ways to address the biological implausibility of predictive coding (e.g., Spratling, 2008), perhaps the larger issue is that it predicts that degradations that underspecify stimuli like the ones used in the present research should generate higher response levels than relatively clear stimuli, because they will tend to differ greatly from stored memories. In contrast, all the neural data we are aware of show that the effect of occlusion and contrast reduction is an overall decrease in response levels across brain areas (Nielsen et al., 2006; Williford & Maunsell, 2006; Kovacs et al., 1995; Sclar, Maunsell, & Lennie, 1990). Furthermore, processing over time increases response levels for degraded stimuli (Kovacs et al., 1995), opposed to the decrease that predictive coding would suggest as the residual error associated with a degraded stimulus decreases. Thus, an excitatory model of recurrent processing that amplifies weak signals over time provides a better overall fit with the biology of recurrent processing in the brain and is more consistent with the neural data on degraded object recognition.

We also consider the possibility that the interaction between the mask and degradation can be explained by purely feedforward mechanisms. Feedforward accounts of masking assume the existence of at least two distinct processing channels—one for processing the original stimulus and one for processing the mask—with different speeds of information transmission (Breitmeyer & Ganz, 1976). The relatively faster transmission speed of the mask channel allows a mask to "catch up" with the original stimulus and impair its processing via interchannel inhibitory mechanisms. Although the present experiment was not explicitly designed to rule out this feedforward explanation, this explanation is unlikely, especially considering there is little neural evidence that interchannel inhibition exists as a masking mechanism (Fahrenfort et al., 2007; Enns & Di Lollo, 2000). Furthermore, the feedforward explanation of masking was originally posed to explain metacontrast masking, in which the masking stimulus does not spatially overlap the original stimulus yet still impairs its recognition. The pattern masks used in our experiment, however, contain considerable spatial overlap with the preceding stimulus, and thus, spatial-based inhibition should be minimal. At the highest levels of the ventral stream where pattern masking has been reported to impair object recognition (Grill-Spector et al., 2000; Rolls et al., 1999), receptive fields subtend large portions of the visual field (generally 10–20°; Rust & Dicarlo, 2010; Kobatake & Tanaka, 1994), and thus, spatial-based inhibition at this level is virtually nonexistent. Altogether, these points cast serious doubt on feedforward-based masking as a viable explanation of our results.

Despite the robustness conferred by recurrent processing, precise timing data from neural recordings impose the strict temporal constraint of as little as 100–150 msec on the computations that can take place before category-tuned brain areas become active, providing strong support for models based on a single feedforward sweep of responses (VanRullen & Koch, 2003; Li, VanRullen, Koch, & Perona, 2002; VanRullen & Thorpe, 2001; Thorpe et al., 1996). These tasks generally only involve a binary decision about whether an image contains a target object (such as an animal), compared with the six-way categorization task used in the present research. Thus, it might be the case that target detection tasks can be solved relatively well with only feedforward responses (Serre, Oliva, & Poggio, 2007), whereas recognition tasks like ours that contain significantly more feature overlap and uncertainty across a larger number of categories require recurrent processing and thus take longer for the brain to resolve. This view is consistent with more conservative estimates of the time course of object recognition (e.g., 150–300 msec reported by Johnson & Olshausen, 2003). Alternatively, it could be the case that target detection tasks already reflect some degree of influence from recurrent processing. In accordance with this latter view, recent reports have cited recurrent processing effects occurring as early as 100–150 msec (Koivisto, Railo, Revonsuo, Vanni, & Salminen-Vaparanta, 2011; Roland, 2010; Bar et al., 2006; Foxe & Simpson, 2002; Lamme & Roelfsema, 2000).

Conclusions

The research described here demonstrates one of the fundamental roles of recurrent processing during object recognition: creating a strong, stable representation that promotes robust recognition. This is especially important when visual stimuli are not viewed under idealized conditions, which is a frequent occurrence during real-world vision. Objects in our field of view are often occluded, shrouded under complex lighting and shadows, and generally suffer from countless other sources of environmental variability. Standard models of object recognition that use only a single series of feedforward computations for

recognizing stimuli (Serre, Kreiman, et al., 2007; Serre, Oliva, & Poggio, 2007; VanRullen, 2007; Riesenhuber & Poggio, 1999) depend on sufficient signal in the visual stimulus for recognition but often break down under these realistic, suboptimal conditions (O'Reilly et al., under review).

Overall, it has become apparent that vision is a highly interactive, dynamic process that depends on multiple brain areas at different levels of the ventral visual hierarchy participating in processing (Roland, 2010; Spivey, 2007). Whether these dynamics are reflected during object recognition in the initial preattentive representations or are due to late, top–down attentional effects remains an open question. Furthermore, the exact time course of feedforward and recurrent computations that give rise to object recognition processes is not well understood, because recent reports are suggesting that recurrent processing might manifest much faster than initially thought. Ultimately, research that focuses on short timescale neural dynamics and biologically realistic computational modeling will be necessary to fully understanding the computations and interactions that give rise to object recognition.

APPENDIX

LVis Model

The LVis model (Leabra Vision model) and the training/testing methods as they relate to the present simulations are briefly described here. See O'Reilly et al. (under review) for a detailed description. The model consists of a hierarchy of feature processing layers that roughly correspond to areas along the ventral stream of the brain—primary visual cortex (V1), extrastriate cortex (V2/V4), inferotemporal cortex (IT), and a categorical output layer that can be conceptualized as either anterior inferotemporal cortex or pFC. The model receives grayscale bitmap images as inputs, which are processed with filtering that captures the relevant computations of the retina and LGN of the thalamus. The results of this filtering are further processed by a chain of V1-like operations—Gabor filtering followed by a spatial "max" operation (e.g., Riesenhuber & Poggio, 1999)—and then used as inputs to the model proper. Subsequent layers in the model contain decreasing numbers of units (V1: 3600 units; V2/V4: 2880 units; IT: 200 units; Output: 200 units, only six of which were used for the six-way categorization task used in the present simulations) as well as increasing receptive field sizes, computed through a series of converging connections.

Overall, the model can be viewed as an extension of a large class of hierarchical feedforward models of visual processing in the brain (e.g., Masquelier & Thorpe, 2007; Serre, Kreiman, et al., 2007; Delorme & Thorpe, 2001; Riesenhuber & Poggio, 1999). The primary innovation of the model is that hierarchically adjacent areas (e.g., V2/V4 and IT) are reciprocally connected, providing an account of the recurrent connectivity observed throughout the ventral stream. Feedforward connections generally contribute 80–90% of the total input to a receiving layer and feedback connections contribute the remaining 10–20% of the total input. As all connections are excitatory, layer activations are controlled using a *k*-winners-take-all (*k*WTA) inhibitory competition rule (O'Reilly & Munakata, 2000; O'Reilly, 1996) that ensures only the *k* most active units remain active over time. The specific *k* value varies for each layer in the model but is generally in the range of 10–20% of the number of units in the layer.

The model is trained using an extension of the Leabra algorithm (O'Reilly & Munakata, 2000; O'Reilly, 1996), which contains both self-organizing and error-driven components. The model learns a sparse distributed representation at the IT level that serves as a translation between the more graded sensory inputs represented at lower levels and categorical outputs. The sparseness of the representation arises from the learning algorithm as well as other Leabra mechanisms (e.g., *k*WTA, recurrent connectivity) that interact over the course of learning.

Simulation Methods

Before simulating the experiment, the model was trained across images from the six categories used in the human experiment. Images were taken from the CU3D-100 data set (http://cu3d.colorado.edu). Two exemplars from each category were reserved to assess generalization performance, which was 91%, averaged across five random combinations of training and testing exemplars (referred to as training/testing splits). There were 980 total training images, distributed roughly equally across the six categories. Each image was presented both during the initial training phase, as well as the subsequent simulations with small variations in foveal position, scale, and planar rotation. These 2-D variations were important for the model's ability to learn an invariant representation similar to that coded by IT neurons (Serre, Kreiman, et al., 2007; Riesenhuber & Poggio, 1999) and also ensured that each presented image was likely to be unique. Although the model was optimized for generalization, the simulations described below were conducted using the training images. This was done to prevent confounding the effects of occlusion and masking with any performance impairment because of generalization. Additional simulations using the testing images indicated that the results were slightly noisier, but qualitatively similar.

Each model was trained for 25,000 trials total using an extension of the Leabra algorithm (O'Reilly & Munakata, 2000; O'Reilly, 1998), and the resulting trained weights were used for the subsequent simulations. Separate simulations were performed for each combination of masking, occlusion, and contrast variables. Each of these simulation types used the same five training/testing splits and corresponding weights. Masks were constructed in the same manner as in the experiment, except that pregenerating a

large number of masks was largely unnecessary because of the model being prevented from learning during the actual simulated experiment. A total of 100 masks were pregenerated for use in the simulations.

The simulated experiment consisted of seven presentations of each of the 980 training images with variations in foveal position, scale, and planar rotation. The presented image was degraded using the same occlusion and contrast manipulations as during the experiment. Normally, the model's inputs are normalized after filtering the image to enhance their dynamic range, but this step was omitted for all simulations, as normalization would undo the effects of contrast reduction. For simulations during which the mask was absent, the image was clamped into to the model's inputs and the model was iterated for 50 cycles after which the category associated with the most active output unit was recorded as the model's response. For simulations during which the mask was present, presentation of the image proceeded in the same way, but the model's inputs were subsequently reclamped with a randomly chosen mask before the 50th model cycle. Any activation associated with the original image that had been established before the remapping was preserved and allowed to interact with any new activation that was established as a result of processing the mask. The mask image remained clamped into to the model's inputs for the remainder of the 50 cycles, after which the model's response was recorded.

No assumptions were made about how the time integration represented by a single cycle of the model's processing mapped onto the passage of time in the physical world. Thus, to find the equivalent of presenting a mask with the latency of 100 msec used in the experiment, the processing cycle that the mask was mapped into the model's inputs was varied as a free parameter to find the best fit of subjects' data. The model was fit to subjects' data by minimizing the sum squared error of the effect of the mask, combined across the three experimental conditions (Control, Occlusion, and Contrast). For the data fits in depicted in Figure 3B, occlusion and contrast were held constant at the levels used in the experiment. Because subjects' accuracy scores were higher overall compared with the model, we computed model outputs using a majority vote across the seven 2-D variations in foveal position, scale, and planar rotation while holding the occlusion manipulation fixed (i.e., if the model's outputs for the seven 2-D variations were *fish, fish, car, fish, car, fish, fish*, the final voted output would be computed as *fish*). This voting procedure reflects variability because of fixation error and head position during the experiment, which can be used in an aggregate manner during the recognition process (Ratcliff & McKoon, 2008; Bradski & Grossberg, 1995; Ratcliff, 1978). The amount of improvement from the voting procedure was around 3–4% across conditions. For the more general predictions depicted in Figure 3C that did not require absolute fits, this voting procedure was omitted.

The plots in Figure 4 were created by first presenting an unoccluded image to the model and recording the activation patterns in each of the annotated model areas. The model was then presented with the same image, but with an occlusion pattern applied. During each cycle of processing, the similarity between the current activation pattern in each of the model layers and the corresponding unoccluded activation pattern was calculated by taking the cosine of the angle between the two activation vectors. A value of 1.0 indicated that the two vectors were identical, and for model units that represent category-tuned neurons (denoted with "Cat" on the plots), this indicated a correct response. To explore the effect of removing feedback on the model's activation patterns, feedback weights between category-tuned units and IT units as well as IT units and extrastriate units were multiplied by zero when computing the activations across each layer.

Acknowledgments

The authors would like to thank Matt Jones for his helpful comments and suggestions. This research was funded by NSF grant SBE-0542013 to the Temporal Dynamics of Learning Center (an NSF Science of Learning Center) and ONR grant N00014-10-1-0177 to Randall O'Reilly.

Reprint requests should be sent to Dean Wyatte, Department of Psychology and Neuroscience, University of Colorado Boulder, Muenzinger D244, 345 UCB, Boulder, CO 80309, or via e-mail: dean.wyatte@colorado.edu.

REFERENCES

Bacon-Mace, N., Mace, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research, 45,* 1459–1469.

Bar, M., Kassam, K., Ghuman, A., Boshyan, J., & Schmidt, A. (2006). Top–down facilitation of visual recognition. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 449–454.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57,* 289–300.

Boehler, C. N., Schoenfeld, M. A., Heinze, H. J., & Hopf, J. M. (2008). Rapid recurrent processing gates awareness in primary visual cortex. *Proceedings of the National Academy of Sciences, U.S.A., 105,* 8742–8747.

Bradski, G., & Grossberg, S. (1995). Fast-learning viewnet architectures for recognizing three-dimensional objects from multiple two-dimensional views. *Neural Networks, 8,* 1053–1080.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Breitmeyer, B. G., & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review, 83,* 1–36.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., & Poggio, T. (2007). A model of v4 shape selectivity and invariance. *Journal of Neurophysiology, 98,* 1733–1750.

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J.-F., Poline, J.-B., et al. (2001). Cerebral mechanisms of work

masking and unconscious repetition priming. *Nature Neuroscience, 4,* 752–758.
Delorme, A., & Thorpe, S. (2001). Face identification using one spike per neuron: Resistance to image degradations. *Neural Networks, 14,* 795–803.
Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology, 129,* 481–507.
DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73,* 415–434.
Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4,* 345–352.
Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience, 19,* 1488–1497.
Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2008). The spatiotemporal profile of cortical processing leading up to visual perception. *Journal of Vision, 8,* 12.1–12.12.
Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1,* 1–47.
Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from v1 to frontal cortex in humans. A framework for defining "early" visual processing. *Experimental Brain Research, 142,* 139–150.
Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience, 23,* 5235–5246.
Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13,* 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11,* 127–138.
Geisser, S., & Greenhouse, S. W. (1958). An extension of box's results on the use of the f distribution in multivariate analysis. *Annals of Mathematical Statistics, 29,* 885–891.
Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience, 3,* 837–843.
Haynes, J.-D., Driver, J., & Rees, G. (2005). Visibility reflects dynamic changes of effective connectivity between v1 and fusiform cortex. *Neuron, 46,* 811–821.
Heller, J., Hertz, J. A., Kjaer, T. W., & Richmond, B. J. (1995). Information flow and temporal coding in primate pattern vision. *Journal of Computational Neuroscience, 2,* 175–193.
Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36,* 791–804.
Hupe, J. M., James, A. C., Girard, P., & Bullier, J. (2001). Response modulations by static texture surround in area v1 of the macaque monkey do not depend on feedback connections from v2. *Journal of Neurophysiology, 85,* 146–163.
Hupe, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by v1 v2 and v3 neurons. *Nature, 394,* 784–787.
Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3,* 499–512.
Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway. *Journal of Neurophysiology, 71,* 856–867.
Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in v1/v2 contributes to categorization of natural scenes. *The Journal of Neuroscience, 31,* 2488–2492.
Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science (New York, N.Y.), 293,* 1506–1509.
Kovacs, G., Vogels, R., & Orban, G. A. (1995). Selectivity of macaque inferior temporal neurons for partially occluded shapes. *The Journal of Neuroscience, 15,* 1984–1997.
Lamme, V., & Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23,* 571–579.
Lamme, V., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology, 8,* 529–535.
Lamme, V., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in v1. *Journal of Cognitive Neuroscience, 14,* 1044–1053.
Lerner, Y., Harel, M., & Malach, R. (2004). Rapid completion effects in human high-order visual areas. *Neuroimage, 21,* 516–526.
Lerner, Y., Hendler, T., & Malach, R. (2002). Object-completion effects in the human lateral occipital complex. *Cerebral Cortex, 12,* 163–177.
Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, U.S.A., 99,* 9596–9601.
Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing timing timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron, 62,* 281–290.
Masquelier, T., & Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology, 3,* 247–257.
Nielsen, K. J., Logothetis, N. K., & Rainer, G. (2006). Dissociation between local field potentials and spiking activity in macaque inferior temporal cortex reveals diagnosticity-based encoding of complex objects. *The Journal of Neuroscience, 26,* 9639–9645.
Noguchi, Y., & Kakigi, R. (2005). Neural mechanisms of visual backward masking revealed by high temporal resolution imaging of human brain. *Neuroimage, 27,* 178–187.
O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8,* 895–938.
O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences, 2,* 455–462.
O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
O'Reilly, R., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. (under review). Recurrent processing during object recognition.
Pelli, D. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.
Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. *Cognition, 67,* 111–145.
Rao, R. P. (1999). An optimal estimation approach to visual perception and learning. *Vision Research, 39,* 1963–1989.
Rao, R. P. N., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation, 9,* 721–763.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–107.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922.

Rieger, J. W., Braun, C., Bulthoff, H. H., & Gegenfurtner, K. R. (2005). The dynamics of visual pattern masking in natural scene processing: A magnetoencephalography study. *Journal of Vision, 5,* 275–286.
Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 3,* 1199–1204.
Roland, P. (2010). Six principles of visual cortical dynamics. *Frontiers in Systems Neuroscience, 4,* 1–21.
Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology, 73,* 713–726.
Rolls, E. T., Tovee, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: Information analysis. *Journal of Cognitive Neuroscience, 11,* 300–311.
Rust, N. C., & Dicarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *The Journal of Neuroscience, 30,* 12978–12995.
Sclar, G., Maunsell, J. H., & Lennie, P. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vision Research, 30,* 1–10.
Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research, 165,* 33–56.
Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, U.S.A., 104,* 6424–6429.
Spivey, M. J. (2007). *The continuity of mind*. New York: Oxford University Press.
Sporns, O., & Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics, 2,* 145–162.
Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience, 2,* 1–8.
Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522.
VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology, 3,* 167–176.
VanRullen, R., & Koch, C. (2003). Visual selective behavior can be triggered by a feed-forward process. *Journal of Cognitive Neuroscience, 15,* 209–217.
VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13,* 454–461.
Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The shine toolbox. *Behavior Research Methods, 42,* 671–684.
Williford, T., & Maunsell, J. H. R. (2006). Effects of spatial attention on contrast response functions in macaque area v4. *Journal of Neurophysiology, 96,* 40–54.