Behavioral/Cognitive

The Neural Dynamics of Attentional Selection in Natural Scenes

Daniel Kaiser, Nikolaas N. Oosterhof, and Marius V. Peelen
Center for Mind/Brain Sciences, University of Trento, 38068 Rovereto, Italy

The human visual system can only represent a small subset of the many objects present in cluttered scenes at any given time, such that objects compete for representation. Despite these processing limitations, the detection of object categories in cluttered natural scenes is remarkably rapid. How does the brain efficiently select goal-relevant objects from cluttered scenes? In the present study, we used multivariate decoding of magneto-encephalography (MEG) data to track the neural representation of within-scene objects as a function of top-down attentional set. Participants detected categorical targets (cars or people) in natural scenes. The presence of these categories within a scene was decoded from MEG sensor patterns by training linear classifiers on differentiating cars and people in isolation and testing these classifiers on scenes containing one of the two categories. The presence of a specific category in a scene could be reliably decoded from MEG response patterns as early as 160 ms, despite substantial scene clutter and variation in the visual appearance of each category. Strikingly, we find that these early categorical representations fully depend on the match between visual input and top-down attentional set: only objects that matched the current attentional set were processed to the category level within the first 200 ms after scene onset. A sensor-space searchlight analysis revealed that this early attention bias was localized to lateral occipitotemporal cortex, reflecting top-down modulation of visual processing. These results show that attention quickly resolves competition between objects in cluttered natural scenes, allowing for the rapid neural representation of goal-relevant objects.

Key words: biased competition; category-based attention; MEG decoding; natural scene categorization; visual search

Significance Statement

Efficient attentional selection is crucial in many everyday situations. For example, when driving a car, we need to quickly detect obstacles, such as pedestrians crossing the street, while ignoring irrelevant objects. How can humans efficiently perform such tasks, given the multitude of objects contained in real-world scenes? Here we used multivariate decoding of magnetoencephalogaphy data to characterize the neural underpinnings of attentional selection in natural scenes with high temporal precision. We show that brain activity quickly tracks the presence of objects in scenes, but crucially only for those objects that were immediately relevant for the participant. These results provide evidence for fast and efficient attentional selection that mediates the rapid detection of goal-relevant objects in real-world environments.

Introduction

Our daily-life visual environments, such as city streets and living rooms, contain a multitude of objects. Out of this overwhelming amount of sensory information, we must efficiently select those objects that are relevant for current goals. Visual and attention systems have developed and evolved to optimally perform real-world tasks like these (Barlow, 1961; Felsen and Dan, 2005; Wolfe et al., 2011), as demonstrated by the finding that human observers can rapidly detect the presence of familiar object categories in natural scenes (Thorpe et al., 1996), even when concurrently performing another attention-demanding task (Li et al., 2002). In the present study, we used multivariate decoding of MEG data to show that the rapid extraction of categorical information from cluttered natural scenes is mediated by a highly efficient category-based attention mechanism that biases the neural representation of cluttered scenes in favor of the attended category within 200 ms after scene onset.

It has long been recognized that the visual system can only represent a subset of the many objects present in cluttered scenes at any given time, such that objects compete for representation (Neisser, 1967; Treisman et al., 1983; Duncan, 1984). According to the biased competition model of attention (Desimone and Duncan, 1995), top-down attention acts to resolve this competition, biasing competitive interactions in favor of currently rele-

Received April 27, 2016; revised July 12, 2016; accepted Aug. 4, 2016.
Author contributions: D.K. and M.V.P. designed research; D.K. performed research; N.N.O. contributed unpublished reagents/analytic tools; D.K. and N.N.O. analyzed data; D.K. and M.V.P. wrote the paper.
This work was supported by the Autonomous Province of Trento, Grandi Progetti 2012 project Characterizing and improving brain mechanisms of attention (ATTEND).
The authors declare no competing financial interests.
Correspondence should be addressed to Dr. Marius V. Peelen, Center for Mind/Brain Sciences, University of Trento Corso Bettini 31, 38068 Rovereto, Italy. E-mail: marius.peelen@unitn.it.
DOI:10.1523/JNEUROSCI.1385-16.2016
 0270-6474/16/3610522-07$15.00/0

vant stimuli (targets) relative to currently irrelevant stimuli (distracters). Previous EEG/MEG research has characterized the temporal dynamics of attentional selection based on spatial location (Mangun and Hillyard, 1991; Eimer, 1996; Crist et al., 2008), simple features (Luck and Hillyard, 1994; Hopf et al., 2004; Zhang and Luck, 2009), or objects in highly simplified and artificial displays (Downing et al., 2001; Furey et al., 2006; Baldauf and Desimone, 2014). However, because these studies used simple stimuli or objects in isolation, it remains unclear how these results apply to the selection of familiar objects in naturalistic environments. For example, Zhang and Luck (2009) showed that the time course of feature-based attention critically depends on the complexity of the stimulus display. Given the many differences between artificial scenes and natural scenes, and between features and categories, this finding illustrates that it is difficult to generalize previous EEG/MEG findings to category-based attentional selection in natural scenes. Recent fMRI studies have started investigating attention in naturalistic stimuli, providing evidence for category-based attentional modulation of scene processing in high-level visual cortex (Peelen et al., 2009; Peelen and Kastner, 2011; Çukur et al., 2013). However, because of the slow temporal resolution of fMRI, these studies do not provide information about the time course of category-based attentional selection in natural scenes.

One possibility is that top-down category-based attention biases the initial processing of visual input at higher levels of the visual system. This account predicts a difference between the representation of attended and unattended object categories as soon as visual processing reaches the categorical stage (150–200 ms). Alternatively, the initial categorization of objects in scenes may be automatic and largely unaffected by top-down attentional set (Goddard et al., 2016; Groen et al., 2016). According to this view, attention may act as a feedback mechanism, biasing processing of task-relevant objects after categorization has taken place. This account thus predicts that attended and unattended object categories are initially represented similarly, with attention effects arising later in time (>250 ms) (Bansal et al., 2014).

In the present study, we used MEG decoding to track the neural representation of objects naturally present in scenes as a function of top-down attentional set. We provide evidence for rapid (<200 ms) category-level representations of within-scene objects. Crucially, we find that these early categorical representations depend on the match between visual input and top-down attentional set: only objects that matched the current attentional set were processed to the category level within the first 200 ms after scene onset. These results provide evidence for an early attentional biasing mechanism that facilitates the rapid detection of objects in cluttered scenes.

Materials and Methods

Participants. Forty-nine healthy adults (19 male; mean age 24.5 years, SD = 4.1 years) were recruited at the Center for Mind/Brain Sciences at the University of Trento. All participants gave informed consent, and all procedures were approved by the ethical committee of the University of Trento and performed in accordance with the Declaration of Helsinki. One participant was excluded from all analyses because of technical problems, and another one because of excessive movement during MEG recording.

Main experiment procedure. Participants performed a category search task, where they had to indicate whether the target category (either cars or people) was present in a briefly presented scene stimulus (see Fig. 1*A*). The stimulus set was composed of natural scene photographs that could include (one or multiple) exemplars of two categories: cars and people. This led to four different scene types: scenes with cars, scenes with people, scenes with cars and people, and scenes without any of the two categories. For each of these four types, 40 unique stimuli were used; during the experiment, each stimulus was presented once in its veridical version and once mirrored horizontally, leading to a total of 80 stimuli per scene category, and 320 stimuli in total. The scenes (13.5° × 10.1° visual angle) were back-projected onto a translucent screen in front of the participant (110 cm viewing distance). Stimulus presentation was controlled using the Psychophysics Toolbox (Brainard, 1997) (RRID:SCR_002881).

The experiment consisted of a total of eight blocks of 80 trials each, with the target category being constant within a block, and alternating between cars and people across blocks. The target category in the first block was counterbalanced across participants. The first four blocks contained the 160 veridical stimuli, with each stimulus appearing once in the car target block and once in the people target block; the second four blocks contained the 160 mirrored versions of the stimuli. Thus, every unique scene was shown once in every condition (i.e., the car or people task). Trial order within blocks was randomized.

On every trial, a cue (a pink fixation cross, displayed for 800 ms) indicated that the stimulus would come up shortly. Then a scene stimulus was presented for 83 ms, immediately followed by a perceptual mask for 800 ms. After a randomly jittered intertrial interval (between 2200 and 3000 ms), the next trial started. Participants were instructed to indicate as fast and as accurately as possible whether the target category was present in the scene by pressing one of two response buttons (button assignment was counterbalanced across participants). Participants detected the target presence or absence correctly in 86% of trials (SE = 0.7%), with a mean response time of 576 ms (SE = 7.1 ms). Trials with incorrect responses, no responses within the response window (i.e., until the start of the next trial), and response times faster than 200 ms were excluded from all MEG analyses.

Isolated object experiment procedure. To characterize response patterns for cars and people in isolation, participants performed two blocks of an additional experiment after completing the fourth run of the main experiment. During each of these two blocks, participants viewed images of cars and headless bodies centrally (8° approximate visual angle), with the trial structure being identical to the main experiment (800 ms prestimulus cue, 83 ms car/body stimulus, 800 ms mask, 2200–3000 ms intertrial interval; see Fig. 1*E*). Each block consisted of 80 trials (40 cars, 40 bodies; randomly intermixed; see Fig. 1*F* for stimulus examples). Participants were instructed to detect upside-down targets that occurred at eight random times during every block; these trials were excluded from all analyses.

MEG acquisition and preprocessing. Electromagnetic brain activity was recorded using a 306-channel Elekta Neuromag System. Signals were sampled continuously at 1000 Hz and bandpass filtered online between 0.1 and 300 Hz. Offline preprocessing was done using MATLAB (The MathWorks; RRID:SCR_001622) and the fieldtrip analysis package (Oostenveld et al., 2011) (RRID:SCR_004849). Data were concatenated for all blocks of the main experiment and for both blocks of the Isolated Object experiment, high-pass filtered at 1 Hz (to avoid temporal signal blur, a one-pass minimum-phase FIR-filter with Kaiser window was used), and epoched into trials ranging from −200 to 500 ms with respect to stimulus onset. Based on visual inspection, trials containing eye blinks and other movement-related artifacts were discarded from all analyses. Similarly, sensors with consistently high noise levels were discarded. The epoched data were then baseline-corrected from −200 ms to stimulus onset, and downsampled to 100 Hz to increase the signal-to-noise ratio of the multivariate classification analysis (Carlson et al., 2013).

MEG decoding analysis. All multivariate classification analyses were performed using MATLAB (RRID:SCR_001622) and the CoSMoMVPA analysis package (www.cosmomvpa.org) (Oosterhof et al., 2016; RRID:SCR_014519). Classification was performed separately for every 10 ms time bin. Only data from the magnetometers was used, as these sensors offered more reliable overall classification performance both within the Isolated Object experiment and the main experiment. Linear discriminant analysis (LDA) classifiers were trained to discriminate the patterns across sensors for two conditions of interest in one subset of the data (subset of trials), and subsequently tested on another, independent subset of the data (disjoint subset of trials). For cross-validation analyses within the Isolated Objects experiment and within the main experiment,

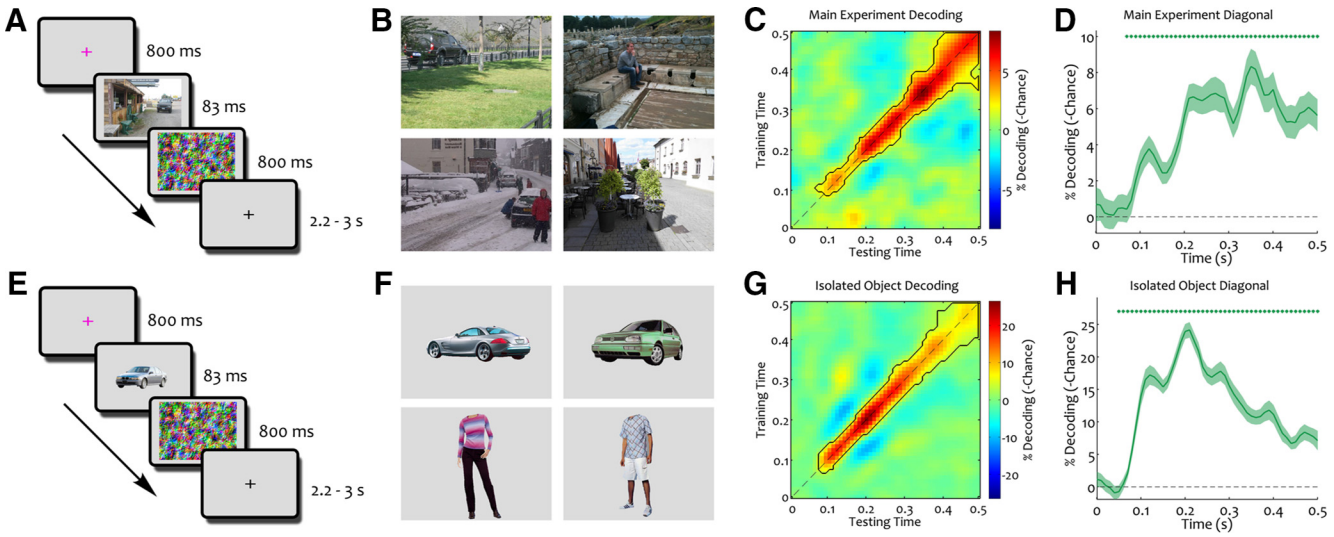

Figure 1. ***A***, Main experiment procedure. Participants indicated as fast and accurately as possible whether the target category (cars or people, alternating every block of 5 min) was present in a briefly presented natural scene (83 ms), which was followed by a perceptual mask (800 ms). ***B***, Main experiment example stimuli. Natural scenes could contain people, cars, both categories, or neither of the two categories. ***C***, Time × time MEG decoding matrix showing decoding performance relative to chance for the discrimination between scenes containing cars and scenes containing people, regardless of task. Black outline indicates clusters of above-chance decoding ($p < 0.05$, corrected for multiple comparisons). ***D***, Decoding accuracy along the diagonal of the cross-decoding matrix for objects embedded in scenes. Shaded area represents SEM. Dots indicate time points of above-chance decoding performance ($p < 0.05$, corrected for multiple comparisons). ***E***, Isolated Object experiment procedure. Participants viewed images of cars and headless bodies in isolation while reporting occasional upside-down targets. Timing of presentation was identical to that of the main experiment. ***F***, Isolated Object experiment example stimuli. ***G***, Time × time MEG decoding matrix showing decoding performance relative to chance for the discrimination between cars and people in isolation. Black outline indicates clusters of above-chance decoding ($p < 0.05$, corrected for multiple comparisons). ***H***, Decoding accuracy along the diagonal of the cross-decoding matrix for isolated objects. Shaded area represents SEM. Dots indicate time points of above-chance decoding performance ($p < 0.05$, corrected for multiple comparisons).

the data were divided into two subsets of trials by randomly assigning labels to the data, with the constraint of an equal amount of trials in every subset; classifiers were then trained on one of these subsets and tested on the other subset (data were averaged for both train/test directions). For the cross-decoding analysis, classifiers were trained on all the trials of the Isolated Objects experiment, and all trials of interest from the main experiment were used as the testing set. To reduce trial-by-trial noise and thus increase the reliability of the data supplied to the classifier, new, "synthetic" trial data were created, which consisted of an average of five independent trials: for every data subset and condition separately, five trials were chosen randomly and averaged together, to generate a new "synthetic" trial for classification; this procedure was repeated 100 times (with the constraint that no original trial was used more than one time more often than any other trial), so that for every condition and for both the training and test sets, exactly 100 of these synthetic trials were available. Classification accuracy was assessed as the percentage of correct predictions of the classifier. The classification procedure was repeated for every possible combination of training and testing time, leading to a 50 × 50 points (i.e., 500 ms × 500 ms with 100 Hz resolution) classification accuracy map for every comparison in every participant. Individual subject accuracy maps were smoothed with an averaging box filter spanning 3 × 3 time points (i.e., 30 ms in both training and testing time).

Cross-decoding and searchlight analyses. To test for attentional modulation in the cross-decoding analysis, classification accuracy was computed separately for target and distracter trials. To increase the sensitivity of this analysis, time clusters of interest were selected from the overall decoding accuracy maps (averaged across targets and distracters) by taking all points that led to significant above-chance overall decoding in three time windows: from stimulus onset to 220 ms ("Early"), from 230 ms to 340 ms ("Mid"), and from 350 ms to 500 ms ("Late"; see Fig. 2*A*). Decoding performance for targets and distracters was then calculated by separately averaging classification accuracies for target and distracter decoding for all time × time combinations belonging to a specific time cluster; subsequently, one-tailed *t* tests of the target-distracter difference against zero were performed (see Fig. 2*C*). To approximately assess the spatial distribution of a target-distracter difference, a sensor-space searchlight analysis within each of these three time clusters was performed: the cross-classification analysis was repeated using sensor neighborhoods of 20 sensors each. Each of these neighborhoods was constructed by defining a sphere of 10 sensors in the left hemisphere that was symmetrically mirrored to the right hemisphere; these two neighborhoods were then collapsed, and the resulting searchlight map was displayed on one hemisphere (see Fig. 3*A*). For each neighborhood, a separate analysis was performed for each time point in each one of the three clusters, and results were averaged within each cluster to obtain the searchlight topography for the respective cluster. Subsequently, attention effects were quantified by comparing the topographies for target and distracter decoding separately for each time cluster. Additionally, a searchlight analysis without symmetry constraints was performed, where spherical neighborhoods of 15 sensors each were used.

Statistical testing. To identify time-periods or sensors yielding above-chance classification, we used a threshold-free cluster-estimation procedure (Smith and Nichols, 2009) with default parameters, using multiple-comparison correction based on a sign-permutation test (with null distributions created from 10,000 bootstrapping iterations) as implemented in CoSMoMVPA (Oosterhof et al., 2016). Statistical maps were then thresholded at $Z > 1.64$ (i.e., $p < 0.05$, one-tailed) to reveal significant decoding performance. The topographical searchlight maps were thresholded at $Z > 2.13$ (i.e., $p < 0.017$, one-tailed) to Bonferroni correct for the three time clusters we tested for. Additionally, for all tests, uncorrected *t* values from conventional *t* tests are reported for the peaks of decoding accuracy.

Results

Scene decoding

In a first analysis, we tested whether scenes containing cars and scenes containing people (Fig. 1*B*, top row) evoke reliably different MEG response patterns regardless of task. Linear classifiers were trained to discriminate scenes containing cars versus scenes containing people using a cross-validation approach (see Materials and Methods). This classification analysis was performed for all combinations of training and testing time points within a time window of 500 ms after stimulus onset and with a temporal resolution of 100 Hz, thus leading to a 50 × 50 time points matrix of decoding accuracy (with chance level at 50%). Classifiers reliably

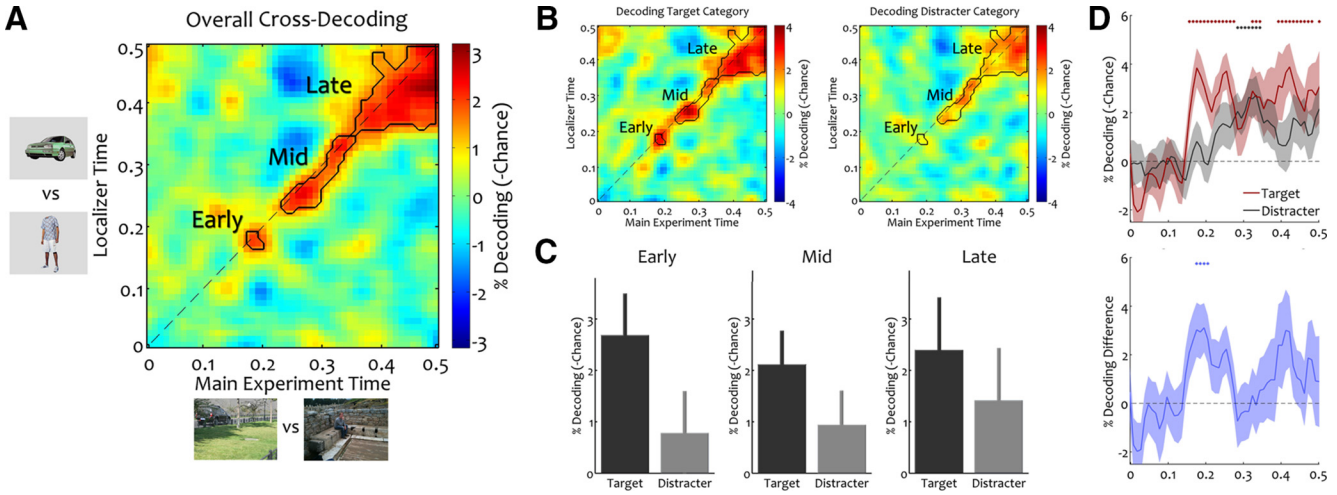

Figure 2. Cross-decoding results. ***A***, Classifiers trained on discriminating cars and people in isolation successfully discriminated cars and people when embedded in natural scenes in the main experiment. Collapsed across targets and distracters, cross-classification was possible in three time clusters: between 180 and 220 ms ("Early"), between 230 and 340 ms ("Mid"), and between 350 and 500 ms ("Late"). Black outlines indicate clusters of above-chance decoding ($p < 0.05$, corrected for multiple comparisons). ***B***, Cross-decoding matrices for targets and distracters separately. ***C***, Mean decoding performance for targets and distracters in the three time clusters. ***D***, Decoding accuracy along the diagonal of the cross-decoding matrices (***B***), separately for targets (red line) and distracters (gray line; top). Directly comparing the decoding accuracy for targets and distracters revealed a significant difference between 180 and 210 ms (blue line; bottom). Shaded area represents SEM. Dots indicate time points of above-chance decoding performance ($p < 0.05$, corrected for multiple comparisons).

discriminated scenes containing cars and people between 80 and 500 ms (464 time points in total, peak t value $t_{(46)} = 7.89$; Fig. 1*C*), with local maxima along the diagonal of the time × time matrix, at 110, 210, and 350 ms (Fig. 1*D*).

Decoding category processing

The successful discrimination of car scenes and person scenes shows that these scene types evoked reliably different sensor patterns starting early in time. However, multiple aspects may have contributed to this decoding in addition to the processing of the cars and people. For example, the person and car scenes may have systematically differed in terms of scene layout or the presence of category-associated objects (e.g., traffic signs in car scenes). Therefore, to unequivocally reveal the time course of object category processing in cluttered scenes, we next used a cross-decoding approach, training classifiers on discriminating cars and people presented in isolation (Fig. 1*E*) and testing these classifiers on discriminating scenes containing cars or people. To identify MEG response patterns to cars and people in isolation, participants completed a separate Isolated Object experiment (Fig. 1*E*), in which they viewed cars and people in isolation (Fig. 1*F*) while performing an orthogonal task (see Materials and Methods). For this experiment, the two categories were reliably discriminable between 60 and 500 ms after stimulus onset (397 time points in total, peak t value $t_{(46)} = 35.3$; Fig. 1*G*), with peak classification accuracy at 200 ms (Fig. 1*H*), consistent with recent MEG category decoding studies (Carlson et al., 2013; Cichy et al., 2014). The finding that isolated objects could be decoded earlier and with higher accuracy than objects embedded in scenes likely reflects a cost of the visual complexity of the scenes and the diverse characteristics of the objects they contained (e.g., in location, size, and viewpoint).

Interestingly, classifiers trained on discriminating isolated cars and people were able to reliably discriminate scenes containing these categories. This effect was significant at three distinct time points (Fig. 2*A*): an early time cluster from 180 to 220 ms (13 time points, peak t value $t_{(46)} = 3.85$), an intermediate time cluster from 230 to 340 ms (72 time points, peak t value $t_{(46)} = 4.01$), and a late time cluster from 350 to 500 ms (185 time points, peak t value $t_{(46)} = 4.52$). These results demonstrate that MEG response patterns as early as 180 ms after scene onset carry reliable category information about small and highly variable images of cars and people embedded in cluttered natural scenes.

Attentional selection in natural scenes

Next, we moved to our main question regarding the temporal dynamics of attentional selection in natural scenes: when does the categorical representation of target objects differ from that of distracter objects? To increase sensitivity in uncovering such attentional modulations, we used a method frequently used in fMRI studies (Poldrack, 2007): we defined regions of interest (here: in time, i.e., time clusters of interest) based on the three clusters identified in the overall decoding analysis (Fig. 2*A*), and tested for target-distracter differences within these time clusters of interest. Classifiers were again trained on data from the Isolated Object experiment, but now two different test sets were used: the test data either consisted of scenes where the contained category was the target (i.e., car scenes in the car task and people scenes in the people task) or scenes where the contained category was a distracter (i.e., car scenes in the people task and people scenes in the car task). Figure 2*B* shows the cross-classification matrices for targets and distracters separately. To quantify the target-distracter difference within the three time clusters of interest, classification performance was then averaged for all time points falling within each cluster. A time cluster × attention ANOVA revealed higher decoding accuracy for targets than distracters ($F_{(1,46)} = 6.73$, $p = 0.013$), which was similarly strong for the three time clusters (interaction: $F_{(2,92)} = 0.35$, $p = 0.708$). Most importantly for addressing the competing hypotheses outlined in the Introduction, this attention effect was also observed separately for the early time cluster ($t_{(46)} = 2.31$, $p = 0.013$; $p < 0.05$, Bonferroni-corrected for three comparisons).

These results were confirmed by comparing the diagonals of the target and distracter decoding matrices (Fig. 2*D*). This analysis revealed that the category of targets could be decoded as early as 160 ms after stimulus onset (peak t value $t_{(46)} = 4.98$), whereas

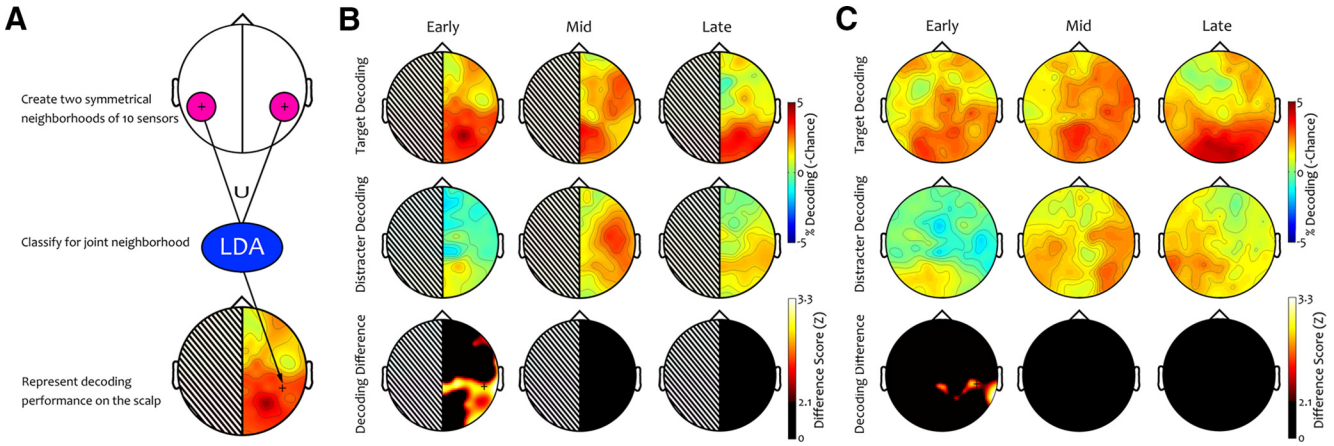

Figure 3. Cross-decoding searchlight. ***A***, For each of the three time clusters, we performed a searchlight analysis in sensor space, decoding object category separately for targets (cars in the car task vs people in the people task) and distracters (cars in the people task vs people in the car task). Searchlight neighborhoods were defined as symmetric sensor neighborhoods of 10 sensors in each hemisphere. Each pair of symmetric neighborhoods was concatenated to form a 20-sensor neighborhood. For each of these joint neighborhoods, we then computed LDA classification accuracies. Results are plotted on the right hemisphere but represent both hemispheres in a mirror-symmetric way. ***B***, Two top rows represent the decoding topographies for targets and distracters for the three time clusters. Bottom row represents statistical maps of the difference between targets and distracters. A significant target-distracter difference was observed in the first time window over occipitotemporal cortex. Black cross represents the sensor location exhibiting the greatest difference. *Z* scores >2.13 correspond to statistically significant results ($p < 0.05$, corrected for multiple comparisons, and Bonferroni-corrected for the three time clusters). ***C***, Cross-decoding searchlight with nonmirrored neighborhoods. A significant target-distracter difference was observed in the first time window, and in sensor locations over right occipitotemporal cortex. Black cross represents the sensor location exhibiting the greatest difference. *Z* scores >2.13 correspond to statistically significant results ($p < 0.05$, corrected for multiple comparisons, and Bonferroni-corrected for the three time clusters).

the decoding of the distracter category emerged much later, starting at 290 ms (peak *t* value $t_{(46)} = 3.46$). The difference between target and distracter decoding was significant between 180 and 210 ms (peak *t* value $t_{(46)} = 3.08$).

Searchlight analysis

To identify the peak location of this attentional enhancement in sensor space, a multivariate searchlight analysis was conducted, analogous to the searchlight analysis developed for fMRI (Kriegeskorte et al., 2006) (see Materials and Methods). For each time cluster separately, category information for targets and distracters was compared within local sensor neighborhoods. As category information is partly contained in asymmetries between hemispheres (e.g., caused by lateralized processing of the human body) (Willems et al., 2010), searchlight neighborhoods were constructed in a mirror-symmetric way that retained this information, thus resulting in mirror-symmetric topographical maps (Fig. 3*A*). The comparison of topographical searchlight maps for target and distracter decoding revealed a significant difference for the early time cluster (180–220 ms), again confirming the early attention effect obtained in the previous analyses. This attention effect was localized to the lateral occipitotemporal cortex (22 sensor locations in total, peak *t* value $t_{(46)} = 3.42$; Fig. 3*B*). We additionally performed a searchlight analysis without symmetry constraints. This analysis similarly revealed an early effect of attention and indicated a right-hemispheric lateralization of the effect (10 sensor locations in total, peak *t* value $t_{(46)} = 4.69$; Fig. 3*B*).

Discussion

To recover category-level neural processing from patterns of MEG activity, we trained multivariate classifiers on discriminating cars and people in isolation (presented in a separate experiment) and tested these classifiers on scenes containing either cars or people. This cross-decoding approach allowed for tracking category-level representations over time. Averaging across target and distracter objects, we found that the category of within-scene objects was represented as early as 180 ms after stimulus onset, despite scene clutter and large variation in object features and locations across scenes. Importantly, we find that this early category representation depends on the behavioral relevance of a category: MEG sensor patterns before 200 ms carried information about the target category but not the distracter category. Because the presented scenes were identical in the two conditions, this effect must reflect a top-down bias toward the processing of the task-relevant category.

A multivariate searchlight analysis revealed that the attentional bias at 180–210 ms was localized to the lateral occipitotemporal cortex (Fig. 3), particularly in the right hemisphere. This finding provides support for the interpretation that our findings reflect attentional biases of visual category processing rather than, for example, effects related to decision-making or verbalization. The right lateral occipitotemporal cortex was also the key region implicated in category-based attentional selection in previous fMRI studies (Peelen et al., 2009; Peelen and Kastner, 2011; Seidl et al., 2012; Soon et al., 2013). The current MEG results complement these findings by providing a first temporal characterization of attentional selection in natural scenes, showing that categorical attentional set rapidly biases within-scene object processing in lateral occipitotemporal cortex.

The current experiment investigated category-based attentional selection in a large and diverse set of natural scenes. Considering the variability in appearance of the objects across scenes and the large degree and variable nature of scene clutter, it is highly unlikely that our results reflect attentional modulation of low-level feature processing, such as line orientation or color, as investigated previously with EEG (Luck and Hillyard, 1994; Zhang and Luck, 2009). Instead, we interpret results as reflecting attentional biases at higher levels of the visual processing hierarchy, with attention directed to mid- or high-level features that are diagnostic of the presence of a category in variable and cluttered scenes (Ullman et al., 2002; Evans and Treisman, 2005; Delorme et al., 2010; Reeder and Peelen, 2013; Hickey et al., 2015). On this account, attentional templates would be implemented at higher

levels of the visual hierarchy and thus expected to bias visual processing only once scene processing reaches these stages of the visual processing hierarchy. The earliest decoding of the category of attended objects observed here (160 ms) matches the latency found for the categorization of isolated images of people versus objects when low-level stimulus differences are controlled for (Stekelenburg and de Gelder, 2004; Thierry et al., 2006; Kaiser et al., 2016). Accordingly, in the current study, attention effects emerged as soon as object category could be extracted from the scenes (Fig. 2), highlighting the efficiency of category-based attentional selection in naturalistic environments.

Natural scenes not only add undesired complexity and clutter, but scene structure can also facilitate object detection and identification. Previous studies have shown that objects that are placed congruently within a scene are processed more efficiently than objects that are placed incongruently (Biederman, 1972; Bar and Ullman, 1996; Neider and Zelinsky, 2006). Electrophysiological studies have demonstrated that such scene-object consistencies impact waveforms from ~300 ms after stimulus onset (Mudrik et al., 2010; Võ and Wolfe, 2013). These comparably late effects likely reflect differences in semantic processing rather than attentional selection (in these studies, the location of the target was cued before scene onset). But scene structure may also facilitate efficient attentional selection: scene context guides attention toward likely target locations (Torralba et al., 2006) and constrains the possible appearance of target objects at different locations in the scene (e.g., as a function of depth; Wolfe et al., 2011). Furthermore, efficient scene parsing is supported by the regular arrangement of objects, such that objects are grouped based on typical spatial dependencies among them, leading to more efficient visual search (Kaiser et al., 2014). Further research is needed to directly relate these processes to rapid category detection and to the early categorical attentional modulation reported here.

Previous EEG studies have measured the time course of target detection in natural scenes by measuring evoked potentials to the presence versus absence of targets (Thorpe et al., 1996; VanRullen and Thorpe, 2001; Johnson and Olshausen, 2003, 2005; Delorme et al., 2004; Codispoti et al., 2006). These studies differed from the current study in several ways. Most importantly, the aim of these studies was to measure the earliest time point at which the brain signaled the presence of a target scene, to provide evidence for fast feedforward processing of natural scenes (Thorpe et al., 1996; VanRullen and Thorpe, 2001). This differs from the aim of the current study, which was to track category-level representations of both targets and distracters, to compare the time courses of these representations as an index of attentional modulation. It should also be noted that the scenes and tasks used in these previous studies may not have required selective attention to the same degree: target objects were large foreground objects, reducing the need for early attentional selection (Desimone and Duncan, 1995; Lavie, 1995; Luck et al., 2000; Zhang and Luck, 2009); previous findings suggest that top-down attentional modulation may arise later in the absence of attentional competition (Zhang and Luck, 2009; Bansal et al., 2014). Furthermore, participants in these studies performed superordinate categorization tasks (detecting animals or vehicles), for which attentional templates may be less effective (Delorme et al., 2004; Vickery et al., 2005; Schmidt and Zelinsky, 2009). Importantly, while previous findings of target-selective evoked potentials are consistent with the current findings of early attentional modulation, they would be equally consistent with the absence of such a target-distracter difference at this latency. Indeed, target-selective EEG activity in previous studies has been attributed to decision-related processing occurring after visual processing is completed (Thorpe et al., 1996; Johnson and Olshausen, 2005) as well as to target-related processing in occipitotemporal cortex (VanRullen and Thorpe, 2001; Delorme et al., 2004; Codispoti et al., 2006). By comparing target and distracter representations using a cross-decoding approach that eliminates the influence of task-related decision processes, the present study provides the first evidence for early attentional modulation during naturalistic visual search.

The current results fit well within the biased competition model of attention (Desimone and Duncan, 1995), in which top-down attention biases competitive interactions between stimuli in favor of currently relevant stimuli. An important aspect of the biased competition model is the concept of attentional templates: internal descriptions of task-relevant information (Duncan and Humphreys, 1989). Attentional templates are activated before visual processing, biasing the processing of incoming visual information (Chelazzi et al., 1993). Importantly, attentional templates are not restricted to one type of visual property, such as a target's location or its low-level features, but may equally include properties encoded at higher stages of the visual processing hierarchy when these properties best distinguish targets from nontargets. For example, a previous fMRI study investigating category-based attention in natural scenes revealed that preparatory activity patterns in high-level visual cortex carried information about the category of the top-down attentional template (Peelen et al., 2011). This template subsequently biased the processing of the scene in favor of the attended category, thereby facilitating detection performance (Peelen et al., 2011; Soon et al., 2013), a finding that was confirmed with TMS (Reeder et al., 2015). The current results suggest that these category-based attentional templates mediate efficient selection in complex natural scenes already at the level of early category-selective responses.

Together, our findings provide novel insights into the temporal dynamics and neural mechanisms of object detection in photographs of everyday scenes. Our results show a strong influence of top-down attention on the category-level representation of objects in cluttered scenes. More generally, the current findings demonstrate how MEG can be used to track and localize, in real time, the neural representation of objects during naturalistic vision.

References

Baldauf D, Desimone R (2014) Neural mechanisms of object-based attention. Science 344:424–427. CrossRef Medline

Bansal AK, Madhavan R, Agam Y, Golby A, Madsen JR, Kreiman G (2014) Neural dynamics underlying target detection in the human brain. J Neurosci 34:3042–3055. CrossRef Medline

Bar M, Ullman S (1996) Spatial context in recognition. Perception 25:343–352. CrossRef Medline

Barlow H (1961) Possible principles underlying the transformation of sensory messages. In: Sensory communication (Rosenbirth W, ed), pp 217–234. Cambridge, MA: Massachusetts Institute of Technology.

Biederman I (1972) Perceiving real-world scenes. Science 177:77–80. CrossRef Medline

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436. CrossRef Medline

Carlson T, Tovar DA, Alink A, Kriegeskorte N (2013) Representational dynamics of object vision: the first 1000 ms. J Vis 13:1–19. CrossRef Medline

Chelazzi L, Miller EK, Duncan J, Desimone R (1993) A neural basis for visual search in inferior temporal cortex. Nature 363:345–347. CrossRef Medline

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. Nat Neurosci 17:455–462. CrossRef Medline

Codispoti M, Ferrari V, Junghöfer M, Schupp HT (2006) The categorization of natural scenes: brain attention networks revealed by dense sensor ERPs. Neuroimage 32:583–591. CrossRef Medline

Crist RE, Wu CT, Karp C, Woldorff MG (2008) Face processing is gated by visual spatial attention. Front Hum Neurosci 1:10. CrossRef Medline

Çukur T, Nishimoto S, Huth AG, Gallant JL (2013) Attention during natural vision warps semantic representation across the human brain. Nat Neurosci 16:763–770. CrossRef Medline

Delorme A, Rousselet GA, Macé MJ, Fabre-Thorpe M (2004) Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. Brain Res Cogn Brain Res 19:103–113. CrossRef Medline

Delorme A, Richard G, Fabre-Thorpe M (2010) Key visual features for rapid categorization of animals in natural scenes. Front Psychol 1:21. CrossRef Medline

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193–222. CrossRef Medline

Downing P, Liu J, Kanwisher N (2001) Testing cognitive models of visual attention with fMRI and MEG. Neuropsychologia 39:1329–1342. CrossRef Medline

Duncan J (1984) Selective attention and the organization of visual information. J Exp Psychol Gen 113:501–517. CrossRef Medline

Duncan J, Humphreys GW (1989) Visual search and stimulus similarity. Psychol Rev 96:433–458. CrossRef Medline

Eimer M (1996) ERP modulations indicate the selective processing of visual stimuli as a result of transient and sustained spatial attention. Psychophysiology 33:13–21. CrossRef Medline

Evans KK, Treisman A (2005) Perception of objects in natural scenes: is it really attention free? J Exp Psychol Hum Percept Perform 31:1476–1492. CrossRef Medline

Felsen G, Dan Y (2005) A natural approach to studying vision. Nat Neurosci 8:1643–1646. CrossRef Medline

Furey ML, Tanskanen T, Beauchamp MS, Avikainen S, Uutela K, Hari R, Haxby JV (2006) Dissociation of face-selective cortical responses by attention. Proc Natl Acad Sci U S A 103:1065–1070. CrossRef Medline

Goddard E, Carlson TA, Dermody N, Woolgar A (2016) Representational dynamics of object recognition: feedforward and feedback information flows. Neuroimage 128:385–397. CrossRef Medline

Groen II, Ghebreab S, Lamme VA, Scholte HS (2016) The time course of natural scene processing with reduced attention. J Neurophysiol 115:931–946. CrossRef Medline

Hickey C, Kaiser D, Peelen MV (2015) Reward guides attention to object categories in real-world scenes. J Exp Psychol Gen 144:264–273. CrossRef Medline

Hopf JM, Boelmans K, Schoenfeld MA, Luck SJ, Heinze HJ (2004) Attention to features precedes attention to locations in visual search: evidence from electromagnetic brain responses in humans. J Neurosci 24:1822–1832. CrossRef Medline

Johnson JS, Olshausen BA (2003) Timecourse of neural signatures of object recognition. J Vis 3:499–512. CrossRef Medline

Johnson JS, Olshausen BA (2005) The earliest EEG signatures of object recognition in a cued-target task are postsensory. J Vis 5:299–312. CrossRef Medline

Kaiser D, Stein T, Peelen MV (2014) Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. Proc Natl Acad Sci U S A 111:11217–11222. CrossRef Medline

Kaiser D, Azzalini DC, Peelen MV (2016) Shape-independent object category responses revealed by MEG and fMRI decoding. J Neurophysiol 115:2246–2250. CrossRef Medline

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868. CrossRef Medline

Lavie N (1995) Perceptual load as a necessary condition for selective attention. J Exp Psychol Hum Percept Perform 21:451–468. CrossRef Medline

Li FF, VanRullen R, Koch C, Perona P (2002) Rapid natural scene categorization in the near absence of attention. Proc Natl Acad Sci U S A 99:9596–9601. CrossRef Medline

Luck SJ, Hillyard SA (1994) Electrophysiological correlates of feature analysis during visual search. Psychophysiology 31:291–308. CrossRef Medline

Luck SJ, Woodman GF, Vogel EK (2000) Event-related potential studies of attention. Trends Cogn Sci 4:432–440. CrossRef Medline

Mangun GR, Hillyard SA (1991) Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. J Exp Psychol Hum Percept Perform 17:1057–1074. CrossRef Medline

Mudrik L, Lamy D, Deouell LY (2010) ERP evidence for context congruity effects during simultaneous object-scene processing. Neuropsychologia 48:507–517. CrossRef Medline

Neider MB, Zelinsky GJ (2006) Scene context guides eye movements during visual search. Vision Res 46:614–621. CrossRef Medline

Neisser U (1967) Cognitive psychology. New York: Appleton-Century-Crofts.

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011:156869. CrossRef Medline

Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. Front Neuroinform 10:27. CrossRef Medline

Peelen MV, Kastner S (2011) A neural basis for real-world visual search in human occipitotemporal cortex. Proc Natl Acad Sci U S A 108:12125–12130. CrossRef Medline

Peelen MV, Fei-Fei L, Kastner S (2009) Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460:94–97. CrossRef Medline

Poldrack RA (2007) Region of interest analysis for fMRI. Soc Cogn Affect Neurosci 2:67–70. CrossRef Medline

Reeder RR, Peelen MV (2013) The contents of the search template for category-level search in natural scenes. J Vis 13(3):13. CrossRef Medline

Reeder RR, Perini F, Peelen MV (2015) Preparatory activity in posterior parietal cortex causally contributes to object detection in scenes. J Cogn Neurosci 27:2117–2125. CrossRef Medline

Schmidt J, Zelinsky GJ (2009) Search guidance is proportional to the categorical specificity of a target cue. Q J Exp Psychol (Hove) 62:1904–1914. CrossRef Medline

Seidl KN, Peelen MV, Kastner S (2012) Neural evidence for distracter suppression during visual search in real-world scenes. J Neurosci 32:11812–11819. CrossRef Medline

Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83–98. CrossRef Medline

Soon CS, Namburi P, Chee MW (2013) Preparatory patterns of neural activity predict visual category search speed. Neuroimage 66:215–222. CrossRef Medline

Stekelenburg JJ, de Gelder B (2004) The neural correlates of perceiving human bodies: an ERP study on the body-inversion effect. Neuroreport 15:777–780. CrossRef Medline

Thierry G, Pegna AJ, Dodds C, Roberts M, Basan S, Downing P (2006) An event-related potential component sensitive to images of the human body. Neuroimage 32:871–879. CrossRef Medline

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520–522. CrossRef Medline

Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev 113:766–786. CrossRef Medline

Treisman A, Kahneman D, Burkell J (1983) Perceptual objects and the cost of filtering. Percept Psychophys 33:527–532. CrossRef Medline

Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5:682–687. CrossRef Medline

VanRullen R, Thorpe SJ (2001) The time course of visual processing: from early perception to decision-making. J Cogn Neurosci 13:454–461. CrossRef Medline

Vickery TJ, King LW, Jiang Y (2005) Setting up the target template in visual search. J Vis 5:81–92. CrossRef Medline

Võ ML, Wolfe JM (2013) Differential electrophysiological signatures of semantic and syntactic scene processing. Psychol Sci 24:1816–1823. CrossRef Medline

Willems RM, Peelen MV, Hagoort P (2010) Cerebral lateralization of face-selective and body-selective visual areas depends on handedness. Cereb Cortex 20:1719–1725. CrossRef Medline

Wolfe JM, Võ ML, Evans KK, Greene MR (2011) Visual search in scenes involves selective and non-selective pathways. Trends Cogn Sci 15:77–84. CrossRef Medline

Zhang W, Luck SJ (2009) Feature-based attention modulates feedforward visual processing. Nat Neurosci 12:24–25. CrossRef Medline