Report

Decoding the Representation of Multiple Simultaneous Objects in Human Occipitotemporal Cortex

Sean P. MacEvoy[1,2,*] and Russell A. Epstein[1,2]
[1]Department of Psychology
[2]Center for Cognitive Neuroscience
University of Pennsylvania
Philadelphia, PA 19104
USA

Summary

Previous work using functional magnetic resonance imaging has shown that the identities of isolated objects viewed by human subjects can be extracted from distributed patterns of brain activity [1]. Outside the laboratory, however, objects almost never appear in isolation; thus it is important to understand how multiple simultaneously occurring objects are encoded by the visual system. We used multivoxel pattern analysis to examine this issue, testing whether activity patterns in the lateral occipital complex (LOC) evoked by object pairs showed an ordered relationship to patterns evoked by their constituent objects. Applying a searchlight analysis [2] to identify voxels with the highest signal-to-noise ratios, we found that responses to object pairs within these informative voxels were well predicted by the averages of responses to their constituent objects. Consistent with this result, we were able to classify object pairs by using synthetic patterns created by averaging single-object patterns. These results indicate that the representation of multiple objects in LOC is governed by a response normalization mechanism similar to that reported in visual areas of several nonhuman species [3–6]. They also suggest a population coding scheme that preserves information about multiple objects under conditions of distributed attention, facilitating fast object and scene recognition during natural vision.

Results

Single- and Paired-Object Classification

In a block design, subjects viewed single objects in four categories (shoes, chairs, cars, or brushes), as well as object pairs containing objects from two of these categories. Several previous studies have shown that information about the category of viewed objects is present in distributed patterns of activity measured with functional magnetic resonance imaging (fMRI) [1, 7] and we first wished to replicate this finding as a means of validating the quality of our data. Figure 1A shows classification performance for single objects within standard functionally defined regions of interest (ROIs) (see Supplemental Experimental Procedures available online). Consistent with previous work, classification accuracy was significantly above chance in lateral occipital complex [LOC; two-tailed t test, $t(11) = 4.95$, $p = 0.0004$]. Classification accuracy was also above chance in the parahippocampal place area [PPA; $t(11) = 2.77$, $p = 0.018$] but not in the fusiform face area

*Correspondence: macevoy@gmail.com

[FFA; $t(11) = 1.56$, $p = 0.15$] or a nonbrain ROI [$t(11) = 0.13$, $p = 0.89$]. (See Supplemental Results for additional classification analyses, including the impact of changes in stimulus position upon accuracy.)

We next assessed the accuracy of the classifier in distinguishing among object pairs (Figure 1B). For classification purposes, each unique object pair was treated as a distinct stimulus (e.g., chair+brush and car+brush were treated as different stimulus categories), producing six pairs from the pool of four object categories. Classification accuracy for pairs was significantly above chance in LOC [$t(11) = 4.68$, $p = 0.0007$] but not in the PPA [$t(11) = 1.04$, $p = 0.31$], FFA [$t(11) = 1.68$, $p = 0.12$], or the nonbrain ROI [$t(11) = 0.86$, $p = 0.40$]. These results, along with whole-brain maps of local classification accuracy (Figure S1), indicate that activity patterns in LOC reliably discriminate among object pairs as well as among single objects.

Relationship of Paired-Object to Single-Object Responses

Do LOC patterns evoked by pairs bear any relationship to patterns evoked by their constituent objects? To answer this question, we first assessed the ability of a linear model to explain responses to object pairs [3, 5]. For each voxel, we performed a linear regression of the responses to pairs against the sum of responses to their constituent objects. This procedure is illustrated in Figure 2A for a voxel with a strong linear relationship between responses to pairs and to single objects ($R^2 = 0.96$).

Many voxels had much lower R^2 values, which could have reflected the impact of noise on a linear relationship, a nonlinear relationship, or no relationship at all. To differentiate between these possibilities, we used a searchlight classification technique to identify local voxel clusters that carried information about stimulus identity (see Experimental Procedures). We reasoned that searchlight clusters that most accurately differentiated among object pairs would contain voxels that were the most instructive of the "true" relationship between responses to pairs and constituent single objects. Therefore, if a linear model provides a good description of this relationship, we would expect to see R^2 increase as a function of searchlight classification accuracy. (See Supplemental Results and Discussion for a detailed treatment of this approach.)

Figure 2B plots median R^2 within each LOC searchlight cluster as a function of cluster classification rank for one subject. (We used classification rank, rather than raw classification accuracy, as the independent variable in order to facilitate averaging data across subjects, among whom overall classification accuracy varied.) For this subject, there was a clear trend toward higher R^2 values as classification accuracy improved. This relationship was also apparent in R^2 averaged across subjects (Figure 2C). To quantify this trend, we computed correlation coefficients between R^2 and classification rank within LOC for each subject. All subjects had positive correlation coefficients and all but two were significantly greater than zero at a $p < 0.05$ threshold. Across subjects, mean correlation coefficients were significantly above zero [mean = 0.33, $t(11) = 6.32$, $p = 0.00006$]. From the positive relationship between R^2 and classification rank, we infer that
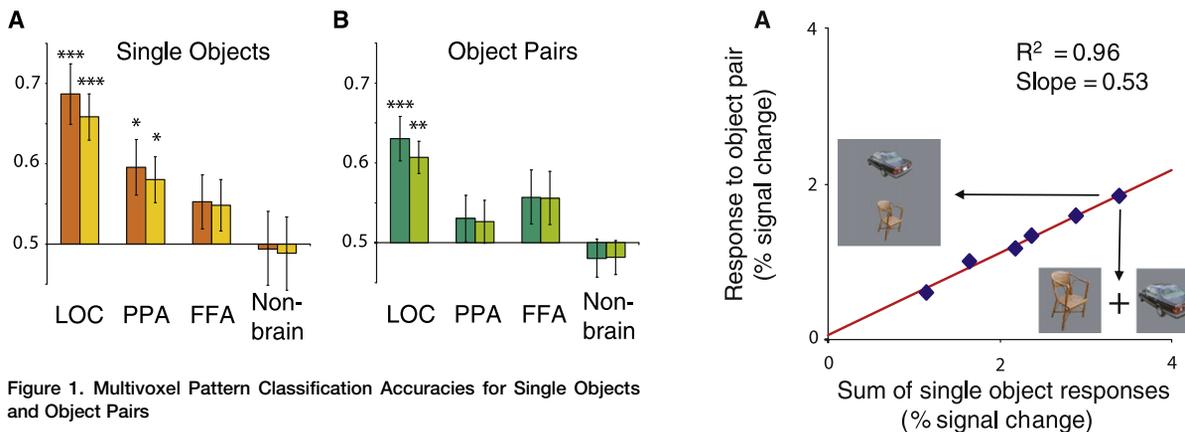
Figure 1. Multivoxel Pattern Classification Accuracies for Single Objects and Object Pairs

(A) Classification accuracy was significantly above chance (0.5) for the four single-object categories in both LOC and the PPA.

(B) Classification accuracy for the six possible category pairs was significantly above chance only in LOC.

Patterns were averaged across stimulus position (singles) or configuration (pairs) prior to classification. Dark-hued bars represent classification accuracy based on all voxels within each ROI, and lighter-hued bars represent accuracy for ROIs matched in voxel count to the smallest ROI for each subject. Asterisks denote significance of difference from chance performance (*p < 0.05; ***p < 0.001). Data are for 12 subjects. Error bars represent standard error of the mean (SEM).

responses to object pairs are well approximated by a linear combination of responses to single objects. (Similar analyses for the PPA, FFA, and retinotopic cortex can be found in the Supplemental Results.) A permutation-based control analysis demonstrated that this relationship was not a trivial outcome of voxel selection (i.e., "peeking;" see Supplemental Results).

Although R^2 captures the quality of a linear relationship, it does not specify its parameters. To understand whether voxels in LOC obeyed any specific linear relationship between pair and single-object responses, we examined the slope terms returned by the linear regressions described above. Figure 2D illustrates the relationship between classification rank and median slope for each searchlight position for one subject, and Figure 2E plots the same relationship averaged over all subjects. As with R^2, median searchlight slopes increased as classification accuracy improved. More importantly, slope values among high performing clusters fell close to 0.5, indicating that pair responses were approximately the average of responses to their constituent single objects. This result echoes a previous finding by Zoccolan et al. [5] that neuronal responses in macaque inferotemporal (IT) cortex to pairs of objects are well predicted by the average of the single-object responses. Although the terminal slope value in Figure 2E was 0.62, this value was not significantly different from 0.5. Terminal slope values for LOC were fairly consistent across subjects, with 8 of 12 subjects' values falling between 0.35 and 0.65. Furthermore, analysis of the distribution of residual error between actual pair responses and regression lines indicated that these results are more consistent pair responses that were simple, rather than weighted, averages of responses to constituent single objects (see Supplemental Results).

Linear regression returns an intercept term in addition to slope, which was not significantly different from zero in the top 30 LOC searchlight clusters [$t(11) = 0.76$, $p = 0.45$]. Thus, the responses to object pairs were truly the averages of responses to single objects, without any additional offset reflecting systematic differences in overall activity evoked by pairs and single objects.
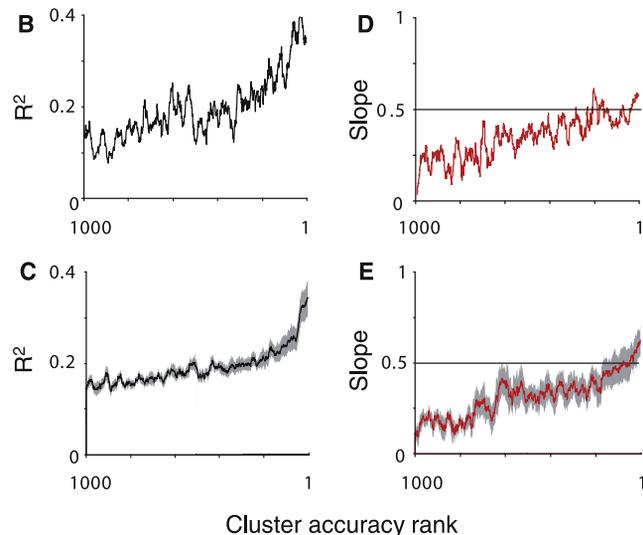

Figure 2. Relationship between Single- and Paired-Object Responses in LOC

(A) Responses of a single voxel to each of the six object pairs, plotted against the responses to the sum of responses evoked by each pair's constituent objects. For this voxel, pair responses showed a tight linear relationship ($R^2 = 0.96$) to single-object responses.

(B) Median R^2 within LOC searchlight clusters plotted as a function of each cluster's rank in classifying object pairs, for one subject. Data were smoothed with a 20-bin mean filter. R^2 increased with classification rank, suggesting that a linear model provides a good prediction of pair responses as noise is reduced.

(C) Same as (B), averaged across all subjects. Missing data for subjects with fewer than 1000 searchlights were ignored when the computing the average curve. Shaded regions fall within SEM.

(D) Median slope within searchlight clusters for one subject plotted as function of searchlight accuracy rank, smoothed with a 20 bin mean filter. Regression slopes fell close to 0.5 for the highest ranked searchlight clusters.

(E) Same data as in (D), averaged across subjects, with the same conventions as in (C).

Classification Using Synthetic Patterns

The preceding analyses suggest that we may approximate the responses of LOC voxels to object pairs as the averages of responses evoked by their constituent objects. To test this assertion, we repeated the pair pattern classification procedure but replaced pair patterns in one half of the data with "synthetic" patterns that were the averages of patterns

evoked by the corresponding single objects. Patterns were limited to voxels that fell within the 30 highest performing searchlights in terms of pair classification, which typically afforded the highest average classification performance across subjects (Figure S4). It is critical to note that although these voxels were selected on the basis of high pair classification in their searchlights, this criterion was completely independent of the responses to single objects that were used to construct synthetic patterns.

Classification accuracies using synthetic patterns are shown in Figure 3A. At a rate well above chance [$t(11) = 8.54$, $p < 0.00001$], the classifier was able to correctly identify patterns evoked by object pairs based on comparison to synthetic response patterns derived by averaging the single-object responses within each voxel. Although classification based on these synthetic patterns was not as accurate as classification based on actual pair patterns, it was significantly more accurate [$t(11) = 5.45$, $p = 0.0002$] than classification based on a set of "MAX" function synthetic patterns generated by taking the higher of each voxel's responses to the two single objects comprising each pair [8]. This is consistent with the idea that pair responses reflect linear rather than nonlinear combinations of single-object responses.

Our ability to classify pairs from single-object patterns suggests that inverting the operation should allow us to decode the identities of single objects from the pattern evoked by a pair. Reddy and Kanwisher [7] found that classification accuracy for single objects was markedly degraded in LOC when a second object was present. The origin and nature of this "clutter cost" was unclear, however. Was information about the identity of objects actually lost? Or did the "cost" simply reflect the joint representation of both objects? Under the second scenario, we should be able to recoup clutter costs through appropriate decoding of patterns evoked by pairs.

We first assessed the impact of clutter in our own data by measuring classification accuracy for single objects within pairs. A correct classification decision was recorded when the Euclidean distance between the pattern evoked by a pair and the pattern evoked by one of its component objects (the "target" object for the purposes of classification only) was less than the distance between the pair pattern and the pattern for a comparison object not in the pair. Consistent with Reddy and Kanwisher [7], accuracy for single objects in pairs was significantly lower than accuracy for single objects by themselves (Figure 3B) in LOC [$t(11) = 4.27$, $p = 0.0013$], reflecting a substantial clutter cost.

To recover this clutter cost, we assumed that patterns evoked by object pairs were the averages of patterns evoked by their constituent objects. Accordingly, to extract the pattern evoked by a target object from a pair response, we subtracted a half-scaled version of the pattern evoked by the nontarget object and multiplied the resulting pattern by two. Applying this treatment produced a significant improvement in classification [Figure 3B; $t(11) = 4.02$, $p = 0.002$]. Linear decoding via this approach recovered an average of 48% of clutter costs associated with the presence of a second object. This result confirms the claim that information about each individual object is embedded in patterns evoked by object pairs.

Discussion

The principal finding of this study is that under conditions of distributed attention, voxelwise patterns of activity in object-
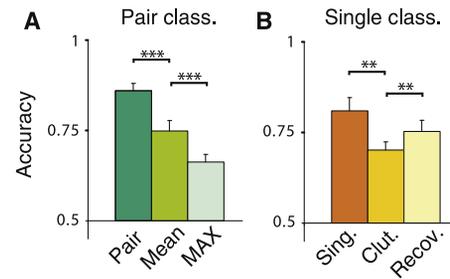

Figure 3. Synthetic Patterns Decode Patterns Evoked by Object Pairs in LOC

(A) Performance in classifying patterns evoked by pairs using actual pair patterns (Pair), synthetic patterns derived the means of single-objects patterns (Mean), or a MAX-function combination of single-object patterns (Max). Although classification accuracy for the mean predictor was not as high as for actual pairs, it was significantly higher than accuracy for the nonlinear MAX predictor.

(B) Performance for classifying single-object patterns (Sing.) was significantly reduced when a second, cluttering object appeared simultaneously (Clut.). A large portion of this clutter cost was recovered by assuming that pair patterns were the average of their component object patterns and by linearly decomposing responses to pairs accordingly (Recov.). Patterns included all voxels that fell within the top 30 clusters ranked in terms of pair classification rank. Across subjects, this corresponded to an average of 154 voxels. Error bars represent SEM.

selective cortex evoked by pairs of objects are the average of the patterns evoked by the individual component objects. Consistent with this result, pair patterns could be decoded with high accuracy by reference to synthetic patterns generated by averaging the single-object responses. Conversely, subtraction of an appropriately-scaled version of the voxel pattern evoked by one object of a pair recovered the pattern evoked by the second object.

This work builds on and extends two previous findings. First, Zoccolan et al. [5] demonstrated that responses of object-selective neurons of macaque area IT to pairs of objects were precisely predicted by the average of responses to their constituent objects. Our results demonstrate that a similar averaging rule applies to human LOC. Second, Reddy and Kanwisher [7] demonstrated a clutter cost for classification of single, focally attended objects when a second, unattended stimulus was present. Here we demonstrate that when the two objects are equally attended, a substantial portion of this cost for one object can be recouped if the response pattern to the second object is known.

These results potentially provide important insights into how visual recognition might proceed in the real world. The fact that objects in natural scenes almost always appear amidst other objects presents both a challenge and an opportunity for the visual system. The challenge is to identify single objects even when they are surrounded by the clutter of other stimuli. Attentional mechanisms might help solve this problem by boosting up the neural response to attended objects while suppressing the neural response to unattended objects [9–11]. However, this suppression of unattended object response can potentially negate an important informational opportunity. Specifically, the multiple objects within the scene might, if considered together, convey information about the "gist" or "context" of the scene [12–15]. Behavioral studies indicate that humans can indeed extract this gist information very rapidly [12, 16]; furthermore, observers can report the identities of objects within a scene even after very brief

presentation times that are unlikely to permit attention to be moved serially from object to object [17]. Our results suggest a way in which the visual system might accomplish this feat. In particular, if the pattern evoked by a multiple-object scene is linearly related to the patterns evoked by its constituent objects, then the gist might correspond simply to an initial hypothesis about the set of objects contributing to this overall pattern and a judgment about the category of scene that is most likely to contain such objects. Indeed, if there is a lawful relationship between the representations of a whole scene and of its component objects, then the same neural system can be use to represent both.

This reasoning explains why it would be advantageous for the visual system to maintain a linear relationship between single- and multiple-object responses, but it does not explain why the voxel patterns evoked by pairs resemble the average of single-object patterns. In its adherence to the mean, LOC appears to obey rules similar to those that have been described previously in a variety of visual areas in nonhuman species [3, 4, 6, 18–20] and which have traditionally been explained as an outcome of competition between stimuli for limited neural bandwidth [10, 11, 19, 21]. Our results suggest an alternative framing of this phenomenon in which response averaging reflects a normalization process that actively supports the coding of multiple simultaneous objects by avoiding the problems presented by saturation of neural responses [22]. Because individual neurons have finite firing rates, pure summation of responses to multiple objects runs the risk of driving some neurons to saturation, particularly those that respond vigorously to both objects. Once this happens, the population response to a pair of objects is no longer a linear combination of the patterns evoked by each object by itself and information about the identity of each object is lost. By scaling population responses by the number of stimuli present, normalization helps avoid this problem by ensuring that response saturation cannot be reached.

The presence of multistimulus normalization in LOC might also provide a window into its functional organization. Whenever response normalization has been found in nonhuman visual areas, such as with multiple oriented contours in V1, V2, and V4 [3, 4, 18, 19, 23], directions of motion in middle temporal and medial superior temporal cortex [6, 20], or shape in IT [5], the simultaneously presented stimuli have differed along some dimension that is "mapped" across the surface of the area under study (i.e., the individual stimuli presented by themselves activate spatially distinct clusters of neurons [24–27]). We speculate that this sort of mosaic-like organization might be a prerequisite to multistimulus normalization. If so, our results provide additional evidence that LOC neurons are clustered according to shape or category. (Indeed, such functional clustering might be necessary for multivoxel pattern analyses to work in the first place [28, 29].)

Finally, our data revealed two additional novel and somewhat surprising phenomena. First, the LOC territory that best encoded object pairs was largely identical to the LOC territory that best encoded single objects (Figure S1). In contrast, the PPA did not encode object pairs as reliably as LOC even though it did encode information about single objects. These findings are consistent with previous claims that LOC rather than the PPA is the primary region involved in encoding object identity information, even when more than one object is present in a scene [30]. Second, LOC response patterns did not distinguish between different spatial configurations of a pair (i.e., shoe over brush was indistinguishable from brush over shoe). This suggests that when attention is distributed evenly across a scene, object identity is encoded independently of object location in the ventral stream [31].

Experimental Procedures

Stimulus and Task

Stimuli were 60 photographic images (1.7° square) of common objects from four categories (brushes, cars, chairs, and shoes) with all background elements removed. Stimuli were presented in 15 s blocks (see Supplemental Experimental Procedures). In single-object blocks, 15 exemplars from the same object category were presented one at a time at a single screen position that was centered either 1.7° above or below the fixation point. In paired-object blocks, 15 exemplars from two categories (30 in total) were presented two at a time, with exemplars from one category appearing in the top screen position and exemplars from the other category appearing in the bottom screen position. Within each scan run, each object category was presented twice in the single-object condition (once in the upper screen position and once in the lower screen position) and each category pairing was shown twice (corresponding to the two possible spatial configurations; e.g., top brush/bottom chair and top chair/bottom brush).

To ensure that attention was paid equally to all objects, we required subjects (n = 12) to perform a one-back repetition detection task while maintaining central fixation. In paired-object blocks the repetition could occur at either stimulus location, forcing subjects to attend to both.

Data Analysis

Following standard preprocessing, fMRI data were passed to a general linear model implemented in VoxBo, from which voxelwise beta values associated with each stimulus condition were extracted (see Supplemental Experimental Procedures). Multivoxel pattern classification was implemented with custom code written in MATLAB and using an algorithm similar to Haxby et al. [1]. In brief, response patterns were extracted for each ROI from each of the six experimental scans. Data were then divided into halves (e.g., even runs versus odd runs) and the patterns within each half were averaged. A "cocktail" mean pattern (consisting of the average pattern across all stimuli) was calculated separately for each half of the data and then subtracted from each of the individual patterns before classification. Separate cocktails were computed for single objects and paired objects. No pattern normalization was applied at any point.

Pattern classification proceeded through a series of pairwise comparisons between stimulus conditions. Correct classification decisions were recorded when the Euclidean distance between the patterns evoked by condition A in opposite halves of the data was shorter than between condition A and condition B in opposite halves of the data. This procedure was repeated for every possible stimulus pairing, and correct decisions were accumulated across every possible binary split of the six scan runs. Additional analysis showed that the Euclidean distance metric produced classification accuracies similar to a correlation-based classifier [1].

Searchlight voxel selection [2] was implemented with custom MATLAB code. For each voxel, we defined a spherical mask that included all other voxels within a 5 mm radius. Searchlight clusters near the cortical surface were truncated to ensure that only voxels within the brain were included. Similarly, when searchlights were used on predefined ROIs, searchlight masks were truncated where necessary so that only voxels within the ROI were included.

Supplemental Data

Supplemental Data include Supplemental Results, Supplemental Discussion, Supplemental Experimental Procedures, and four figures and can be found with this article online at http://www.cell.com/current-biology/supplemental/S0960-9822(09)00980-4.

References

1. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

2. Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA 103, 3863–3868.

3. MacEvoy, S.P., Tucker, T.R., and Fitzpatrick, D. (2009). A precise form of divisive suppression supports population coding in primary visual cortex. Nat. Neurosci. 12, 637–645.

4. Luck, S.J., Chelazzi, L., Hillyard, S.A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J. Neurophysiol. 77, 24–42.

5. Zoccolan, D., Cox, D.D., and DiCarlo, J.J. (2005). Multiple object response normalization in monkey inferotemporal cortex. J. Neurosci. 25, 8150–8164.

6. Recanzone, G.H., Wurtz, R.H., and Schwarz, U. (1997). Responses of MT and MST neurons to one and two moving objects in the receptive field. J. Neurophysiol. 78, 2904–2915.

7. Reddy, L., and Kanwisher, N. (2007). Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. Curr. Biol. 17, 2067–2072.

8. Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025.

9. Posner, M.I. (1980). Orienting of attention. Q. J. Exp. Psychol. 32, 3–25.

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

11. Kastner, S., De Weerd, P., Desimone, R., and Ungerleider, L.C. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. Science 282, 108–111.

12. Potter, M.C. (1975). Meaning in visual search. Science 187, 965–966.

13. Treisman, A. (2006). How the deployment of attention determines what we see. Vis. Cogn. 14, 411–443.

14. Oliva, A., and Torralba, A. (2007). The role of context in object recognition. Trends Cogn. Sci. 11, 520–527.

15. Schyns, P.G., and Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. Psychol. Sci. 5, 195–200.

16. Biederman, I. (1972). Perceiving real-world scenes. Science 177, 77–80.

17. Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. (2007). What do we perceive in a glance of a real-world scene? J. Vis. 7, 10.

18. Moran, J., and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. Science 229, 782–784.

19. Reynolds, J.H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. J. Neurosci. 19, 1736–1753.

20. Britten, K.H., and Heuer, H.W. (1999). Spatial summation in the receptive fields of MT neurons. J. Neurosci. 19, 5074–5084.

21. Miller, E.K., Gochin, P.M., and Gross, C.G. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. Brain Res. 616, 25–29.

22. Heeger, D.J. (1992). Normalization of cell responses in cat striate cortex. Vis. Neurosci. 9, 181–197.

23. Carandini, M., Heeger, D.J., and Movshon, J.A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. J. Neurosci 17, 8621–8644.

24. Ghose, G.M., and Ts'o, D.Y. (1997). Form processing modules in primate area V4. J. Neurophysiol. 77, 2191–2196.

25. Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. Nature 360, 343–346.

26. Albright, T.D., Desimone, R., and Gross, C.G. (1984). Columnar organization of directionally selective cells in visual area MT of the macaque. J. Neurophysiol. 51, 16–31.

27. Hubel, D.H., and Livingstone, M.S. (1987). Segregation of form, color, and stereopsis in primate area 18. J. Neurosci. 7, 3378–3415.

28. Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.

29. Haynes, J.D., and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8, 686–691.

30. Epstein, R.A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. Trends Cogn. Sci. 12, 388–396.

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