The following document is the final, peer-reviewed version of the manuscript:

fMRI "Brain reading": Detecting and Classifying Distributed Patterns of fMRI Activity in Human Visual Cortex

David Cox and Robert Savoy (2003) NeuroImage 19 (2): 261-270

The full, typeset version can be obtained at the NeuroImage website (or by searching on Google… but I didn't just tell you that).

Elsevier prohibits the posting of the final typeset version, and they have recently begun issuing DCMA take-down notices to enforce this policy. Because they are assholes.

Apologies for any inconvenience.

fMRI "Brain reading": Detecting and Classifying Distributed Patterns of fMRI Activity in Human Visual Cortex

David Cox [1,2]
Robert Savoy [1,2]

1. The Rowland Institute for Science, Cambridge, MA 02142
2. The MGH/MIT/HMS Athinoula A. Martinos Center for Structural and Functional Biomedical Imaging, Charlestown, MA 02129

Please address correspondence to:

David Cox
davidcox@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Avenue E25-242
Cambridge, MA 02139
(617) 452-2908
fax: (617) 253-2964

Abstract

Traditional (univariate) analysis of Functional MRI data relies exclusively on the information contained in the time course of individual voxels. Multivariate analyses can take advantage of the information contained across space, from multiple voxels. Such analyses have the potential to greatly expand the amount of information extracted from fMRI data sets.

In the present study, multivariate statistical pattern recognition methods, including Linear Discriminant Analysis and Support Vector Machines, were used to classify patterns of fMRI activation evoked by the visual presentation of various categories of objects. Classifiers were trained using data from voxels in predefined regions of interest during a subset of trials for each subject individually. Classification of subsequently collected fMRI data was attempted according to the similarity of activation patterns to prior training examples. Classification was done using only small amounts of data (20 seconds worth) at a time, so such a technique could, in principle, be used to extract information about a subject's percept on a near real-time basis.

Classifiers trained on data acquired during one session were equally accurate in classifying data collected within the same session and across sessions separated by more than a week, in the same subject. Though the highest classification accuracies were obtained using patterns of activity in lower visual areas as input, classification accuracies well above chance were achieved using regions of interest restricted to higher-order object-selective visual areas.

In contrast to typical fMRI data analysis, in which hours of data across many subjects are averaged to reveal slight differences in activation, the use of pattern recognition methods allows a subtle ten-way discrimination to be performed on an essentially trial-by-trial basis within individuals, demonstrating that fMRI data contains far more information than is typically appreciated.

Introduction

The idea that the activity of a population of neurons in the brain represents some aspect of the external sensory world is as old as neuroscience itself. From the earliest single unit recording experiments through the relatively recent invention of fMRI, one of the predominant themes in neuroscience has been the development and understanding of the relationship between the activity of neurons and the sensory world.

The human brain is capable of representing an almost limitless collection of complex visual objects. This ability extends from the most common of everyday objects to objects that have never before been seen or imagined. However, while some of the details of the basic representational architecture of early visual cortex are known (e.g. retinotopy, hypercolumns, etc.(Hubel and Wiesel, 1968; Hubel and Wiesel, 1969)), relatively little is known about how the higher-order visual cortex represents complex real-world visual objects and the conjunctions of features that comprise them.

There exists a continuum of possible coding schemes that could be used to represent complex objects in the brain, ranging from highly localized architectures ("grandmother" coding) where individual functional units are used to represent individual classes of stimuli, to fully distributed schemes, where all functional units participate in representation, and it is the relative pattern of activity that counts. It is not clear where along this continuum human extrastriate cortex lies, and it is possible that the answer may be different at different spatial scales (e.g. a representation could be mostly localized when considered at the scale of large region of cortex, but distributed at finer scales, or vice versa). A debate surrounding this question has emerged recently based on fMRI evidence for the relative modularity (Downing et al., 2001; Kanwisher et al., 1997) or distributed-ness Haxby et al., 2001; Ishai et al., 1999 of activity in human ventral extrastriate visual cortex.

Functional MRI as a is a technology well-suited to asking questions about representation in the human brain, as it offers a noninvasive window onto brain function with whole-brain coverage and reasonable spatial resolution. However, most commonly used analysis methods for fMRI data are ill-suited to dealing with distributed patterns of activity. fMRI data is fundamentally multivariate (that is, a single fMRI acquisition in time contains information about the local brain hemodynamics at thousands of locations), yet fMRI data is almost always analyzed in an essentially univariate way, treating each voxel as a separate entity as far as statistical analysis is concerned. While this is a natural way to seek functional localization, such an approach by definition ignores the interrelationships between the activity at different locations and the possibility that the variables and organization of interest may not have a one-to-one correspondence with the voxels in an fMRI data set. A variety of multivariate techniques have been applied to fMRI data (Friston et al., 1999; McIntosh et al., 1996; McKeown et al., 1998), but to date, relatively few of these efforts have been aimed at studying fine-grained questions about how the brain represents different classes of stimuli (with Haxby et al., 2001, and more recently, Spiridon and Kanwisher, 2002 , being notable exceptions).

The purpose of the present investigation is to apply a new family of multivariate techniques directly to the problem of object representation in fMRI. Statistical Pattern Recognition algorithms are designed to learn and later classify multivariate data points based on statistical regularities in the data set. Fundamentally, pattern recognition algorithms operate by dividing up a high dimensional space into regions corresponding to different classes of data. This and other multivariate approaches are powerful because

they can potentially discriminate between different classes of multivariate data even when the data, as projected along any individual dimension, are statistically indistinguishable (see Figure 1 for examples). Pattern recognition techniques have been applied to a wide range of practical problems, from face recognition (Papgeorgiou and Poggio, 1999) to the analysis of DNA microarray data (Brown et al., 2000). In this study, we applied several pattern recognition techniques, including relatively new Support Vector Machine Classifiers, to a body of fMRI data in order to begin to gain insight into how the brain represents complex visual objects.

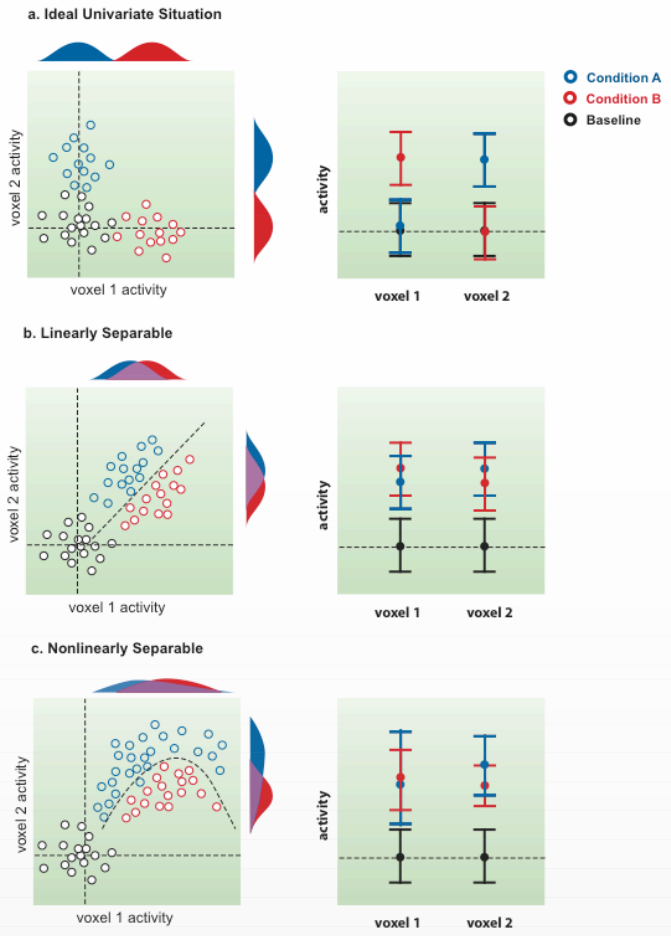


Figure 1. Multivariate Analysis Basics: Most fMRI studies use univariate analyses, in which the data at each voxel (across the many image acquisitions in a given run or session) are treated independently from the data at other voxels. In Figure 1a, the hypothetical activity from two voxels in an ideal univariate situation is shown. On the left, the activity for voxel 1 is plotted against the activity for voxel 2, and each circle corresponds to one acquisition, with the color indicating to which condition each acquisition corresponds. The colored, filled curves on the right and above this plot represent projections of the distribution of the data onto the vertical (voxel 2) and horizontal (voxel 1) axes, respectively. Note that these projected distributions do not overlap, and thus a univariate analysis discriminates between the two conditions perfectly. On the right, the same data is summarized in a traditional univariate manner, with the activity at each voxel plotted separately. Error bars indicate some measure of the spread of the data in this hypothetical example. Voxel 1 is significantly more active during Condition A than baseline, but is not active during Condition B. Similarly, Voxel 2 is active during Condition B, but not during Condition A. There is a clear mapping between location (voxel) and condition, and if one looks at the distributions of the data from each condition at either voxel (depicted in blue and red at the margins of the left plot), they are clearly separable.

However, in Figure 1b the situation is slightly more complicated. Both voxels are now active relative to baseline during both Condition A and B, and there is no clear mapping between voxel and condition. Furthermore, if one looks at the distributions of data for each condition for each, there is significant overlap, and the conditions might be difficult to distinguish without large amounts of data. However, if one looks at both dimensions simultaneously, it is obvious that the data from each condition occupies a distinct region of the two dimensional space and that one need only draw a line between the data clouds to distinguish them. Linear classifiers can perform this basic partitioning of space with arbitrarily many dimensions (or in this context, voxels) and arbitrarily many classes of data (in this context, conditions), drawing hyperplanes between different classes of data. Nonlinear classifiers take this basic idea one step further by allowing decision boundaries to take forms other than straight lines or hyperplanes. Figure 1c shows an example of a nonlinear decision boundary.

In the context of brain imaging, a pattern recognition problem is at its root a "brain-reading" problem. That is, given a pattern of brain activity across space at a given point in time as measured by fMRI, a pattern recognition approach seeks to infer what percept a subject was experiencing. The basic scheme for classification is as follows (see Figure 2). A subject is shown blocks of various categories of visually presented objects while in the scanner. fMRI volumes are acquired while the subject looks at the objects, and the pattern of activity over an independently selected set of voxels is extracted (see *Feature Selection* in *Materials and Methods*). This pattern is then given to a classifier, along with a label that identifies the category corresponding to the stimulus the subject was viewing, and the classifier learns a mapping between patterns of brain activity and stimulus categories Then, in an independent imaging session, (separated by as much as several weeks) the same subject views the same categories of objects with either the same (Experiment 1) or different (Experiment 2) exemplars. Functional MRI volumes are collected with nearly identical spatial sampling, and the same voxel subset is extracted. These patterns of activity are then given to the trained classifier that attempts to infer the category of objects the subject was viewing.

In the present investigation, we attempted to classify the category of object that a subject was looking at (out of 10 possible categories, including similar categories, such as horses and cows) using only very small amounts of data (20 seconds worth, roughly corresponding to the timescale of the hemodynamic response function).

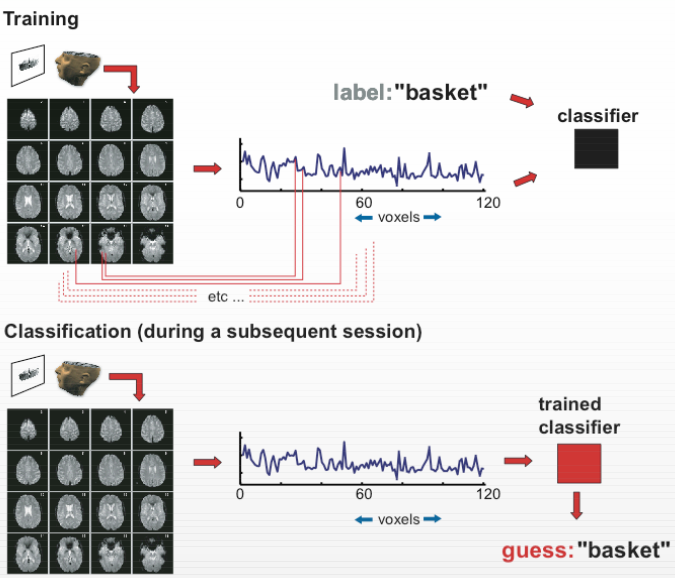


Figure 2. General classification scheme: Each subject viewed 20-second blocks of images belonging to one of ten categories. The pattern of activity over a previously selected subset of voxels is treated as a high-dimensional vector, shown here as a profile plot (the mapping between the voxels in the brain and points in the profile plot is symbolized by red lines). These vectors, along with labels corresponding to the category, are given to a classifier, which learns statistical regularities in the patterns. In a subsequent session (separated in time by as much as a month), the same subject is shown the same category of objects. fMRI data is collected with the same spatial sampling, and the pattern over the same voxel subset is extracted. This pattern is given to the trained classifier that infers the category the subject was viewing.

Materials and Methods

fMRI Data Acquisition and Subjects

Data were collected using a 3 Tesla Siemens Allegra head-only MRI scanner (21-24 trans-axial slices for whole-brain coverage, 3.125 x 3.125 mm in-plane, 5 mm thick, TR=2sec, TE=30msec, GRE EPI) using a custom-designed thermoplastic head-restraint

system that permitted repeatable subject placement to within a few millimeters over months of scanning. Four subjects participated in a varying numbers of sessions (S1: male aged 52, 8 sessions, S2: male aged 23, 8 sessions, S3: male aged 44, 2 sessions, S4: female aged 22, 3 sessions). The number of subjects was intentionally small because we wished to emphasize the power of massively repeated imaging of individuals to elucidate principles of organization, rather than generalizations of structure-function relationships across subjects.

Basic Experimental Paradigm

The fundamental strategy of the present study is to use one set of fMRI responses to stimuli (the "training" fMRI data) to generate classifiers that are used to group subsequent fMRI responses evoked by the same or similar stimuli (the "test" fMRI data). In Experiments 1-3, subjects were shown images belonging to various categories organized into 20-second blocks, with 20-second fixation blocks at the beginning, middle and end of each run. During each block, 10 stimuli were shown for 2 seconds each. Subjects were instructed to view the objects and to covertly name the object category as each individual object was presented. All images contained a fixation cross, and subjects were instructed to maintain fixation throughout.

In Experiment 1, subjects were shown blocks of gray scale images of baskets, birds, butterflies, chairs, cows, tropical fish, garden gnomes, horses, African masks, and teapots. This set of categories of objects was chosen to span a wide variety of different kinds of objects: living/nonliving, common/uncommon, large/small etc. Each block consisted of 10 presentations drawn randomly without replacement from a set of 12 exemplars of each category. The same sets of objects were used for both the training and test sets.

In Experiment 2, blocks of the same categories of objects were used as in Experiment 1, but the set of 12 exemplars was divided into two 6-exemplar sets, and different sets were used in training and test runs.

Data Preprocessing

Images were motion-corrected using the image registration software from SPM99. In all cases, subject motion was less than 0.5 mm in any given imaging run. No explicit temporal or spatial smoothing was performed, nor were data transformed into Talairach space prior to analysis. All analyses were performed within individual subjects.

Individual runs were normalized by subtracting the mean activity during the three 20 second fixation blocks included in each run in order to help mitigate overall differences in fMRI signal across runs and sessions.

Data from individual 20 seconds blocks were averaged together over a window starting 2 acquisitions after the beginning of the block and ending at the end of the block, to account for the hemodynamic delay. Each averaged block was treated as a single "trial" for the purposes of training or testing the classifiers.

Feature (Voxel) Selection

The performance of pattern recognition applications typically depends a great deal on the number and quality of the variables (in this case, voxels) that are given to the

classifier. Variables that contain little information about the discrimination being made only add unrelated noise to the classifier and degrade performance. Similarly, variables that contain information that is redundant with other variables already being considered add little to accuracy of classification and can even seriously impair classification (an effect known as the "Hughes Phenomenon", Hughes, 1968). As a result, many pattern recognition applications contain a "feature selection" step in which a subset or composite set of all available variables is selected that contains enough information to perform the classification, but not so many as to degrade classifier performance.

In the present study, "feature selection" is equivalent to voxel selection. A 21 slice, 64x64 matrix volume contains 86,016 voxels, roughly 27,000 of which actually contain brain tissue. In order to perform pattern recognition more effectively, one can find a reasonable subset of these voxels to feed to the classifier. Ideally, we should have some a priori reason for believing that a given set of criteria for voxel selection will provide information about a given discrimination. Two voxel (feature) selection algorithms were independently tested in the present study.

In the first method, a univariate one-way ANOVA was performed on the data from the training set only. This procedure identifies voxels that vary significantly across at least one of the categories of stimuli ($p < 0.05$, Bonferroni corrected for multiple comparisons). Voxels selected via this method were found throughout visual cortex. In particular, clusters of voxels were found in early visual areas, near the calcarine fissure (average center of mass across all subjects in Talairach coordinates: 3.25 mm right, -84.1 mm anterior, -4.3 mm superior; average size: 7297 mm^3), spread across the lateral occipital gyri (right: 43.1, -67.8, -1.9, 2571 mm^3; left: -44.8, -73.6, -2.7, 2193 mm^3), and in inferotemporal cortex (right: 29.6, -46.9, -17.6, 946 mm^3; left: -36.9, -39.3, -18.8, 1224 mm^3). Though other brain areas were likely involved in this task (e.g. left inferior frontal cortex since the task involved covert naming), we did not detect significant variation *between* categories of stimuli in these areas, so they did not enter into the following analysis.In the second method, separate functional localizer runs were performed in which subjects were shown 20-second blocks of whole objects interleaved with 20-second blocks of scrambled versions of the same objects. A correlation analysis was performed using the BrainVoyager 4.0 software package, with a single boxcar predictor ("high" for whole objects and "low" for scrambled objects) convolved with a simulated hemodynamic response function (Boynton et al., 1996). Voxels that were significantly correlated with this predictor ($R < 0.4$, $p < 7.2 \times 10^{-5}$) were deemed "object processing areas" and were used as the voxel subset in subsequent analysis. As expected, these voxels were found in known object processing areas in inferotemporal cortex (right: 31.82, -49.63, -18.3, 645 mm^3; left: -41.5, -51.8, -17.3, 845 mm^3), and in lateral occipital cortex (right: 38.8, -75.5, -9.0, 2291 mm^3; left: -42.4, -69.0, -12.1, 2078 mm^3). Each of these regions likely contains several distinct visual areas, though the boundaries and exact functional roles of these areas is not well understood.

In short, we selected one set of voxels that maximized variance across the 10 categories, and another set of voxels based on independent functional grounds (i.e. whole object selectivity). A hybrid variant on the above (including only voxels which met both criteria) and a variant in which voxels from inferotemporal cortex were selected manually on anatomical grounds were also tried, with similar results.

Classifiers

The choice of classifier in this paradigm is largely a matter of efficiency. Any mathematical machinery that can extract information reliably from a set of data proves that information is present. In all cases, classifier accuracy allows one to place lower bounds on the information content in a data set. Three types of classifiers were independently tested in the present study: a linear discriminant classifier, a linear support vector classifier, and a cubic polynomial classifier. A standard (SVD based) linear discriminant classifier algorithm from the Statistical Toolbox in MATLAB was used (Mathworks, Natick, MA). The other two classifiers belong to a family of statistical learning techniques known collectively as Support Vector Machines. These classifiers have emerged in the past decade as a promising new tool for statistical pattern recognition, due to their excellent performance and generalization abilities(Mueller et al., 2001). The OSU SVM toolbox (http://eewww.eng.ohio-state.edu/~maj/osu_svm/) based on the LIBSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm) was used to perform these classifications. Multi-class classification was achieved through the training of k(k-1)/2 pair-wise binary classifiers which each contributed to the final decision by a simple voting mechanism. Default kernel settings were used for both the linear and polynomial SVM classifiers.

Results

In all cases, it was possible, with accuracies far above chance, to determine what object a subject was looking at, based purely on isolated collections of just 20 seconds worth of fMRI data at a time. Accuracy remained high even when training and test data sets were separated by days or weeks (as in Experiment 1).

Figure 3. Summary of Classification Accuracy: Classification accuracy is shown for all subjects in both experiments according to which classifier was used (linear discriminant classifier, linear support vector machine, or cubic polynomial support vector machine) and whether the classifier was given free access to voxels throughout the brain, or restricted access to voxels in object-selective cortex (see *Feature Selection* in *Methods*). Except for a few results obtained with a linear discriminant classifier, all results were highly significant. P-values were computed relative to the null hypothesis that the classifier was operating at chance level. * $p < .01$, ** $p<10^{-10}$, *** $p < 10^{-20}$

	free voxel subset selection			restricted voxel selection		
10-way, same exemplars	LDC	linear SVM	poly SVM	LDC	linear SVM	poly SVM
S1 (5 sessions, 56 blocks/cond)	51 %***	85 %***	83 %***	15 %*	41 %***	41 %***
S2 (5 sessions, 40 blocks/cond)	29 %***	58 %***	52 %***	14 %*	33 %***	28 %**
10-way, different exemplars						
S1 (4 sessions, 40 blocks/cond)	29 %**	59 %***	53 %***	13 %	29 %**	19 %*
S2 (2 sessions, 20 blocks/cond)	42 %***	82 %***	80 %***	14 %	37 %**	29 %**
S3 (2 sessions, 20 blocks/cond)	52 %***	90 %***	93 %***	21 %*	55 %***	49 %***
S4 (2 sessions, 20 blocks/cond)	56 %***	97%***	88 %***	21 %*	55 %***	49 %***

A summary of the classifier performance for all subjects is shown in Figure 3 (it should be noted that the accuracies presented here are computed in a more conservative manner than in Haxby et al., 2001 and Spiridon and Kanwisher, 2002, see Discussion). Highest classification accuracies were obtained when the ANOVA-based feature selection algorithm was allowed to pick voxels from anywhere in the brain, including low-level visual areas. This is potentially problematic, because high classification accuracy could simple simply be a product of "reading" the images from the strongly retinotopically organized early visual cortex. To partially address this issue, we repeated the analysis using a voxel subset restricted to the non-retinotopic areas (via anatomical criteria or via an independent functional localizer), and classification accuracies remained well above chance.

The general classification procedure described here also succeeds when different stimuli are used in the training and test sets (Experiment 2). The overall classifier performance in this paradigm is shown in Figure 3, lower table. Thus, classifier performance does generalize, at least in part, across different exemplars of a given category of objects. The effects of possible retinotopic organization in higher-order "object" areas, and other caveats are addressed below.

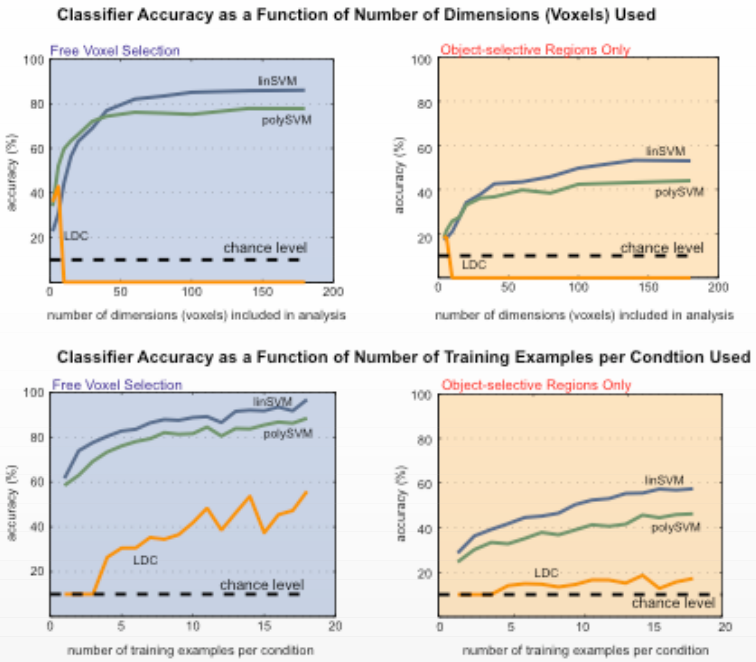


Figure 4. Classifier accuracy as a function of voxel subset size and number of training examples: Classifier performance is shown for one representative subject for three different types of classifiers: Linear Discriminant Classifier (LDC), Linear Support Vector Machine (linSVM), and Polynomial Support Vector Machine (polySVM). The top graphs show accuracy as a function of the number of voxels included in the analysis. The values shown are averages of accuracy for 10 randomly selected subsets of each size. Linear Discriminant Classifier accuracy goes to zero for larger voxel subsets as the covariance matrix becomes singular and classification cannot be computed. The bottom graphs show accuracy as a function of the number of training examples per category available to the classifier. Each value is an average of the accuracies obtained with 10 random subsets of the available training examples. Blue plots on the left were based on voxels drawn from anywhere in the brain, while orange plots on the right were based on voxels restricted to object selective areas (see *Feature Selection* in *Methods and Materials*).

Classifier accuracy as a function of dimensionality (number of voxels) and number of training examples is shown for one representative subject (S4) in Figure 4. In most subjects, an asymptote in performance was observed with ~100 voxels and 10 training examples. Support Vector Machines (Mueller et al. 2001) performed much better than

the linear discriminant classifier used here, particularly when large numbers of dimensions (voxels) were used. This is not surprising since our estimates of the covariance matrix (whose inverse is used to calculate a Mahalanobis distance) became increasingly less general and eventually became singular (however, improved methods of estimating a covariance matrix such as in (Long et al., 2002). Interestingly, however, in spite of many conceivable sources of nonlinearity in neural signals, the nonlinear (cubic polynomial) SVMs used did not significantly outperform their linear counterparts. Whether this apparent linear separability is due to the underlying neural signal itself, its coupling to hemodynamics, or a failure to properly capture the true nonlinear character of the decision boundaries using a cubic polynomial remains to be seen.

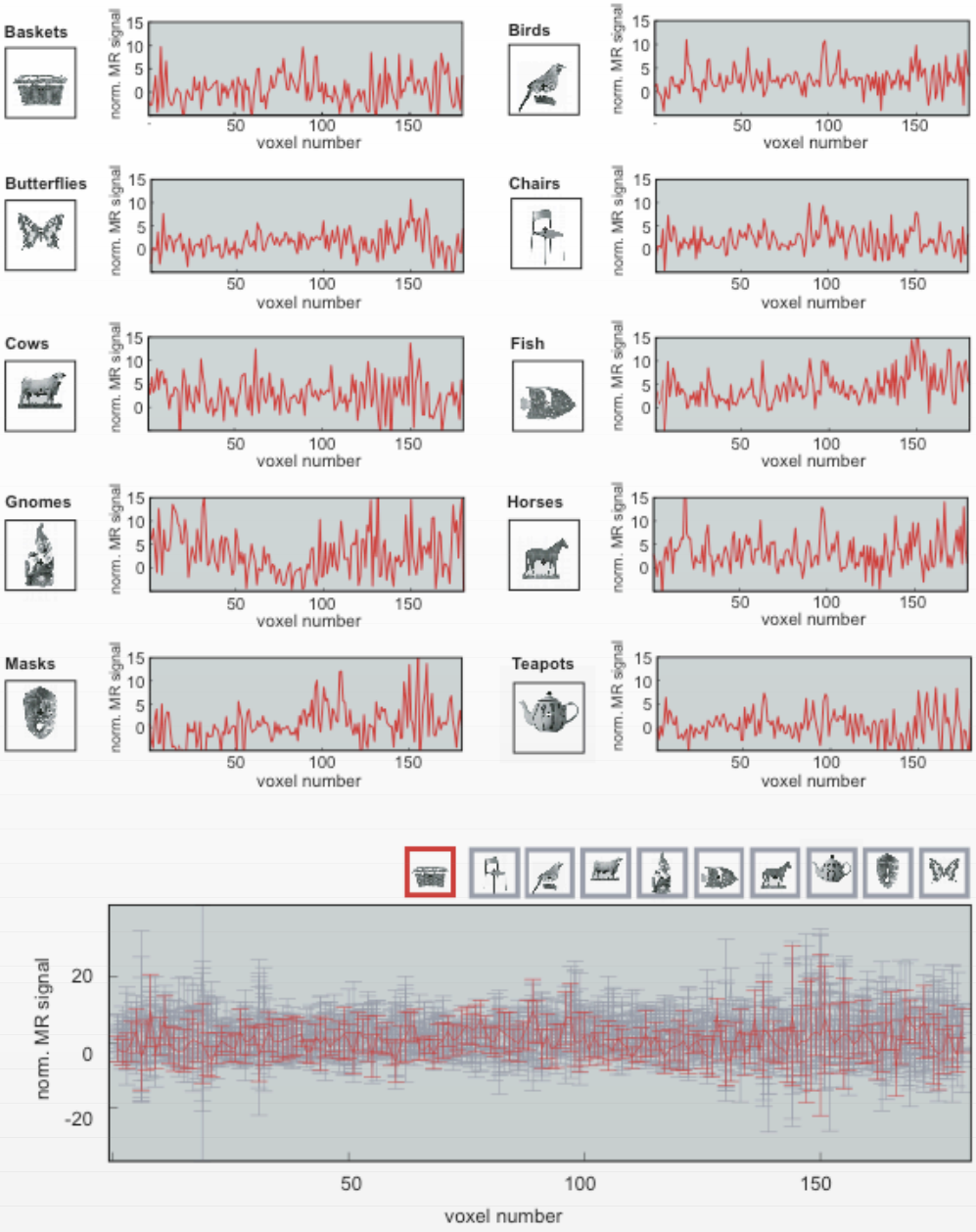
Information Appears to Be Distributed

Figure 5 shows, for each of the ten object categories, the pattern of activity in the 180-voxel subset used for one subject (S4, Experiment 2, using a voxel subset restricted to object selective regions) averaged over all the testing blocks. The bottom portion of Figure 5 shows the same averages for all object categories with their standard deviations across testing blocks, and with the date for one object category (baskets) shown in red. Note that the mean of each category at each voxel is contained within one standard deviation of the mean of every other category. Thus, it is not possible that the classifiers are simply using a single voxel or group of voxels assigned to each class to make the discrimination. Rather, the classifiers must be pooling information across a number of voxels to make the discriminations. It should be reiterated that the classifier is making decisions based on individual 20-second blocks of data at a time, so it does not have the luxury of time-averaging to reduce the noise level.

While it is clear that the classifiers used here are exploiting information distributed across many voxels, this does not by itself necessarily mean that there is *signal* present in all of these voxels. In theory, it is possible that only one or a few voxels contain signal that disambiguates the categories, and the remainder merely contain noise that is correlated with the noise in the signal-bearing voxels. Linear classifiers classify data by orienting hyperplanes that separate different categories of data, and since drawing these hyperplanes amounts to taking linear combinations of dimensions (voxels), the classifiers are capable of subtracting (correlated) noise from signal-bearing voxels (nonlinear classifiers allow for even more complex such arrangements). While it is difficult to completely rule out the possibility that signal is primarily present in a subset of the voxels used by the classifier, we can at least rule out this concern in its most extreme form. When the accuracies of the sort plotted in Figure 4 were computed with hundreds of randomly selected voxel subsets (many of which were non-overlapping), classification accuracy remained fairly stable across different subsets (30% +/- 5% s.d. accurate for subsets of 30 voxels). In all randomly selected subsets, classification accuracy was well above chance. Thus, since relatively high classification accuracies can be achieved from many different small non-overlapping voxel subsets, it is unlikely that only a small subset of voxels is driving the classification performance observed here.

It should also be noted that we did not test some of the most hotly debated categories of objects in this study (such as faces or body parts), so our results do not bring new evidence to bear on the relative distributedness or modularity of these particular classes of stimuli. Nevertheless, for a wide range of other stimuli, distributed information appears to be present in fMRI data.

Figure 5. Average patterns of activity over object selective areas: For each category of objects, the mean pattern of activity over 180 object selective voxels (see *Feature Selection* in *Methods and Materials*) is shown for one representative subject. The average pattern of activity for each category is plotted in red in upper plots. The same average pattern is shown in lower plots, but with standard deviation error bars showing the spread of 20 second averaged blocks (which were treated as individual examples in classification), and with mean and standard deviation for all other categories plotted behind in blue. All voxels have mean activity that is within one standard deviation of all other categories. Since classification was performed on a trial by trial basis (i.e. no averaging across examples could be done to reduce noise), there was no way that the classifiers could have used the activity at individual voxels or regions (i.e. localized "modules") to discriminate between the categories at the accuracy level observed in this study. Instead, the classifiers must leverage entire patterns of activity to perform the discrimination.



Discussion: *The Information Content of Small Quantities of fMRI Data*

The fact that it is possible to gain large amounts of information about which category of objects a subject is viewing from such small quantities of fMRI data is surprising for several reasons. First, while most fMRI experiments pool data collected over many minutes for each of many subjects to find subtle differences in activation across tasks, the present method can infer what category of object a subject is viewing (including categories as similar as horses and cows) using just 20 seconds worth of data. This suggests that, treated as a multivariate entity, fMRI data contains a great deal more information than is typically extracted from univariate analyses.

Second, it is interesting to note that each fMRI voxel is very large relative to the size of a neuron, with each 3x3x5 mm voxel encompassing millions of neurons. Thus, if it is possible to extract information even at this poor resolution, this suggests that patterns of activation across cortex are organized at a relatively coarse scale, as suggested by Haxby and colleagues in what they called "object form topology" (Haxby et al., 2001). It is easy to imagine a scenario where this would not be the case: if nearby neurons or groups of neurons were organized in a "salt-and-pepper" manner, with nearby units doing very different things, fMRI would be unable to detect anything, because any differences in activity would be washed out by the intrinsic averaging over space that takes place in one fMRI voxel. This is not to say that all neural populations within a given fMRI voxel are behaving in a homogenous manner; rather, neurons across extrastriate cortex are organized, at least in part, at a fairly coarse spatial scale.

While this study bears some superficial similarities to a 2001 study by Haxby and colleagues (in both studies, subjects were shown blocks of images, distributed patterns of activation were studied, and "accuracies" were computed) the present work differs in several critical ways from that study. For one, much smaller amounts of data were used as the basic units for classification: 20 seconds in the present study as compared to half of an entire imaging session in Haxby's study. Since 20 seconds corresponds roughly to the time scale of a single hemodynamic event, this rate of classification is arguably close to limit allowed by the low-pass nature of the hemodynamic response. Also, importantly, in Experiment 2, different stimuli were used for training and classification (whereas in Haxby's experiment the same set of stimuli were used throughout), partially allaying fears that results of Haxby et al. were simply artifacts of low-level stimulus similarity (but see Discussion below).

It should also be noted that while Haxby and colleagues reported very high accuracies (> 90%) in identifying which class of stimuli a subject was viewing, a different accounting of percent accuracy was used in each study. The 53% asymptotic accuracy (chance = 10%) shown in Figure 4 (upper right) for object selective areas actually corresponds to 85% accuracy using Haxby's accounting scheme (chance = 50%). Given that the present analysis operates on individual blocks of data, rather than averaging over half of an entire fMRI session (30 minutes worth of data), this represents a substantial increase in information extraction. A comparison of the mutual information extracted (prior entropy minus conditional entropy given the output of the classification scheme) of both paradigms verifies that the present classification scheme can extract information at a bit-rate more than 15 times greater than was possible with Haxby's method. Thus the present study demonstrates not only that distributed information about object perception is present in fMRI data, but that with the right mathematical tools, relatively large quantities of information can be extracted from relatively small quantities of fMRI data.

While the present study focused on distributed patterns of activity in visual cortex, the techniques described here could, in principle, be used to study any cognitive domain where distributed representations might be present. In particular, the techniques laid out here provide a framework for detecting whether information is present in the patterns of activity in a given region of the brain. For instance, one could imagine using similar analyses to seek out distributed information in language areas during language tasks.

Since many fMRI studies are aimed at making inferences about and between groups of individuals, one might reasonably ask whether the classification paradigm described here is useful for the study of groups of subjects. In addition to the within-subject analyses described above, we also attempted using a classifier trained using Talairach-normalized data from one subject to classify blocks of Talairach-normalized data from another subject. However, all such attempts failed.

There are a number of reasons why this might be the case. One possibility is that the relevant regions of cortex in the subjects' brains are not brought into adequate alignment by the Talairach transform, which is notoriously imperfect at precisely aligning cortical structures (Grachev et al., 1999). If this is the case, the application of more sophisticated inter-subject alignment algorithms might help (Fischl et al., 1999). Another possibility is that the classifiers are taking advantage of structure in the data for one subject that is idiosyncratic to that subject, even though there may be additional structure present that is generalizeable across subjects. To remedy this, one could imagine training a classifier on data from many subjects (rather than just one), thus avoiding overtraining on the details of one subject and arriving at a more general distillation of what patterns correspond to different percepts *across* subjects. Yet another possibility is the spatial organization of cortex at this scale of analysis *is* mostly idiosyncratic to individuals, perhaps dependent on each individual's experience with the visual world. Further studies are needed to disambiguate these possibilities.

Even if it is possible to use these kinds of classifiers across subjects, it is not immediately clear what such an exercise would teach us. For the researcher interested in exploring distributed patterns of activity within and across groups of subjects, perhaps a better strategy is to take the present study as a demonstration of the power of multivariate approaches in extracting large quantities of information from fMRI data, and to develop new domain-appropriate analysis techniques to exploit this information.

Information Content and Early Visual Areas

While a pattern recognition approach shows great promise for extracting large amounts of information from fMRI data and for guiding multivariate exploration of representation in the human brain, one must always remain cautious about the nature of the information that a classifier is using to distinguish different classes of stimuli. The fact that information can be extracted by our analysis does not necessarily mean that this information is used by the brain or that the information is used in the way that we think it is. One must always remain conscious of this concern for all analysis techniques that are fundamentally correlational (including traditional univariate fMRI data analysis).

One particular concern in this study is the role of retinotopic information. There is recent evidence that even high-level extrastriate cortex may have some retinotopic organization in humans (Hasson et al., 2002). This could mean that the classifiers described in the current investigation are simply picking up on trivial visual

commonalities between members of a given category of objects. This concern is even graver when we consider that retinotopic information need not be preserved in higher-order visual cortex in an obvious spatially regular way for it to nevertheless provide fundamentally retinotopic information to the kinds of classifiers described in the present paper. For instance, if one considers simply rotating a spatially regular retinotopic representation through an arbitrary axis in high dimensional fMRI image-space, the information can be completely obfuscated to a human observer while still being completely available to a suitably flexible classifier. Statistical pattern recognition methods provide a new way of thinking about fMRI data and can potentially provide a great deal of information about hidden structure in fMRI data (particularly in conjunction with other multivariate methods), but retinotopy and related concerns will ultimately need to be addressed before we can fully understand what classifiers are telling us.

With these caveats in mind, however, the present investigation underscores a deeper level of complexity present in fMRI data than is often appreciated, and it points to directions by which this complexity might be studied and understood. Any complete model of object recognition in extrastriate cortex will need to come to terms with the coarse-scale distributed nature of activity within the ventral visual pathway, and the implications of such organization for underlying neural representation. Probing the nature of distributed patterns of activity will require further development and application of multivariate techniques in order understand how the information contained within patterns of fMRI activity is organized.

References

Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. J Neurosci *16*, 4207-4221.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A *97*, 262-267.

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. Science *293*, 2470-2473.

Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum Brain Mapp *8*, 272-284.

Friston, K., Phillips, J., Chawla, D., and Buchel, C. (1999). Revealing interactions among brain systems with nonlinear PCA. Hum Brain Mapping *8*, 92-97.

Grachev, I. D., Berdichevsky, D., Rauch, S. L., Heckers, S., Kennedy, D. N., Caviness, V. S., and Alpert, N. M. (1999). A method for assessing the accuracy of intersubject registration of the human brain using anatomic landmarks. Neuroimage *9*, 250-268.

Hasson, U., Levy, I., Behrmann, M., Hendler, T., and Malach, R. (2002). Eccentricity bias as an organizing principle for human high-order object areas. Neuron *34*, 479-490.

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

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. J Physiol *195*, 215-243.

Hubel, D. H., and Wiesel, T. N. (1969). Anatomical demonstration of columns in the monkey striate cortex. Nature *221*, 747-750.

Hughes, G. F. (1968). On The Mean Accuracy Of Statistical Pattern Recognizers. IEEE Trans Infor Theory *14*, 55-63.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., and Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. Proc Natl Acad Sci U S A *96*, 9379-9384.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci *17*, 4302-4311.

Long, C. J., Solo, V., Purdon, P., Sperling, R., Dale, A., Greve, D., Lange, N., Albert, M., and Brown, E. N. (2002). A New Framework for the Analysis of Multiple Subject fMRI; Nonsingular Random Effects Modeling in the Presence of Nonstationary Noise. Paper presented at: 8th International Conference on Functional Mapping of the Human Brain (Sendai, Japan, Neuroimage).

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., and Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. Neuroimage *3*, 143-157.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J. (1998). Analysis of fMRI data by blind separation into independent spatial components. Hum Brain Mapp *6*, 160-188.

Mueller, K. R., Mika, S., Raetsch, G., Tsuda, K., and Schoelkopf, B. (2001). An Introduction to kernel-based learning algorithms. IEEE Neural Networks *12*, 181-201.

Papgeorgiou, C., and Poggio, T. (1999). A Pattern Classification Approach to Dynamical Object Detection. Proceedings of ICCV, 1223-1228.
Spiridon, M., and Kanwisher, N. (2002). How Distributed Is Visual Category Information in Human Occipito-Temporal Cortex? An fMRI Study. Neuron *35*, 1157.

Acknowledgments
We thank Michael Burns and Michael Beauchamp for helpful discussions and comments. This work was supported by the Rowland Institute for Science (now "The Rowland Institute at Harvard University") and the Athinoula A. Martinos Center for Biomedical Imaging.