Title

Timecourse of neural signatures of object recognition

Journal

Journal of Vision, 3(7)

ISSN

1534-7362

Authors

Johnson, Jeffrey S

Olshausen, Bruno A

Publication Date

2003-08-28

DOI

10.1167/3.7.4

Peer reviewed

Timecourse of neural signatures of object recognition

Jeffrey S. Johnson Center For Neuroscience, UC Davis, Davis, CA, USA

Bruno A. Olshausen Center For Neuroscience, UC Davis, Davis, CA, USA and Redwood Neuroscience Institute, Menlo Park, CA, USA

How long does it take for the human visual system to recognize objects? This issue is important for understanding visual cortical function as it places constraints on models of the information processing underlying recognition. We designed a series of event-related potential (ERP) experiments to measure the timecourse of electrophysiological correlates of object recognition. We find two distinct types of components in the ERP recorded during categorization of natural images. One is an early presentation-locked signal arising around 135 ms that is present when there are low-level feature differences between images. The other is a later, recognition-related component arising between 150-300 ms. Unlike the early component, the latency of the later component covaries with the subsequent reaction time. In contrast to previous studies suggesting that the early, presentation-locked component of neural activity is correlated to recognition, these results imply that the neural signatures of recognition have a substantially later and variable time of onset.

Keywords: object recognition, visual cortex, electrophysiology, ERP, ERPimage

Introduction

In the real world, survival is contingent upon the rapid, accurate recognition of objects. Despite extensive research in both neuroscience and computer science, how this task is accomplished remains a mystery.

One view is that object recognition constitutes a chicken-egg type problem between lower and higher levels of image analysis. The low-level shape features that are useful for identifying an object - edges, contours, surface curvature and the like - are typically ambiguous in natural scenes, so they cannot be computed directly based on a local analysis of the image. Rather, they must be inferred based on global context and higher-level knowledge. However, the global context itself will not be clear until there is some degree of certainty about the presence of low-level shape features. A number of theorists have thus argued that recognition depends on information circulating through cortico-cortical feedback loops in order to disambiguate representations at both lower and higher levels in parallel (Mumford, 1994; Ullman, 1996; Lewicki & Sejnowski, 1997; Rao & Ballard, 1999; Lee & Mumford, In Press).

However, a large body of both neurophysiological and psychophysical data suggests that recognition is performed so rapidly that it must be accomplished in one feedforward sweep of activity propagated through the visual system (Fukushima, 1980; Mel, 1997; Riesenhuber & Poggio, 1999; VanRullen & Thorpe, 2001a; VanRullen & Thorpe, 2002). For example, rapid serial visual presentation (RSVP) techniques (Intraub, 1999) and other masking paradigms (Breitmeyer, 1984) reveal behavioral selectivity to images presented for 100 ms or less. Single unit recordings in macaque superior temporal sulcus show that face-selective activity can be elicited by masked images presented for as little as 14 ms (Keysers, Xiao, Földiák & Perrett, 2001) and human functional magnetic resonance imaging (fMRI) studies have shown activations in object recognition areas to be correlated with masked image presentations as brief as 40 ms (Grill-Spector, Kushnir, Hendler & Malach, 2000). This sort of result is difficult to reconcile with models requiring feedback because the original image should be replaced by a subsequent one by the time signals from higher level areas are fed back to lower levels.

Further, it has been argued based on known biophysical and anatomical constraints (e.g. speed of spike propagation, number of cortical processing stages) that the earliest responses in macaque higher-level, object recognition areas must be based on one or at most two spikes per neuron in the pathways leading up to them (Thorpe & Imbert, 1989). This hypothesis is supported by latencies of about 100 ms reported for neurons in macaque inferotemporal cortex (IT) (Nowak & Bullier, 1997). Response latencies and reaction times in recognition tasks are a bit later in humans than macaques (Fabre-Thorpe, Richard & Thorpe, 1998), so that in humans, intracranial EEG studies show the earliest response latencies in the facial recognition areas of the fusiform gyrus to be 130-140 ms (Allison, Puce, Spencer & McCarthy, 1999). This 30-40 ms delay is presumably due to the increased conduction times associated with larger head size.

A study of the timecourse of object recognition in humans has revealed a component in the electroencephalogram (EEG) arising approximately 150 ms after image presentation that is related to the detection of objects in complex natural scenes (Thorpe,

DOI 10:1167/3.7.4 Received March 31, 2003; published August 28, 2003 ISSN 1534-7362

Fize & Marlot, 1996). Importantly, this signal is presentation-locked, meaning that the timing of its onset is constant across trials and is not correlated with the subsequent reaction time. The latency of this component is also constant for both novel and previously learned images (Fabre-Thorpe, Delorme, Marlot & Thorpe, 2001). Thus, EEG differences seen in object recognition tasks are nearly as fast as the earliest known responses in higher-level areas. These results seem to leave little or no opportunity for recurrent activity in cortical feedback loops to have an effect on the earliest responses in IT, and suggest that object recognition is performed in a rapid, feedforward manner without regard to the particular object present, the context, or familiarity.

One difficulty in using methods such as EEG or other physiological signals to measure the timecourse of object recognition arises from the fact that various visual features are not equally probable across image categories. For example, certain spatial frequencies, textures, colors, and simple spatial patterns may be associated with a given category of images without necessarily being diagnostic for the objects themselves. Low-level features such as these could be detected in parallel and might give rise to a purely stimulus-related pre-recognition EEG signal that varies across categories of images. This sort of signal could be easily confused with target-related signals that depend upon the completion of object recognition. As such, when comparing the average EEG waveforms obtained from different categories of images, any differences could be due to non-specific visual processing as well as the task relevance of the stimuli. Interestingly, experiments that control for featural differences by mixing together images from different categories (VanRullen & Thorpe, 2001b) also show differences in the EEG which rise to significance around 150 ms, but the topographical distribution and overall timecourse of these difference signals are not the same as those obtained when different categories are compared. This raises the question as to whether there are different underlying neural processes at work in the two cases.

Here we address this question by examining the nature of the EEG signals obtained in two types of categorization tasks. In the first task, subjects categorize stimuli into animal and nonanimal categories, so that there are potential low-level feature differences between target and nontarget images, similar to previous experiments (Thorpe, Fize & Marlot, 1996; Fabre-Thorpe et al., 1998; Fabre-Thorpe et al., 2001). In the second task, the target categories change from trial to trial and the images assigned to these categories are arranged so as to ensure that the pool of target and nontarget images is identical across subjects. Thus, any EEG differences between targets and nontargets in the second task must be contingent upon recognition rather than due to image features. We show that the earliest differences reliably contingent upon recognition appear between 152 and 300 ms after presentation and have an onset which covaries with the subsequent reaction time. Earlier differences, which appear by 137 ms and are presentation-locked, only occur when comparing signals evoked by images with featural differences, suggesting that they are due to differences in early visual processing rather than the result of recognition. These results suggest that the timecourse of object recognition, while still quite rapid, is not as fast as previously thought. We discuss the implications of these results for models of recognition.

Methods

Participants

A total of thirty-nine adult subjects (10 males and 29 females, aged 18 to 41 years) participated in the four experiments reported in this study. Seven subjects participated in Experiment 1, twelve in Experiment 2, twelve in Experiment 3, and eight in Experiment 4. All participants had normal or corrected to normal vision. All participants gave informed consent and the UC Davis Human Subjects IRB approved all studies.

Stimuli

Sample images used in the four experiments described here are shown below (Figure 1). Images used in Experiment 1 were taken from a commercially available collection (Corel Photo CDs, out of production) and included images of animals, flowers, and other outdoor scenes. The images used in Experiments 2 and 4 were created by digitally centering a cutout image of an object (Hemera Photo-Objects) against one of thirty artificially created backgrounds. The backgrounds were created to match the images from Experiment 1 in terms of both their spatial and color second order statistics (pairwise correlations) by generating 1/f noise for each of the principal components in color space and scaling each by the square root of the variance along each component. Experiment 2 also used these background images without a superimposed object ("Background-Only"). Images from Experiment 3 were collected from the internet. The following links provide additional, full-sized examples of the images used in Experiment 1, Experiments 2 and 4, and Experiment 3.

Experimental Procedures

Two types of experiments were used in this study, a "single-category" experiment and three "cued-target" experiments.

In the single-category case, subjects were given one category ("animal") which served as the target for the duration of the experiment. Images were presented for 40 ms and were not masked. There was a delay of 3300-3700 ms between images (Figure 2). The timing of the image

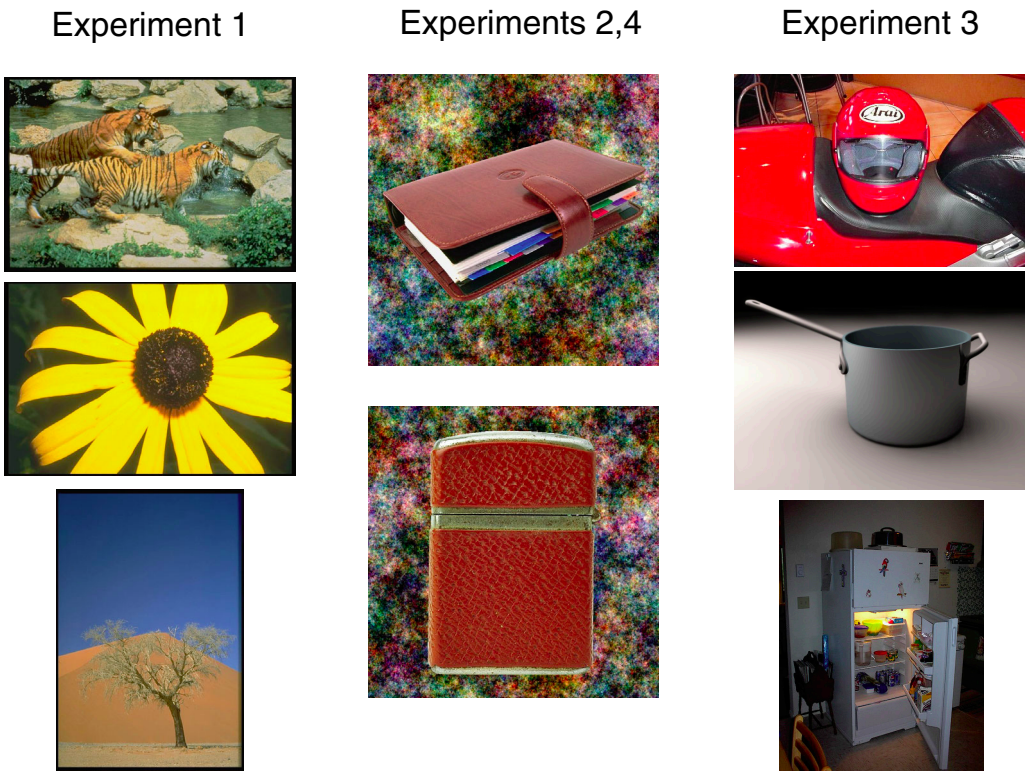

Figure 1. Sample images. Images in Experiment 1 subtended 15x10˚ or 10x15˚ of visual angle. Images from Experiments 2 and 4 subtended 15x15˚. Images from Experiment 3 subtended an average of 11.4x13.0˚. The hyperlinks in the legend lead to web pages with full-sized examples.

sequence did not depend on the subject's responses. Subjects were asked to make button-press responses as quickly as possible and to delay their blinks for about 2 seconds after each image presentation.

In the cued-target case, a word (target cue) was presented on the screen before each image to inform subjects of the target category for that trial. The target cue remained on the screen until the subject was ready to go and performed a button press to initiate the trial. After an additional 500-700 ms delay, the image was presented for 40 ms, not masked. The next target cue appeared 1700 ms after the previous image (Figure 2). Subjects were asked to make button-press responses as quickly as possible and to delay their blinks after each image until the subsequent target cue appeared.

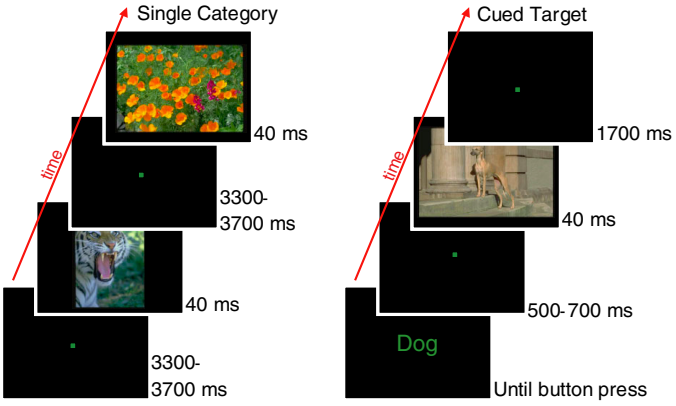

Figure 2. Schematic of task. Single-category Task: Participants were informed of a target category ("animal") before the experiment. Cued-target Task: Participants were informed of a target category, which varied widely, before each image presentation by means of a word presented on screen.

All images were centrally presented on a CRT monitor. Image presentation was controlled by a PC running the software Presentation (NeuroBehavioral Systems). Viewing distance was 75 cm except for Experiment 3 where the viewing distance was 65 cm.

Experiment 1: Single-category experiment. This experiment had two conditions: "forced choice", where subjects pressed one button if the image contained a target and another button if the image contained no target, and "go/no-go", where subjects pressed one button if the image contained a target and did nothing if the image contained no target. Counterbalances were made across subjects for button presses and for given image appearance in the forced choice or go/no-go condition. Six subjects viewed 500 images in the forced choice condition and 500 images in the go/no-go condition. One subject viewed 1000 images in the forced choice condition only. Targets and nontargets appeared equally in both forced choice and go/no-go conditions. Results

from nontarget images containing flowers were analyzed separately and are not included in this study.

Experiment 2: Cued-target, 1/f Backgrounds experiment. Target cues were all formulated for entry level classification (Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976; Jolicoeur, Gluck & Kosslyn, 1984). Subjects viewed 990 images and made a forced choice response. Target objects, nontarget objects, and background-only images (which required a 'nontarget' response) appeared equally frequently. Counterbalances were made across subjects for button presses and target status of individual images (except background-only images). Results from background-only images were analyzed separately and are not included in this study.

Experiment 3: Cued-target, Natural Backgrounds experiment. Target cues were presented at two levels of abstraction, entry level and superordinate level. Cues of both levels of abstraction were randomly interleaved. Subjects viewed a total of 1000 images. Superordinate level targets, superordinate level nontargets, entry level targets, and entry level nontargets appeared equally frequently. Because some superordinate categories were less intuitive than others, subjects were briefed on the superordinate categories before the experiment. Counterbalances were made across subjects for target status and required level of classification of individual images.

Experiment 4: Cued-target, Motor Related Control experiment. 1/f background images used in Experiment 4 all contained objects; no background-only images were used. This experiment had two conditions, Respond-Target and Respond-Nontarget. Both conditions were go/no-go and consisted of 500 images. Targets and nontargets appeared equally frequently. The Respond-Target condition was run in its entirety before the Respond-Nontarget condition. Subjects were not informed of the upcoming change in condition until it occurred. Counterbalances were made across subjects for target status and response mode condition (forced choice or go/no-go) of individual images.

The following links provide full lists of the target cues used in Experiment 2 (entry level) and Experiment 3 (entry level and superordinate level), along with some associated data including number of times used, and average reaction time. The target cues used in Experiment 4 were identical to those in Experiment 2, and are not included.

EEG Recording and Data Analysis

Subjects were fitted with a 19-channel electrode cap (Electro-Cap International, Eaton, Ohio) and were prepared for EEG recording according to standard techniques. Recorded channels (FP1, FP2, F7, F3, FZ, F4, F8, T7, C3, CZ, C4, T8, P7, P3, PZ, P4, P8, O1, O2) were selected from the International 10-20 set of electrode positions (American Electroencephalographic Society, 1994). In addition to the cap electrodes, facial electrodes were attached to record horizontal and vertical electrooculogram (EOG). All recordings were referenced to the right mastoid. Subjects performed the experiment in a darkened, sound-dampened, electrically shielded booth. EEG signals were amplified (SA Instrumentation, San Diego) with a high-pass cutoff of 100 Hz and a low-pass cutoff of 0.01 Hz, then sent through an analog-to-digital converter before being recorded at 256 samples/sec on a PC running Digitize (Arthur Jones, LBNL).

Raw data were normalized, artifact rejected, and analyzed using Matlab software developed in-house. Software for the display of scalp topographies was developed by Scott Makeig (Salk Institute, San Diego). Data were artifact rejected on a trial-by-trial basis for eyeblink and on a channel-by-channel basis for drift, blocking and excessive alpha wave.

To create the ERPimages, individual correct-response EEG trials were assigned to 3.9 ms (one sample) wide bins on the basis of reaction time (RT) and an average was calculated for each bin. Bins with RTs between 300 and 600 ms were sorted by RT. Each averaged bin was then vertically expanded, with an expansion factor proportional to the number of EEG trials underlying the average. In the case of the difference ERPimages, the averaged nontarget bin was subtracted from the averaged target bin to create one difference wave at each RT. These difference waves were then sorted by RT and expanded, with an expansion factor proportional to the lesser of the number of EEG trials, target or nontarget, underlying the original RT bins before subtraction. All ERPimages were then smoothed vertically with a Gaussian filter having a standard deviation of 1/50 the height (number of expanded trials) of the plot.

To perform the spatial frequency image analysis, the center 512x512 pixel region of each image was extracted and windowed with a 1-cycle cosine function (to remove boundary artifacts). The power spectrum of each windowed image was calculated by the discrete Fourier transform, and then averaged over all images for a particular class. The average power spectrum was then bandpass filtered into 9 one-octave bands spaced by 1/2 octave, at 8 orientations each. This resulted in an 8x9 element array containing the power in each band, which was then interpolated in 2D with a cubic spline.

Results

Image Statistics

In the "single-category" paradigm, target and nontarget categories were composed of different sets of images – "animal" images and "nature" images. These two sets of images could potentially differ in terms of low-level image statistics. For instance, textures or shapes associated

with fur, eyes and other animal parts, the horizon, sky, trees, land terrain, water and other image features are likely to be unevenly distributed across the animal and nature categories. This could potentially give rise to measurable neural processing differences unrelated to object recognition in striate or early extrastriate cortex.

While there are a host of potential differences in low-level features and textures between the two sets of images, differences in the power spectrum are the easiest to demonstrate. A spatial-frequency analysis of the animal and nature images (excluding images containing flowers) used in our single-category experiment reveals substantial differences in the power spectrum – namely, scenes containing animals have approximately equal power in all directions whereas those without animals have much more power in the vertical spatial frequencies (Figure 3). Similar differences in the average power spectra between major categories of images have been shown previously (Oliva & Torralba, 2001).

The design of the "cued-target" paradigm ensured that each image appeared equally in both the target and nontarget conditions, across subjects. Since the set of target images and the set of nontarget images were identical, there were no differences in low-level features, spatial frequency or otherwise, between the images. The only difference between target and nontarget images was in their conceptual status, not in their featural content. Two types of images were used in the cued-target experiments, one with cutout photos of objects digitally placed atop an artificial 1/f background ("1/f BG"), and another with objects in their natural context (Figure 1).

Target Minus Nontarget Signals

The trial-averaged EEG waveform – known as the event-related potential (ERP) – was computed separately for target and nontarget stimuli. These two ERPs were directly compared by subtracting the nontarget ERP from the target ERP. For both single-category and cued-target tasks, the resulting difference waveform shows an early positive divergence centered around frontal and central electrodes.

The grand average ERP waveforms for the forced choice, single-category task at electrode FZ are plotted in Figure 4a. The first differences between targets and nontargets arise in the first positive deflection, which begins about 100 ms after image onset. The ERP difference waveform for this task (Figure 4f, black) has two separate peaks, an early peak rising before 150 ms and a late peak starting around 300 ms. Both peaks have a maximum amplitude of about 4 μV. To determine when the ERPs recorded for targets and nontargets first significantly differ, we used a two-sample t-test for difference of means. In order to avoid counting spurious deflections from baseline, we considered the onset of difference to be the first sample point for which the calculated *p*-value was less than 0.01 and for which the following 9 samples (10 consecutive points) also reached the same criterion. By this definition, targets and nontargets differed as early as 137 ms after presentation. The timecourse of this difference signal is similar, but slightly earlier than that reported in previous single-category studies (Thorpe, Fize & Marlot, 1996).

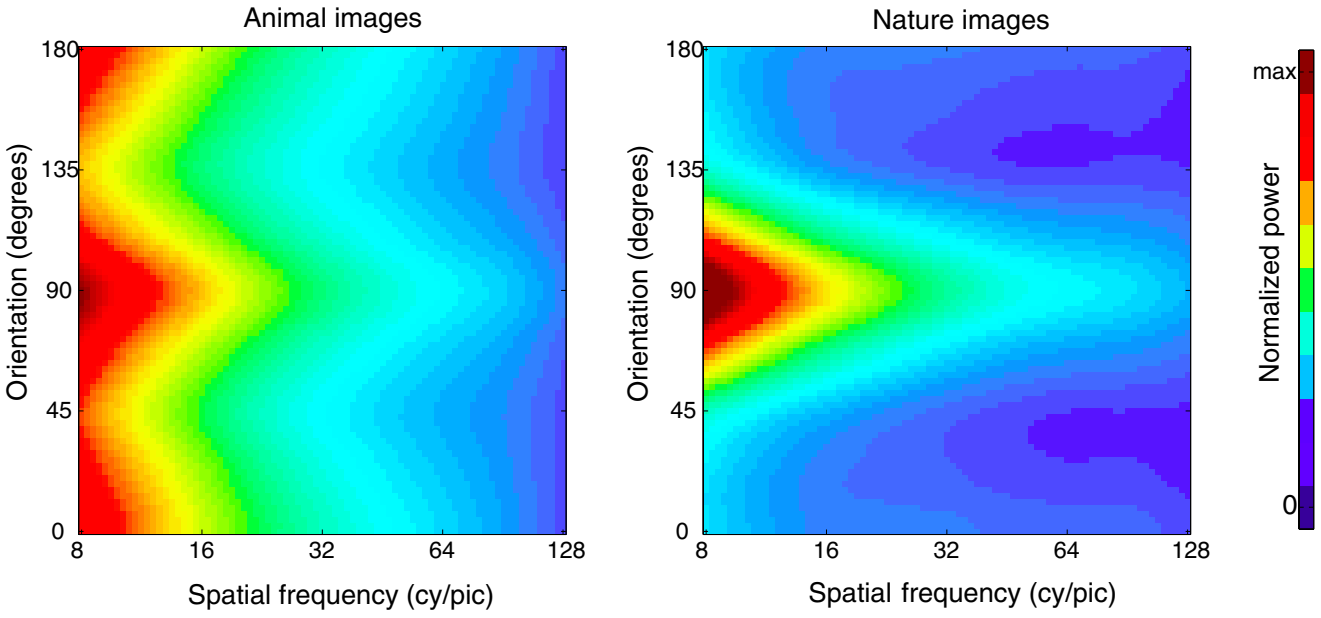

Figure 3. Power spectrum of animal and nature images. Each plot shows the average power spectrum (in log-polar coordinates) for 100 randomly selected images from each class (see methods). The animal images have a more even distribution of power among different orientations. The strong anisotropy in the nature images is most likely due to the presence of the horizon, which produces strong power at 90 degrees orientation in the spatial-frequency domain. Both plots are normalized to the same scale.

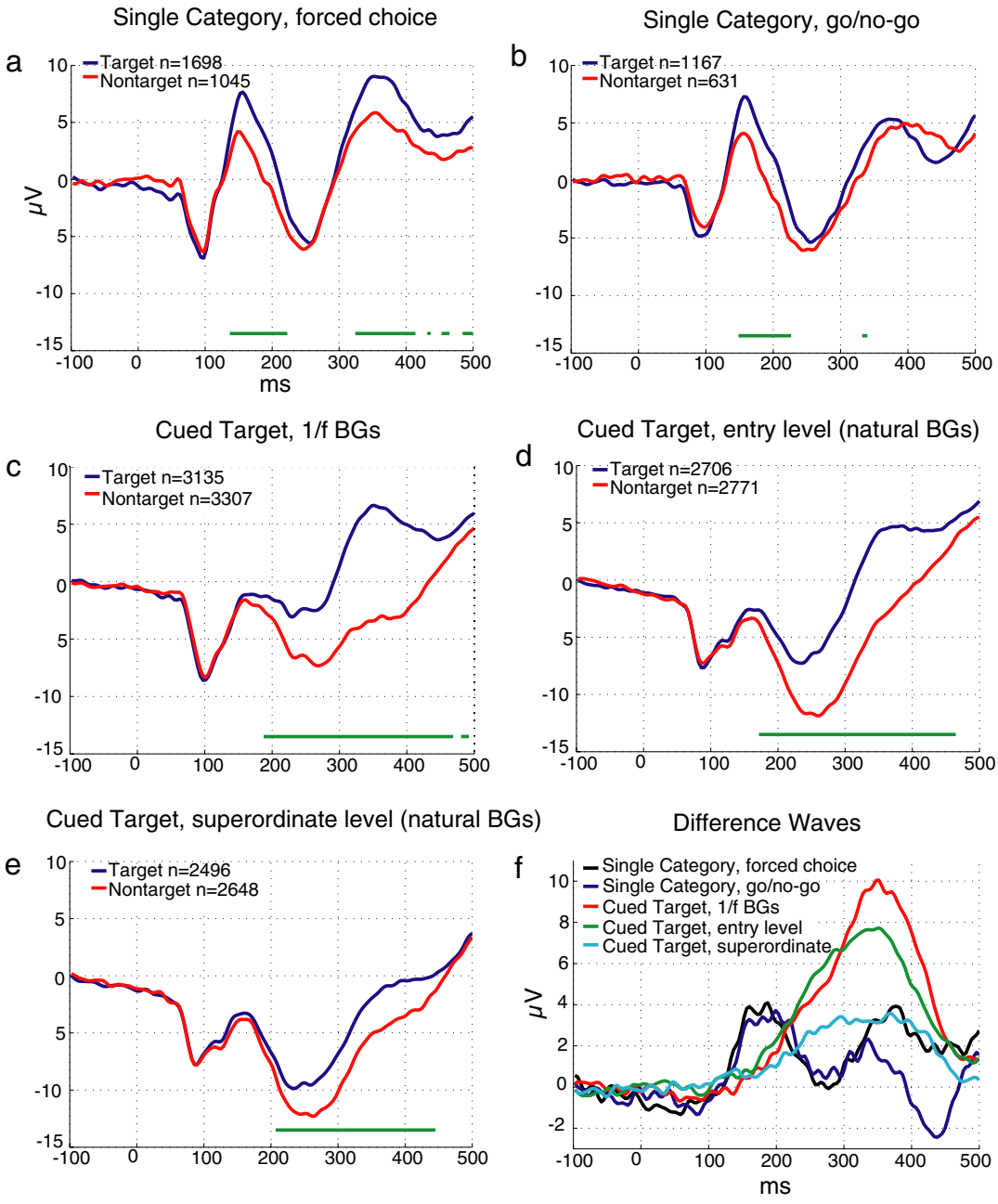

Figure 4. Difference ERPs for Cued-target tasks have a slower rise and higher amplitude than difference ERPs for Single-category tasks. All potentials calculated at electrode FZ. In (a)-(e), green bar below indicates timepoints where waveforms differ significantly ($p < 0.01$). (a) Single-category task, forced choice. (b) Single-category task, go/no-go. (c) Cued-target task, 1/f BGs. (d) Cued-target, entry-level task, natural BGs. (e) Cued-target, superordinate-level task, natural BGs. (f) Difference waves. Black trace (target minus nontarget, single-category, forced choice task) has two peaks, one early (~150-250 ms) and one late (>300 ms). Dark blue trace (single-category, go/no-go task) also has two peaks. The earliest difference is quite similar for both single-category conditions. Red trace, (cued-target task, 1/f BGs), green trace (cued-target, natural BGs, entry-level task) and light blue trace (cued-target, natural BGs, superordinate-level task) have only one, later peak. The amplitude of the differences during entry level categorization are much greater than those in the single-category task (a superordinate categorization), but the amplitude of difference in the superordinate condition of the cued-target task is similar to the amplitude of second difference peak in the single-category task. Time of first differences ($p < 0.01$, two sample t-test, ten consecutive samples): Black trace = 137 ms, dark blue trace = 148 ms, red trace = 187 ms, green trace = 171 ms, light blue trace = 207 ms.

Because previous single-category studies were done with go/no-go responses instead of forced choice responses, we also ran the single-category experiment in a go/no-go response mode. The waveforms for the go/no-go experiment are shown in Figure 4b and the corresponding difference between targets and nontargets in Figure 4f, dark blue. The ERPs and difference waveforms are quite similar, especially in the first 300 ms, to those from the forced choice version of the single-category task. Here, targets and nontargets first differed 148 ms after presentation.

The ERPs for the cued-target task using 1/f backgrounds (Figure 4c) show that the first positive deflection has a much lower amplitude than in the single-category condition. Moreover, while targets and nontargets are clearly different by the peak of the first positivity in the single-category task, they are only beginning to differ at the corresponding time in the cued-target task. The difference waveform obtained in this task (Figure 4f, red trace) rises much more slowly (first significance 187 ms) than those in the single-category task, and has only one clear peak with a much higher late maximal amplitude of 10 μV. This positive, posterior signal during target detection is similar in latency, though not in amplitude, to previous reports of target-related signals (VanRullen & Thorpe, 2001b), and bears similarity to those found in many other target detection studies (Sutton, Braren, Zubin & John, 1965; Picton, 1992).

Aside from the fundamental change in task, there were two other changes between the single-category experiment and the 1/f BG cued-target experiment which might have affected the nature of the difference waveforms. The first was that in the 1/f background cued-target experiment, pictures of cutout objects appeared against artificial backgrounds rather than in their natural context. To determine if this caused the changes in the difference waveforms, a subsequent cued-target experiment was run with objects in their natural context (see Figure 1). In this case, both the ERPs (Figure 4d) and the difference waveforms (Figure 4f, green) resemble the 1/f BG cued-target task. They feature a single peak with a maximal amplitude of 7.75 μV and the difference is first statistically significant 172 ms after presentation. Despite the disparity of the image backgrounds in the two versions of the cued-target task no qualitative differences are evident between the two, suggesting that the signals that we are recording are relatively insensitive to the nature of the image background.

A second change between the single-category and 1/f BG cued-target experiments was the level of abstraction of the requested categorization. In the single-category experiment, the target category was at a superordinate level of abstraction ("animal"), whereas in the 1/f BG cued-target experiment, the target categories were at a lower, entry level of abstraction (e.g. "cat", "chair") (Rosch, et al. 1976, Jolicoeur, Gluck & Kosslyn, 1984). In the natural image cued-target experiment, participants also categorized some images at the superordinate level (e.g. "animal", "furniture") to match the categorizations made in the single-category task. Again, the ERPs for these trials resembled those in the 1/f BG cued-target condition (Figure 4e). The difference waveforms (Figure 4f, light blue) maintain the single peak typical of cued-target tasks, and have an even slower onset (first significance 207 ms). However, the maximal amplitude of the difference drops to 3.5 μV, similar to that of the single-category difference.

Scalp topographies reveal dramatic differences between the distribution of these potentials across both space and time. The scalp topography of the single-category (forced choice) difference captures the two peaks seen in the ERP difference waveform (Figure 5a). The first peak consists of a frontal positivity coupled with a strong

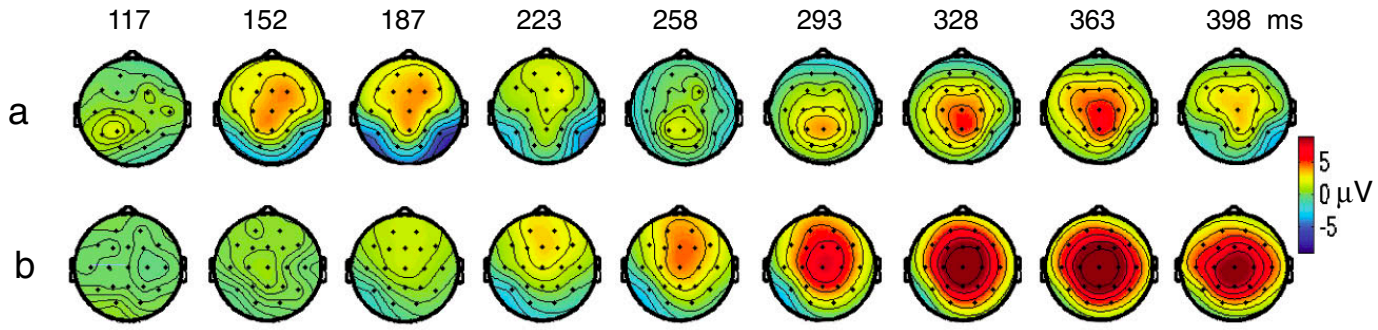

Figure 5. Earliest differences in Single-category Task and Cued-target task differ in ERP scalp topography. Time series of topography plots from 117-398 ms. (a) Single-category task, target minus nontarget. The early difference peak shows a frontal positivity and occipito-temporal negativity. The late difference peak shows central positivity with little occipito-temporal negativity. (b) Cued-target task, 1/f BGs, target minus nontarget. One peak, with a timecourse similar to the late peak in (a) shows central positivity with little occipito-temporal negativity.

bilateral occipito-temporal negativity, while the second peak is a central positivity and lacks the negative occipito-temporal component. The single peak of the 1/f BG cued-target experiment (Figure 5b) has a time course and scalp distribution which closely resemble those of the second peak of the single-category experiment, and lacks the bilateral occipito-temporal negativity characteristic of the first peak.

Presentation-Locked and Reaction-Time Dependent Signals

To further investigate the differences between the single-category and cued-target conditions, we utilized a modified version of the ERPimage (Jung, Makeig, Westerfield, Townsend, Courchesne & Sejnowski, 1999; Makeig, Westerfield, Jung, Enghoff, Townsend, Courchesne & Sejnowski, 2002). The ERPimage shows the EEG waveforms from all trials sorted by reaction time, and thus clearly reveals the presentation-locked and reaction-time related components contributing to the ERP. Individual EEG trials were separated by target status, and each group was sorted and binned by reaction time to create an ERPimage. We then compared the binned ERPimages for targets and nontargets by subtracting the latter from the former. This difference ERPimage reveals the trial-by-trial differences in the EEG for trials having the same reaction time (see methods for more detail on ERPimage creation). ERPimages were created for both the single-category (forced choice) experiment and the 1/f BG cued-target experiment.

The target and nontarget ERPimages at electrode FZ (Figures 6a, 6b) show presentation-locked activity as vertical structure. Activity that is correlated with the reaction time has a diagonal structure, often closely following the displayed reaction time (RT) curve (solid black line).

The difference ERPimage for the single-category task (Figure 6c, left) shows that the earliest differences arise from presentation-locked components. The dashed vertical line indicates the onset of statistical significance for these trials (see Figure 4) and the target minus nontarget positivity is evident at the onset of statistical significance regardless of reaction time. Difference ERPs (Figure 6d, left) created for fast-RT trials (300-400 ms response) and slow-RT trials (400-500 ms response) do not differ greatly in amplitude or onset time in the single-category task.

The difference ERPimage for the 1/f BG cued-target task (Figure 6c, right) shows that the earliest differences are not only slower but are also dependent on the reaction time for a given trial rather than presentation-locked. The target minus nontarget differences on fast-RT trials rise more quickly than on slow-RT trials. In contrast to the single-category task where all trials appeared to contribute equally to the statistical significance, in the cued-target task only the fastest trials appear to be responsible for the earliest statistically significant differences. In fact, in the slowest trials shown in Figure 6c (right), the difference does not appear to arise before 300 ms. Difference ERPs for fast- and slow-RT trials (Figure 6d, right) also reflect the RT-dependent component of this difference.

We calculated the time of first statistical significance for fast- and slow-RT trials for both the single-category and cued-target tasks shown in Figure 6d. Using the same statistical criterion as before ($p < 0.01$, ten consecutive samples) we see a time lag for first significance in the cued-target task, with fast-RT trials first differing at 187 ms and the slow-RT trials first differing at 219 ms, a delay of 32 ms. If we adopt a stricter statistical criterion ($p < 10^{-4}$, five consecutive samples), this delay extends to 63 ms (fast-RT = 195 ms, slow-RT = 258 ms). There was a small 11 ms delay in the single-category results under the original criterion (fast-RT = 137 ms, slow-RT = 148 ms) that disappeared using the stricter criterion (fast-RT = slow-RT = 152 ms).

The onsets of the earliest differences seen in the cued-target case are correlated with the subsequent reaction time, while the onsets of those in the single-category case are, if anything, only marginally correlated with RT. The development of RT-dependence, taken together with changes in overall latency, number of peaks, peak amplitude, and scalp topography, very strongly suggest that the earliest ERP differences in the cued-target case are not simply a delayed form of the same signal seen in the single-category case. Since the fast, presentation-locked differences are seen only in the single-category case - where the images contain demonstrable low-level feature differences - but not in the balanced cued-target case, it seems likely that they are due to differences in early visual processing rather than the completion of object recognition.

Motor-Related Control

In order to determine if the target-related signal we found could be due to motor effects (motor planning, motor execution, motor countermand) rather than target awareness, we performed a task swapping experiment. Subjects were given a cued-target task with 1/f BG images, but asked to respond in a go/no-go fashion only to targets for the first half of the experiment, then only to nontargets for the second half. Subjects were not informed of the impending reversal in response mode until the time came to change. The target minus nontarget difference in the first half of the experiment corresponded to a response minus no-response condition. In the second half of the experiment, the target minus nontarget difference corresponded to the reverse case: a no-response minus response condition. If the signals we see are due to differences in motor preparation or output

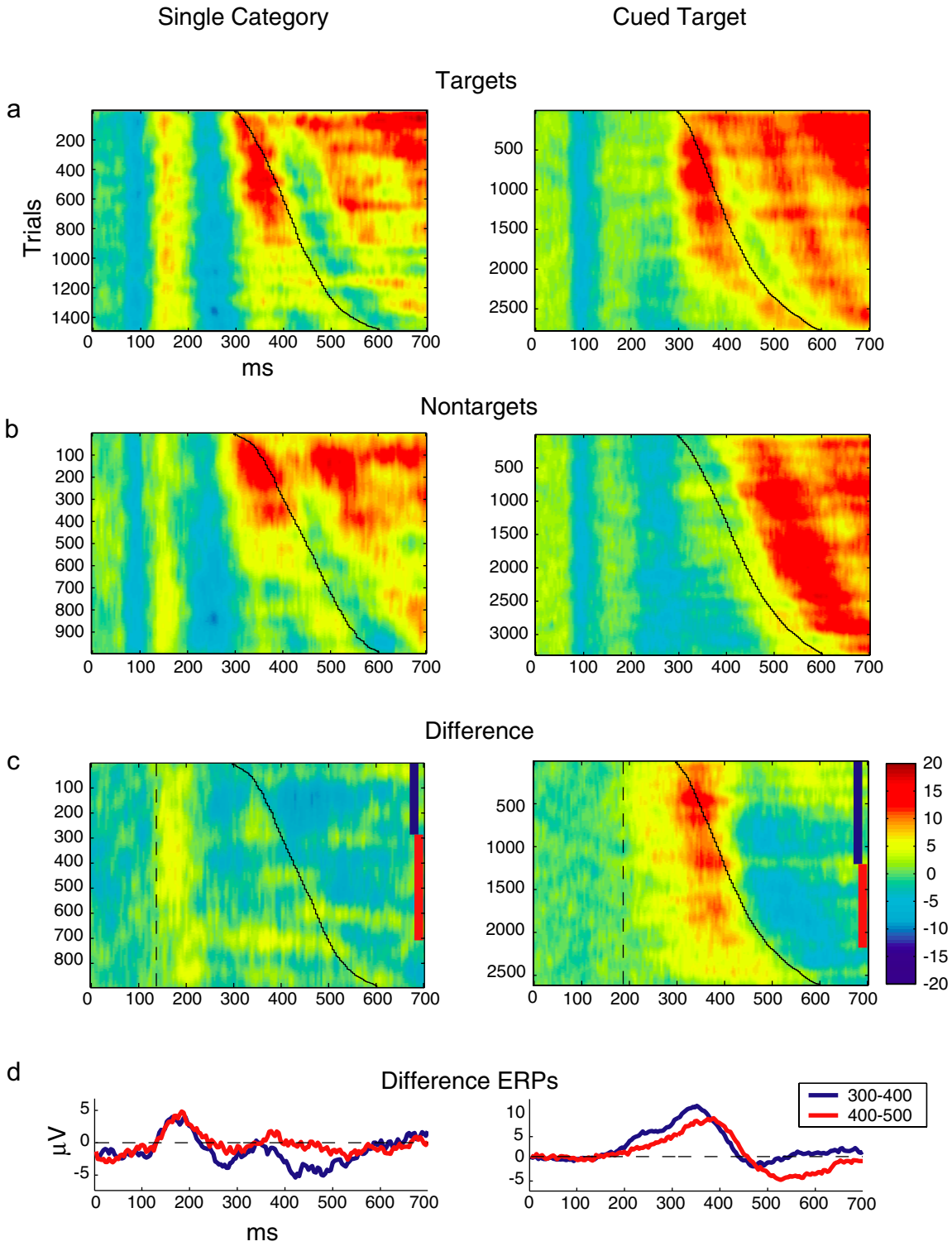

Figure 6. ERP Images show presentation-locked and RT-dependent activity in target minus nontarget comparisons. Black curve indicates reaction time on each trial. All plots are at electrode FZ. (a) Target ERP Images. Early activity is presentation-locked, later activity has some correlation with RT. (b) Nontarget ERP Images. (c) Target minus Nontarget differences. Vertical dashed line shows first time of significance (see Figure 3). Vertical solid lines (blue = 300-400 ms, red = 400-500 ms) identify trials used in fast-RT and slow-RT ERPs in (d) below. The first difference in the single-category condition is presentation-locked. The first difference in the cued-target condition rises more quickly on fast-RT trials than slow-RT trials. (d) Difference ERPs, split by RT. Blue traces, RT = 300 – 400 ms. Red traces, RT = 400 – 500 ms. The first differences for fast-RT and slow-RT trials are identical for the single-category condition. In the cued-target condition, the fast-RT differences rise more quickly than slow-RT differences.

between target and nontarget responses, we should expect them to reverse in sign and follow the motor response rather than the target status of the image in the second half of the experiment. The ERP waveforms and difference waveforms (Figure 7) are similar to those for cued-target conditions in duration, amplitude, and scalp topography. They clearly do not reverse sign as would be expected for a motor-related signal suggesting that they are not due to motor effects but rather to the conceptual status of the image as target or nontarget.

In the respond-target condition, the difference becomes significant 152 ms after presentation. This is the earliest difference seen in any cued-target task in this study. The difference waveforms rise somewhat later during the more difficult respond-nontarget condition, becoming significant at 176 ms.

Reaction Times and Accuracy

One possible concern in comparing results from the single-category and cued-target tasks is that the cued-target task, with its switching targets, may be inherently more difficult that the single-category task. A comparison of the reaction times and accuracy for targets in the single-category task (Table 1, 427 ms, 94.1%) and the corresponding superordinate categorizations in the cued-target task (520 ms, 86.9%) might suggest that the cued-target response mode is more difficult than the single-category response mode. While this could be due to the switching of target category, another possibility is that the required categorizations are more difficult in the cued-target case. However, if we look only at cued-target trials for which the target cue was "animal", target responses (438 ms, 95.8%) are slightly more accurate and only 11 ms slower than those seen in the single-category experiment. This suggests that the cued-target paradigm per se is not significantly more difficult than the single-category paradigm, but rather that the single-category target of "animal", though nominally superordinate, is easier than the other superordinate categorizations it was grouped with in Experiment 3. In addition, when the same images are categorized at both superordinate and entry levels, we find a 56 ms delay for superordinate categorization, similar to reports of a 50 ms delay in a similar task using line drawings (Rosch, et al., 1976). This suggests that the superordinate categories we chose were not unusually difficult ones.

Table 1. Accuracy and Reaction Time.

Experiment	Target %	Target RT	Nontarget %	Nontarget RT
1, Forced Choice	94.1	427	96.3	467
1, Go/No-Go	93.7	399	92.3	n/a
2	90.1	432	96.2	450
3, Entry	93.9	464	96.9	507
3, Superordinate	86.9	520	92.5	563
3, "Animal"	95.8	438	97.4	516
4, Go Target	97.3	423	96.5	n/a
4, Go Nontarget	92.4	n/a	98.9	520

Separated by target and nontarget. Experiment 1 separated into forced choice and go/no-go conditions. Experiment 3 separated into entry-level and superordinate-level categorizations. Trials from Experiment 3 where the superordinate category was "animal", most similar to Experiment 1, are included in Experiment 3, superordinate but also shown separately. Experiment 4 separated into go-on-target and go-on-nontarget conditions.

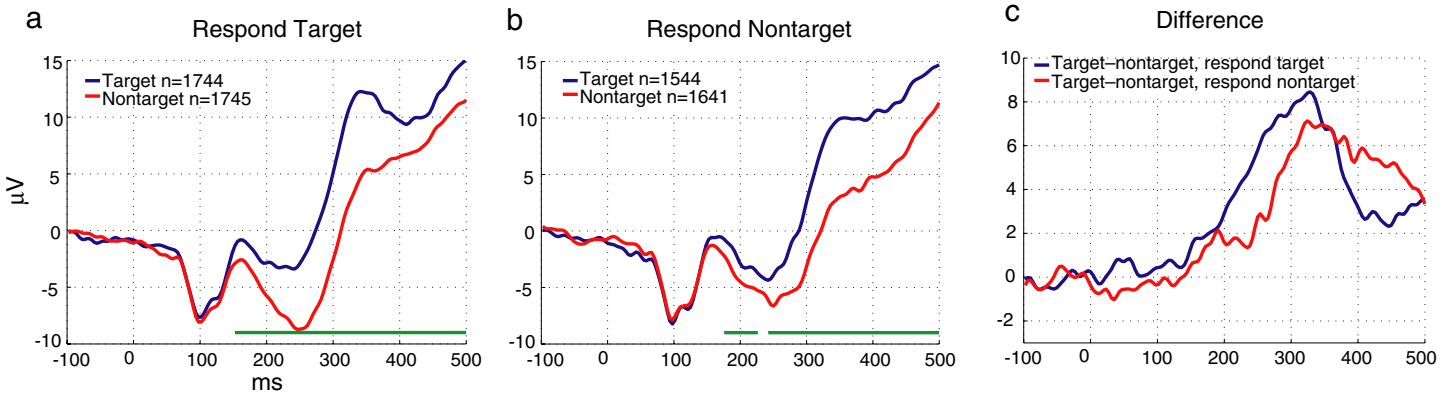

Figure 7. Differences in cued-target task are not due to motor or motor-related activity. ERPs for motor control experiment, electrode FZ. In (a) and (b), green bar below indicates timepoints where waveforms differ significantly ($p < 0.01$). Subjects responded with a button press to only targets or nontargets, depending on condition. (a) Respond-target task. (b) Respond-nontarget task. (c) Difference waves, target minus nontarget. Blue trace, respond-target task. From a motor point of view, the blue trace is response minus no-response. Red trace, respond-nontarget task. From a motor point of view, the red trace is no-response minus response. The major activity does not invert in the no-response minus response condition, as would be the case for motor-related activity. Time of first differences: Respond-target = 152 ms, respond-nontarget = 176 ms ($p < 0.01$, two sample T-test, ten consecutive samples).

Two other observations merit comment. First, go/no-go categorizations appear to be easier than forced choice categorizations, as we would expect from the easier mapping of task to motor output. This change in difficulty seems to be manifested mostly in reaction time in the single-category case (399 ms vs. 427 ms) and in accuracy in the cued-target case (97.3% vs. 90.1%). Second, there is a difference between performance in the entry level cued-target task for natural images (464 ms, 93.9%) and images with artificial backgrounds (432 ms, 90.1%). This might be accounted for by a speed/accuracy tradeoff, but certain characteristics of the natural images (potential 'distractor' objects, object not always centered at fixation, etc.) may also play a part in the slower reaction times in this case.

Discussion

We have shown that the earliest components in the EEG that are correlated with recognition have an onset that varies between 150 and 300 ms and is correlated with the subsequent reaction time. These results stand in stark contrast to previous EEG studies showing early, fixed latency components correlated with object recognition arising around 150 ms (Thorpe, Fize & Marlot, 1996; Fabre-Thorpe et al., 1998; Fabre-Thorpe et al., 2001). These previous studies are equivalent to our "single-category" paradigm, in which we show that the early, fixed-latency components (in our case, arising ~135 ms) are most likely due to low-level feature differences in the images. We therefore conclude that the early 135-150 ms onset, fixed-latency component in the EEG corresponds to differences in early visual processing among images, and that the neural signatures actually corresponding to object recognition occur later and with variable latency.

These studies also demonstrate that a population-wide *p*-value of the ERP is not necessarily representative of the average signal latency. While the population-wide *p*-value is useful for determining the latency of a stimulus-locked signal, as seen in the single-category experiment, it appears from our data to be highly biased by the fastest reaction time trials when estimating the latency of an RT-dependent signal. As such, it is difficult to draw conclusions about the correspondence of ERP differences across experiments on the basis of *p* -value latencies alone. The ERPimage (Jung, et al., 1999) has the advantage of showing both stimulus-locked and RT-dependent components simultaneously, making it more valuable than a simple global *p*-value for the analysis and interpretation of EEG components that may not have a constant latency.

It should be noted that although we have attributed the earliest differences in neural activity in the single-category case to low-level feature differences, this does not preclude these features playing a role in recognition. For instance, Ullman has argued (Ullman, Vidal-Naquet, & Sali, 2002) that pictorial image features of "intermediate complexity" are the most useful features for the discrimination of objects. In addition, it has been shown that natural images can be classified into animal and nonanimal categories at a success rate of 80% using nothing but measures of global image statistics such as the power spectrum (Torralba & Oliva, In Press), suggesting that some low-level features are fairly consistent within categories of images. Thus, the visual system might be able to build a good template of the features associated with a category and use this template to make preemptive categorizations with reasonable accuracy. This sort of template could range from a simple list of features which are expected to be present to a complex mental image of the object in question. If such a template is created, the early differences we ascribe to low-level visual processing may in fact represent an interaction between the bottom-up processing of the image and a top-down target template.

The results from this study can be interpreted to support a top-down template model if it is assumed that in the cued-target case, subjects were not able to adequately create the necessary template. The subjects were forced to change their target for every image in the cued-target experiments, but had only one target for the duration of the single-category experiment. Although the cued-target trials were self-paced (i.e., the subjects were not rushed into a subsequent image presentation before they deemed themselves ready), it is possible that top-down template priming requires a longer time - on the order of tens to hundreds of seconds - to establish. Alternatively, in the single-category case the visual system might learn, over the course of many animal and nature scenes, its own template of features which are commonly associated with each category. This sort of learning would be difficult in the cued-target case because of the large number of different target categories and the low number and non-consecutive presentations of each. If either of these scenarios holds, it is possible that the earliest target minus nontarget differences in single-category tasks represent a special form of recognition based on low-level features of the images rather than simply processing of the image features themselves. However, this sort of feature-based recognition would require a long-term preparatory strategy (as evidenced by the lack of differences in the cued-target case) that is unlikely to be of use in everyday object recognition.

The experiments reported here also provide new insights into the neural signatures of superordinate level categorization. We find that the amplitude of target-related differences in the cued-target, natural images experiment depends greatly on the requested level of categorization, with the amplitude of entry level differences on the order of twice as great as that of superordinate level differences. The amplitudes of these differences compare well with those seen at the same

levels of categorization in Experiments 1 and 2 (Figure 4f), and may reflect a greater level of certainty of categorization. Another possibility is that the entry level cues allow a more precise target template to be created, leading to a better match (as discussed above, albeit this time applying to the later, RT-dependent signal, not the earlier stimulus-locked signal). Previous electrophysiological studies (Tanaka, Luu, Weisbrod & Kiefer, 1999) have suggested that superordinate level categorization occurs as late as 340 ms after image presentation. Given that entry level categorization appears to occur in as little as 152 ms, this would imply that the additional processing required for superordinate categorization might take as much as 200 ms. However, we have found evidence of target/nontarget differences at the superordinate level arising as early as 207 ms, which suggests that superordinate processing takes much less time than previously reported.

While is it not possible to discern whether the recognition-related component we observe in the EEG reflects the act of object recognition, the target decision or yet another process, it seems clear that this component can arise only after the target status of the image is known. Since the knowledge of target status is dependent on successful recognition, our measurements serve as an upper bound for the time of recognition itself. Thus, there is a variable 20-170 ms delay between the hypothetical physiological lower bound of 130 ms in a purely feedforward scheme (Thorpe & Imbert, 1989; Allison, et al., 1999) and our measured upper bound of 150-300 ms.

One hypothesis for this delay is that it reflects the extra time needed for recurrent activity to circulate in cortico-cortical feedback loops. A number of theorists have argued that recurrent processing between higher and lower levels of visual cortex is a necessary aspect of perception. Some have emphasized the role of binding of features (Grossberg, 2001; Knoblauch & Palm, 2001), while others have proposed that it sends the predictions of higher levels to lower levels to "explain away" (Rao & Ballard, 1999), or disambiguate (Lewicki & Sejnowski, 1997; Lee & Mumford, 2003) representations at lower levels. There are numerous examples from neurophysiological studies in animals (Hupe, James, Girard, Lomber, Payne & Bullier 2001; Supèr, Spekreijse & Lamme, 2001), human event related potentials (Murray, Wylie, Higgins, Javitt, Schroeder & Foxe 2002; Noesselt, Hillyard, Woldorff, Schoenfeld, Hagner, Jancke, Tempelmann, Hinrichs & Heinze, 2002) and fMRI (Murray, Kersten, Olshausen, Schrater & Woods, 2002) suggesting that cortico-cortical feedback loops play an active role in perception. Our results, while not a direct demonstration of feedback processes at work, imply that there is at least enough time for considerable recurrent activity to occur prior to recognition. The fact that target-related differences are correlated with the subsequent reaction time suggests that the first feedforward wave of activity through cortex may not always be sufficient to perform reliable recognition and that the routine use of feedback information could be a fundamental component of everyday visual processing.

Yet another possibility is that the delay reflects the extra time required to simply integrate weak or ambiguous signals at various stages along the feedforward pathway. For example, evidence has been shown for neural signals that ramp up during the decision process with a rate that depends on the strength of the stimulus (Gold & Shadlen, 2000), suggesting a neural mechanism for integrating evidence. Although such integration would involve recurrent activity, it is typically hypothesized to occur only within each level of cortical hierarchy, not between them.

One problem for both of these temporal-integration hypotheses, however, is the fact that behavioral performance remains good under strictly masked conditions (Breitmeyer, 1984; Intraub, 1999; Grill-Spector, et al., 2000; Keysers, et al., 2001). A mask that wipes clean the cortical or LGN image should disrupt any obligatory feedback or integrative processes. However, if representations are sufficiently sparse (Olshausen & Field, 1996; Vinje & Gallant, 2000), it may be possible for the representation of the mask or subsequent image to co-exist with the previous activity pattern without destructive interference. This idea is supported by the fact that masks must be individually tailored to different types of images for maximum effectiveness. Thus, it is possible that feedback processes could still operate on persistent, sparse activity left untouched by subsequent visual information.

Despite the somewhat slower timecourse for recognition measured in our experiments, it should be emphasized that the speed with which recognition occurs is still quite impressive. If cortico-cortical feedback processes are indeed a necessary prerequisite to recognition, then the number of iterations in these loops must be fairly small. The real challenge, however, is to understand what exactly is happening in these feedback loops at the level of neuronal representation and how they operate in more realistic, dynamic situations where both the eye and objects in the world are moving. In order to properly design experiments to investigate this, it will be necessary to develop more detailed, neurobiologically-based models of feedback and its role in scene analysis.

Acknowledgments

This work was supported by grant MH57921 (B.A.O.) from the National Institutes of Health. A preliminary version of this work was presented at the 2002 Annual Meeting of the Vision Sciences Society, May 10-15, 2002, Sarasota, FL. Thanks to Charan Ranganath and Scott Murray for draft comments.

Commercial relationships: none.

References

Allison, T., Puce, A., Spencer, D. D., & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral Cortex, 9*, 415-430. [PubMed] [Article]

American Electroencephalographic Society. Guideline thirteen: guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology, 11*, 111-113. [PubMed]

Breitmeyer, B. G. (1984). *Visual Masking: An Integrative Approach*. New York: Oxford University Press.

Fabre-Thorpe, M., Delorme, A., Marlot, C. & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*, 171-180. [PubMed]

Fabre-Thorpe, M., Richard, G. & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9*, 303-308. [PubMed]

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193-202. [PubMed]

Gold, J. I. & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature, 404*, 390-394. [PubMed]

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. [PubMed]

Grossberg, S. (2001). Linking the laminar circuits of visual cortex to visual perception: development, grouping, and attention. *Neuroscience and Biobehavioral Reviews, 25*, 513-526. [PubMed]

Hupe, J.-M., James, A. C., Girard, P., Lomber, S. G., Payne, B.R. & Bullier, J. (2001). Feedback connections act on the early part of the responses in monkey visual cortex. *Journal of Neurophysiology, 85*, 134-145. [PubMed] [Article]

Intraub, H. (1999). Understanding and remembering briefly glimpsed pictures: Implications for visual scanning and memory. In Coltheart, V. (Ed.), *Fleeting Memories: Cognition of Brief Visual Stimuli*, (pp. 47-70), Cambridge, Massachusetts: MIT Press.

Jolicoeur, P., Gluck, M. A. & Kosslyn, S. M. (1984). Pictures and names: making the connection. *Cognitive Psychology, 16*, 243-275. [PubMed]

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E. & Sejnowski, T. J. (1999). Analyzing and Visualizing Single-trial Event-related Potentials. In Kearns, M. S., Solla, S. A. & Cohn, D. A. (Eds.), *Advances in Neural Information Processing Systems 11* (pp. 118-124). Cambridge, Massachusetts: MIT Press.

Keysers, C., Xiao, D.-K., Földiák, P. & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience, 13*, 90-101. [PubMed] [Article]

Knoblauch, A. & Palm, G. (2001). Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Networks, 14*, 763-780. [PubMed]

Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*, 1434-1448.[PubMed]

Lewicki, M. S. & Sejnowski, T. J. (1997). Bayesian learning of higher order structure. In Mozer, M. C., Jordan, M. I. & Petsche, T. (Eds.), *Advances in Neural Information Processing Systems, 9* (pp. 529-535). Cambridge, Massachusetts: MIT Press.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E. & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science, 295*, 690-694. [PubMed]

Mel, B.W. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation, 9*, 777-804. [PubMed]

Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In Koch, C. & Davis, J. L. (Eds.), *Large Scale Neuronal Theories of the Brain*, (pp. 125-152), Cambridge, Massachusetts: MIT Press.

Murray, M. M. Wylie, G. R., Higgins, B. A., Javitt, D. C., Schroeder, C. E. & Foxe, J. J. (2002). The spatiotemporal dynamics of illusory contour processing: combined high-density electrical mapping, source analysis, and functional magnetic resonance imaging. *Journal of Neuroscience, 22*, 5055-5073. [PubMed]

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P. & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences, U.S.A., 99*, 15164-15169. [PubMed]

Noesselt, T., Hillyard, S. A., Woldorff, M. G., Schoenfeld, A., Hagner, T., Jancke, L., Tempelmann, C., Hinrichs, H. & Heinze, H. J. (2002). Delayed striate cortical activation during spatial attention. *Neuron, 35*, 575-587. [PubMed]

Nowak, L. G. & Bullier, J. (1997). The timing of information transfer in the visual system. In Rockland, K. S., Kaas, J. H. & Peters, A. (Eds.), *Cerebral Cortex, 12* (pp. 205-241). New York: Plenum Press.

Oliva, A. & Torralba, A. (2001). Modeling the shape of a scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*, 145-175.

Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607-609. [PubMed]

Picton, T. W. (1992). The P300 wave of the human event-related potential. *Journal of Clinical Neurophysiology, 9*, 456-479. [PubMed]

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79-87. [PubMed]

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019-1025. [PubMed]

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

Supèr, H., Spekreijse, H. & Lamme, V. A. F. (2001).Two distinct modes of sensory processing observed in monkey primary visual cortex (V1). *Nature Neuroscience, 4*, 304-310. [PubMed]

Sutton, S., Braren, M., Zubin, J. & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science, 150*, 1187-1188. [PubMed]

Tanaka, J., Luu, P., Weisbrod, M. & Kiefer, M. (1999). Tracking the time course of object categorization using event-related potentials. *Neuroreport, 10*, 829-835. [PubMed]

Thorpe, S., Fize, D. & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520-522. [PubMed]

Thorpe, S. J. & Imbert, M. (1989) Biological Constraints on Connectionist Modelling. In Pfeifer, R., Schreter, Z., Fogelman-Soulié, F. & Steels, L. (Eds.), *Connectionism in Perspective* (pp. 63-92). New York: Elsevier Science.

Torralba, A. & Oliva, A. (in press). Statistics of natural images categories. *Network: Computation in Neural Systems*.

Ullman, S. (1996). *High-level Vision: Object Recognition and Visual Cognition* Cambridge, Massachusetts: MIT Press.

Ullman, S., Vidal-Naquet, M. & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*, 682-687. [PubMed]

VanRullen, R. & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42*, 2593-2615. [PubMed]

VanRullen, R. & Thorpe, S. (2001a). Rate coding vs temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Computation, 13*, 1255-1283.

VanRullen, R. & Thorpe, S. J. (2001b). The time course of visual processing: from early perception to decision-making. *Journal of Cognitive Neuroscience, 13*, 454-461. [PubMed]

Vinje, W. E. & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science, 287*, 1273-1276. [PubMed]