Patterns of activity in the categorical representations of objects

Thomas A. Carlson*, Paul Schrater, Sheng He
Department of Psychology
University of Minnesota
75 East River Road
Minneapolis, MN 55455

* corresponding author
Email: carl0395@tc.umn.edu
Phone: 612-743-1127
Fax: 612-626-7253
Abstract:

Object perception has been a subject of extensive fMRI studies in recent years. Yet, the nature of the cortical representation of objects in the human brain remains controversial. Analyses of fMRI data have traditionally focused on the activation of individual voxels associated with presentation of various stimuli. The current analysis approaches functional imaging data as collective of information about the stimulus. Linking activity in the brain to a stimulus is treated as a pattern classification problem. Linear discriminant analysis was used to re-analyze a set of data originally published by Ishai and her colleagues (Ishai, A., Ungerleider, L. G., Martin, A., Haxby, J. V. (2000) “The Representation of Objects in the Human Occipital and Temporal Cortex.” *Journal of Cognitive Neuroscience, U.S.A., 12 Supplement 2*, pp. 35-51), available from fMRIDC (accession no. 2-2000-1113D). Results of the analysis found that patterns of activity that distinguish one category of objects from other categories is largely independent of one another, both in terms of the activity and spatial overlap. The information used to detect objects from phase scrambled control stimulus is not essential in distinguishing one object category from another. Furthermore, performing an object matching task during scan significantly improved the ability to predict objects from controls, but had minimal effect on object classification, suggesting that the task based attentional benefit was non-specific to object categories.
INTRODUCTION

The human brain is arguably the most powerful computational architecture known to man and represents one of nature’s crowning achievements. One of the most striking aspects of the brain is its seemingly limitless capacity for representing information. Deserving of considerable effort has been put forth to understand its functional and neuroanatomical architectures. Here, we apply pattern classification methods to fMRI data in an effort to further investigate the nature of how categorical information is represented in the brain.

The debate about how categorical information is represented in the cortex is often summarized into two competing theories (distributed vs. modular). The modular viewpoint proposes that the cortex can be divided into distinct modules dedicated to processing and representing particular types of information (Fodor, 1983). This view has been supported by evolutionary considerations and by findings that some neuropsychological patients exhibited category specific deficits. In recent years, more evidence supporting this claim has come from functional neuroimaging. Cortical regions were identified that showed preferential activation to specific categories of information, such as the fusiform face area (FFA), the para-hippocampal place area (PPA), and more recently the extrastriate body area (EBA) (Kanwisher et al., 1997, Epstein & Kanwisher 1999, Downing et al., 2001). However, there remain strong challenges to this domain-specificity view. For example, it was shown that FFA also responded robustly to non-face stimuli in the form of novel synthetic objects called greebles, leaving questions as to whether these modules were indeed category specific (Gauthier et al., 1999). Conversely, the distributed hypothesis submits that information is diffusely represented in the cortex. This viewpoint, until recently, lacked strong evidence in human neuroimaging and principally relied on evidence drawn from non-human primate studies (Tanaka, 1996, 1997; Wang et al., 1996) and computational modeling (Hinton 1981, Rumelhart & McClelland 1986). fMRI evidence supporting the distributed perspective came when it was reported that previously identified modules in the ventral temporal cortex for faces, houses, and chairs responded robustly to all three categories of objects (Ishai et al., 1999). While the response in these areas to non-preferred object categories was smaller than that to the preferred category, it nevertheless was significant. Based on these findings, the authors proposed the object form topology hypothesis -- that object form is represented in the ventral temporal cortex continuously in a distributed and overlapping arrangement.

More recently, it was argued that the smaller response in ventral temporal cortex to non-preferred stimuli has a functionally important role as the response conveys information sufficient to determine the stimulus category (Haxby et al., 2001). The argument was based on a novel classification analysis that predicts the object category by computing correlations between regression coefficients for each of the stimulus categories. The analysis showed that object categories could be predicted robustly based on responses from the ventral temporal cortex even after removing areas that responded maximally to that category. For example, face stimulus could be distinguished from others based on activities from ventral areas without FFA. Furthermore, activity in regions such as FFA could predict non-preferred stimuli such as houses and chairs from other categories of objects.

Addressing the question of how object categories are represented in the cortex requires establishing a link between the object category being viewed by the observer and brain activity. Analysis methods such as the general linear model (GLM) establish this link by determining how well a model can account for the activity in a given voxel. Haxby et al’s analysis establishes this
link by creating a template of activity for object categories via regression coefficients, in which the coefficients are simply the mean activity across time in a subset of scans for each voxel\(^1\), and using them to predict subsequent functional imaging runs. The current analysis sought to establish this link in a novel way through the application of linear discriminant analysis.

Linear Discriminant Analysis\(^2\), a traditional form of statistical classification analysis, was used to understand what features in the activation vector distinguish different object categories. Instead of using stimulus category to identify voxels as significantly active, we seek to find which voxels contribute to the pattern of activity that is indicative of the stimulus category the observer is viewing. The analysis produces a spatial map called a discriminant or canonical variant that is a spatial weighting of voxels that can be used to reliably predict the category of object presented to the subject based on a single fMRI time acquisition.

One of the key steps in linking an activation pattern to object categories is to find areas that respond well for all members of an object category. This raises a problem for using mean activation across large spans of time to characterize object category representation. Means are notoriously sensitive to outliers, and very strong responses within a voxel to just a few of the stimuli within a class can be enough to generate a significant mean, even though this voxel would have almost no predictive power for an object category on a time-point by time-point basis. Thus, mean activations cannot always distinguish between voxels that code information about an object categories and voxels that code information about particular stimuli within a class (feature encoders). To address this, the current analysis uses individual time acquisitions to test its predictions.

In addition, in contrast to voxel-based methods, categories of objects will be linked to “patterns of activity” across the cortex. This is only natural given that neuronal structures in the brain are highly interconnected and given that the object form topology hypothesis suggests that objects are coded as pattern of activity in a large scale distributed network. The current analysis treats each functional acquisition as a spatially structured high-dimensional activation vector that resides in an activity space. Figure 1a shows the transformation between activity space and voxel space. In activity space, the activity in the cortex evoked by a stimulus is treated as a whole and the significance of a voxel is based on its importance to the pattern of activation that distinguishes object categories, as opposed to how well it correlates with a model of the expected response.

An additional concern is that the results from traditional analysis are highly dependant on the control stimulus. As shown in figure 1b, two hypothetical areas are coding two classes of stimuli. In the traditional sense, the first model is completely modular with the respective regions becoming active only when the preferred stimulus is present; the second model is distributed in that both areas are simultaneously coding each of the two classes of stimulus. Interestingly, the only difference between these two models is the location of the origin, which is defined by the

---

\(^1\) Least square regression coefficients were determined via the pseudo-inverse of the matrix of regressors. When the regressors are non-overlapping temporal functions, which is true of most block fMRI designs, then this pseudo-inverse solution is a weighted average of the time-points where the weights are produced by the temporal filter used to model the hemodynamics. The weighted average reduces to a simple average when the regressors are box-functions. See Appendix.

\(^2\) For those who are not familiar with linear discriminant analysis, a good introduction can be found in Duda, Hart & Stork (2000).
control stimulus! The current analysis principally focuses on activity that defines object categories relative to other categories of objects. This activity, we would argue, is more indicative of the categorical representation of objects as it is defined by their relationship to other object categories. In the example of fig 1b, the relationship between the two classes of stimulus remains constant for both models regardless of the location of the origin as shown by the hypothetical d’ measure.

Prior to the application of the LDA, two preprocessing steps, in addition to those performed by Ishai et al (Ishai et al., 2000), were taken to maximize the information associated with individual time points, and to minimize outliers that would influence the analysis. First, a deconvolution operation was performed on the data to remove the lag due to hemodynamic response and de-correlate the time series data relative to a previously described estimate of the hemodynamic response function. The second preprocessing step removed outliers from the dataset using a previously described method (Rousseeuw & Van Driessen 1999).

The application of linear discriminant analysis to fMRI data is carried out in two distinct steps: dimensionality reduction and classification (Figure 2). Dimensionality reduction consisted of removing sub-cortical regions and areas outside of the brain, and extracting the top 40 principle components in the data. Next, classification was implemented on the reduced data. Individual time acquisitions were labeled according to stimulus condition (i.e. chairs, faces, and
patterns of activity were identified that support several types of discriminations among the three categories of stimuli (chairs, faces, houses, and phase scrambled controls).

**Figure 1.** Linear discriminant analysis applied to fMRI data. There are two basic steps for the application of discriminant analysis to functional imaging data. **A.** Data Reduction. A necessary condition for discriminant analysis is that the number of predictors be equal to or less than the number of data points. Data reduction is performed in two steps. First, regions outside the brain and subcortical regions are masked and removed. Second, Principal component analysis is used to further summarize the data into component processes. **B.** Discriminant Analysis. Principal component eigenvalues for individual time acquisitions are labeled by condition (i.e., chairs, faces, and houses) and projected onto an N dimensional space, with N equal to the number of components (2D example shown). Within this space, an optimal discriminant axis and decision boundary is computed and subsequently tested.

The fits were validated by measuring the predictive performance of the linear classifier using the 632+ bootstrap (Efron et al., 1997), a leave-one-out resampling procedure.
Discriminants for classifying one category of objects from other categories (e.g., faces vs. chairs and houses) will be referred to as category-specific discriminants because they rely on category-specific activity to discriminate between classes of objects. Discriminants that are effective at classifying objects from their phase scrambled controls (e.g., chairs vs. phase scrambled chairs) are termed object-control discriminants because they utilize activities that separates objects from their controls. In addition, a set of pairwise discriminants (e.g., faces vs. chairs), which also rely on category specific activity, were evaluated to facilitate comparisons with the results from another classification analysis (Haxby et al., 2001).

RESULTS

In the following sections, we report results from the linear discriminant analysis of the Ishai et al data, addressing these three questions: 1. How well can the fMRI data be used to classify the stimuli? 2. How much interdependence is there in the different groups of discriminants? 3. How do the discriminants interact spatially?

Performances of Linear Discriminants

Classification performances for the discriminants are shown in Table 1. The discriminants for classifying one object category vs. the other two categories had a mean performance of 74.7% correct classification with chance performance at 50% accuracy. Classification for houses was significantly lower than the other two category-specific discriminants in both the delayed matching condition and the passive viewing condition. Mean performance for the pairwise discriminants was slightly higher at 76.3% accuracy. Discriminant performance can be interpreted as the degree of “uniqueness” between object categories. The lower performance in distinguishing houses from the other two categories suggests that houses share more common activity with the other two classes of objects and therefore less dissociable. This is also reflected in the performance of pairwise discriminants with more accurate discrimination between chairs and faces than that of the two pairwise comparisons involving houses. Interestingly, the increased cognitive demands of the delayed matching condition did not result in any significant enhancement in the discriminant performance relative to passive viewing for both the category-specific and the pairwise discriminants; the two discriminants that rely on category specific activity.

Performance for the object-control discriminants was notably higher than the category specific discriminants. This result is expected given that the object-control discriminants can benefit from shared processes across categories of objects. General responses to objects compared to phase scrambled controls, such as those found in the lateral occipital areas (Malach et al., 1995), can contribute to the object-control discriminants. Classification performance for these discriminants was very similar across categories. There was a large effect of task demand (delayed matching task vs. passive viewing) for all three categories of objects. This is contrasted

---

3 A brief description of the experimental paradigm, data collection and preprocessing procedures can be found in the methods section. However, since this is a re-analysis of published data, readers are referred to the original paper for more detailed descriptions (Ishai et al., 2000).
4 Significance accessed by comparing overlap in 95% confidence intervals
with the lack of a task demand effect on the category specific discriminants. Since the stimuli used in the original study were all well above threshold for detection, we do not believe that the subjects had difficulty distinguishing faces, chairs, and houses from their phase scrambled controls in the passive viewing condition. However, if there were a homunculus sitting in these subjects’ head and trying to decide whether the subject just saw a object or a control based on cortical activity patterns, then the delayed matching task certainly made the Homunculus’ job easier, as it did for the object-control discriminant. Curiously, the delayed matching task did not make the homunculus’s job any easier in deciding whether the subject just saw a chair or a face or a house.

**Linear Discriminant Knockouts**
Linear discriminant analysis identifies the patterns of activity that are important for object classification. It also provides an opportunity to examine how much of the information are common between different object categories. Here we introduce the novel idea of a discriminant knockout (Figure 3). Linear discriminant analysis provides an axis in a multidimensional activity space, along which the two conditions can be best separated. What happens when all the information in the cortex along this axis is destroyed, effectively removing this axis? If this process is done recursively, one can remove all the information in the data that can potentially contribute to the discrimination of one category of stimuli from others. The expected result of such a manipulation would be the reduction in performance of the discriminant to the chance levels. More interesting to know is the effect of knocking out one discriminant on the performance of other discriminants, as they are not necessarily orthogonal to one another. For example, we can knockout activity in the data that supports the discrimination of faces against other categories of objects, creating a “virtual lesion”, and then evaluate the effect of this manipulation on the performance of other discriminants.

Results for linear discriminant knockouts for both the passive viewing and delayed matching conditions are shown in figure 4. The dependant measure is relative loss in performance, which is the change in performance due to the knockout scaled by the baseline performance of the discriminant. Results for a discriminant knockout on the same discriminant represent the amount of information that needs to be removed to reduce the discriminant to chance levels (plots marked with stars). This magnitude provides a baseline that can be used to compare the effects of the other knockouts on the discriminant. The difference between knockouts within a discriminant and knockout of other object categories indicates the amount of unique activity between the two categories. For example, in the passive viewing condition, a chair category-specific knockout for the chair specific discriminant requires a 23% relative loss in accuracy to bring the discriminant to chance performance. In comparison, knocking out the

**Figure 3.** Discriminant Knockouts. A. Linear discriminant knockout procedure and evaluation. Overlap in activity space is evaluated via linear discriminant knockouts. Three categories of stimuli activity (open circles, filled circles and X’s) are plotted in a two dimensional space. Initially discriminant performance for classifying open vs. closed circles is 80%. A linear discriminant knockout is then performed by removing the axis that separates open circles and X’s. The original discriminant (open vs. closed circles) is then retested yielding 72% classification accuracy. The effect of linear discriminant knockout is evaluated in terms of relative loss of accuracy, which is the loss in performance scaled by the baseline performance of the discriminant. B. Result of linear discriminant knockouts for object specific and object control discriminants in both the delayed matching and passive viewing conditions.
face category-specific discriminant results in only 6% loss in the performance of chair specific discriminant. The ratio of these two numbers (6% vs. 23%) can be viewed as a measure of shared activity.

Significant interactions were found for 16 of the 18 category specific knockouts, and the remaining two were near significant. The interactions between the category specific discriminants show that there are indeed aspects of the categorical representation, in terms of activity, that are shared across the three categories of objects. The differences between the passive viewing and delayed matching conditions were minimal for category specific discriminants, consistent with the previous result that increased task demands does not affect the discrimination of category specific discriminants. Interestingly, making one type of objects undiscriminable from the other objects did not affect the discriminability of the objects from its scrambled controls. More surprisingly, the reverse is true too. For example, one can lesion the data to the extent that faces are no longer discriminable from scrambled faces, but faces can still be discriminated from chairs or houses based on the same data. This result is interesting in that it suggests that activity in the cortex that enables the categorical discrimination of objects is independent of the information that separates objects from the phase scrambled control stimulus.

Although object-control discriminants are highly independent of the object-specific discriminants, there seems to be a significant amount of shared information in detecting different types of objects from their respective controls. In all cases, knocking out one object-control discriminant had a significant effect on all the other object-control discriminants. The strong interaction between object-control discriminants reflects the significant amount of overlapping processes in object detection. This is in contrast with the weaker interactions found between the category-specific knockouts. Taken together, this pattern of results seems to suggest that category-specific information is not an essential part of the process that distinguishes objects from non-objects.

**Spatial aspects of categorical representations**

**Category Specific Representation in the Cortex**

The projection of the category specific discriminants onto the cortex reveal that the majority of significant voxels were found in the ventral temporal cortex, confirming the strong contribution of the ventral temporal cortex for representing object categorical information. Significant voxels were distributed across the ventral temporal areas in relatively small clusters corresponding with previously identified category selective areas. Interestingly, while the voxels were distributed there appeared to be very little overlap in the significant voxels across categories.

The problem with the aforementioned observation is that the amount of spatial overlap is a function of the threshold applied to the map. The two extreme examples being setting a very low threshold in which the whole brain is significant, thus there would be a large amount of overlap between categories; and setting a very high threshold in which significant voxels would be more likely to cluster in a modular arrangement. To avoid setting an arbitrary threshold, we investigated the relationship between the performance of discriminants and the amount spatial overlap independent of threshold. The number overlapping voxels and the discriminant performance are both a function of the threshold, thus the common variable can bridge a direct relationship between classification performance and the amount of overlapping activation.
Discriminant performance and the percentage of voxels overlapping, relative to the total number of significant voxels, were computed for small increments in threshold. Figure 5 shows the percentage of the number of active voxels participating in multiple discriminants as a function of the discriminant performance. The critical aspect in these plots is the loss of information as the amount of spatial overlap decreases. A decrease in performance coupled with a reduction in the percentage of overlapping voxels indicates the thresholding of voxels that are both important to the prediction and shared across categories. The first two rows in the figure show that the number of shared voxels critical to all three object categories discriminants is relatively small.

There was some variation across subjects. Subjects 2 and 3 show only a minimal loss in performance as the amount of overlap drops (indicating minimal information in the shared voxels), while other subjects show sharp decreases in performance when the overlap fell below between 4% to 12% (indicating critical information are carried in those shared voxels). The amount of overlap in voxels for pairwise object specific discriminants was larger. Again, these thresholds did vary across subjects and comparisons, but the percentage of overlapping voxels was generally quite small.

---

5 Spatial maps were scaled with minimum threshold of 0 and a maximum threshold of 1. Discriminant maps, using the aforementioned scale, were computed for 0.03 increments in threshold.
Voxels contributing significantly to the object-control discriminants were found in occipital, temporal and parietal lobes. Given that these discriminants essentially detect differences between objects and noise, and are likely to encompass many cortical processes, a complete interpretation of them requires more constraints than available in this dataset. Thus, we focus our attention on the differences between the passive viewing and delayed matching tasks that was shown earlier to have an important effect on this discrimination. Under the same threshold, the discriminants for the passive viewing condition had considerably fewer significant voxels than that in the delayed matching condition. Discriminants in the delayed matching condition both recruited new areas of the cortex, and expanded regions identified by the passive viewing task. This is not surprising given the increased cognitive demands presumably would require additional processes and enhance processes already occurring in the passive viewing condition. This result, however, highlights the difficulty in interpreting data collected using tasks such as the delayed matching task. The use of an attention task will often result in more robust response to the stimuli, but can only be interpreted as the combined effect of stimulus and the task.

**DISCUSSION**

A well designed functional imaging study provides a wealth of information. The dataset submitted by Ishai and colleagues serves an excellent example of this, as the group has already reported two analyses based on this dataset (Ishai et al 2000, Ishai et al 1999), the current one would be the third. One of the benefits of performing a re-analysis on a previously published dataset is the opportunity to contrast the two analyses and compare the results.

In this reanalysis, we take seriously the idea that the analyses should explicitly look for patterns of activity that are meaningful, treating voxels as axes in a multidimensional activity space, rather than independently. In doing so, our analysis can more directly investigate hypotheses involving distributed and focal representations. The distributed hypothesis suggests that patterns of activation across the cortex can be used to support different kinds of information. The modular perspective argues that these patterns largely be confined to particular areas. Using linear discriminant analysis in activity space, we can simultaneously and empirically investigate the evidence for both kinds of hypothesis.

The current analysis focused on two types of linear discriminants: category specific and object-control. First, with category specific discriminants, activity specific to object categories was isolated by contrasting one object class with the other two classes of objects. It is important to note, however, that the activity we define as category specific is limited by the number of categories. For example, a pair-wise discriminant for two object categories can not be identified as specific to either object category. In the current study, there are only three categories of objects. Additional research must be performed to further constrain the detection of object specific activity. Second, with object-control discriminants, activity critical for detecting objects was determined by contrasting an object class and its phase scrambled control. A key difference between these two types of discriminants is that shared activities across object categories will not contribute to the category specific discriminants, but they could be important to the object-control discriminants. Our analyses found that these two types of processing evoke patterns of
activity in the cortex that are relatively independent of one another, and are differentially influenced by the task demands.

**Overlapping representation between different object categories**

Category specific discriminants can be used to find patterns of activity unique to a category of objects, and as such can be used to address the issue of how object categories are represented in the cortex. The ability to predict the stimulus presented to subjects based on derived category specific discriminants shows that there are patterns of activity in the cortex unique to each of the three categories of objects that are linearly separable. The weak interactions found after the discriminant knockouts, however, reveal that only a small proportion of the categorical representation of these three objects is shared. Our analysis also found that the category specific activity is largely confined to the specific regions with very little spatial overlap across categories. On the same issue, Ishai and Haxby, among others, have demonstrated that object selective regions do respond to non-preferred stimuli and proposed the object form topology hypothesis to account for these results (Ishai et al., 1999). The results of our analysis showing independent category specific information, on the surface may seem to contradict their hypothesis. How can this be reconciled, especially given that both analyses were performed on the same data? The answer is that the two analyses are looking at two different aspects of the data. In our analysis, activities shared across object categories do not contribute to the object specific discriminants, as it contains no useful information to differentiate between categories of objects. Ishai’s analysis includes this activity and labels it as a component of the distributed representation. The interpretation this activity is a question that we feel remains open and could be addressed in future research.

**Object detection vs. object category classification**

The object-control discriminants reflect the activities associated with the general processes of detecting an object. The behavior of object detection discriminants was very different from the category specific discriminants. First, classification performance was much higher for object-control discriminants. The result is expected given that many areas are responsive to objects but not to scrambled controls (Malach et al., 1995; Grill-Spector et al., 1998; Grill-Spector et al., 2001), and of all of the activity in these areas can contribute to the object-control discriminant. On the other hand, a relatively smaller number of voxels have a differential response for the three categories and are capable of contributing to the category specific discriminants.

Interestingly, task demands and attention had very little effect on the performance of category specific discriminants and also had negligible effect on the knockout’s across the category specific discriminants. Previous studies have shown that attention could enhance activity in object responsive regions (O’Craven et al., 1999), and enhancement of response magnitude from the matching task is evident in the current data set (Ishai et al 2000). The result of our reanalysis suggest that the delayed matching task simply boosted object related activities, but the benefit was not specific to any particular object category. Given this, it would seem that the use of an attention task would give not be useful for studying the categorical representation of objects, as would only make the interpretation of the results more complex without emphasizing critical aspects between object categories.
**Difference between linear discriminants and correlational analysis**

The classification performances of the derived linear discriminants were found to be robust, but substantially lower than that found by Haxby et al (2001) using a correlation classification method. The performance for our pair-wise discriminants ranged from 73.2% to 83% in the delayed matching condition, while Haxby reported near perfect classification for these three classes of objects. This difference is to be expected for several reasons. First, our analysis attempts a much more difficult prediction: decisions were made for each individual time acquisition rather than for regression coefficients derived from half the total acquisitions. For block designs, least square regression coefficients are category means across time for each voxel. In Haxby’s analysis, the prediction performance is limited by the variability of these means, rather than the variability of individual time acquisitions. If we replaced individual scans with means across half the points (41 time points) in a group, and we assume for the sake of argument that all of these scans are independent, then a $d'\$ measure of performance would increase by a factor of $\sqrt{41}$. For all of our comparisons, such an increase would put performance in terms of percent correct at ceiling. Adjusting for the fact that the time points are not independent predicts a smaller gain, however it would require only 4 independent time points to drive all our performance values over 95%. Thus, the near perfect performance reported in Haxby certainly in part owes to the statistical advantage of classifying averages rather than individual acquisitions. However, there is an addition difference as well. In many ways, their classification procedure is a different kind of prediction. The prediction in Haxby’s analysis uses a previous run as a memorized template, their procedure is a type of matched filter detection, appropriate for when the target to be detected is a signal exactly known. In our prediction, we do not know a priori what the relevant pattern of activation is, so it is derived empirically.

**General Discussion**

In summary, the current paper applied LDA to address the question of categorical object representations. We found that activity in the cortex that defines object categories relative to other categories of objects is independent of activity that defines an object category relative to a control stimulus. The attentional benefit of using a delayed matching task acts principally on non-specific object processes. We found modest amounts of overlap, both in terms of activity and spatial overlap, between object-specific activity, however the magnitude of the overlap seems to be insufficient to support a the continuous topological arrangement that has been proposed (Ishai et al., 1999).

In addition, we sought to highlight the benefits of this new approach in examining functional imaging data. One key idea in our analysis is that activation patterns can be decomposed into components, some of which are more informative with respect to particular stimulus categories, and linear discriminant analysis can be used to extract these informative components. A second important concept is that of activity space in which voxels, not acting alone and without spatial constraint, can participate in concert for representing and processing of information. Given the highly interconnected nature of the brain, it is our belief that this brain-based analysis method holds promise for future functional imaging research.
METHODS

Data
Data was acquired from the FMRIDC database (accession no. 2-2000-1113D. [Ishai, A., Ungerleider, L. G., Martin, A., Haxby, J. V. (2000) “The Representation of Objects in the Human Occipital and Temporal Cortex.” Journal of Cognitive Neuroscience, U.S.A., 12 Supplement 2, pp. 35-51.]) A summary of the relevant information from the original study will be presented below to facilitate interpretation of the results (for detailed methods see Ishai et al.)

Subjects
Six right-handed subjects with normal vision participated in the experiment. One subject's data, Subject 9 had compression errors that resulted in a significant loss of data. The remaining five subjects represent the data subjected to the analysis presented in this report.

Stimuli
Stimuli were gray scale photographs of faces, houses, and chairs presented on a gray background. Control stimuli for the photographs were phase-scrambled images of the respective stimuli that preserved spatial frequency information.

Experimental Procedure
Subjects performed one of two tasks, a delayed matching and passive viewing task. Stimuli presented during passive viewing were presented at a rate of 2 photographs/second. In the delayed matching task, target stimuli were presented for 1.5 seconds, then after delay of .5 seconds, were presented two alternative stimuli for 2 seconds. Subjects made a key press response to indicate their choice. Task difficulty was equated across the object categories, which was confirmed both during functional imaging acquisition and psychophysically. The subjects task, delayed matching and passive viewing, alternated between runs. Each run consisted of alternating 21-second blocks of a stimulus category (houses, faces, chairs) followed by a 21 second block of the respective phase scrambled control. Two blocks of each of the three classes of objects were presented in each run, and were counter balanced across a total of six runs for each of the experimental conditions.

Data Acquisition
A 1.5-T General Electric Signa scanner with a whole head RFD coil was used for data acquisition. 18 contiguous coronal slices were obtained during functional acquisitions (TR = 3 sec, TE = 40 msec, FOV = 20cm, 64 X 64 matrix, voxels size 3.125 X 3.125 X 5mm). For each subject, a high-resolution whole head anatomical was acquired in a separate session (124, 1.5 thick sagittal slices, TR = 13.9, TE = 5.3, FOV = 24 cm, 256 X 256 matrix)

Data Preprocessing by Ishai et. al.
The acquired dataset were EPI scan volumes registered with an iterative method (Woods, Cherry, & Mazziotta, 1992), spatially smoothed in plane with a gaussian filter (FWHM was 3.75mm along the x and y axis), and ratio normalized to the same mean global intensity.
Additional Data Preprocessing

To prepare the data for linear discriminant analysis two additional pre-processing steps were taken. A voxel-by-voxel deconvolution operation was performed on individual run acquisitions to remove the effects of hemodynamic blurring and facilitate labeling of time points. The deconvolution filter was given by a gamma function:

\[ h(t) = \frac{(t/\tau)^{n-1}}{\Gamma(n-1)} \exp(-t/\tau) \]

using parameter values \( n = 3, \tau = 1.25 \) in good agreement with empirical measures (Friston et al., 1998). It is important to note that this operation is certain to be imperfect given the complexities of the relationship between the MR signal and the neural response; thus local correlations between adjacent time acquisitions may still exist. Each of the runs was then normalized and the voxel-by-voxel mean was removed. The six experimental runs, for each of the experimental tasks, were then appended into a single volume time course. Two data reduction methods were used to reduce the number of variables below the number of recorded time samples. Initially, the dimensionality of the data equals the number of voxels. Thus, each time acquisition has 73,728 (64 X 64 X 18) dimensions. Voxels in sub-cortical regions and areas outside the brain were masked and excluded from further analysis, which removed approximately 80% of the total voxels. Next, principle component analysis was performed using singular value decomposition (SVD) to further reduce the dimensionality of the data into a number of component variables less than the 84 time samples (7 acquisitions per block X 2 blocks per run X 6 runs = 84 acquisitions per object category). Friston et al. have suggested retention of all components whose variance is greater than the mean variance (Friston et al., 1996). However, this procedure was found to retain too many components for this data set, yielding overfit and lack of generalization. We used 40 components, approximately half the number of time samples for a class of objects, which was sufficient to capture 80-85% of the total variance in the data.

Linear Discriminant Analysis:

Each time acquisition was given one of 6 labels according to its corresponding stimulus type (Chairs, Faces, Houses, Chairs-noise, Faces-Noise, House-Noise). Let \( X \) denote the data matrix, with rows given by the masked voxels and columns representing the time samples. As noted above, we performed dimensionality reduction by projecting \( X \) onto the principal components basis \( B \) yielding a reduced data matrix:

\[ Y = B X \]

LDA produces a set of orthogonal axes in the data space termed canonical variates or discriminants that best separate class means given the within-class covariance.

The fit is only meaningful if it can be validated against overfit. We validated the fits by measuring the predictive performance of the linear classifier using the 632+ bootstrap (Efron et al., 1997), a state-of-the-art leave-one-out resampling procedure. In this procedure, each time point is removed from the dataset. The discriminant is fit using a random subset of the remaining data, and the time point of interest is then subsequently classified. By repeating this procedure
for numerous random subsets and all the time points, we obtain a measure of the generalizability of the classifier that mitigates against overfitting. It is important to note that individual time acquisitions are tested using data from within the same block, data from other blocks within the same run, and data from separate run acquisitions, which results in 83% of the data coming from separate run acquisitions. This is important in that the aforementioned temporal correlations remaining from the de-convolution operation are minimized because the majority of the information used to fit the discriminant comes from acquisitions in separate runs.

The number of distinct discriminants is one less than the number of classes compared. For two class comparisons, the discriminant $L$ has a particularly simple form:

$$L = - (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2)$$

where $\mu_i$ and $\Sigma_i$ are the mean and covariance of the $i^{th}$ group. Thus, the best direction to distinguish two classes is the direction along the difference between the means, rotated by the inverse sum of the within class variances. Because mean and covariance estimates are extremely susceptible to contamination by outliers, we computed robust estimates of the mean and covariance using Rousseeuw’s minimum covariance determinant method (Rousseeuw & Van Driessen 1999). The robust estimates resulted in approximately 10% of the time points in each class being rejected as outliers.

For each of the conditions in the experiment, delayed matching and passive viewing discriminants were derived to identify patterns of activation significant to objects classes. Comparisons included testing each object class against the respective phase scrambled control, each object class vs. the other two object classes (houses vs. faces and chairs, faces vs. houses and chairs, and chairs vs. houses and faces), and pairwise contrasts containing all possible permutations of the three object categories.

**Prediction Error**

Although linear discriminant analysis provides axes that best distinguish classes, it does not directly provide a measure of the classification performance. However, if we assume that the data are well described by a multivariate Gaussian distribution\(^6\), we can perform classification within a Bayesian framework.

Given we project the data onto the linear discriminant vector $L$, the data given the class-conditional distribution of the projected data is given by:

$$P(z_i | L, \mu_i, \Sigma_i) = \frac{1}{C} \exp\left(\frac{-1}{2} z_i^T (L^T \Sigma_i L)^{-1} z_i\right)$$

\(^6\) Data was found to be well approximated by multivariate Gaussian distribution. In multidimensional space, the distance of individual samples from the mean follow a chi-square distribution. We found that these distances were well described by the chi-square distributed with a mean correlation coefficient of 0.897 and a standard deviation of 0.029 between the distances and the chi-square distribution.
Where $C$ is a normalization constant, and $z_i = L^T (y_i - \mu_i)$. Classification response is then based on the ratio of posterior distributions

$$r = \frac{P(z_i | i, \mu_i, \Sigma_i)\pi_i}{P(z_j | j, \mu_j, \Sigma_j)\pi_j}$$

where $\pi_i$ represents the prior probability of the class, and class $i$ is chosen over class $j$ when $r$ is greater than one. Because we used data to derive our discriminant, it is important to have a procedure to test the data that is not part of the training set, to test predictability. We estimated the prediction error using the 632+ bootstrap procedure (Efron & Tibshirani, 1997), in which testing is performed on a set of data points excluded from the training via resampling. Prediction (classification) accuracy is reported as the ratio of the correct classification of labeled time points to the total number of time points tested.

**Discriminant knockout**

In order to test how much of the discrimination for one pair of classes is affected by the removal of the directions required to discriminate another pair, we systematically “knocked-out” via projection the best discriminants for a pair of classes until prediction performance was within 5% of chance. Knockouts were performed via a simple matrix, constructed as follows. An orthogonal basis was constructed from the discriminant axis using SVD:

$$L = USV^T$$

If $L$ is an $N$ length column vector, the matrix $U$ is an $N$ by $N$ orthonormal rotation matrix that has $L$ as its first column. We construct a new matrix $U_d$ by replacing the first column with zeros. Then the knockout matrix $K_L$ is given by:

$$K_L = U^T U_d$$

Multiplying $y$ by the reduced data matrix $K_L$ resulted in a new data set in which the discriminant axis is removed via projection.

**Spatial Overlap**

Spatial overlap between discriminants was computed as the ratio of the intersection and the union of non-zero voxels that survive a thresholding procedure. The voxels $v_m$ significant for a discriminant were computed as:

$$v_m = \{v | d(v)_{\max(d)} > \gamma \},$$

where $d = BL$ is the discriminant $L$ transformed back into voxel space and $\gamma$ is the threshold. In words, the discriminants were scaled between –1 and 1 and the voxels greater than the threshold were selected. Using the same threshold for all the discriminants, the proportion overlap is
computed as the ratio of the number of surviving voxels in the intersection to the union of the discriminants. Thus, for discriminants $j$ and $k$, the proportion overlap $p_v$ is given by:

$$p_v = \frac{\#(v_j^i \cap v_m^k)}{\#(v_j^i \cup v_m^k)}$$

In addition, for each $\gamma$ the prediction accuracy could be computed, by thresholding the discriminant (setting to zero all voxels whose absolute value is less than the threshold) and recomputing the prediction accuracy for each value. The overlap plots were generated by plotting proportion overlap against the recomputed prediction accuracy for corresponding values of the threshold as the threshold varied from 0 to 1.

**Appendix: Relations between the analysis in Haxby et.al (2001) and LDA.**

Let $X$ denote the data matrix as above. For a voxel-wise regression analysis, a model of the form

$$X_n = G\beta_n$$

is fit, where $X_n$ is the $n^{th}$ voxel’s time series, $G$ is a matrix of regressors in which each column is a vector that models the expected temporal response to one of the stimuli, and $\beta$ is the vector of regression coefficients. The least square estimate of $\beta_n$ is given by:

$$\beta_n = (G^TG)^{-1}G^TX_n$$

When the regressors in $G$ are non-overlapping blocks, then the regression coefficients are weighted averages of the time points associated with each class. For example, in an experiment with two stimulus classes, stimulus class $a$ presented in the first half of $2n$ time points and stimulus $b$ in the second half, then the regression matrix would look like:

$$G = \begin{pmatrix}
a_1 & 0 \\
a_2 & 0 \\
\vdots & \vdots \\
0 & b_{n-1} \\
0 & b_n
\end{pmatrix} = \begin{pmatrix}
a & 0 \\
0 & b
\end{pmatrix}$$

where $a$ and $b$ are length $n$ vectors of ones. In this case, the matrix pseudo-inverse reduces to

---

Footnote: Block overlap makes results in the overlapping time points to contribute to several categories. For a small number of overlapping points, the discussion that is true for non-overlapping blocks still holds approximately.
\[
(G^T G)^{-1} G^T \begin{pmatrix} a^T \\ \|a\|^2 \\ 0 \\ 0 \\ b^T \\ \|b\|^2 \end{pmatrix} = \begin{pmatrix} w_a \\ 0 \\ 0 \\ w_b \end{pmatrix}
\]

where \( w_a = (1/n, 1/n, \ldots, 1/n, 0, \ldots, 0) \) & \( w_b = (0, \ldots, 0, 1/n, 1/n, \ldots, 1/n) \) when \( a \) and \( b \) are vectors of ones.

For non-overlapping blocks,

\[
\beta_n^a = (w_a \cdot X_n) = w_a \cdot X_n = \mu_n^a
\]

where the last equality only holds for \( w_a = (1/n, 1/n, \ldots, 1/n, 0, \ldots, 0) \). However, the result is that the regression coefficient for the \( n \)th voxel for a stimulus class \( a \) is just the (weighted) average of the time points in which stimulus class \( a \) was presented.

The correlation based prediction used by Haxby et. al. (2001) can be related to the LDA procedure. Specifically, the correlation measure they use can be viewed as the expectation of a discriminant that is given by the difference between normalized class means. The correlation measure, where angle brackets denote an inner product across voxels, the spatial correlation between the regression coefficients for classes \( a \) and \( b \) is given by:

\[
\rho_{ab} = \frac{\langle \beta_a, \beta_b \rangle}{\sqrt{\langle \beta_a, \beta_a \rangle \langle \beta_b, \beta_b \rangle}} = \frac{\langle \mu_a, \mu_b \rangle}{\|\mu_a\| \|\mu_b\|}
\]

For pairwise comparisons, their classification procedure involves comparing the correlations between classes \( a \) and \( b \) vs. within classes on odd and even runs. Specifically, they compute four correlations: \( \rho_{aa}^{oe}, \rho_{ab}^{oe}, \rho_{ab}^{oe}, \rho_{bb}^{oe} \), where the subscripts denote the first and second classes, and the superscripts whether the regression coefficients were computed from odd or even runs (\( o \) or \( e \)). For example, \( \rho_{ab}^{oe} \) is the correlation coefficient between regression coefficients computed for class \( a \) from odd runs, and class \( b \) from even runs. They consider four comparisons, \( \rho_{aa}^{oe} > \rho_{ab}^{oe}, \rho_{aa}^{oe} > \rho_{ab}^{oe}, \rho_{ab}^{oe} < \rho_{bb}^{oe}, \rho_{ab}^{oe} < \rho_{bb}^{oe} \), and the predictive performance for each subject was computed as the proportion of these pairwise comparisons that had the relationship indicated by the inequality signs. Each one of these comparisons is equivalent to comparing whether the difference between the correlations is greater or less than zero:
\[ \rho_{\alpha a}^{oc} > \rho_{\alpha b}^{oc} = \rho_{\alpha a}^{oc} - \rho_{\alpha b}^{oc} > 0 \]

\[ = \frac{\mu_{\alpha a}^c}{\| \mu_{\alpha a}^c \|} \cdot \frac{\mu_{\alpha a}^c}{\| \mu_{\alpha a}^c \|} - \frac{\mu_{\alpha b}^c}{\| \mu_{\alpha b}^c \|} \times \frac{\mu_{\alpha b}^c}{\| \mu_{\alpha b}^c \|} > 0 \]

\[ = \frac{\mu_{\alpha a}^c}{\| \mu_{\alpha a}^c \|} \cdot \frac{\mu_{\alpha b}^c}{\| \mu_{\alpha b}^c \|} > 0 \]

\[ = \frac{\mu_{\alpha a}^c}{\| \mu_{\alpha a}^c \|} \cdot w_{ab} > 0 \]

\[ = \frac{1}{\| \mu_{\alpha a}^c \|} E_t[\langle w_{ab}, X_{\alpha}^o \rangle] > 0 \]

where \( E_t \) denotes a temporal average. When the class means are similar in magnitude, the normalization constants factor out and the procedure is equivalent to using the difference between class means as a discriminant axis.
References


