NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 23 (2004) 156–166

Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area?

Stephen José Hanson,[a,*] Toshihiko Matsuka,[a] and James V. Haxby[b]

[a] *Rutgers University, Newark, NJ 07102, USA*
[b] *Princeton University, Princeton, NJ 08540, USA*

Received 13 January 2004; revised 5 May 2004; accepted 18 May 2004

Haxby et al. [Science 293 (2001) 2425] recently argued that category-related responses in the ventral temporal (VT) lobe during visual object identification were overlapping and distributed in topography. This observation contrasts with prevailing views that object codes are focal and localized to specific areas such as the fusiform and para-hippocampal gyri. We provide a critical test of Haxby's hypothesis using a neural network (NN) classifier that can detect more general topographic representations and achieves 83% correct generalization performance on patterns of voxel responses in out-of-sample tests. Using voxel-wise sensitivity analysis we show that substantially the same VT lobe voxels contribute to the classification of all object categories, suggesting the code is combinatorial. Moreover, we found no evidence for local single category representations. The neural network representations of the voxel codes were sensitive to both category and superordinate level features that were only available implicitly in the object categories.

Keywords: Combinatorial codes; Ventral temporal lobe; Object recognition

Introduction

How does the brain encode and represent objects? Functional brain imaging has revealed that the human ventral object vision pathway has a complex functional architecture. Different categories of objects evoke different patterns of response in these cortices. Based on standard methods for analyzing and interpreting functional brain imaging results, these patterns are usually described in terms of the locations of regions that respond more strongly to one category, for example, faces, than to all others (Aguirre et al., 1998; Downing et al., 2001; Epstein and Kanwisher, 1998; Hasson et al., 2003; Ishai et al., 1999; Kanwisher et al., 1997; McCarthy et al., 1997). In previous work, however, Haxby et al. (2001) showed that category-related information is also carried by weaker responses in these patterns of response and proposed that strong and weak responses may all play an integral role in the representation of objects. Thus, the representations for multiple categories overlap because a strong response to one category and intermediate or weak responses to other categories in the same piece of cortex are all parts of the representations for these categories. Such representations have an essentially unlimited carrying capacity by virtue of the number of combinatorial possibilities. By contrast, representations based on localized processors or modules, identified by maximal response to the objects for which they are specialized, are limited by the number of category-dedicated regions that can fit into a cortical space.

The similarity method of Haxby et al. (2001) was intended as a demonstration of a concept, designed to attempt to measure category-related, distributed patterns of response, but it was inefficient and insensitive to the range of possible distributed coding possibilities. It and other analyses (Spiridon and Kanwisher, 2002) have also confused category identification with feature (cortical response) sensitivity making it unlikely that functional areas could be uniquely identified (cf. Bartels and Zeki, 2004). Others have since applied various multivariate methods for analyzing distributed patterns of response in functional magnetic resonance imaging (fMRI) data sets, such as linear discriminant analysis (Carlson et al., 2003) and support vector machines (Cox and Savoy, 2003). All of these methods examine a form of information in fMRI data that is overlooked in standard methods of analysis (Friston et al., 1994). The usual statistical methods analyze the temporal course of response in each voxel independently of all other voxels then search for clusters of voxels with similar responses. By contrast, these multivariate methods explicitly analyze how the response varies across clusters of voxels and how these patterns of response, or landscapes, change with cognitive or perceptual state (see Haxby, in press). These types of methods could be used to detect representations that involve specific local codes that index a compact region (cf. Fodor, 1983), perhaps varying in shape or size, or for probabilistic maps that vary in intensity over the region in a distributed and possibly overlapping way. There are actually four logical possibilities for such coding schemes: (1) spatially local or compact codes that indicate the presence or absence of a type of object, (2) spatially local or compact codes that also indicate "likelihood" of the object type, (3) distributed codes that are non-overlapping and hence act as a potential local code but are distributed through the region in a unique pattern (these types of codes could also vary in intensity), finally, (4) distributed codes that are either partially or completely overlapping and vary in intensity. The case of completely overlapping distributed code is often called a combinatorial

[*] Corresponding author.
E-mail address: jose@psychology.rutgers.edu (S.J. Hanson).

Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2004.05.020

code. They only depend on the pattern of activity in which each subregion of the landscape responds in a continuous way to create an object code. Activity in a subregion is therefore more similar to the kind of coding such as specific values that a variable can take on, rather than a likelihood or intensity measure that could indicate strength of a response in a specific patch or even set of patches.

The method of Haxby et al. (2001) measured the similarity of a pattern of response to a template, defined individually for each subject, using a correlation coefficient as the index of similarity. Briefly, the data are divided into statistically independent halves, the patterns of response in each half of the data to each category are calculated, and correlations between these patterns are used as indices of the replicability of the pattern of response to each category (within-category correlations) and the confusability of patterns of response to different categories (between-category correlations). This correlation method is a test of whether a replicable pattern of response in one experimental condition exists that is significantly different from the pattern of response in another experimental condition. To test whether the information carried by a pattern of response resided only in the cortex that responded maximally to one category, the patterns of response to two categories were compared with the cortex that responded maximally to either category excluded from the analysis.

Previous methods of topographic pattern analysis, however, have not provided an unbiased test of whether the patterns of response are most consistent with a distributed or a localist code for the representation of faces and objects. We decided, therefore, to reanalyze data from the experiment of Haxby et al. (2001) with a neural net (NN) classifier that could detect either a localized or a distributed code with no initial bias toward either. Neural networks are nonlinear response functions that consist of "nodes", which possess both an activation function and an input function. An input function defines the integration of inputs to the node, typically this function is a weighted average (dot product) over the input values (in this case voxel values). An activation function or output function defines the transformation of net input through the integration function to "rate of firing" function. Often, such a function is sigmoidal in nature, such as a logistic function, such low net input is transformed to low response rates and high net input is transformed to high response rates. These outputs, which typically vary between zero and one, can also be used to indicate the "likelihood" of a given input vector. Feed-forward neural networks often have layers of nodes with intermediate nodes that are known to make them universal approximators (Hanson and Burr, 1990; Hornik et al., 1989). Because of their broad approximation powers, NNs have the ability to detect locally contiguous inputs, "patches", that are consistent across training examples or widely dispersed inputs that may have no obvious spatially contiguity.

In addition to providing an unbiased comparison of distributed versus localist models for category-related patterns of response, NN classifiers also offer a more general method in detecting topographic patterns than the correlation method. Because the method of Haxby et al. used correlation as the measure of pattern similarity, the weight given to a single voxel is based on the deviation of the response in that voxel from the mean response across voxels rather than on the discriminating power of that voxel. By contrast, NN classifiers adjust the weight assigned to each voxel to maximize discriminatory power. Therefore, NN classifiers have the potential to detect the more exact form of the topographic pattern.

NN classifiers also address another shortcoming of the correlation method, namely the uncertainty about the precise extent of response pattern overlap. Haxby et al. showed that the pattern of response to an object category was highly specific to that category even when the analysis was restricted to cortex that responded maximally to other object categories. Also apparent from the correlation analysis were extensive negative correlations between categories, suggesting a potential network of associations between object categories that were primarily inverse relationships in activation, ones that could form an associative basis. These results suggested that information about multiple categories is distributed in overlapping representations, but it is not an exhaustive test of whether each voxel contributes information to the representation of all categories. It is possible that no maximal responses in a piece of cortex only carry information about one or two categories in addition to the category that elicits the maximal response. Such a representational scheme, therefore, would be localized to scattered, small cortical patches that have some degree of category-specificity. With NN classifiers, we can apply a sensitivity analysis to determine whether each individual voxel contributed to the classifier for each category and, thus, make an exact quantitative estimate of the extent of response pattern overlap. This kind of analysis adds noise to the input voxel after training the NN to optimal generalization performance. As noise increases for each specific voxel input, the classification error of the trained NN is monitored for significant increases in error given small perturbations of noise indicating that that voxel is contributing to the overall classification performance. In this way, each voxel can be "queried" as to its contribution to the specific object identity.

In the present research, we therefore ask two basic questions: whether we can show improvement in out-of-sample generalization and further can we identify the object code in temporal lobe more precisely? Specifically, the kinds of codes that we investigate in this paper are a special case of more general topographic codes; ones in which differential intensities in some fixed spatial patterns code for objects; similar to a piano where the same set of keys are played but with different amplitude modulation; thus producing unique output with the same keys. From a computational point of view, this might be the simplest type of code to implement that is efficient, high capacity, and rapidly extensible. In the next sections, we examine this specific coding hypothesis and provide results for the Neural Network Classifiers.

Methods

Data acquisition

The data consisted of 64 slices 64 × 40 BOLD collected from a GE 3T (repetition time = 2500 ms, forty 3.5-mm-thick sagittal images, field of view = 24 cm, echo time = 30 ms, flip angle = 90°). We used 7–10 slices from this set and used Haxby's feature masks that he had used for his correlations. Haxby had done feature selection using thresholded high variance voxels that created slice masks for 7–10 slices with 5–150 voxels per slice (500–600 voxels per volume) depending on the subject.

Experimental procedures (Haxby's original procedure from Science 2001)

Patterns of neural response were measured with functional magnetic resonance imaging (fMRI) in six subjects while they viewed pictures of faces, cats, five categories of manmade objects (houses, chairs, scissors, shoes, and bottles), and control, non-sense images. Stimuli were gray-scale images of faces, house, cats, bottles, scissors, shoes, chair, and nonsense random patterns. The categories were chosen so that all stimuli from a given category would have the same base level name. Control nonsense patterns were phase-scramble images of the intact objects. Twelve time series were obtained in each subject. Each time series was begun and ended with 12-s rests and contained eight stimulus blocks of 24-s duration, one for each category, separated by 12-s interval of rest. Stimuli were presented for 500 ms with an interstimulus interval of 1500 ms. Repetitions of meaningful stimuli were pictures of the same face or object photographed from different angles. Stimuli for each meaningful category were four images each of 12 different exemplars. Volumes of interest (VOIs) were drawn on high-resolution structural images to identify ventral temporal (VT), lateral temporal, and ventrolateral occipital cortex. The VOI for ventral temporal cortex extended from 70 to 20 mm posterior to the anterior commissure in Talairach brain atlas coordinates and consisted of the lingual, parahippocampal, fusiform, and inferior temporal gyri. The VOI for lateral temporal cortex also extended from 70 to 20 mm posterior to the anterior gyrus and both banks of the superior temporal sulcus. The VOI for ventrolateral occipital cortex extended from the occipital pole to 70 mm posterior to the anterior commissure and consisted of the lingual, fusiform, inferior occipital, and middle occipital gyri. Voxels within these VOIs that were significantly object-selective were used for the analysis. To identify the object-selective voxels, Haxby et al. (2001) used an eight-regressor model. The first regressor was the contrast between stimulus blocks and rest. The remaining seven regressors modeled the response to each meaningful category.

Bootstrap

Out-of-sample generalization refers to a test that uses mutually exclusive data sets that provide an unbiased test of a classifier. Split half tests are often used for out-of-sample tests; in this case, the out-of-sample test is based on a held out half of the whole sample while the classifier is trained on one of the halves and the second (“unseen”) half is used for testing. Split half out-of-sample tests are inefficient and we will use throughout these analysis an $N - 1$ bootstrap test, which efficiently uses all the data by training on $N - 1$ part of the data and testing on the single left out case and then replace that single sample and remove another independent sample for testing.

Neural network settings (softmax and cross entropy)

For each subject, we created a 10-hidden node, eight-way multi-class NN classifier. We used the hyperbolic tangent activation transfer function for its hidden nodes, that is,

$$h_j = \frac{\exp(a_j^H) - \exp(-a_j^H)}{\exp(a_j^H) + \exp(-a_j^H)}$$

where $a_j^H = \sum_i x_i w_{ij}$.

Here, x_i was the activity of voxel i. The activations of the hidden nodes were then fed forwarded to the mutually exclusive output nodes, where we used the softmax function (also known as a smooth version of winner-take-all activation function) for obtaining their activations:

$$O_k = \frac{\exp(a_k^O)}{\sum_m \exp(a_m^O)}$$

where $a_k^O = \sum_j h_j w_{jk}$.

This softmax function normalizes outputs (i.e., output lies between 0 and 1 and sum up to unity). The error function for our NN classifier was the cross entropy function or:

$$E = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_k^n \ln\left(\frac{O_k^n}{t_k^n}\right)$$

Scaled conjugate gradient

The scaled conjugate gradient (SCG) method is a variant of a conjugate gradient method that uses Levenberg–Marquardt approach for finding appropriate step size (Moller, 1993). Instead of using computation-intensive line search procedure, SCG uses approximated Hessian matrix (multiplied by the direction vector) to scale the step size α_j. To find the appropriate step size, only Hessian matrix multiplied by a conjugate direction vector $\boldsymbol{d}$ is need. This Hessian matrix product can be approximately computed rather efficiently for the multilayer perceptrons by using central differences (Bishop, 1995). However, to maintain definiteness of the Hessian, a scalar λ is included in the computation:

$$s_j = \frac{f'\,(w_j + \varepsilon_j \boldsymbol{d}_j) - f'\,(w_j)}{\varepsilon_j} + \lambda_J \boldsymbol{d}_j$$

where ε is a small number.

The step size for SCP is then obtained by

$$\frac{a_j = -\boldsymbol{d}_j^T f'\,(w_j)}{\boldsymbol{d}_j^T \times s_j + \lambda_j \|\boldsymbol{d}_j\|^2}.$$

If the comparison parameter given by

$$\Delta_j = \frac{2\{E(w_j) - E(w_j + a_j \boldsymbol{d}_j)\}}{a_j \boldsymbol{d}_j^T \boldsymbol{g}_j}$$

results in bigger than 0, then new conjugate direction and weight are obtained, namely,

$$w_{j+1} = w_j + a_j \boldsymbol{d}_j$$

$$d_{j+1} = \boldsymbol{g}_{j+1} + \beta_j \boldsymbol{d}_j$$

where $\beta_j = \frac{|\boldsymbol{g}_{j+1}| - \boldsymbol{g}_{j+1}\boldsymbol{g}_j}{\boldsymbol{d}_j^T \boldsymbol{g}_j}$.

Here, $\boldsymbol{g}$ is a gradient vector.

Sensitivity analysis

Each input to the NN represented a particular voxel. Each input line was perturbed with random noise by adding a sufficient Gaussian source to each voxel until it reached threshold at the hidden layer, thus producing a response at the output category layer. Due to the normal operation of the Neural Network that typically had 300–600 active weighted inputs, the variance of the Gaussian noise source had to be increased to a scalar value (order 100) so that a single input could produce a significant response in the NN. This value was fixed for all voxels and specific to each subject. Output errors (responding "X"| given input was "Y") that exceeded 30% overall error were considered significant given they were near the median of the sensitivity distributions (see Fig. 1).

Results

Voxel or feature properties

As described above in the Methods section, Haxby defined voxel masks in the ventral temporal area that resulted in approximately 300–600 voxels (features) depending on each subject. We converted the voxel intensities in these sets to *z* scores (demeaned and normalized to standard deviation in each time series) and examined their distributions. As shown in Fig. 2, we have a typical subject's frequency distribution over the voxel set. For all six subjects, we found no evidence of significant modes or obvious mixtures in the underlying distribution, and despite a strong skew in all distributions, they appeared single peaked and smooth. Similarly, a principle components analysis (PCA) of the voxel set showed strong first order influence (30%) with a long and shallow tail indicating that the last nine components of the 10 uniformly extracted a majority of the variance in voxel space (93%). This type of PCA property can often suggest that there is nonlinear structure available to exploit.

Linear classifiers

The voxel features were submitted to various linear classifier methods and complete cross-validation was done for $N - 1$ (Jackknife) and $N - 2$ to assess value and stability of the out of sample generalization (performed over blocks; see Methods or cross-validation section). Haxby had defined a type of "prototype" classifier based on average spatial pattern derived from independent split half samples ($N - 6$). The highest correlation over all average patterns for "face", "house", and so forth, was used to classify the other split half sample to one of the eight categories. We replicated Haxby's correlation method using raw voxels (as opposed to Beta weights in order to be comparable to the classifiers discussed next; but still selected voxels using his original Beta weight masks) and used cross-validation to determine stable out of sample performance. The Haxby correlation method achieved 66.8% correctly classification in $N - 1$ (and slightly less for $N - 2$). Because the number of variables in this problem greatly exceeds the number of samples, the linear discriminate analysis is ill-conditioned and fails to produce a

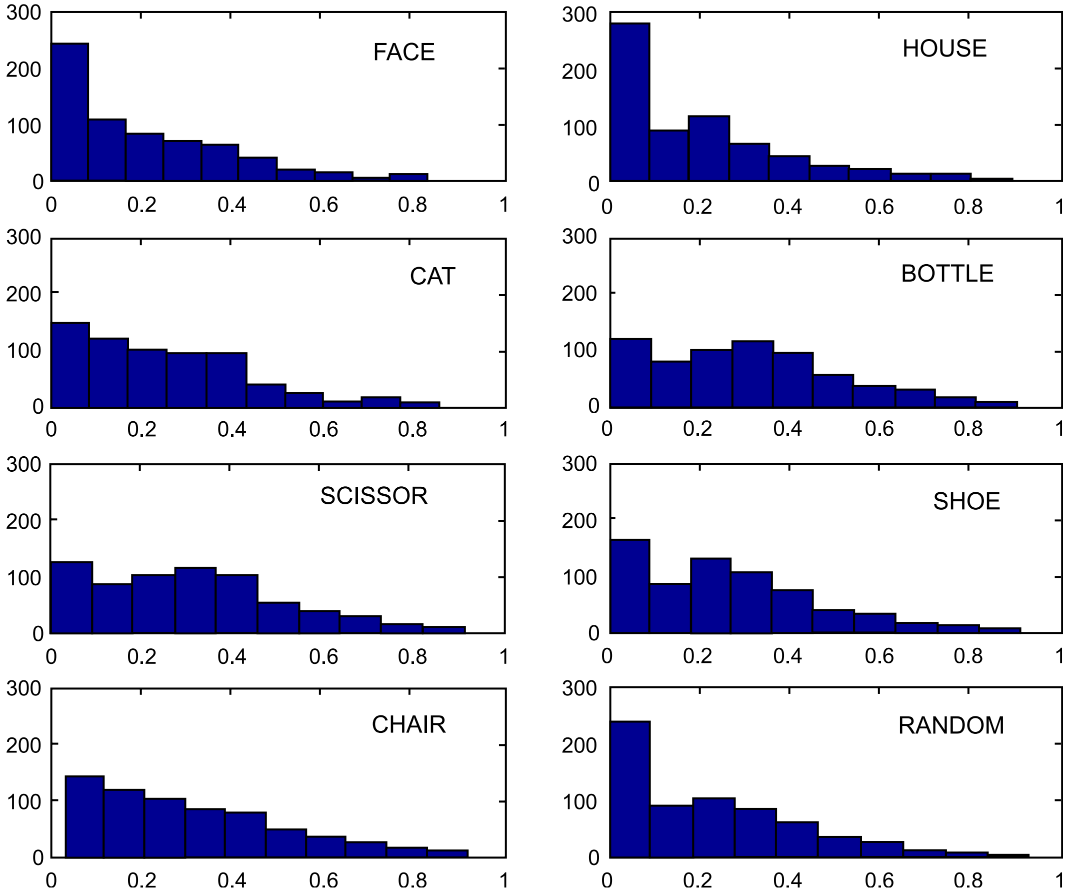

Fig. 1. Sensitivity distributions for Subject 4 across all categories, the median for all distributions was 30% error.

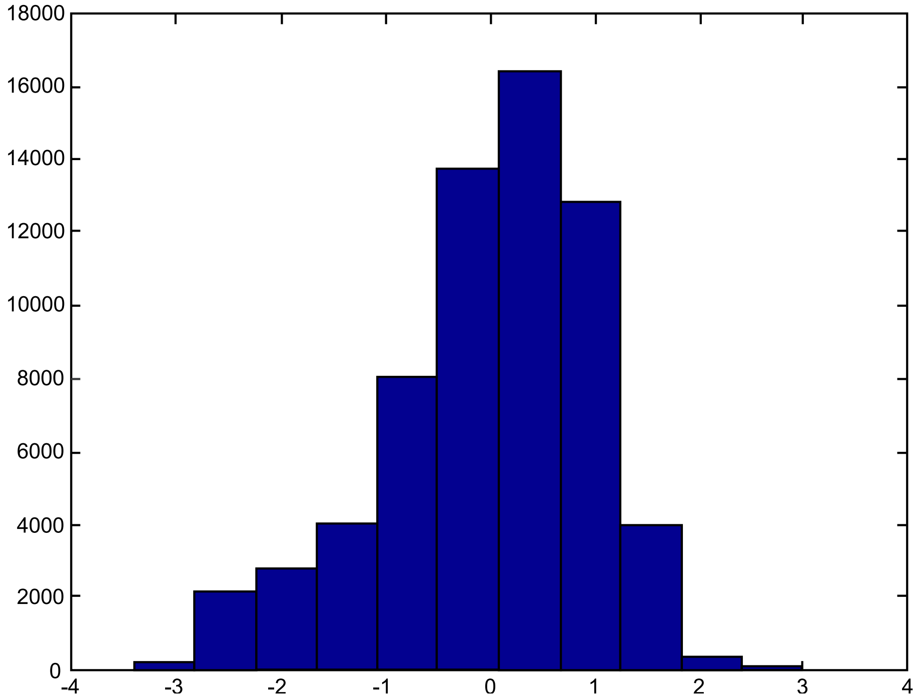

Fig. 2. Distributions of voxel intensities in the "Haxby Masks" from three typical subjects.

result. However, if one first does singular value decomposition (SVD) on the original variables and uses these derived variables to compute within group variance, the LDA of that new feature space achieves as high as 78% in $N - 1$ bootstrap. This type of data compression before classification is similar in operation to nonlinear classifiers, such as neural networks (NN) that attempt to optimize lower dimensional projections of the original variables as they learn to classify.

Neural networks

We used simple feed-forward neural networks that are known to have general classification capabilities. We considered various architectures (modular, multiple layers, etc.), but in preliminary tests with a smaller data set, were able to find good generalization with a single layer network but not significant improvements in out-of-sample generalization with more complex architectures or variations in weight estimation procedures. Weight parameters in all networks were found using scaled conjugate gradient search (Moller, 1993), also known to be efficient in relatively large search spaces. $N - 1$ bootstrap (in this case we used separate blocks) was used to evaluate the classifier, which allowed for 88 (i.e., 11 exemplars × 8 categories) in-sample training and eight (i.e., 1 exemplar × 8 categories) out-of-sample opportunities.

To minimize voxel pattern overlap due to the extended time scale of the hemodynamic response, we used whole blocks as exemplars for the out-of-transfer transfer point (seven scans). We also created a "REST" category and in preliminary analysis trained the networks with REST voxels (from the same Haxby voxel mask) to provide the classifier a background baseline for contrast against the category voxel patterns. In subsequent tests, in fact, REST was not required for significant transfer results and hence was not included in the final analysis (in this case, we only used the original Haxby mask voxels). Also somewhat surprising was the critical nature of the choice of the output function with the error metric. Shown in Table 1 are some of the various output functions and parallel error metrics we used with and without the REST condition. In fact, only one of these many conditions showed significant transfer for ALL subjects (we could achieve reasonable transfer for Subject 2 for most output functions or error metrics, but this did not generalize across all subjects). Weighting relative category errors using softmax and measuring the similarity of the distribution of errors as in a cross-entropy measure provided significantly stronger generalization than any other case (with scaled conjugate gradient (SCG), see supplemental material for more detail on this error metrics and learning functions). Although we will discuss this in a later section, apparently the voxel contributions required weighting against a background of other,

Table 1
Nonexclusive, partial list of the neural network classifier configurations implemented and tested in the present study

Error metric	Output function	Gradient estimate	Input transformation	Background
SSE, MSE	Logistic	BP	−1,1	Rest
ABSOLUTE	Logistic	BP w/Momentum	Min−max	Rest
SSE, MSE	Linear	SCG	Z-norm, scan-wise	No rest
Cross-entropy	SOFTMAX	SCG	Means	No rest
Cross-entropy	SOFTMAX	SCG	Z-norm. scan-wise	No rest

Abbreviations: SSE, sum of squared error; MSE, mean squared error; BP, back propagation; SCG, scaled conjugate gradient.

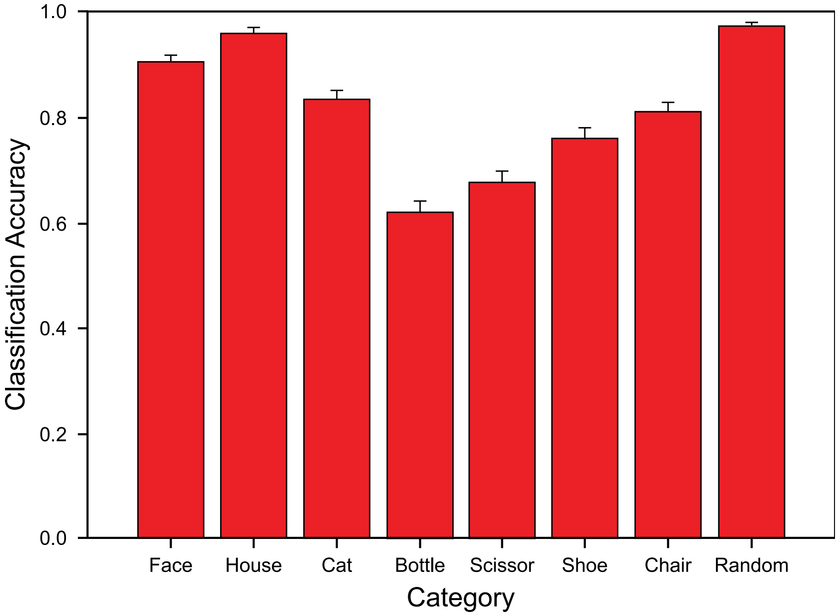

Fig. 3. Classification and $N-1$ bootstrap generalization rates for all categories averaged for all subjects. Note that overall category out-of-sample generalization is 82.5%.

potentially very subtle changes in voxels associated with other category judgments. Shown in Fig. 3 are the mean bar graphs for transfer averaged for all subjects in each of the eight categories. Overall, we are accounting for a mean of 99.5% in training and mean of 82.5% in transfer. The range of transfer is from a mean value of 92% for face" and "house" tokens to 63% for "scissor" tokens with other tokens falling in between these cases. Model selection results shown in Fig. 4 tested transfer at seven different hidden unit values finding the best case to be 10 hidden units, similar to the what the PCA indicated previously about the structure of the voxel intensity. Consequently, all results reported were done with neural network classifiers of 10 hidden units. We should note that this does not imply the network was using principle components for projections. In fact, as it will be shown

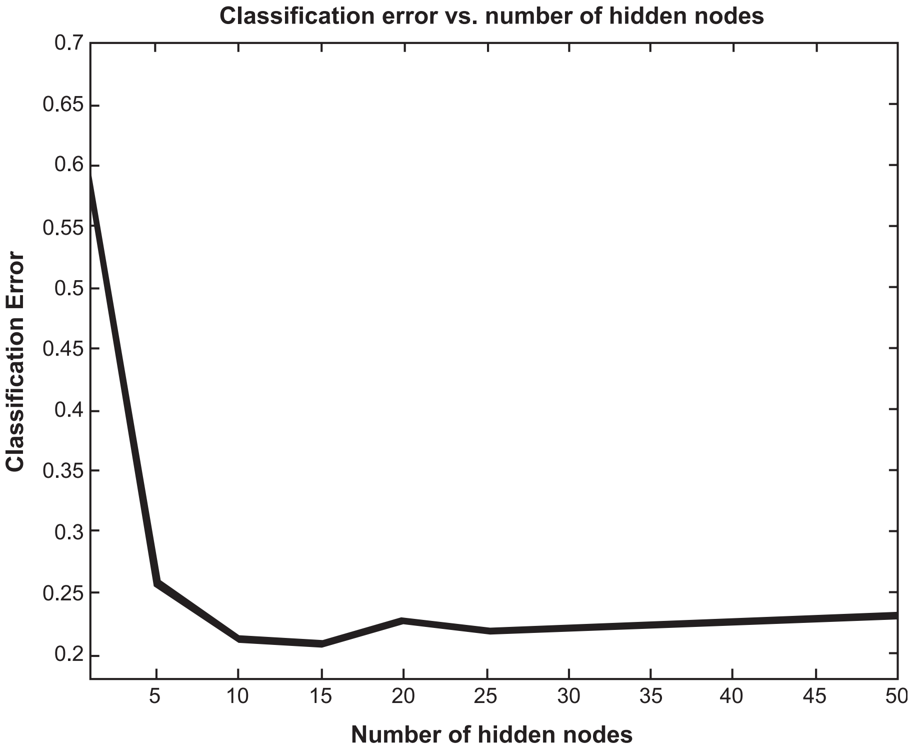

Fig. 4. Model selection results indicating that between 9 and 15 hidden units are best for classification.

below that combinations of hidden unit patterns were critical for identification (for example, see also Japkowicz et al., 2000).

Hidden unit analysis

The trained neural network's hidden unit states can be analyzed to indicate some aspects of the underlying representation that support the classifier (Hanson and Burr, 1990). Submitting the hidden unit activities over all exemplar scans to an agglomerative clustering analysis shows the distances between exemplars and categories as represented in hidden unit space. In Fig. 5 below, we show the result of such a cluster analysis on 10 hidden units from the trained networks. From the dendrogram, it is clear that the network has produced a 10-dimensional embedding of the original exemplars as distinct categories. From left to right, the hidden space shows ''faces'', ''cats'', ''houses'', and so forth, and apparently makes at the next level of the dendrogram a distinction between the group ''faces'' and ''cats'' and all other categories. This type of ''animate or inanimate'' distinction is evidence—for the first time—that fMRI signals could encode an implicit semantic distinction based only on learning categories from specific exemplars sampled from those categories. As we see below, this type of distinction is apparently part of a larger code that indexes a specific exemplar, while at the same time coding for the entire category.

Sensitivity analysis

Although it is possible that all feature or voxels are used in the classifier for it to achieve its transfer results, in practice this outcome is unlikely. All features do not have equal weight in the analysis and as outlined earlier, several outcomes are possible. First, as hypothesized by many in the field, there could be a relatively local code for these object types that are segregated by category type (e.g., ''faces'', ''places'', ''body parts'', etc.). Second, there could be as Haxby appeared to show, a distributed code that was relatively unique to each category type, nonetheless completely nonlocal in its coding properties. This result contradicts the previously discussed research apparently showing there are specific areas of the brain with specific extent and volume, in effect focal and volume limited, that uniquely code for specific object types or categories. The third possibility demonstrated here is that the codes are combinatorial in the sense that the same voxels or features are reused in an efficient way for object type category codes. One way to test this hypothesis is by performing the following sensitivity analysis of the trained classifiers. To determine the contribution of each voxel to the overall classification and generalization results, Gaussian noise of sufficient width (in this case to scale the weight from a single voxel with a background input of 500–600 other voxels; see Methods or Sensitivity analysis) is added to each voxel, one at a time, and the generalization error is again recalculated for the new classifier. Noise is sampled and added hundreds of times to get a stable estimate of the error contribution. If the error increases significantly, this indicates the voxel is showing a contribution to the classification performance. If, on the other hand, increases in noise to that voxel provide little or no significant change in classification error, then we will index it as having little contribution in the classification performance. In this way, we effectively assay the voxel's classification contribution by ''lesioning'' it with the perturbing noise source. If we threshold the voxels sensitivity at 30% change in the classification error and plot those voxels and those voxels showing a change

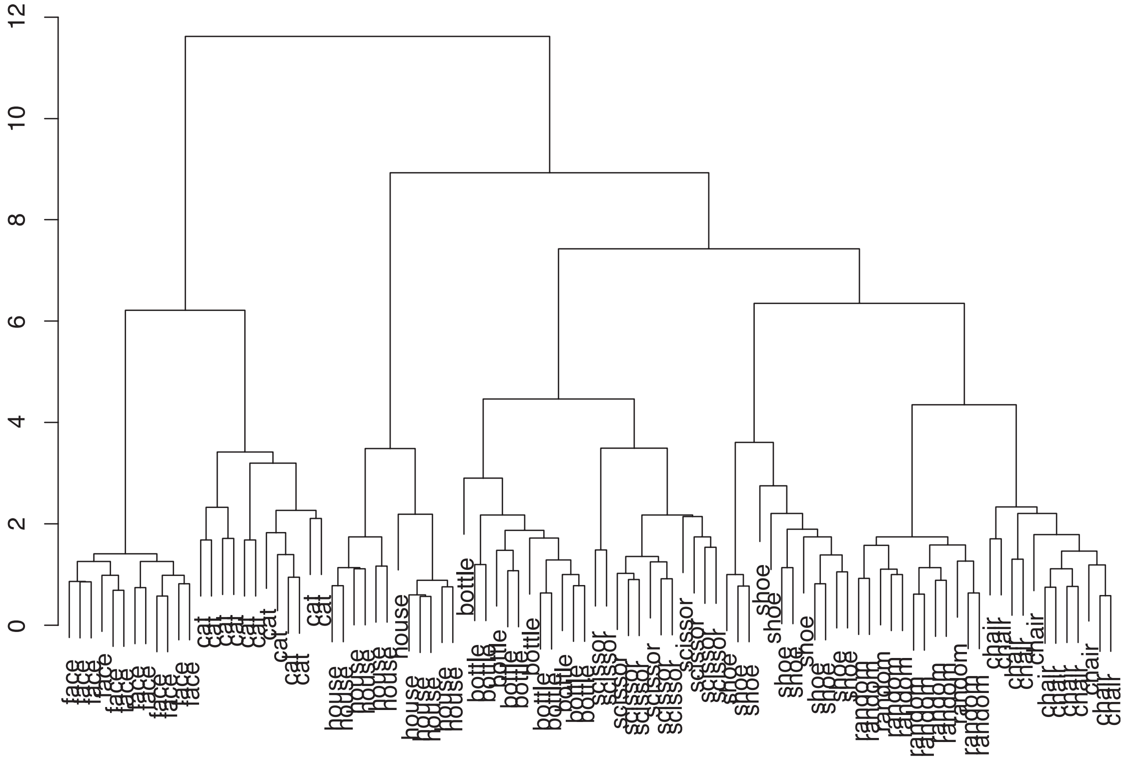

Fig. 5. Cluster dendrogram showing the responses of the hidden units to all exemplar scans. Note that each category set is represented in hidden unit space and that there appears to be an ''animate or inanimate'' distinction learned from examples of object exemplar scans by the neural network.

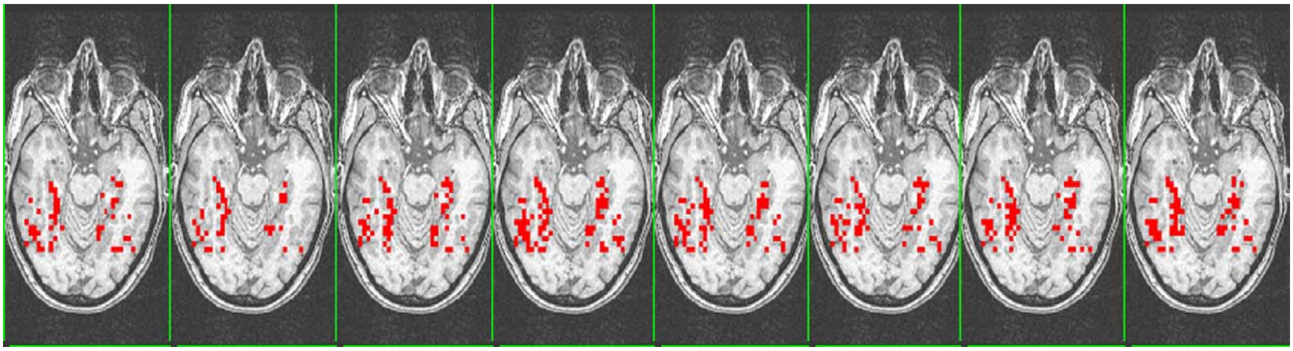

Fig. 6. Sensitivity analysis in one slice containing ventral medial temporal lobe. Red voxel patterns show the voxels that had more than a 30% increase in classification error due to noise perturbation at each voxel. Note that each subject has a slightly different pattern of sensitivity response (not shown here), although this may be due to difference in anatomy or registration and not necessarily due to the specific visual voxel pattern shown.

above 30% in sensitivity error (although the maximum rarely exceeded 60% error increase and typically was around 40%), we see in the following Fig. 6 a typical subject's sensitivity values plotted in a mid-level slice for all eight object categories. There are two observations to note (not all shown): (1) intersubject variability in terms of object coding is high, notwithstanding that the VMT mask is relatively small, (2) the voxels that are most sensitive across all object types within each subject are practically identical. The overlap for all subjects (calculated within subjects) and between category type voxels (500–600) is 88.4%.

In Table 2 below, we show the pairwise sensitivity overlap of the significant voxels of one category with all other categories. Note that "random" category has the lowest overlap with other categories while "face", "house", and "cat" have the highest with all other categories. These results suggest that there is very little local response whatsoever for a category input.

Do any voxels show sensitivity to specific objects (house, face)?

Finally, to provide the strongest possible test of the localization of object identity in temporal lobe, highly selective masks were identified for the "fusiform face area" and the "parahippocampal place area" that also did not overlap in voxel space. These voxels were then probed for their responses to "face" and "house" based on the sensitivities previously computed. Distributions of sensitivities for all voxels and subjects were calculated for the four possible cases of voxel mask and object sensitivity. Specifically, we show in Fig. 7 the FACE sensitivity response given HOUSE voxels, the HOUSE sensitivity response given FACE voxels, the HOUSE sensitivity response given HOUSE voxels, and finally the FACE sensitivity response given FACE voxels. If there were any special voxel selectivity for object type, the distributions for the matched object sensitivity and object voxel type (Sens(X)| Vox(X)) would skew toward the right (the rightmost column in Fig. 7) indicating higher sensitivity. For object response sensitivities for different object voxel types (the leftmost column in Fig. 7), we would also expect the distributions to skew to the left (as they are doing), indicating that there is no special response of FACE by FACE voxels, HOUSE by HOUSE, FACE by HOUSE, or HOUSE by FACE. In the present case, all distributions are skewed to the left and have roughly the same range, and show the same median response to either object by either the FFA or the PPA voxels. In effect, the distributions for each voxel type overlap in their responses, indicating no particular local response to object type.

Table 2
Overlap of voxels by category as determined from the sensitivity analysis

House	Cat	Bottle	Scissor	Shoe	Chair	Random	
0.8885	0.9230	0.8127	0.8376	0.8829	0.8740	0.8969	Face
0.8641	0.7382	0.7717	0.8256	0.8089	0.9277		Random
0.8656	0.8991	0.9530	0.9363	0.9197			Chair
0.9665	0.9126	0.9293	0.8105				Shoe
0.9461	0.9602	0.8440					Scissor
0.9664	0.8979						Bottle
0.8812							Cat

Discussion

We have reanalyzed the Haxby et al. (2001) object recognition data using feed-forward neural networks and showed significant out-of-sample generalization performance (82.5%) on scans between blocks of stimulus trials. Networks performing a potential compression of 50:1 of voxels to hidden units were able to correctly classify and recognize all (672) tokens based only on individual scans, indicating that voxel variation alone can be use to code for objects that human subjects are visually observing. Most interestingly, the sensitivity analysis of voxels showed very high overlap of the same voxels being recruited across all object categories and exemplars. As discussed previously, such codes are often considered combinatorial, since they take advantage of the possible combinations of values that could arise from the same variables. These types of codes are not uncommon in biological coding. For example, in the context of odor coding, Malnic et al. (1999) show that unique combinations of the same odorant receptors can code for different odorants. In the present case, the combinatorial codes are expressed at the voxel level in terms of millions of cells whereas previous cases are measuring tens or dozens of cells at most. In general, combinatorial codes are one of the most efficient that could exist for coding a large set of responses in redundant and lossless way. Suppose there are about 100 voxels in VMT coding for object category. If, for example, each voxel has only a fidelity of just three different values, the number of types that could be stored and recognized with such a scheme is 3^100 or equivalently 10^50 (trillions and trillions; in effect an unlimited numbers of potential object exemplar or categories as opposed to the hard bound that

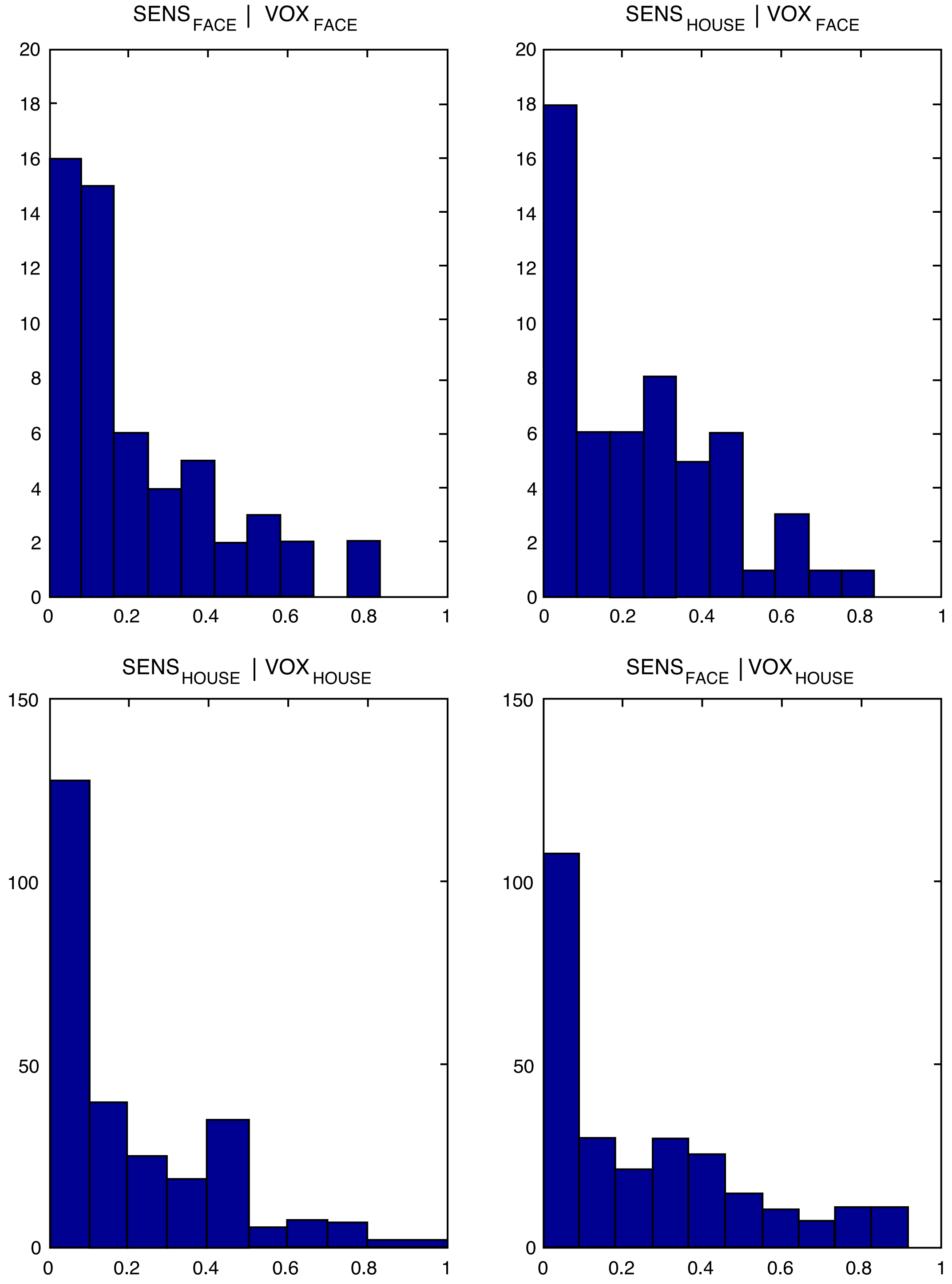

Fig. 7. Distributions of sensitivities of FFA or PPA voxels to either face or house. Note the lack of right skew in any distribution, indicating no special response of voxel to object type.

seems to be implied by the prevailing theories of assigning cortical space to object codes)!

How are FFA and PPA insensitivities for face and house consistent with past research?

In Fig. 6 and especially Fig. 7, we show that there is a lack of object specificity in the PPA (so-called "place" area) and the FFA (the solaced "face" area). There has been considerable research in both visual neuroscience and neuroimaging showing specificity or responses in these specific areas of ventral temporal lobe. How can we reconcile recent work that seems to provide evidence for differential responses of the FFA and PPA for faces and places, respectively (e.g., Spiridon and Kanwisher, 2002)? In the present case, the inconsistency in the neuroimaging literature often reflects a confusion between identification and similarity. For example, measures of “topographic overlap” used in the Haxby (2001) and as well in the Spiridon and Kanwisher (2002) study actually measure the similarity or distance between the prototype and target patterns. The procedure of ranking the similarity measure to pick the closest prototype as the category label ignores similarity of other object types. The similarity between a set of exemplars and a category prototype is not equivalent to the probability of identification of that exemplar with a given category. Consider the two following correlation patterns that may be similar to Spiridon and Kanwisher (2002) shown in Table 3.

In each case, the FACE prototype pattern is most similar to the target FACE stimulus and according to the Haxby (2001) method would be labeled FACE. Unfortunately, if we take into account the high similarity of the other possible categories, using Luce’s Choice ratio, which constrains the similarity of the target to all possible prototype responses:

$$P(C_{\text{face}}) = \frac{\text{Similarity}(C_{\text{face}})}{\sum \text{Similarity}(C_{ij})}$$

This produces the following estimated probabilities, 0.29 for the first row and 0.71 for the second row. In other words, the first row produces a relative probability much lower than the rank-measure similarity that Spiridon and Kanwisher (2002) used to determine “preference” of voxels (whether they selected FFA or used temporal lobe masks) for object identification. Given that most of the data patterns using these correlation methods look like this first row, it is not surprising that such methods may inflate the probability of identification by ignoring nearby category responses. The neural network in the present work in fact, used SOFTMAX, which is a version of Luce’s Choice ratio. The second confusion involves the difference between identification and sensitivity. As shown in this paper, it may be possible to correctly classify voxel intensities to a given category label; however, this does not also mean that all voxels or features are utilized for the classification. VOI selection can only confirm the category label that the VOI maximally responds to without at the same time indicating what other VOIs not selected or what part of the VOI chosen actually is most responsible for the classification success. Bartels and Zeki (2004) recently argued a similar view concerning functional identification “we should emphasize, however that the results using neuroimaging can never be used to show the noninvolvement of areas, and that positive results are only of a correlational nature.” But in fact, using sensitivity measures through model-based classifiers as shown here can in fact provide causal evidence for area involvement and using more comprehensive masks can indeed show the noninvolvement of areas.

Categorization, distributed representations, and the basic level

In many past neuroimaging studies, object, category, and exemplar are often used interchangeably. However, it is important to make a distinction between category (object type) and exemplar (object token) because responses in the temporal lobe are not directly dependent on the “face” or “chair” category but on exemplars drawn from these categories. More critically, it may be important to distinguish between different levels of reference in a category hierarchy (Rosch and Mervis, 1975). “Face” for example is part of the human head that is part of the human body that in turn is a type organic form, and so on. Given the usage in this literature, it is often hard tell which level of reference is being invoked. Consider “face” as a category, despite the inability for prosopagnosics in recognizing a particular face exemplar (“is this George Bush?”), they also do not have problems discriminating a face from a chair. In this case, “face” may be at a more subordinate level (in terms of a category hierarchy) than “chair”, in that subjects have more “expertise” with the face (cf. Gauthier et al., 2000) category as opposed to chair. Hence a particular face exemplar may be more similar to the knowledge that a subject has with the chair in their living room they often sit in, than a particular kind of chair.

In any case, the kind of encoding strategy that the present result suggests is available in temporal lobe is consistent with this type of hierarchical representation. Assuming a configural feature encoding of categories and exemplars from categories, a combinatorial code can allow a mixture of general and specific features. If a stimulus varies in its position in a category knowledge hierarchy, we would predict that the density of the code would covary with the level in the hierarchy, less dense distributed activity reflecting more general category reference. In fact, it is hard to see how the structure of category knowledge could be represented in localist accounts of categories like “face” and “house”. Given the inability of these kinds of accounts to deal with fundamental properties of category structure, it calls into question the proposal of a specific area in the brain that could be considered a “face” area, for example.

Distributed representation accounts of category knowledge are consistent with the observations of temporal lobe representation in this paper. Connectionist models, for example (Rumelhart et al., 1986), provide a theoretical account of present results. Networks of neurons or areas could be activated to signal the presence or absence of a particular category or exemplar from a category. What is particularly intriguing is the possibility that in our analysis we are observing knowledge representation at the voxel level in a distributed computational network. Experiments that focus on the variation of activity in the temporal lobe as a function of category structure and level of reference should help determine the aspects of this type of representational network.

Although at first glance, it may make sense to try to assign specific functions to specific structures in the brain, it may be that the brain is not actually organized in such a way to cooperate with this type of analysis. The history of object recognition and category learning focuses on lesions, single cell measurements, and more recently, neuroimaging. Only in the last few years have researchers asked critical questions about coding at this high-level system description, while most have continued in the agenda of recognizing one-to-one structure and function. Our classifier was completely agnostic to the coding and was only required to produce a good classification based on some voxel, voxels, or even as it did in this case, voxel pattern. Although such codes are difficult to discover, it makes it particularly hard when the prevailing methods, as they are in neuroimaging, actually bias the researcher against finding distributed patterns. The signal

Table 3
Hypothetical correlation patterns resembling Spiridon and Kanwisher (2002)

	Face	House	Cat	Shoe
Target pattern	0.95	0.82	0.75	0.65
Target pattern	0.20	0.01	0.02	0.05

detection paradigm, as it is normatively practiced in neuroimaging, tends to lead to finding specific voxels at specific locations, although there had been work on using the general linear model (GLM) for detecting distributed patterns (Friston et al., 1994), and in using multivariate pattern recognition methods (Cox and Savoy, 2003), there has been little mainstream interest, use, or acceptance of these types of analysis. On a more hopeful note, however, there has been recent proposals about interpreting fMRI data as neural distributed function (Shaw et al., 2003). McIntosh (2000), in particular, discusses the response of brain areas dependent on a type of "neural context" and has been a proponent for multivariate spatial analysis for sometime.

Several new directions arise from the current results and analysis. First, what is the nature of the voxel combinatorial features themselves: what is the nature of the "code-book"? One possible way to answer this question is to train networks to classify voxels to arbitrary feature elements (that exhaustively cover the original stimulus space), and then to query the voxels using sensitivity analysis to see what codes are most productive in classification error with which specific voxel regions. A second kind of analysis asks the question of how nonregionally distributed are the codes? Clearly, these kinds of inquiries and the present analysis tend toward a picture of brain imaging analysis as better served with distributed and combinatorial tools (like Neural Networks), rather than the dominant analysis that tacitly assumes brain function can be discovered through homogenous, unimodal signal detection methods.

Acknowledgments

This research was supported by a McDonnell Foundation Grant to S. Hanson and NSF ITR Grant EIA-0205178. We wish to thank Maggie Shiffrar and Catherine Hanson for providing feedback on earlier versions of this paper.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. Neuron 21, 373–383.

Bartels, A., Zeki, S., 2004. Functional brain mapping during free viewing of natural scenes. Hum. Brain Mapp. 21, 75–85.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, New York.

Carlson, T.A., Schrater, P., He, S., 2003. Pattern of activity in the categorical representations of objects. J. Cogn. Neurosci. 15, 704–717.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270.

Downing, P., Jiang, Y., Shuman, M., Kanwisher, N., 2001. A cortical area selective for visual processing of the human body. Science 293, 2470–2473.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598–601.

Fodor, J., 1983. The Modularity of Mind, MIT Press, Cambridge.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 214–220.

Gauthier, I., Skudlarski, P., Gore, J.C., Anderson, A.W., 2000. Expertise for cars and birds recruits brain areas involved in face recognition. Nat. Neurosci. 3 (2), 191–197.

Hanson, S.J., Burr, D.J., 1990. What connectionist models learn: toward a theory of representation in connectionist networks. Behav. Brain Sci. 13, 471–518.

Hasson, U., Avidan, G., Deouell, L., Bentin, S., Malach, R., 2003. Face-selective activation in a congenital prosopagnosic subject. J. Cogn. Neurosci. 15, 419–431.

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

Haxby, J.V., Analysis of topographically organized patterns of response in FMRI data: distributed representations of objects ventral temporal cortex. In: Kanwisher N. and Duncan J. (Eds.), Functional Neuroimaging of Visual Cognition: Attention and Performance XX. Oxford Univ. Press. In press.

Hornik, K., Stinchocombe, M., White, H., 1989. Multilayer feedforward network are universal approximators. Neural Netw. 2, 359–366.

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., Haxby, J.V., 1999. Distributed Representation of Objects in the Human Ventral Visual Pathway. Proc. Natl. Acad. Sci. USA 96 (16), 9379–9384.

Japkowicz, N., Hanson, S.J., Gluck, M., 2000. Nonlinear autoassociation is not equivalent to PCA. Neural Comput. 12, 531–545.

Kanwisher, K., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Malnic, B., Hirono, J., Sato, T., Buck, L.B., 1999. Combinatorial receptor codes for odors. Cell 96 (5), 713–723.

McCarthy, G., Puce, A., Gore, J.C., Allison, T., 1997. Face specific processing in the human fusiform gyrus. J. Cogn. Neurosci. 9, 605–610.

McIntosh, R., 2000. Towards a network theory of cognition. Neural Netw. 13, 861–870.

Moller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. 6, 525–533.

Rosch, E., Mervis, C.B., 1975. Family resemblances: studies in the internal structure of categories. Cogn. Psychol. 7, 573–605.

Rumelhart, D.E., McClelland, J.L.and the PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volumes 1 and 2. MIT Press, Cambridge, MA.

Shaw, M.E., Strother, S.C., Gavrilescu, M., Podzebenko, K., Waites, A., Watson, J., Anderson, J., Jackson, G., Egan, G., 2003. Evaluating subject specific preprocessing choices in multi-subject BOLD fMRI data sets using data driven performance metrics. NeuroImage 19, 988–1001.

Spiridon, M., Kanwisher, N., 2002. How distributed is visual category information in human occipital–temporal cortex? An fMRI study. Neuron 35, 1157–1165.