Partially Distributed Representations of Objects and Faces in Ventral Temporal Cortex

Alice J. O'Toole[1], Fang Jiang[1], Hervé Abdi[1], and James V. Haxby[2]

Abstract

■ Object and face representations in ventral temporal (VT) cortex were investigated by combining object confusability data from a computational model of object classification with neural response confusability data from a functional neuroimaging experiment. A pattern-based classification algorithm learned to categorize individual brain maps according to the object category being viewed by the subject. An identical algorithm learned to classify an image-based, view-dependent representation of the stimuli. High correlations were found between the confusability of object categories and the confusability of brain activity maps. This occurred even with the inclusion of multiple views of objects, and when the object classification model was tested with high spatial frequency ''line drawings'' of the stimuli. Consistent with a distributed representation of objects in VT cortex, the data indicate that object categories with shared image-based attributes have shared neural structure. ■

INTRODUCTION

Neuroimaging studies of ventral temporal (VT) cortex responses to objects have concluded variously in favor of modular (Spiridon & Kanwisher, 2002; Kanwisher, McDermott, & Chun, 1997), distributed (Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Haxby et al., 2001), and process-based (Gauthier, Skudlarski, Gore, & Anderson, 2000; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999) accounts of visually based object recognition. Evidence for both the modular and distributed hypotheses comes from the application of novel analyses of the patterns of neural activity that result when viewing objects. Although there is general agreement about the locations of maximal responses to certain classes of objects in VT cortex, the analysis of brain activity patterns has led to an active debate on the nature of the neural representations that underlie these responses.

Diverging conclusions about the distributed versus modular nature of object representations in VT cortex have been reached by researchers using reasonable, but nonconvergent, quantifications of modular/distributed codes. Resolving a debate about the nature of object representations in the cortex requires a quantitatively precise description of the pattern parameters that define the different hypotheses. A more precise definition of ''distributed'' versus ''modular'' patterns of activation will give us a standard for determining the degree to which individual patterns vary on this dimension. It will not, however, help us understand why certain areas are more or less modular/distributed or how object category representations are organized. We argue here that this requires a computational analysis of the structure of object categories that can account for the structure of the brain activity patterns. The representational parameters of this analysis can constrain hypotheses about the kinds of representations that may underlie the neural response patterns.

One difficulty in interpreting pattern-based data in the current debate is that ''modular'' and ''distributed'' have been characterized often as qualitatively discrete kinds of representations rather than as the endpoints of a continuous variable—with partially distributed codes lying between these extremes. The variable connecting modular with distributed is ''voxel information content.'' At the modular extreme, each voxel carries information relevant for only one category of objects. At the distributed extreme, all voxels carry information for all categories. In between, voxels vary in the quality of information they carry about different object categories (e.g., a voxel might contribute to classification accuracy for most, some, or no other object categories).

We propose that understanding partially distributed codes in the context of stimulus parameters is the key to linking neural responses with the physical world they represent. This is because distributed and modular patterns of neural responses provide clues for the more interesting question of how we represent and recognize objects neurally. The theoretical viewpoints suggested by modular versus distributed codes provide predictions about the how distributed individual categories of objects should be and about which categories should share voxels.

What does a distributed activity pattern indicate about the representation of objects? According to the object-

[1]The University of Texas, [2]Princeton University

form topography model, the representations of faces and other objects are widely distributed and overlapping because VT cortex contains a topographically organized representation of the attributes that underlie object and face recognition (Haxby et al., 2001). This model predicts, therefore, that voxel information content should be shared or distributed as a function of the shared attributes of objects. Similar object categories would share more voxels than dissimilar categories because the objects in these categories share more attributes. Alternatively, dissimilar categories share few attributes and would appear ''modular,'' because their encodings will involve mutually exclusive sets of voxels. By this account, the modularity of some preferred areas does not, in and of itself, invalidate the object-form topography hypothesis.

In this article, our first goal was to employ a pattern-based classification analysis to generate data on the confusability of the neural response profiles for pairs of object categories. The pattern confusability data provide a measure of shared neural resources among categories. These data set a standard for evaluating an analogously derived computational assessment of object category structure and add to the single category identification accuracy measures and summary contrast data reported in previous works (Spiridon & Kanwisher, 2002; Haxby et al., 2001).

In the process, we replicate previous findings on modular versus distributed representations in VT cortex that have been interpreted as divergent, using data from a single experiment. Previous studies have employed different operational definitions of distributed/modular and have also defined the extent of preferred regions in different ways. To validate our methods, we report the primary comparisons examined in these studies using a consistent set of operational definitions applied to a single data set. The replication component of this study is particularly important because it serves to illustrate the relatively convergent nature of the functional neuroimaging data in the context of the less convergent operational definitions of the theoretical constructs.

Our second goal was to implement a computational analysis of object recognition that operates on the stimuli from the fMRI experiment and to evaluate the relationship between the neural responses and the stimulus structure. The use of a computational model of stimulus structure in conjunction with a computational model of functional neural activity expands the repertoire of tools available to researchers for probing the neural representations that underlie object perception. Previous studies show that it is possible to classify brain scans by visual object category with a high level of accuracy (Carlson et al., 2003; Cox & Savoy, 2003; Spiridon & Kanwisher, 2002; Haxby et al., 2001) but stop short of linking the neural classification to predictions generated by the physical structure of the objects themselves. To the best of our knowledge, the present study is the first to combine an analysis of stimulus structure with an analysis of neural activation patterns from a functional neuroimaging study (although, for combined analysis of the perceptual and neural structure of objects, see Edelman, Grill-Spector, Kushnir, & Malach, 1999). This stimulus analysis brings the previously divergent neuroimaging results under a single framework that accommodates most of the previous findings and offers new insight into the reasons why object category responses vary in the extent to which they are distributed.

RESULTS

In addressing the first goal, as noted previously (Spiridon & Kanwisher, 2002), the existence of partial responses by brain areas that prefer certain object categories to other (nonpreferred) object categories is open to 2 interpretations. Either the nonpreferred area contributes to the representation of the object or the activation reflects an epiphenomenal engagement of the visual system in response to any potentially relevant stimulus (Spiridon & Kanwisher, 2002). To differentiate these alternatives, we used the standard of whether the inclusion of additional voxels to the preferred and nonpreferred areas increased the accuracy with which objects can be classified based on the brain activation patterns. This is a strong and conservative test of voxel contribution, because even if the preferred area contains sufficient information for categorizing an object, a redundant representation in the nonpreferred area may be relevant. Redundant representations may simply encode other aspects of the stimulus that are needed depending on the perceptual demands of the task.

We applied pattern-based classification analyses to determine the discriminability of the brain-map patterns available from a previously published study (Haxby et al., 2001). Participants in this study viewed 8 categories of objects (faces, houses, cats, shoes, chairs, scissors, bottles, and scrambled objects). Pattern-based classification analyses treat brain scans as patterns of interdependent voxels, rather than as collections of independent voxels (see also Carlson et al., 2003; Cox & Savoy, 2003; Petersson, Nichols, Poline, & Holmes, 1999). They also provide a quantitative measure of the separability of brain scan data in the context of the original experimental conditions (Phillips, Moon, Rizvi, & Rauss, 2000; Petersson et al., 1999). Pattern-based classification analyses address the question, ''Given a functional brain map, what is the likelihood that the pattern of activation corresponds to a brain state indicating Condition A versus Condition B?'' This translates into determining the likelihood that a particular brain activity map indicates, for example, that a subject is looking at a face or house. This is a kind of ''brain reading'' (Cox & Savoy, 2003; Thomas, VanHulle, & Vogels, 2001; Edelman et al., 1999), where the discrimination index measures

success in determining the condition of origin for the brain map.

The goal of the analysis was to determine the pairwise ''neural discriminability'' of the object categories using the brain scans collected while a subject viewed different categories of objects. We applied the procedure to the fMRI data from each subject separately (cf. Haxby et al., 2001) and report the discriminability results averaged over the subjects. Odd and even runs of trials served alternately as the training and testing sets to yield 2 measures of performance for each subject on each pair of object categories. For simplicity, we describe the analysis for the face–house discrimination. The other object category pairs were treated analogously.

We proceeded as follows. First, half of brain maps from face condition and half of the maps from the house condition (i.e., the training set maps) were submitted to a PCA. This provided a multidimensional space of the scans defined by orthogonal axes or PCs. These axes are ordered by the amount of variance each explains in the data. This variance includes, but is not limited to, voxel activation changes that are due to changes in the experimental condition. Because PCA was applied to brain scans, individual PCs are themselves interpretable as brain scans that can be projected back onto the anatomy of the subject and viewed. Figure 1 shows a PC from the neuroimaging data projected back onto the anatomy of a subject.

The next step was to determine the ''positions'' of individual brain maps in the PCA space by computing their coordinates on each of the PCs. Coordinates represent the similarity of individual brain scans to the PCs. These coordinates can contain information about object category contrasts. Information about a category contrast might, for example, be seen in the opposition of positive versus negative coordinate values for scans from the two categories. Figure 1 shows an example of this kind of PC-based contrast for the face and house categories. Scans taken while this subject viewed houses tend to have negative coordinates on this PC, whereas scans taken while the subject viewed faces tend to have positive coordinates. To illustrate the activation profile represented by this PC, Figure 1 shows the areas that are relatively more activated for faces (orange) versus the areas that are less activated for faces (blue). The reverse pattern occurs for houses, with more active areas in blue and less active areas in orange.

The third step was to measure formally the discriminability of scans from the two categories. To do this, linear discriminant classifiers were trained to classify scans from the two object categories using the coordinates of the scans in space. Classification accuracy was tested using the scans from the test set. Finally, we created an ''optimal classifier'' by evaluating classification accuracy on individual PCs and combining the most accurate PCs into a single classifier. Discrimination performance was measured with the signal detection measure of d', which penalizes misclassifications that might reflect a model response bias (e.g., the model classifying all scans as houses, which would lead to 100% correct house classifications and 0% correct face classifications). Thus, the measure assesses the separability of the 2 classes of scans by combining classification success on both categories of objects simultaneously.

Brain Scan Discriminability by Category

The moderate to high classification accuracy for all possible pairs of object categories (Table 1) replicates the performance levels reported previously for these data using a correlation-based classification algorithm with individual category identification measures (Haxby et al., 2001). Discrimination of the object category pairs was not homogeneous, $F(27,135) = 6.62$, $p < .0001$. Consistent with previous work (Spiridon & Kanwisher, 2002; Haxby et al., 2001), houses and faces were discriminated most accurately. The range of d' we found for category pair discriminations indicates sufficient variation to make use of the confusability matrix in Table 1 as a standard for evaluating an analogously derived computational assessment of object category structure.

The PC displayed in Figure 1 achieved a high degree of separation between face and house brain maps ($d' = 3.3$). This map contrasts the activation areas for houses and faces, with the contrasted regions approximating the fusiform face area and parahippocampal place area, respectively.

Voxel Sharing by Category

To determine the extent to which voxels share in the coding of object representations, we compared classification accuracy for all available voxels with accuracy using voxels from preferred and nonpreferred areas. For all 8 categories of objects, classification accuracy was better for the all-voxel condition than for either the preferred on nonpreferred areas [mean $d'_{\text{all voxels}} = 2.30$, mean $d'_{\text{preferred}} = 1.95$, mean $d'_{\text{nonpreferred}} = 1.96$, $F(1,10) = 17.1$, $p < .01$, see Table 2]. When we included all 28 pairs of category discriminations (i.e., face vs. house, face vs. cat, etc.), 23 of 28 comparisons yielded their best performance in the all-voxel condition. Thus, voxels from across VT cortex improve classification accuracy over and above that achievable with either the preferred or the nonpreferred regions.

Notwithstanding, discrimination using only the preferred or nonpreferred regions was still highly accurate (mean $d'_{\text{preferred}} = 1.95$, mean $d'_{\text{nonpreferred}} = 1.96$). In collapsing across the 8 categories of objects, there was no significant difference in classification accuracy between the preferred and nonpreferred regions. However, for 3 categories of objects this pattern differed. For categorization of the houses versus other categories, there was a trend toward more accurate classifica-

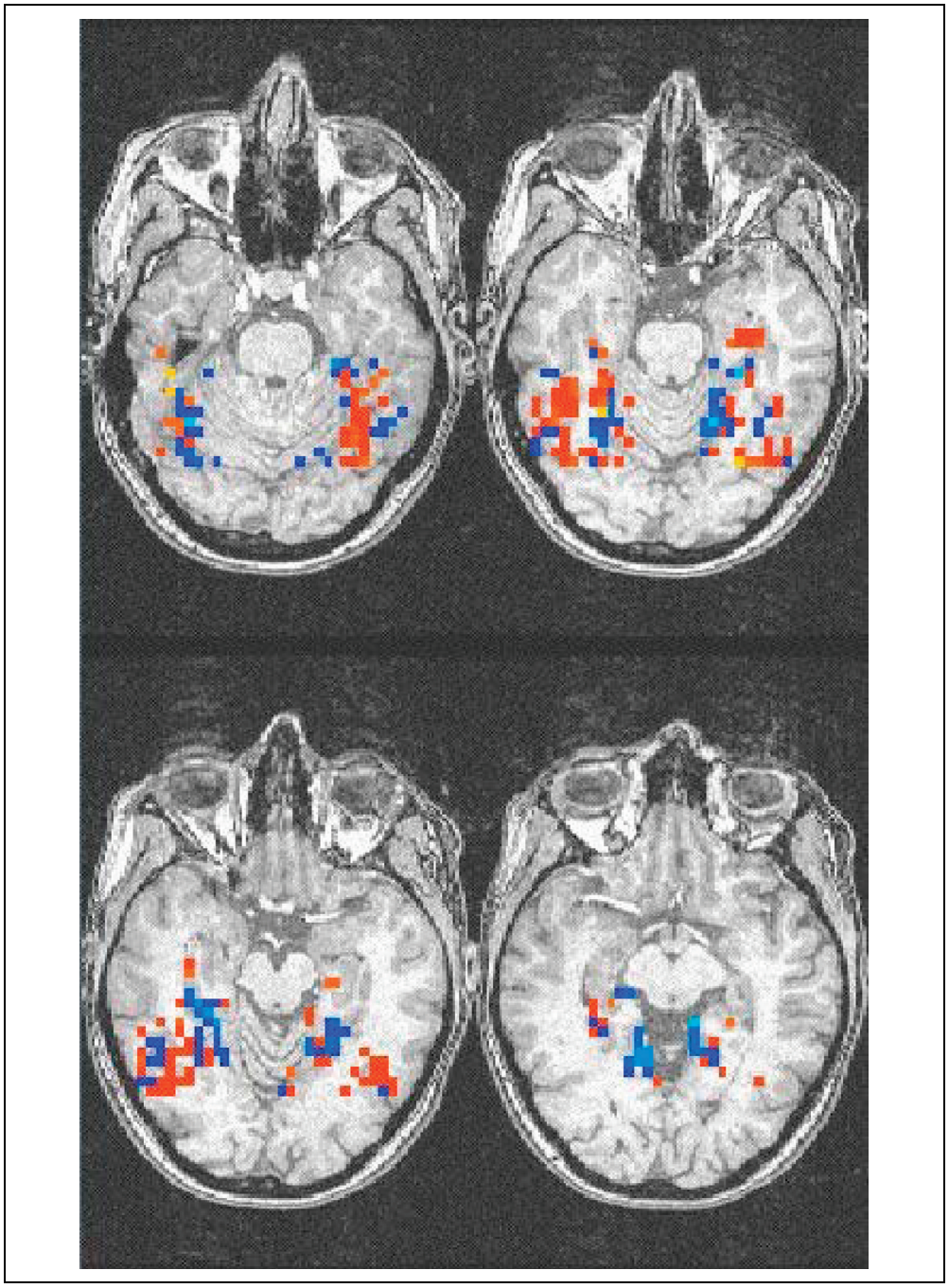

Figure 1. Example of a PC that separates faces and houses (d' = 3.3). Face area in orange and house area in blue. Intensity indicates the weighting of each voxel on this component. Data are shown on 4 contiguous axial slices.

tion with the preferred area (mean $d'_{\text{preferred}}$ = 2.76, mean $d'_{\text{nonpreferred}}$ = 2.03), whereas categorization of the chairs and scrambled stimuli tended to be more accurate with the nonpreferred regions (Table 2).

Finally, using only the preferred region for faces, we measured the discrimination index for all possible pairs of small objects (chair, shoe, scissors, and bottles) (cf., Spiridon & Kanwisher, 2002). Consistent with a modular organization of faces in VT cortex, discrimination for small objects was less accurate (d' = .77) than the face area's ability with comparisons involving faces (d' = 1.63). Using the preferred house area, small object discrimination was likewise less accurate (d' = 1.21) than the house area's ability with comparisons involving houses (d' = 2.30). These data fall between previous findings of chance (Spiridon & Kanwisher, 2002) and good (Haxby et al., 2001) performance for face and house areas with object discrimination. Differences between the definitions of the areas may account for the variability of this finding. Our data suggest that although there is information for object comparisons in the face and house areas, the areas are better suited to discriminations from their respective preferred object classes (Spiridon & Kanwisher, 2002). This replicates, at least

Table 1. Classification Accuracy for Brain Maps and Stimuli

	Face	*House*	*Cat*	*Chair*	*Shoe*	*Scissors*	*Bottle*	*Scramble*
Face		3.47	1.79	3.00	2.67	2.58	2.22	3.08
House	4.52		3.39	2.18	2.86	2.69	2.89	2.62
Cat	4.08	2.85		2.18	2.34	2.09	2.31	2.88
Chair	4.08	4.52	1.61		1.73	1.55	1.23	2.07
Shoe	4.52	4.52	2.92	2.82		1.44	1.29	2.38
Scissors	3.97	4.52	2.81	2.89	3.55		1.19	2.15
Bottle	3.87	4.08	1.96	2.91	3.26	2.09		2.07
Scramble	3.73	4.52	3.17	3.97	4.52	3.26	1.49	

Upper triangle = brain-map accuracy; lower triangle = stimulus category accuracy.

qualitatively, findings used in support of a modular account of object representation in VT cortex (Spiridon & Kanwisher, 2002).

The data from the first two analyses indicate that highly accurate classification can be achieved by using the statistical structure of brain activity patterns with a scan-by-scan ''brain-reading'' approach (Cox & Savoy, 2003; Thomas et al., 2001; Edelman et al., 1999). Accuracy was highest when all available voxels contributed to the classification, rejecting the strictest definition of modular encoding for all 8 categories. Voxels additional to those in the preferred areas have a tangible and quantifiable role to play in representing objects. These findings are also consistent with the previous data showing that the preferred regions for faces and houses may be impaired at classifying nonpreferred objects (Spiridon & Kanwisher, 2002).

In summary, we replicate previous findings that have been interpreted to support a distributed encoding of objects. Specifically, both preferred and nonpreferred regions can provide good, and in some cases, comparable, information for object classification. We also replicate, qualitatively, previous findings that have been interpreted to support a modular encoding of objects. Specifically, preferred regions for faces and houses are not well suited to object classifications that do not involve faces and houses, respectively. The replication of this constellation of findings within a single data set validates our methods and illustrates the convergence of data on this issue in the context of the somewhat less convergent operational definitions of the theoretical constructs.

Next, we measured the discriminability of the object categories themselves. This extends previous work (Carlson et al., 2003; Cox & Savoy, 2003; Spiridon & Kanwisher, 2002; Haxby et al., 2001) by generating predictions about the degree to which individual categories of objects could share voxels effectively in a distributed neural representation, assuming a visually based encoding of the structure and features of the stimuli.

Stimulus Similarity by Category

The structure of the stimulus categories was analyzed using methods identical to those applied in the brain scan discriminability analysis. Optimal classifiers were created for all possible pairs of object categories. These classifiers operated on the PC coordinates derived from the stimuli presented to subjects in the fMRI experiment. This representation is the basis of many current computational models of face recognition (Petersson et al., 1999). These models typically make use of 2-D view-based representations, which can be used to recognize objects from multiple viewpoints when the model training includes examples of varying views of the objects (Riesenhuber & Poggio, 2000). These view-based models are consistent with neurophysiological

Table 2. Classification Accuracy with Preferred and Nonpreferred Areas

	Face	*House*	*Cat*	*Chair*	*Shoe*	*Scissors*	*Bottle*	*Scramble*
All voxels	2.69	2.87	2.43	1.99	2.1	1.95	1.89	2.46
Preferred	2.32	2.76	1.94	1.55	1.78	1.71	1.58	1.97
Nonpreferred	2.30	2.03	2.04	1.85	1.82	1.69	1.69	2.27

In all cases, the all-voxel condition accuracy exceeds the preferred and nonpreferred accuracy.

(Logothetis, Pauls, & Poggio, 1995; Perrett, Hietanen, Oram, & Benson, 1992) and psychological (O'Toole, Edelman, & Bülthoff, 1998; Tarr, Williams, Hayward, & Gauthier, 1998; Logothetis, Pauls, Bülthoff, & Poggio, 1994) data on object and face recognition. As implemented here, the PCA-based model assumes invariances that can be achieved with 2-D affine transformations, or alternatively, in a more biologically plausible fashion with hierarchical models (cf., Riesenhuber & Poggio, 1999, 2000). This assumption was met for this stimulus set, as the objects were sufficiently aligned to achieve highly accurate classification without these transformations.

The results of the simulation indicated that the object categories used in the fMRI experiment, like the brain scan categories, are highly, although not homogeneously, discriminable (see Table 1). As was seen for the brain-map discriminations, faces and houses were the categories discriminated most accurately. These data add to previous findings (Spiridon & Kanwisher, 2002; Haxby et al., 2001) a measure of how dissimilar objects are within the categories tested. Thus, the data generate a prediction about the degree to which voxels for individual pairs of categories should be shared given a distributed encoding of the visual structure of objects.

Comparing Brain Scan and Stimulus Similarity Profiles

The object-form topography model predicts that brain-map response patterns to different categories of objects should be distributed when the attributes needed to represent the categories are shared. Brain maps corresponding to different object categories should be most discriminable, therefore, when the features needed to represent them are not shared between the two categories. We compared the confusability of the object-stimuli and the brain scans by correlating the discrimination indices (d') from the object category simulations with the discrimination indices from the neural response patterns. For the full set of object pair discriminations, the relationship between the stimulus and brain scan discriminability was moderately strong and statistically significant ($r = .42$, $p = .03$, $df = 26$). Using average d' for the 8 object categories, the correlation was .67 ($df = 6$, $p = .07$).

Visual inspection of the data indicated a marked deviation from the stimulus–brain-map relationship only for the scrambled controls and the cats. Removing these two categories from the analysis increased the correlation between brain map and stimulus discriminability to .84, for all possible object discrimination pairs. This explains 71% of the variance in the discriminability of the stimuli and brain scans. For the data averaged across categories, the correlation increased to $r = .92$, explaining 85% of the variance.

Why do the correlations improve when the scrambled controls and cats are removed from the analysis? For the scrambled images, there is minimal form information and so the cortical response may be noisy. It is less obvious why the elimination of the cat stimuli increased the correlation, although one explanation may be found in a recent study that looked at fusiform gyrus response to cat faces (Tong, Nakayama, Moscovitch, Weinrib, & Kanwisher, 2000).

Object Representations Linking Brain-map Structure to Stimulus Structure

The relationship between brain-map structure in VT cortex and object category structure indicates that the shared attributes of object structure, as defined by this simple stimulus-based analysis, are reflected in the similarity of the patterns of brain responses to these attributes. The stimulus analysis was carried out directly on an image-based representation of the objects. This raises the possibility that the relationship between object structure and brain-map structure might be based on low-level visual features. This ''low-level visual'' interpretation of VT cortex is not consistent with previous neuroimaging work showing similar responses to objects over changes in viewpoint and across image format changes (e.g., grayscale photographs and line drawings). We show next that despite the relatively ''low-level'' nature of the image-based code, with appropriate experience, the object categorization model operates successfully across these changes and is therefore consistent with these neuroimaging findings.

First, the stimuli we modeled (Haxby et al., 2001) included images of objects taken from different viewpoints (Figure 2). As seen in Table 1, the algorithm

Figure 2. Example chair stimuli from neuroimaging experiment and simulations show the range of view variations in the data set.

successfully discriminated object categories across a broad range of views. This is not surprising for a 2-D image-based model when its experience includes examples of objects from diverse viewpoints (Riesenhuber & Poggio, 2000). The view-based nature of the model, and its ability to operate across viewpoint when the training set includes a range of viewpoints, is consistent both with psychological (O'Toole et al., 1998; Tarr et al., 1998; Logothetis et al., 1994) data on object and face recognition and with neurophysiological data on the selectivity of neurons in the inferotemporal cortex to objects and faces (Logothetis et al., 1995; Perrett et al., 1992). The object attributes detected in the view-based computational analysis are thus consistent with the view invariant responses found in functional neuroimaging studies.

Second, there is evidence that VT cortex areas respond to both grayscale and line drawing versions of the object stimuli (Spiridon & Kanwisher, 2002; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999). To test the applicability of the present model to these data, we carried out two additional simulations to test our stimulus-based discrimination algorithm on spatial frequency filtered (high-pass), ''line-drawing'' versions of the original stimuli (Figure 3).

The first simulation was trained on the original grayscale images and tested on its ability to discriminate object categories using high-pass images of the objects. We counterbalanced the available stimuli to be certain that during the test, the model was discriminating high-pass images of stimuli not learned previously in their grayscale versions (i.e., novel exemplars). To classify the object categories correctly, therefore, general information about the categories that transcends both the stimulus set (individual pictures) and the low-level features (grayscale–line vs. line drawings) must be learned.

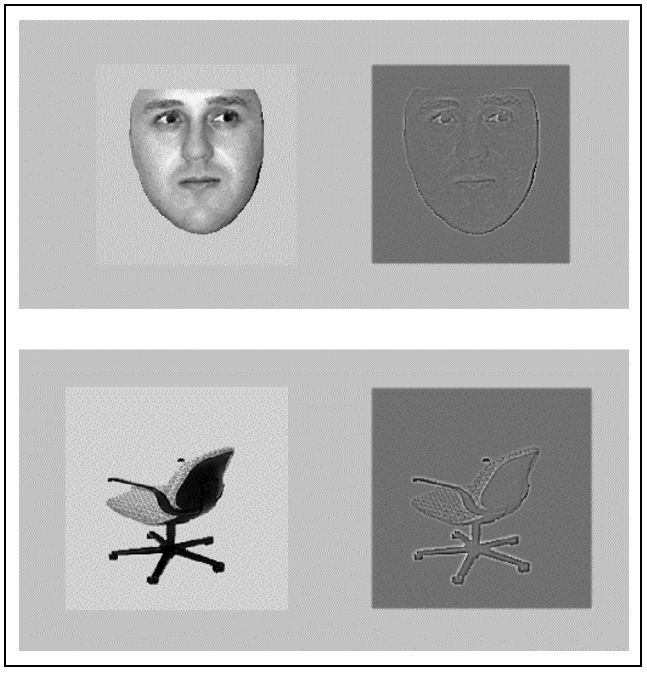

Figure 3. Examples of stimuli and high-pass-filtered versions of these stimuli.

Object classification accuracy for the stimuli tested as high-pass images remained good with an average d' of 1.59. More importantly, the correlation between stimulus and neural response discriminability across all possible pairs of objects remained strong ($r = .57$, $p = .03$, $df = 13$). In a second simulation trained with grayscale and high-pass images, performance improved somewhat ($d' = 1.77$), and the correlation between stimulus and neural response discriminability increased slightly ($r = .63$, $p = .01$, $df = 13$). The decline in performance between the high-pass image and the grayscale classification is consistent with fMRI data, indicating weakened responses to line drawings by comparison to photographs (Spiridon & Kanwisher, 2002).

The relationship between brain-map and stimulus confusability was maintained over changes in low-level features. The features captured in the object classification model are, therefore, consistent with those measured in neuroimaging studies.

DISCUSSION

The debate concerning the representation of objects in VT cortex has focused on the issue of whether the neural response patterns that result from viewing faces and objects are modular or distributed. Less attention has been paid to the nature of object representations that are predicted by distributed versus modular activation patterns, and no attention has been paid to the potential significance of partially distributed activation patterns for representing objects. Rather than treating neural activation patterns that are neither perfectly modular nor perfectly distributed as ''noise'' or artifact, we assume that the degree of overlap in these patterns can contain important information about the nature of object representations in cortex. To constrain the interpretation of neural response patterns, we combined an analysis of these partially distributed neural activation patterns with a computational model of object recognition. The results are supportive of a ''feature-based'' representation of objects in VT cortex. Although the exact nature of these features is unknown, in the context of previous neurophysiological and computational findings, the present data constrain the representation in three ways.

First, a view-dependent, image-based representation accounts well for the confusability of the neural response patterns for objects and faces. Although more complex or abstract representations are plausible, by parsimony, this finding indicates that more complex representations need not be invoked to account for

the structure of the neural responses. Neurophysiological studies support a hierarchical organization of visual processing that begins with retinotopic feature analysis in early visual areas and proceeds to progressively more complex, view-based object codes in the inferotemporal cortex (Kobatake & Tanaka, 1994). View-dependent cells can be pooled to recognized objects independent of viewpoint, given experience with multiple views of the objects (Logothetis et al., 1995). The mixture of view-selective and view-independent cells in the inferotemporal cortex suggests that these codes share neural space (Perrett et al., 1992). View-based object classification models operate analogously. Although these models represent relatively low-level view-specific attributes that do not generalize across viewpoint, they can achieve view independent performance through appropriate experience.

The second representational constraint concerns voxel sharing. If voxel sharing reflected the epiphenomenal engagement of the visual system in response to any potentially relevant stimulus, we would expect no relationship between the confusability of the stimulus categories and the confusability of the neural response patterns. The present result is consistent, therefore, with an unencapsulated, distributed coding of objects, in which the physical properties of the object categories are reflected, in kind, at the level of VT cortex. The object-form topography hypothesis assumes a neural encoding of object attributes that is distributed, because it posits that VT cortex contains a representation of objects in terms of their attributes or features. It follows, therefore, that object categories will share neural space when they share common attributes.

Finally, these data help to separate the concept of representational types (object-form topography, Haxby et al., 2001; view-based, Riesenhuber & Poggio, 1999; Poggio & Edelman, 1990; structural, Biederman & Gerhardstein, 1993; intermediate complexity features, Ullman, Vidal-Naquet, & Sali, 2002) from the concept of pattern response types (modular, distributed). The underlying assumptions of the object-form topography hypothesis are consistent with a distributed coding in principle, but with a wide range of response types in practice. The principle of distributed encoding stands because it relies on the encoding of objects in terms of attributes that may be shared among different categories. The relationship between stimulus and neural confusability supports this principle. In practice, however, the object-form topography model allows for response types that are modular, when the shared neurally represented features between categories are minimal.

Understanding why the activation patterns for certain categories are more or less distributed can leverage more precise information about the nature of object representations in VT cortex and the features on which they may be based. The approach we take with PCA and a linear discriminant classifier is simple and direct. A variety of more sophisticated feature extraction algorithms and classifiers should be explored in future work. In conjunction with a stimulus model, a pattern-based classification approach to the analysis of functional neuroimaging data can constrain the interpretation of neural representations of objects.

METHODS

Pattern Classification of the Brain Scan Data

The brain scan simulations were carried out to determine the neural discriminability of all possible pairs of object categories for each of the 6 subjects. Two counterbalance conditions were constructed using different halves of the scans (odd vs. even runs of trials). These served alternately as learning and test sets for all parts of the procedure. Classification results are based, therefore, on averages over the 6 subjects and the 2 counterbalance runs.

Scan Data

Raw epi scan files from 6 participants viewing 8 categories of stimuli (faces, houses, cats, chairs, shoes, bottles, scissors, and scrambled images) were used in the analysis (Haxby et al., 2001). The data were corrected for movement artifacts but were not preprocessed in other ways. Because the analyses were carried out individually for each subject, brain alignment was not necessary.

For each subject and category of objects, there were 84 scans, totaling 672 scans per subject. For some reason, subject 5 had only 70 usable scans, making a total of 560. The original scans consisted of $64 \times 64 \times 40$ voxels and were masked to include only VT cortex voxels that were significant across objects in the fMRI study (Haxby et al., 2001). The number of significant voxels varied between 307 and 675 for the participants, with an average of 465.5 across the 6 participants.

Procedure

The scans for each subject were divided into the odd and even trial runs for use alternately as the training and test sets. The analysis was applied to all possible pairs of object categories (e.g., face vs. house, face vs. chairs, etc.). In each case, a principal components analysis (PCA) was performed on the scans in the training set, which consisted of 84 scans (42 from each of the 2 categories in the pair) of length N, where N was the number of significant voxels for the subject under consideration.

To represent individual scans for input to the linear discriminant analyses, the coordinates of the scan projections on the principal components (PCs) were computed. The coordinate vectors provided a concise and

complete representation of the scans in the PCA-based space.

The purpose of the linear discriminant classifiers was to determine the discriminability of the scans from pairs of categories. One component of this effort was to determine which PCs were useful for discriminating scans from the two object categories. We assessed this by training two kinds of linear discriminant networks to discriminate pairs of object categories before constructing a third ''optimal'' classifier from the ''useful'' PCs. Information from both ''preliminary'' classifiers was used to select PCs useful for the discrimination task.

The process of constructing the optimal classifier proceeded as follows. First, we trained a linear discriminant classifier to predict object category with input vectors consisting of the full set of coordinates on all available PCs. One measure of the importance of individual PCs can be found in the ''learned'' weights associated with each input coordinate. Weights with large absolute values indicate PC's useful for predicting the object category of the scans in the learning set. These weights were used subsequently as part of the information employed for selecting PCs for the optimal classifier.

Second, a series of single dimension classifiers were implemented to assess classification performance for individual PCs on the test set of brain scans. Coordinates from the training set scans on each PC were used to predict object category. The performance of each PC for classifying test set scans was assessed using the signal detection measure d', computed as z score (hit rate) $- z$ score (false alarm rate). Using the face–house classification task as an example, the hit rate is the proportion of face scans correctly categorized as face scans, whereas the false alarm rate is the proportion of house scans incorrectly categorized as faces. This yielded a measure of the utility of individual PCs for generalizing object recognition to the test set scans.

The optimal classifier was constructed by selecting the most useful PCs for classifying both the learning and test scans. The purpose of using data from both the learning and test scans was to avoid the inclusion of PCs that might have succeeded on one or the other set of scans by chance or by overfitting. Specifically, we chose the 20 PCs in the all-coodinate classifier with the largest weights (absolute values). Next, we chose the 20 PCs from the individual classifiers that produced the largest d'. PCs that appeared in both lists were combined and used for the optimal classifier. These formed a low-dimensional, noncontiguous subspace classifier tailored to discriminating the object categories.

Preferred and Nonpreferred Regions

The preferred region for an individual category was defined as the set of voxels that gave their maximum response to that category, by comparison to the other 8 categories (Haxby et al., 2001).

Pattern Classification of the Object Category Stimuli

Stimulus Data

The original images used to generate the brain-map data were analyzed. For each category, forty-eight 400×400 grayscale images were available. These included 12 exemplar objects photographed from 4 views. The scrambled category contained samples of phase-scrambled exemplars from each of the other 7 categories of objects.

Procedure

The procedure for classifying the stimuli by object category was identical to that employed for brain scan data. Again, stimuli were divided into training and test sets. Images were converted to vectors by concatenating the rows of pixels in the image into a vector. PCA was performed on pairs of object categories using the images in the learning set (e.g., 24 face vectors and 24 house vectors). Linear discriminant classifiers, as described previously, were used to determine which PCs were useful for classifying the objects (e.g., face vs. house). Optimal classifiers were then formed to generate the d' data in the bottom triangle of Table 1.

APPENDIX

Each analysis was performed on data from a single subject viewing two categories of objects. These data formed an I row $\times$ K column matrix, $\mathbf{X}$, where I was the number of voxels and $K = 2 \times J$, with K equal to the number of scans available per subject for the 2 categories. See Methods for the exact values of I and J.

The analysis proceeded as follows. Each column of $\mathbf{X}$ was normalized to length 1. The matrix $\mathbf{X}$ was then divided into an $I \times J$ training matrix, denoted $\mathbf{X}_{\text{train}}$, and an $I \times J$ test matrix, denoted $\mathbf{X}_{\text{test}}$. The training and test scans in each matrix included an equal number of scans from each of the two categories, which were selected according to the counterbalance scheme described previously. Thus,

$$\mathbf{X} = [\mathbf{X}_{\text{train}}\mathbf{X}_{\text{test}}] \tag{1}$$

Next, the matrix for the training set was decomposed according to the singular value decomposition:

$$\mathbf{X}_{\text{train}} = \mathbf{P}\Delta\mathbf{Q}^{\text{T}} \tag{2}$$

where

$$\mathbf{P}^{\text{T}}\mathbf{P} = \mathbf{I} \text{ and } \mathbf{Q}^{\text{T}}\mathbf{Q} = \mathbf{I} \tag{3}$$

are the left and right matrices of singular vectors and Δ is the diagonal matrix of the singular values ranked

from the largest to the smallest (see Abdi, Valentin, & Edelman, 1999, Equation 3.20 *ff.*).

In general, depending upon the simulation, we kept a subset of the singular vectors (*L* vectors) and denote the matrix composed of the *L* selected columns of matrix **P** (respectively, Δ, $\mathbf{Q}^{\mathrm{T}}$) as $\mathbf{P}_{[L]}$ (respectively, $\Delta_{[L]}$, $\boldsymbol{Q}_{[L]}{}^{\mathrm{T}}$). To discriminate between the two categories of objects being viewed, we trained a linear classifier using the projections $\boldsymbol{Q}_{[L]}$. The linear classifier was obtained by computing a weight vector (see Abdi et al., 1999, Equation 4.14 and p. 60, *ff.*)

$$\mathbf{w} = \mathbf{Q}_{[L]}\mathbf{t} \tag{4}$$

where **t** is the target vector whose values are set to 1 for the elements of the first category and −1 for the elements of the second category.

The category membership of scans from the test set was obtained by first computing their projections on the left singular vectors:

$$\mathbf{Q}_{\mathrm{test}} = \mathbf{X}_{\mathrm{test}}^{\mathrm{T}}\mathbf{P}_{[L]}\Delta_{[L]}^{-1} \tag{5}$$

and then multiplying the projections by the weight vector **w** to obtain the predicted category membership $\mathbf{t}_{\mathrm{test}}$

$$\mathbf{t}_{\mathrm{test}} = \mathbf{Q}_{\mathrm{test}}\mathbf{w} \tag{6}$$

Scans with $\mathbf{t}_{\mathrm{test}} \geq 0$ were assigned to the first category and scans with negative values of $\mathbf{t}_{\mathrm{test}}$ were assigned to the second category. The proportion of scans correctly assigned to 1 of the 2 categories in the pair (arbitrarily chosen) gave the hit rate, whereas the proportion of scans incorrectly assigned to that arbitrarily chosen category gave the false alarm rate. The statistic d' was computed for each pair of categories based on the hit and false alarm rate.

The procedure was followed for all 28 pairs of categories for each of the 6 subjects. This gave 6 matrices of d' values, which were averaged to produce Table 1.

Acknowledgments

This work was supported by a grant from ONR to A.J.O. and H.A.

Reprint requests should be sent to Alice J. O'Toole, School of Behavioral and Brain Sciences, The University of Texas at Dallas, GR4.1, Richardson, TX 75083-0688, or via e-mail: otoole@utdallas.edu.

The data reported in this experiment have been deposited with the fMRI Data Center archive (www.fmridc.org). The accession number is 2-2004-1181E.

REFERENCES

Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks.* Thousand Oaks, CA: Sage.

Biederman, I., & Gerhardstein, P. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1162–1182.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience, 15,* 704–717.

Cox, D., & Savoy, R. (2003). Functional magnetic resonance imaging (fMRI) ''brain reading'': Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage, 19,* 261–270.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1999). Towards direct visualization of the internal shape representation space by fMRI. *Psychobiology, 26,* 309–321.

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

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform ''face area'' increases with expertise in recognizing novel objects. *Nature Neuroscience, 2,* 568–573.

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

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences, U.S.A., 96,* 9379–9384.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71,* 856–867.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4,* 401–414.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5,* 552–563.

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research, 55.*

Perrett, D., Hietanen, J., Oram, M., & Benson, P. (1992). Organization and function of cells responsive to faces in temporal cortex. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences, 335,* 23–30.

Petersson, K. M., Nichols, T. E., Poline, J.-B., & Holmes, A. P. (1999). Statistical limitations in functional neuroimaging. I. Non-inferential methods and statistical models. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences, 354,* 1239–1260.

Phillips, P. J., Moon, H., Rizvi, S., & Rauss, P. (2000). The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions: Pattern Analysis and Machine Intelligence, 22,* 1090–1103.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize 3D objects. *Nature, 343,* 263–266.

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

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience Supplement, 3,* 1199–1204.
Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI Study. *Neuron, 35,* 1157–1165.
Tarr, M. J., Williams, P., Hayward, W., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience, 1,* 275–277.
Thomas, E., VanHulle, M. M., & Vogels, R. (2001). Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience, 13,* 190–200.
Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. *Cognitive Neuropsychology, 17,* 257–279.
Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5,* 682–687.