Neural mechanisms of rapid natural scene categorization in human visual cortex

Marius V. Peelen[1,2,3], Li Fei-Fei[1,4] & Sabine Kastner[1,2,3]

The visual system has an extraordinary capability to extract categorical information from complex natural scenes. For example, subjects are able to rapidly detect the presence of object categories such as animals or vehicles in new scenes that are presented very briefly[1,2]. This is even true when subjects do not pay attention to the scenes and simultaneously perform an unrelated attentionally demanding task[3], a stark contrast to the capacity limitations predicted by most theories of visual attention[4,5]. Here we show a neural basis for rapid natural scene categorization in the visual cortex, using functional magnetic resonance imaging and an object categorization task in which subjects detected the presence of people or cars in briefly presented natural scenes. The multi-voxel pattern of neural activity in the object-selective cortex evoked by the natural scenes contained information about the presence of the target category, even when the scenes were task-irrelevant and presented outside the focus of spatial attention. These findings indicate that the rapid detection of categorical information in natural scenes is mediated by a category-specific biasing mechanism in object-selective cortex that operates in parallel across the visual field, and biases information processing in favour of objects belonging to the target object category.

In daily life we often look for particular object categories in our environment that are relevant for ongoing behaviour. For example, before crossing a street we look whether cars are near, perhaps not noticing other objects in the visual scene present at the same time, such as people walking on the other side of the street. Behavioural experiments have shown that such detection of familiar object categories in natural scenes is extremely rapid[1,2], and can be done even without focal attention[3]. These results indicate the existence of selection mechanisms for familiar object categories that operate independently of spatial attention.

In the present study, using functional magnetic resonance imaging (fMRI), we investigated neural mechanisms for extracting object category information from complex natural scenes. We proposed that the rapid detection of categorical information may be mediated by top-down mechanisms that bias processing in favour of the searched-for object category. These biasing mechanisms may lead to a filtering of the scene that effectively limits the visual representation of the scene to objects belonging to the task-relevant object category, thereby facilitating the rapid detection of the target category. To probe this hypothesis, we measured the influence of a category detection task (detecting either people or cars) on fMRI activity evoked by pictures of real-world scenes. Furthermore, given that behavioural studies have found evidence for parallel processing in rapid scene categorization[3,6,7], we asked whether such influences can be observed both inside and outside the focus of spatial attention.

A large (>2,000 pictures) set of photographs of outdoor scenes (cityscapes and landscapes; Supplementary Fig. 1) was selected for the experiment. A subset of these scenes contained people or cars in natural, daily-life situations. As in natural vision, the visual appearances and spatial locations of the people and cars in these scenes were highly variable. For example, a scene could show a single person sitting on a bench in a park close to the camera, or could show a group of people walking on a street at a distance (Supplementary Fig. 1). On each trial, subjects briefly (~130 ms, see Methods) viewed four different simultaneously presented pictures (two along the vertical axis, and two along the horizontal axis[8]), followed immediately by perceptual masks presented at the same locations (Fig. 1). The task was to report the presence of either people ('body task') or cars ('car task') in the horizontally or vertically presented pictures, resulting in four different task combinations. These four combinations were always performed in separate functional runs to prevent confusion about the target category and task-relevant picture locations. We will refer to the task-relevant pictures as 'attended' pictures, and the task-irrelevant pictures as 'unattended' pictures. A separate analysis confirmed that subjects spatially attended the task-relevant pictures more than the task-irrelevant pictures, by showing spatially-specific attentional modulation in retinotopic visual cortex (see Supplementary Fig. 2). The content of the attended pictures was manipulated independently from the content of the unattended pictures, such that all possible pairings of scene types
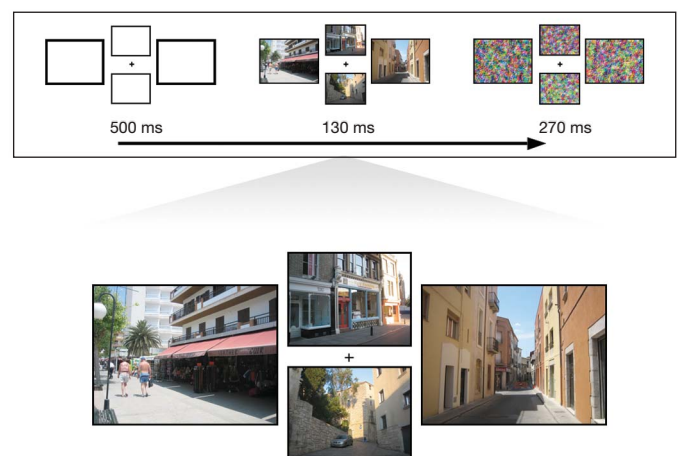

Figure 1 | Schematic overview of trial layout. Each trial started with a spatial cue indicating the relevant stimulus locations (which were held constant within a run). This was followed by the four pictures presented for, on average, 130 ms. Presentation time of the pictures was adjusted for each subject to arrive at ~80% accuracy. The pictures were followed by perceptual masks (270 ms). The next trial started, on average, 1,300 ms after the offset of the masks.

[1]Department of Psychology, Princeton University, Princeton, New Jersey 08540, USA. [2]Center for the Study of Brain, Mind, and Behavior, Princeton University, Princeton, New Jersey 08540, USA. [3]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08540, USA. [4]Department of Computer Science, Princeton University, Princeton, New Jersey 08540, USA.

were equally likely. This design allowed us to measure brain responses to attended and unattended pictures separately by means of selective averaging.

We analysed multi-voxel patterns of activation across the object-selective visual cortex, as these have been shown to be sensitive to object-category information[9,10]. The object-selective visual cortex, often referred to as the lateral-occipital complex, was localized in each subject by contrasting responses to intact versus scrambled objects in a separate localizer scan[11]. The analysis approach was to correlate response patterns evoked by the two scene types (containing either people or cars) in the main experiment with response patterns evoked by a different set of pictures of human bodies and cars presented in a separate experiment (Fig. 2; Methods Summary). Unlike the stimuli in the main experiment, this second set of stimuli consisted of isolated objects without scene context that were centrally presented. We expected to find, in the case of the car stimuli for example, a higher correlation between response patterns evoked by scenes containing cars with response patterns evoked by isolated car pictures (within-category comparison) than with response patterns evoked by isolated body pictures (between-category comparison). We found that response patterns indeed correlated more strongly for within-category comparisons than between-category comparisons ($T_9 = 4.1$, $P < 0.005$; Fig. 3a), a difference that tended to be stronger for the attended than the unattended pictures (main effect of attention: $F_{1,9} = 3.9$, $P = 0.08$). This suggests a high-level representation of object-category information that is invariant to the different viewing conditions, many of which are typically encountered in daily life.

Of central interest to the present study was the effect of search task on this category information. Notably, category information depended fully on search task (category × task: $F_{1,9} = 13.5$, $P < 0.01$; Fig. 3a). During the body task, patterns of activation carried significant category information about bodies ($T_9 = 5.0$, $P < 0.001$), but not cars ($T_9 = -0.5$). Conversely, during the car task significant category information was present for cars ($T_9 = 2.9$, $P < 0.05$), but not for bodies ($T_9 = -1.8$). Furthermore, the category × task interaction was independent of spatial attention (category × task × attention: $F_{1,9} = 0.1$, $P = 0.79$), and significant for both the attended ($F_{1,9} = 11.7$, $P < 0.01$) and the unattended pictures ($F_{1,9} = 10.5$, $P < 0.05$). For these analyses we sorted trials on the basis of the content of either the attended or the unattended pictures. Because the content of these two picture pairs was fully independent, the content of one picture pair was effectively 'averaged out' when
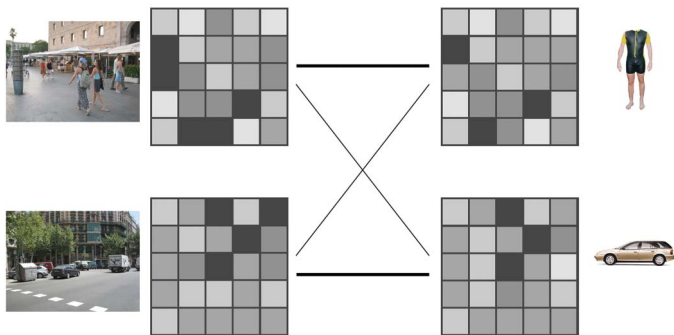
Figure 2 | Schematic overview of analysis approach. The approach of the multi-voxel pattern analysis was to correlate patterns of activation to conditions in the main experiment (depicted on the left) with patterns of activation to conditions in the category localizer (depicted on the right). The thickness of the lines between patterns indicates the hypothesized strengths of the correlations. For example, higher correlations were expected between patterns evoked by scenes containing people and isolated body pictures (within-category comparison) than with isolated car pictures (between-category comparison). This approach allowed us to measure the influence of search task and spatial attention on category information in visual cortex.
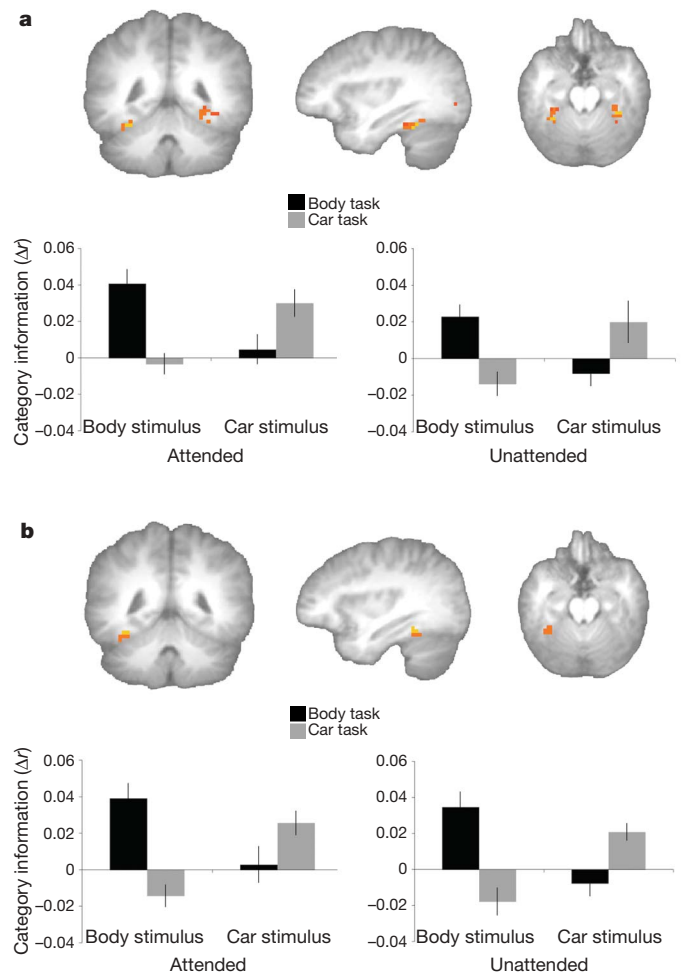
Figure 3 | Results of multi-voxel pattern analysis. a, The top panel shows the ventral cluster of object-selective cortex (intact versus scrambled objects) in a group-average analysis at $P < 0.005$ (Talairach coordinates of peak: $x = 35$, $y = -41$, $z = -18$). The lower panel shows category information as a function of category, task and attention in individually-defined object-selective cortex. Category information was calculated by taking the difference between within-category comparisons and between-category comparisons, and reflects the amount of category information in multi-voxel patterns of activation (see Fig. 2 and Methods Summary). Error bars indicate ± s.e.m. **b**, The top panel shows the result of the category × task contrast in the group-average searchlight analysis at $P < 0.005$ (uncorrected). The lower panel shows category information as a function of category, task and attention in the sphere surrounding the peak voxel of the activation from the group-average searchlight analysis (Talairach coordinates of peak: $x = 35$, $y = -44$, $z = -18$). Error bars indicate ± s.e.m.

measuring responses to the other picture pair, and vice versa. To ensure further that the results we have described were not related to the simultaneously presented other picture pair, we calculated the category × task interaction separately for the different simultaneously presented conditions. As expected, the category × task interaction for the attended pictures did not depend on the content of the unattended pictures, and vice versa ($P > 0.1$, for both tests). Finally, to test whether these results were specific to higher-level visual cortex, we performed the same analyses in retinotopically defined visual areas V1, V2 and V3. None of these regions contained category information, and there were no significant effects of task or category ($P > 0.1$, for all tests).

These results indicate that body and car stimuli embedded in natural scenes were visually processed to a high level only when they were actively searched for. Objects belonging to the target category were processed up to the category level, even when they were presented

2

outside the focus of attention. In contrast, objects belonging to the irrelevant category were not represented at the category level, even when they were presented inside the focus of attention.

To investigate whether other regions in the brain may show similar effects to those observed in object-selective visual cortex, we used an information-based functional brain mapping approach, measuring patterns of activation throughout the brain by means of a spherical searchlight[12]. Within a given sphere, we correlated patterns of activation to the conditions in the main experiment with patterns of activation to the separately presented isolated body and car stimuli, identical to the analysis in object-selective cortex. Individual subject data were spatially normalized, which allowed us to use a random-effects group analysis to test where in the brain search task influenced category information. The most significant category × task interaction was found in the ventral temporal cortex, overlapping the object-selective visual cortex (Fig. 3b). The correlation values in the sphere surrounding the peak voxel of this cluster were very similar to those of the a priori defined object-selective region (Fig. 3). As expected (given the selection of the sphere), category information (within- versus between-category comparison) depended strongly on the search task (category × task: $F_{1,9} = 34.9$, $P < 0.0005$; Fig. 3b). During the body task, patterns of activation carried significant information about bodies ($T_9 = 8.9$, $P < 0.00001$), but not cars ($T_9 = -0.6$), whereas for the car task, significant information was present for cars ($T_9 = 5.8$, $P < 0.0005$), but not (or negative) for bodies ($T_9 = -2.6$). The category × task interaction was independent of attention (category × task × attention: $F_{1,9} = 0.1$, $P = 0.81$), and significant for both the attended ($F_{1,9} = 16.6$, $P < 0.005$) and the unattended pictures ($F_{1,9} = 35.6$, $P < 0.0005$). This analysis demonstrated that the effect of search task on category information was strongest in the ventral temporal cortex, and could be observed in a group-average analysis without defining individual a priori regions-of-interest.

The present results provide evidence for a neural mechanism of attentional selection at the level of object category. Notably, this mechanism in high-level visual cortex seems to operate across the visual field independent of spatial attention, similar to attentional selection mechanisms for simple features (for example, for colour or motion direction) in lower-level visual cortex[13–15]. As such, it provides a neural basis for recent behavioural findings reporting an extraordinary capability of the human visual system to rapidly extract object category information from natural scenes, even outside the focus of spatial attention[3]. The selection mechanism described here may operate through the pre-activation of neurons representing the target category that subsequently biases the processing of the scene in favour of the target category[16]. Such an account of our data are in line with parallel models of visual search, in which a 'search template' biases the processing of a scene across the visual field in favour of objects that match the template[17,18]. Given the variability in the visual characteristics and spatial locations of individual category exemplars in natural scenes, such search templates would need to be rather abstract, invariant to geometric and photometric changes, and spatially unspecific. Furthermore, successful performance in our task required extra, space-based, selection mechanisms to prevent responses to the task-irrelevant pictures (see Supplementary Table 1 and Supplementary Discussion). Our results indicate that these additional selection mechanisms (for example, the retinotopically specific spatial attention effects observed in the retinotopic visual cortex; see Supplementary Fig. 2) operate mostly independently from the object-category mechanism observed in activity patterns in object-selective cortex.

Furthermore, the finding that objects belonging to the irrelevant object category were poorly represented in high-level visual cortex, even when they were spatially attended, is consistent with the phenomenon of 'change blindness', the finding that we are mostly unaware of large changes to objects in natural scenes when the identity of the changed object is unknown in advance[19–21]. Indeed, our results provide evidence that, contrary to our subjective experience of a complete internal representation of the external world, the neural representation of real-world scenes is limited to those objects that are directly relevant for ongoing behaviour[22].

Together, the present results provide a possible neural basis for both the limitations in our perception of real-world scenes and our remarkable ability for categorizing such scenes inside and outside the focus of spatial attention.

METHODS SUMMARY

Ten healthy adults (three females) participated in eight runs of the main experiment across two scanning sessions. In different runs, subjects performed one of two detection tasks on either the horizontal or the vertical pairs of pictures, resulting in four different run types. Within a scanning session, each of the presented pictures was unique. The same set of pictures was used in both sessions, such that across the two sessions the same pictures were presented in both detection tasks.

Activity patterns to the body and car conditions in the category localizer were correlated with activity patterns to the body and car scenes in the main experiment. This resulted in four correlations (for example, between body scenes in the main experiment and bodies in the localizer: r_body(main)_body(localizer)). Within-category correlations are the correlations where categories matched (for example, r_car(main)_car(localizer)), whereas between-category correlations are the correlations between non-matching categories (for example, r_car(main)_body(localizer)). Category information was defined as the difference between within-category and between-category correlations. For example, category information (Δr) for the body stimuli (left two bars in Fig. 3) was calculated by: (r_body(main)_body(localizer)) − (r_body(main)_car(localizer)).

Functional (echoplanar imaging (EPI) sequence; 34 slices per volume; resolution = $3 \times 3 \times 3$ mm with 1 mm gap; repetition time (TR) = 2.0 s; time to echo (TE) = 30 ms; flip angle = $90°$) and anatomical (MPRAGE sequence; 256 matrix; TR = 2.5 s; TE = 4.38 ms; flip angle, $8°$; $1 \times 1 \times 1$ mm resolution) images were acquired with a 3T Allegra MRI scanner (Siemens). Functional data were slice-time corrected, motion corrected, and low-frequency drifts were removed with a temporal high-pass filter (cutoff of 0.006 Hz). No spatial smoothing was applied.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Potter, M. C. Meaning in visual search. *Science* **187**, 965–966 (1975).
2. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).
3. Li, F. F., VanRullen, R., Koch, C. & Perona, P. Rapid natural scene categorization in the near absence of attention. *Proc. Natl Acad. Sci. USA* **99**, 9596–9601 (2002).
4. Broadbent, D. E. *Perception and Communication* (Oxford Univ. Press, 1958).
5. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cognit. Psychol.* **12**, 97–136 (1980).
6. Fei-Fei, L., VanRullen, R., Koch, C. & Perona, P. Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Vis. Cogn.* **12**, 893–924 (2005).
7. Rousselet, G. A., Fabre-Thorpe, M. & Thorpe, S. J. Parallel processing in high-level categorization of natural images. *Nature Neurosci.* **5**, 629–630 (2002).
8. Wojciulik, E., Kanwisher, N. & Driver, J. Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. *J. Neurophysiol.* **79**, 1574–1578 (1998).
9. Cox, D. D. & Savoy, R. L. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* **19**, 261–270 (2003).
10. Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
11. Malach, R. *et al.* Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc. Natl Acad. Sci. USA* **92**, 8135–8139 (1995).
12. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103**, 3863–3868 (2006).
13. Maunsell, J. H. & Treue, S. Feature-based attention in visual cortex. *Trends Neurosci.* **29**, 317–322 (2006).
14. Motter, B. C. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.* **14**, 2178–2189 (1994).
15. Saenz, M., Buracas, G. T. & Boynton, G. M. Global effects of feature-based attention in human visual cortex. *Nature Neurosci.* **5**, 631–632 (2002).
16. Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. A neural basis for visual search in inferior temporal cortex. *Nature* **363**, 345–347 (1993).

17. Duncan, J. EPS mid-career award 2004: brain mechanisms of attention. *Q. J. Exp. Psychol. (Colchester)* **59**, 2–27 (2006).
18. Duncan, J. & Humphreys, G. W. Visual search and stimulus similarity. *Psychol. Rev.* **96**, 433–458 (1989).
19. O'Regan, J. K., Rensink, R. A. & Clark, J. J. Change-blindness as a result of 'mudsplashes'. *Nature* **398**, 34 (1999).
20. Rensink, R. A., O'Regan, J. K. & Clark, J. J. To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* **8**, 368–373 (1997).
21. Simons, D. J. & Rensink, R. A. Change blindness: past, present, and future. *Trends Cogn. Sci.* **9**, 16–20 (2005).
22. O'Regan, J. K. Solving the ''real'' mysteries of visual perception: the world as an outside memory. *Can. J. Psychol.* **46**, 461–488 (1992).

METHODS

Stimuli. Two-thousand-and-fifty-six natural scene pictures were selected from an online database[23]. Five-hundred-and-twelve pictures contained one or more people (but no cars), 512 pictures contained one or more cars (but no people), and 1,024 pictures contained no cars and no people. The pictures were mostly photographs of city streets. The position, viewpoint and size of the people and cars in the pictures were highly variable, mimicking real-world viewing conditions (see Supplementary Fig. 1 for sample pictures). Twelve different perceptual masks were created. Each was a coloured picture of a mixture of white noise at different spatial frequencies on which a naturalistic texture was superimposed[24].

All pictures were full-colour photographs reduced to 480 (vertical) × 640 (horizontal) pixels. The pictures along the horizontal axis (size: $7.9° \times 10.5°$) were presented with an offset of $3.5°$ to the right or left of the central fixation cross. The pictures along the vertical axis (size: $4.5° \times 6°$) were presented with an offset of $0.5°$ above or below the central fixation cross.

General procedure. Each subject participated in one practice session and four scanning sessions. The first scanning session was used for mapping retinotopic visual areas. The second and third session consisted of four runs of the main experiment, two runs of the category pattern localizer, and two runs of the object-selective cortex localizer. The fourth session consisted of a position localizer to measure retinotopic representations corresponding to the picture positions in the main experiment. See Supplementary Methods for details of the localizer experiments. Data were analysed using the AFNI software package[25] and MATLAB (The MathWorks).

Main experiment. Each run started and ended with a fixation-only block of 14 s. The average trial duration was 2.2 s. Each trial started with a 500 ms presentation of four placeholders at the location of the to-be-presented pictures. The placeholders of the relevant locations (either the two horizontal, or the two vertical locations) were highlighted during this time, by increasing the width of the placeholder outline by 2 pixels (~$0.05°$). This was followed by the presentation of the four pictures for 130 ms on average. The presentation time of the pictures was individually determined on the basis of the performance in a separate session, and was held constant within a scanning session (see Supplementary Methods). A presentation of perceptual masks directly followed the picture presentations. The combined duration of picture and mask presentations was always 400 ms. A screen with the four placeholders followed the presentation of the masks for a duration of 950 to 1,650 ms (pseudo-randomly chosen, with intervals of 100 ms), with a mean duration of 1,300 ms.

On 20% of the trials (32 trials per run) only the masks (but no pictures) were presented, for 270 ms on average. On the remaining 128 trials, four simultaneously presented pictures were presented followed by four masks (each randomly selected from the total set of 12 masks). Each of the four scenes could contain either: people but no cars ('body'), cars but no people ('cars'); or no people and no cars ('none'). Within each pair of pictures (horizontal and vertical) there were eight possible combinations of these scene types: body–none, none–body, cars–none, none–cars, body–cars, cars–body, none–none and none–none. Each of these combinations was presented equally often. Trial order was randomized (without replacement).

Our research questions focused on the scene combinations that contained either people but no cars, or cars but no people. The other scene combinations were included in the design so that the irrelevant object category did not predict the presence or absence of the relevant object category.

In different runs, subjects performed one of two detection tasks ('body task' or 'car task') on either the horizontal or the vertical pairs of pictures, resulting in four different run types. The task was to press one button for the presence and another button for the absence of the target category in the relevant picture pair. The mapping of the two buttons (index and middle finger) to present and absent responses was counterbalanced across sessions and subjects. Subjects always performed two runs of the same task in a row, to prevent frequent switching from one target category to another. The order of the tasks was counterbalanced across sessions and subjects. In different runs, subjects performed the tasks on either the horizontal or vertical picture pairs, the order of which was held constant within a session, and which was counterbalanced across the two sessions.

Multi-voxel pattern analysis. For each subject, general linear models were created for the main experiment and the category pattern localizer experiment. One predictor (convolved with a standard model of the haemodynamic response function) modelled each condition. Regressors of no interest were also included to account for differences in the mean magnetic resonance signal across scans and for head motion within scans. Two separate regression analyses were run for the main experiment; with trials grouped based on either the task-relevant or task-irrelevant pairs of scene pictures. These regression analyses resulted, for each voxel, in a *t*-value for each condition in the main experiment and for each condition in the localizer. The *t*-values of conditions in the main experiment were correlated, across the voxels of a region of interest (ROI), with *t*-values of the body and car conditions in the localizer. The analyses were repeated using parameter estimates instead of *t*-values, which yielded highly similar results. The analysis was done for each subject and session separately. Correlations were Fisher-transformed before statistical testing: $0.5\ln[(1 + r)/(1 − r)]$. Correlations of the two sessions were averaged. For each subject, ROI and localizer condition, the average correlation across conditions in the main experiment was subtracted out to remove overall differences in correlations that were not of interest. Differences between the resulting voxelwise correlations were then tested using repeated-measures analysis of variance (ANOVA) and *t*-tests (two-tailed) with subject ($n = 10$) as random factor. All trials (correct and incorrect) were included in the analyses. We repeated all analyses on correct trials only, which yielded qualitatively similar but slightly less significant results (due to the smaller number of trials).

Searchlight analysis. A whole-brain pattern analysis was performed using a spherical searchlight[12]. For each voxel in the brain we computed voxelwise correlations in a sphere of 10-mm radius (corresponding to 121 voxels) centred on this voxel. The voxelwise correlations were computed as described above. The correlation values from each sphere were Fisher-transformed and assigned to the centre voxel of this sphere. The correlations were computed for each subject and session separately. Results were transformed into Talairach space, the correlations of the two sessions were averaged for each subject, and random-effects group analyses were performed.

23. Russell, B. C., Torralba, A., Murphy, K. P. & Freeman, W. T. LabelMe: a database and web-based tool for image annotation. *Int. J. Comput. Vis.* **77**, 157–173 (2008).
24. Walther, D. B., Caddigan, E., Fei-Fei, L. & Beck, D. M. Natural scene categories revealed in distributed patterns of activity in the human brain. *J. Neurosci.* (in the press).
25. Cox, R. W. & Hyde, J. S. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).