

High-level visual object representations in inferior temporal cortex

Marieke Christina Mur

© Copyright Marieke Mur, Maastricht 2011

Production: Datawyse | Universitaire Pers Maastricht
ISBN 978 94 6159 098 5

High-level visual object representations in inferior temporal cortex

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op vrijdag 2 december 2011 om 12.00 uur

door

Marieke Christina Mur



Promotores

Prof. dr. P. de Weerd

Prof. dr. R. Goebel

Copromotor

Dr. N. Kriegeskorte (MRC Cognition and Brain Sciences Unit, United Kingdom)

Beoordelingscommissie

Prof. dr. B. Jansma (voorzitter)

Dr. M. Bonte

Prof. dr. E. Formisano

Prof. dr. J-D. Haynes (Bernstein Center for Computational Neuroscience,
Germany)

Prof. dr. R. Vogels (Katholieke Universiteit Leuven, België)

Financiële ondersteuning

National Institute of Mental Health, USA

Universiteit Maastricht

“When one teaches, two learn”
(Robert Half)

Table of contents

	Introduction	9
Chapter 1	Revealing representational content with pattern-information fMRI – an introductory guide	19
Chapter 2	Face-identity change activation outside the face system: “Release from adaptation” may not always indicate neuronal selectivity	35
Chapter 3	Representational similarity analysis – connecting the branches of systems neuroscience	77
Chapter 4	Single-image activation of category-selective regions in human inferior temporal cortex	129
Chapter 5	Matching categorical object representations in inferior temporal cortex of man and monkey	149
Chapter 6	Human object-similarity judgments reflect and transcend primate-IT categorical object representations	199
	Summary	223
	Samenvatting	229
	References	237
	Acknowledgments	249
	Curriculum Vitae	253

Introduction

Imagine a usual working day. We wake up, take a shower, eat breakfast, and rush out to be in time for a meeting at work. During these everyday activities, we recognize and interact with many different kinds of objects. These objects might include our toothbrush, a banana, our keys, and potentially a housemate or partner. During the perception of these objects, we rely heavily on vision as a source of information. Visual information about the objects in our environment is carried by light that enters the eyes, and is relayed to the brain for further processing. Consistent with the dominant role of vision in perception, a large part of the human brain - about one fifth of the cortical volume - is devoted to vision (Wandell et al., 2007).

The challenges of object recognition

Despite the ease with which we perceive and recognize objects, the computational task performed by the brain is far from trivial. Objects are not usually seen in isolation, but are part of a cluttered scene and might be occluded by other objects. Furthermore, objects need to be recognized across a variety of circumstances, including differences in viewpoint, illumination, and location. Successful interaction with the objects in our environment requires not only recognition of individual objects but also assignment of objects to behaviorally relevant categories (e.g. animate objects). Categorization enables us to display a similar behavioral response to objects from the same category despite considerable differences in visual appearance among category members (e.g. Vogels et al., 1999). Furthermore, appropriate behavioral responses can be extended to novel objects once they have been correctly categorized (e.g. Edelman, 1997). The significance of the computational challenges posed to and met by our visual system is highlighted by the difficulty of developing computational models that can approach human object-recognition performance (e.g. Kietzmann et al., 2008).

The primate visual system for object recognition

How does the brain accomplish the challenging task of object recognition? Previous studies of the monkey and human brain have provided a substantial amount of knowledge on the primate visual system. Object recognition takes place in a series of processing steps implemented along the ventral visual pathway (Ungerleider and Mishkin, 1982; Ungerleider and Haxby, 1994). This pathway starts in the primary visual cortex (V1), which is the first cortical station for incoming visual information, located in the posterior occipital lobe of the brain. The pathway runs from V1 via extrastriate visual areas V2 and V4 to inferior temporal (IT) cortex. Research in monkeys has shown that neurons along the

ventral pathway respond selectively to visual features that are important for object recognition. The complexity of preferred stimuli increases along the ventral pathway, evolving from oriented lines and edges in V1 (Hubel and Wiesel, 1968) to gratings and line combinations in extrastriate visual areas (Hegd e and Van Essen, 2000; Anzai et al., 2007), to partial or complete object views in IT (Gross et al., 1972; Tanaka, 1996). Furthermore, multiple studies have reported category-selective responses in IT, especially for faces (e.g. Desimone et al., 1984; Tsao et al., 2006). Consistent with these findings at the neuronal level, neuroimaging studies in humans and monkeys, which can reveal large-scale functional organizations in the brain, have shown a retinotopic organization (i.e. resembling the reflection of an image on the retina) in early visual regions (Serenio et al., 1995; Orban et al., 2004) and patches of object- and category-selective cortex in lateral occipital cortex and IT (Malach et al., 1995; Kanwisher et al., 1997; Tsao et al., 2003). Along with the increased degree of selectivity at later processing stages of the ventral visual pathway, there is an increase in response invariance to image transformations, i.e. neuronal responses at the level of IT are fairly robust against changes in position and scale (e.g. Tanaka, 1996; Hung et al., 2005; but see Kravitz et al., 2010).

Several mechanisms have been proposed to underlie the emergence of selectivity and invariance along the ventral visual pathway. These mechanisms include filtering and pooling of visual information along the processing hierarchy (Riesenhuber and Poggio, 2002), and coding of information by ensembles of neurons (population coding) (e.g. Logothetis et al., 1995; Hung et al., 2005). Visual experience, e.g. seeing objects move and transform over time, may play an important role in the acquisition of invariance to image transformations (Folidak, 1991; Li and DiCarlo, 2008), and has been proposed to contribute to invariance to visual clutter and occlusion as well (Stringer and Rolls, 2000).

To conclude, the series of processing steps along the ventral visual pathway results in high-level object representations at the level of IT that are fairly invariant to object transformations and might form the basis for categorization. These high-level representations are on the interface between perception and cognition and serve as input to higher-order cognitive functions, including problem solving and action planning, which are essential for intelligent behavior (e.g. Duncan, 2010). The nature of these IT object representations has therefore been the object of intense study during the past decades.

Measuring high-level object representations in inferior temporal cortex

As evident in the previous paragraphs, research on object representations has been performed in both humans and monkeys (the monkey provides the best animal model for human brain functioning due to her close evolutionary relationship). Research methods used in monkeys include the examination of behavioral effects of induced temporary or permanent brain lesions (e.g. Gross, 1973) and recording of electrical signals from (populations of) neurons (e.g. Vogels, 1999; also see Nicolelis et al., 2003). These methods are invasive and are therefore generally not used in humans, except in special cases, e.g. in the treatment of epilepsy (e.g. Quiroga et al., 2005). Brain research in humans is based on lesion studies in patients and on data acquired with neuroimaging methods, which non-invasively measure the temporal and spatial characteristics of brain activity. One of the most widely used neuroimaging methods is functional magnetic resonance imaging (fMRI) (Bandettini et al., 1992; Ogawa et al., 1992), which offers the possibility to non-invasively “look into the brain” and visualize brain activity with spatial resolution in the millimeter range (Goebel, 2007). The blood-oxygen-level-dependent (BOLD) fMRI signal has been shown to reflect stimulus-driven neuronal responses (Logothetis et al., 2001), supporting its use as a measure of neuronal activity in humans (e.g. Kanwisher et al., 1997) and, more recently, in monkeys (e.g. Tsao et al., 2003). Furthermore, recent technological developments are pushing the spatial resolution of fMRI, moving it into the sub-millimeter range, which enables measurement of brain activity at the level of cortical columns. This development opens up a promising avenue for future research on brain function. It should nevertheless be kept in mind that the relationship between neuronal activity and BOLD fMRI signal is complex: the fMRI signal gives a blurred and distorted reflection of mass neuronal activity (Logothetis, 2008; Kriegeskorte et al., 2010), posing constraints on the conclusions that can be drawn from fMRI data about the exact neuronal mechanisms at work in the brain region under study (see Logothetis et al., 2008).

The measurement units of fMRI are voxels, which can be seen as 3D pixels (usually 3x3x3 mm in size) that together cover the entire brain. fMRI data consists of an activation time-series for each voxel and can be analyzed in multiple ways. Conventional activation-based analysis focuses on finding brain regions that are, as a whole, involved in a certain mental activity (Friston et al., 1994; Friston et al., 1995ab), e.g. face perception. This motivates spatial smoothing of the data and averaging of activity across voxels within a functional region of interest. Activation-based analysis has contributed considerably to the understanding of the large-scale functional architecture of the brain. Nevertheless, the functional

regions identified by activation-based analysis remain black boxes with their contents beyond our reach.

In recent years, the field has developed tools to investigate the representational content of regions, including the fMRI adaptation technique and pattern-information analysis. fMRI adaptation (Grill-Spector and Malach, 2001) compares activation to pairs of either different or repeated stimuli and then infers neuronal population selectivity from these activation differences. This approach can potentially resolve sub-voxel representations, but it only offers an indirect way to target these representations, and it relies on assumptions that have been questioned by recent experimental results (Tolias et al., 2005; Sawamura et al., 2006). The fMRI adaptation technique and its limitations are discussed in more detail in Chapter 2 of this thesis. Pattern-information analysis investigates the information carried by multivoxel patterns of activity within a region (Haxby et al., 2001; Cox and Savoy, 2003; Kriegeskorte et al., 2006). This information, most of which goes undetected by activation-based analysis, can significantly contribute to our understanding of neuronal representations of mental content (Haynes and Rees, 2006; Norman et al., 2006; Kriegeskorte and Bandettini, 2007a). Pattern-information analysis has gained momentum in recent years, as indicated by its many successful applications in neuroimaging (e.g. Haxby et al., 2001; Carlson et al., 2003; Kamitani and Tong, 2005; Haynes and Rees, 2005a; Kriegeskorte et al., 2007). An introduction to pattern-information analysis can be found in Chapter 1 of this thesis.

Current knowledge on high-level object representations in inferior temporal cortex

Early lesion studies in monkeys showed that bilateral removal of IT resulted in impaired object discrimination performance (Gross, 1973). Consistent with the behavioral effects of bilateral IT removal, IT neurons were reported to respond selectively to partial or complete object views (e.g. Tanaka, 1996). As mentioned before, many studies reported IT neurons that showed a preference for images from a particular object category (e.g. Vogels, 1999; Tsao et al., 2006; Kiani et al., 2007). Furthermore, neurons with similar (category) preferences tended to cluster together (Tanaka, 1996; Tsao et al., 2006). Category-selectivity of individual neurons was often not perfect, likely due to variability of visual appearance across category members. Invariance to these within-category differences has been proposed to arise by population coding, i.e. each neuron represents a subset of category members or features and a population of these neurons can represent the entire range of objects within a category (e.g. Vogels, 1999). This is the same computational solution as suggested for the problem of invariance to

image transformations, and its validity is supported by experimental findings (Vogels, 1999; Hung et al., 2005; Tsao et al., 2006; Kiani et al., 2007).

Results from monkey studies have been complemented by results from neuroimaging studies in humans, which revealed several macroscopic brain regions within IT that respond preferentially to certain object classes. These regions are the fusiform face area (FFA) which on average responds more strongly to faces than other objects (Kanwisher et al., 1997; Puce et al., 1995; Sergent et al., 1992), the parahippocampal place area (PPA) which on average responds more strongly to places (scenes and buildings) than other objects (Epstein and Kanwisher, 1998; Aguirre et al., 1998), and the extrastriate body area (EBA) which on average responds more strongly to human body(parts) than other objects (Downing et al., 2001). The presence of category-selective regions is consistent with reports of category-specific deficits in human patients after brain damage (e.g. Rossion, 2008). However, information on category membership is not confined to category-selective regions. Consistent with the idea of neuronal population coding, pattern-information analysis has shown that multivoxel activity patterns across IT contain information on category membership, even after exclusion of category-selective regions from analysis (Haxby et al., 2001; Cox and Savoy, 2003).

In sum, high-level object representations are implemented by neuronal population codes in IT, which likely contain information on category membership of natural objects. Furthermore, neurons with a similar (category) preference seem to cluster together, which can result in category-selective regions that are detectable at the spatial scale of activation-based fMRI. Pattern-information studies have shown that information on category membership at the level of IT is not confined to category-selective regions.

This thesis

Despite the substantial progress that has been made during the past decades in our understanding of the neural underpinnings of object vision, several key questions about the representation of objects in IT have remained unanswered.

How are individual objects represented in IT?

Apart from a few exceptions, previous neuroimaging studies have grouped object images into predefined natural categories (e.g. faces, bodyparts, houses) during design and analysis, assessing only category-average activation. It is therefore unknown how individual objects are represented at the level of human IT. This leaves a range of questions unanswered: Are the predefined categories the only categories represented in IT? Can we recover a natural-category

structure from IT without assuming an a priori grouping of stimuli? How strong is category selectivity for individual objects? Are individual objects distinguished at the level of IT? In order to address these questions, we need to measure and analyze single-image responses (i.e. treat each object image as a separate condition).

Are the representations consistent between human and monkey?

Research in the monkey has contributed greatly to our understanding of the primate visual system. The close evolutionary relationship between human and monkey motivates using monkey IT as a model for human IT. However, comparisons between the species at the level of IT have remained of a qualitative nature, due to the difficulty of defining the correspondency between measurement units (i.e. single cells and voxels). In order to investigate whether the object representations are consistent between human and monkey, we need to abstract from the activity patterns and compare representational similarities, e.g. do objects that elicit similar activity patterns in the monkey also elicit similar activity patterns in the human?

What computational models explain the IT representation?

Brain information processing can be simulated by computational models. These models can be used as a tool to better understand brain function, and can be evaluated by comparison to brain data. As for the man-to-monkey comparison, quantitative comparison of models to brain data has been complicated by the need for defining the correspondency between model units and brain-measurement units (e.g. voxels). It is therefore unclear what computational models can explain the IT representation. This question can be addressed by comparing representational similarities between brain and models.

Do high-level conscious object-similarity judgments reflect the IT representation?

Previous research has indicated that perceived similarity of abstract shapes reflects activity-pattern similarities in IT. This suggests IT as a neuronal substrate for the perceptual representations that give rise to shape-similarity judgments. Does this finding extend to similarity judgements of real-world object images? This question can be addressed by comparing similarity judgments of real-world object images to activity-pattern similarities in IT for the same set of images.

The work described in this thesis investigates the representation of individual objects in human IT using fMRI (chapters 2, 4, 5), introduces a new framework for quantitative comparison of data from different branches of systems neuroscience (chapter 3), and compares the observed object representations in human IT to data from monkey IT, computational models, and human behavior (chapters 5, 6).

Thesis overview

This thesis consists of six chapters. Chapters 1 and 3 are methodological in nature; the other four chapters describe results of original research. A brief overview of each chapter is given below. The thesis will be concluded with an overall summary.

Chapter 1 gives an introduction to pattern-information analysis of fMRI data. It compares pattern-information analysis to conventional activation-based analysis methods, including the fMRI adaptation technique. This comparison is followed by an intuitive explanation of the most widespread methods used in pattern-information analysis (i.e. linear classification techniques) and an outline of the basic sequence of analysis steps to be followed.

Chapter 2 uses the fMRI adaptation technique to localize invariant face-identity representations in human visual cortex. Face recognition is an important function of the human visual system and requires the existence of distinct neuronal representations of individual faces. The fMRI adaptation technique has been widely used within the domain of face perception to localize these face-identity representations. Previous fMRI adaptation studies have suggested the presence of face-identity representations in face-selective regions, but these studies did not thoroughly investigate the specificity of these effects to FFA. We investigate whether face-identity adaptation effects are specific to face-selective regions and examine the effects of changes in viewpoint and illumination on the spatial extent of face-identity adaptation effects. The results of these investigations lead to a discussion on the interpretation of fMRI adaptation results.

Chapter 3 introduces a new experimental and data-analytical framework called representational similarity analysis (RSA), which enables quantitative comparison of data from the three different branches of systems neuroscience: brain-activity measurement, behavioral measurement, and computational modeling. RSA builds on a rich psychological and mathematical literature on similarity analysis. Correspondency problems are solved by abstracting from activity patterns: instead of trying to directly compare activity patterns between brain and model, brain and model are compared on the basis of dissimilarities between activity patterns. A significant correlation between brain and model dissimilarities suggests that stimuli that elicit similar brain activity patterns also tend to elicit similar model activity patterns. Behavioral dissimilarity can be based on explicit similarity judgments, or reaction times or confusion errors in comparison tasks. After a general introduction, RSA is demonstrated by relating representations of visual objects measured by fMRI in early visual cortex and FFA to computational models spanning a wide range of complexities.

Chapter 4 investigates the response selectivity of category-selective regions for individual object exemplars. Category-selective regions have been defined based on activation averaged over category exemplars. It is therefore unclear to what extent category selectivity holds for individual objects. Furthermore, it is unknown whether category-selective regions act purely as a binary classifier or whether they show a graded response profile to objects from their preferred category. We address these questions using an ungrouped-events design (i.e. each object is treated as a separate condition). We measure fMRI activation of category-selective regions to visual objects from a wide range of natural categories and analyze the acquired data using a signal-detection approach.

Chapter 5 compares IT representations of the same particular objects between human and monkey using RSA. The fMRI data set described in Chapter 4 is now analyzed for multivoxel pattern-information within the RSA framework. Monkey data consists of cell recordings provided by Roozbeh Kiani (Kiani et al., 2007). First, the human and monkey IT object representations are characterized using multidimensional scaling and hierarchical clustering methods. These explorative methods give an impression of the inherent categorical structure present in the data. Then, the human and monkey IT object representations are quantitatively related by correlating corresponding activity-pattern dissimilarities. Similarities and differences between the two representations are discussed. The IT object representations are also compared to computational models of varying complexity.

Chapter 6 relates the human IT object representation described in Chapter 5 to human object-similarity judgments of the same particular object images using RSA. Given the relatively large stimulus set (96 object images), conventional methods for obtaining pairwise similarity judgments would require many hours of data acquisition per subject. We therefore developed a new multi-arrangement method that enables time-efficient and subject-tailored acquisition of similarity judgments. Similarities and differences between the brain representations and similarity judgments are discussed.

Chapter 1

Revealing representational content with pattern-information fMRI – an introductory guide

Conventional statistical analysis methods for functional magnetic resonance imaging (fMRI) data are very successful at detecting brain regions that are activated as a whole during specific mental activities. The overall activation of a region is usually taken to indicate involvement of the region in the task. However, such activation analysis does not consider the multivoxel patterns of activity within a brain region. These patterns of activity, which are thought to reflect neuronal population codes, can be investigated by pattern-information analysis. In this framework, a region's multivariate pattern information is taken to indicate representational content. This tutorial introduction motivates pattern-information analysis, explains its underlying assumptions, introduces the most widespread methods in an intuitive way, and outlines the basic sequence of analysis steps.

Mur M, Bandettini PA, Kriegeskorte N (2009) Revealing representational content with pattern-information fMRI – an introductory guide. *Soc Cogn Affect Neurosci* 4, 101-109. doi: 10.1093/scan/nsn044.

1.1 Introduction

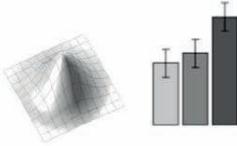
Conventional statistical analysis of functional magnetic resonance imaging (fMRI) data focuses on finding macroscopic brain regions that are involved in specific mental activities (Friston et al., 1994; Friston et al., 1995ab; Worsley and Friston, 1995). In order to find and characterize brain regions that become activated as a whole, data is usually spatially smoothed and activity is averaged across voxels within a region of interest (ROI). These analysis steps increase sensitivity to spatially extended activations, but result in loss of sensitivity to fine-grained spatial-pattern information. In recent years, there has been a growing interest in going beyond *activation* assessment and analyzing fMRI data for the *information* carried by fine-grained patterns of activity within each functional region (Norman et al., 2006; Haynes and Rees, 2006; Kriegeskorte and Bandettini, 2007a). The goal of this tutorial paper is to motivate the use of pattern-information analysis and to provide a step-by-step introduction on how to implement this method.

1.1.1 A region's involvement in task processing versus its representational content

Conventional analysis focuses on regions that become activated as a whole during the performance of a specific task. This motivates spatial smoothing of the data and averaging of activity across an ROI. Since this approach focuses on activations (in the sense of blobs consisting of multiple voxels all showing effects in the same direction) we refer to it as activation-based analysis. Activation-based analysis aims to detect regional-average activation differences and infer *involvement* of the region in a specific mental function. Pattern-information analysis, by contrast, aims to detect activity-pattern differences and infer *representational content* (see Table 1.1, Figure 1.1).

Regional activity patterns can reflect the neural population code (for a striking example, see Kamitani and Tong, 2005). However, fine-grained pattern differences go undetected in activation-based analysis unless the regional-average activation also differs (see Figure 1.1). Pattern-information analysis is suited for detecting pattern changes even if they occur in the absence of regional-average activation changes. For example, a recent study using pattern-information analysis showed that perceptually discriminable speech sounds elicit different patterns of activity in right auditory cortex (Raizada et al., 2010). The speech sounds elicited similar regional-average activation, but the patterns were statistically discriminable.

Table 1.1 Overview of activation-based and pattern-information analysis.

	Activation-based analysis	Pattern-information analysis
		
Goal of the analysis	Investigating the <i>involvement</i> of regions in a specific mental activity	Investigating the <i>representational content</i> of regions
Experimental contrast	Difference between mental activity <i>including</i> component of interest and mental activity <i>excluding</i> component of interest	Difference between representation of object 1 and representation of object 2
Analytical comparison	Compare spatial-average activation across conditions	Compare patterns of activity across conditions
Spatial resolution	Benefits of high-resolution imaging will be limited if data are smoothed	Fine-grained spatial information provided by high-resolution imaging is used effectively
Statistical methods	<ul style="list-style-type: none"> • Spatial smoothing • Combine single-voxel signals by smoothing and averaging activity within ROI • Univariate analysis • Group analysis in common stereotactic space 	<ul style="list-style-type: none"> • No spatial smoothing • Combine single-voxel signals by computing multivariate statistics • Multivariate analysis (typically linear discriminant analysis) • Single-subject analysis in native subject space • Group analysis in common stereotactic space at the pattern-information level

Images in this table are reprinted with permission from Kriegeskorte and Bandettini (2007b).

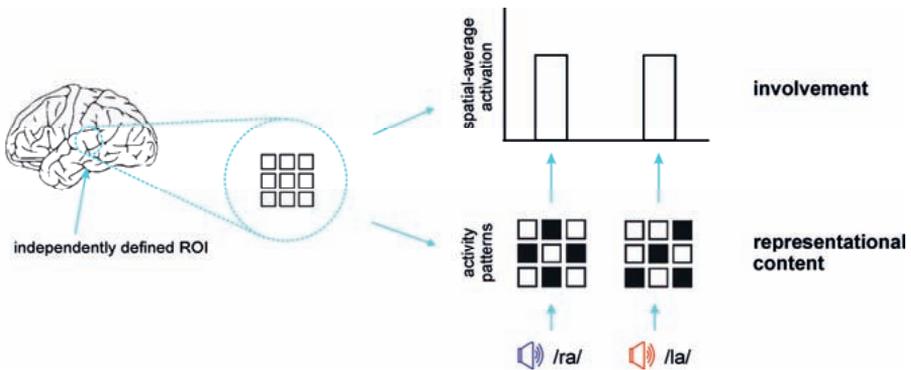


Figure 1.1 Activation indicates *involvement*, pattern-information indicates *representational content*. A specific region of interest (ROI) can show the same spatial-average activation resulting from different patterns encoding different representational content. This figure shows a hypothetical ROI consisting of 9 voxels. The ROI’s multivoxel pattern of activity is different for /ra/ than /la/ speech sounds, but these different patterns result in the same spatial-average activation. This difference will go undetected by conventional activation-based analysis. Pattern-information analysis can be used to show that an ROI’s multivoxel activity pattern differs significantly across conditions, i.e. that the region contains information about the experimental conditions. Differences in multivoxel patterns across conditions can be interpreted as reflecting differences in underlying neuronal population activity. This figure has been adapted with permission from Raizada et al. (2010).

1.1.2 Scope and limitations

The use of pattern-information analysis is not restricted to investigating functional regions defined by activation-based analysis. It can also be used to inves-

tigate patterns of activity across more widely distributed sets of voxels (e.g. Haxby et al., 2001; Carlson et al., 2003) or to *define* functional regions by mapping the whole volume for effects using a multivariate searchlight (“information-based brain mapping”, Kriegeskorte et al., 2006, 2007). The change that activation-based analysis is sensitive to – all voxels changing their activity in *the same direction* – can be viewed as a special case of the changes that pattern-information analysis can detect: any change of the pattern, including spatial-mean activity changes as well as pattern changes where the spatial-mean is unaffected. This general sensitivity makes pattern-information analysis a powerful statistical tool. With many successful applications in neuroimaging, the approach has gained momentum in recent years (e.g. Haxby et al., 2001; Carlson et al., 2003; Cox and Savoy, 2003; Friston et al., 2008; Hanson et al., 2004; Kamitani and Tong, 2005; Haynes and Rees, 2005a; Haynes et al., 2007; Kriegeskorte et al., 2007; Kriegeskorte et al., 2008a; Mourao-Miranda et al., 2005; Mitchell et al., 2008; O’Toole et al., 2005; Pereira et al., 2009; Raizada et al., 2010). Note that related multivariate methods as well as prediction frameworks have been explored before in neuroimaging analysis (Strother et al., 2002; Worsley et al., 1997), but with different conceptual goals.

The blood-oxygen-level-dependent (BOLD) fMRI signal provides a complex reflection of underlying neural activity and is affected by noise (Boynton et al., 1996; Logothetis, 2008). As a consequence, interpretation of the BOLD fMRI signal in terms of underlying neural activity requires caution. The BOLD fMRI contrast has been shown to reflect stimulus-driven neural activity (Logothetis et al., 2001). Although the fine-grained activity patterns measured by fMRI may not precisely reflect neural activity patterns because of hemodynamic blurring and distortion, a change of signal (patterns) across conditions can be interpreted as a change of neural population activity.

Pattern-information fMRI is fundamentally limited by the amount of information about the neural population codes that can be provided by fMRI. Voxel resolution is one such limitation, thus motivating the use of high-resolution fMRI in conjunction with pattern-information analysis (Kriegeskorte and Bandettini, 2007a; Kriegeskorte et al., 2007). A technique that also targets the representational content of functional regions and that is not limited by voxel resolution is fMRI adaptation (Grill-Spector and Malach, 2001). This approach can potentially resolve sub-voxel representations by inferring neural selectivity from fMRI adaptation responses. However, the interpretation of positive findings (“release from adaptation”) in terms of neural population selectivity relies on assumptions that have been questioned by recent experimental results (Tolias et al., 2005; Sawamura et al., 2006; Krekelberg et al., 2006). These results showed that release from adaptation does not necessarily reflect selectivity of the underlying neural population as measured by classical electrophysiological methods. Other

explanations, e.g. attentional effects or carry-over of effects from connected regions (Tolias et al., 2005; Krekelberg et al., 2006), can account for release from adaptation as well. While the fMRI adaptation paradigm compares activation between pairs of either different or repeated stimuli and then *infers* single-stimulus selectivity from these activation differences, pattern-information fMRI follows the simpler logic of contrasting experimental conditions directly to determine if there is an effect on the dependent variable: the activity pattern within an ROI. Although its sensitivity is limited by the measurement technique of fMRI, a positive result, i.e. statistically distinct activity patterns, provides strong evidence for a difference between the underlying neural activity patterns in the region. It has recently been shown that it is possible to combine pattern-information fMRI and fMRI adaptation in a single experiment and simultaneously estimate activity patterns and adaptation effects (Aguirre, 2007).

1.1.3 Study design

Both event-related and block designs can be used in combination with pattern-information analysis. The choice will largely be based on similar considerations as for studies using activation-based analyses. Block designs yield a higher functional contrast-to-noise ratio than event-related designs. This holds both for constant inter-stimulus-interval (ISI) event-related designs (Bandettini and Cox, 2000) and jittered rapid event-related designs (Birn et al., 2002). This implies that block designs will generally yield better estimates of the average response pattern (i.e. the centroid) than event-related designs. This is especially useful for discriminating a small number of conditions (e.g. Haxby et al., 2001). However, event-related designs can be preferable for psychological reasons as they are less predictable and can reduce habituation effects. Moreover, event-related designs can accommodate a larger number of conditions (Kriegeskorte et al., 2008b). Another advantage of particular importance to information-based analysis is that they yield more independent data points than block designs and can therefore yield a better estimate of the shape of each condition's multivariate response distribution. This can improve classification performance and, thus, increase sensitivity in detecting pattern information. On the other hand the condition-mean pattern estimates (centroids) will typically be somewhat noisier. It should also be noted that rapid-event related designs involve temporally overlapping hemodynamic responses. The effects of temporal overlap can be accounted for using the same design optimization techniques that have proven useful for activation-based studies.

1.1.4 Imaging parameters

Most pattern-information analyses so far have utilized lower-resolution fMRI data (see Haxby et al., 2001; Kamitani and Tong, 2005; Haynes and Rees,

2005a), indicating that larger-scale patterns – even if dominated by vascular effects – can contain a considerable amount of information even about quite fine-grained neuronal patterns (consider Kamitani and Tong 2005). If information on a fine spatial scale is of interest, high-resolution fMRI (Kriegeskorte et al., 2007) might be a better choice. However, the tradeoff between the functional-contrast-to-noise ratio and the resolution has to be carefully considered (Kriegeskorte and Bandettini, 2007a). A voxel size of about 2-mm in each dimensions appears to be a reasonable compromise at 3 Tesla.

1.2 Testing for pattern information

In this section, we describe how to test for a multivariate activity-pattern difference. A significant pattern difference implies that the condition can be decoded (with some accuracy above chance level) from the activity patterns. In other words, it implies pattern-information about the experimental condition.

A wide variety of multivariate methods can be used for pattern-information analysis. All these methods aim to determine whether the patterns of activity associated with different conditions are statistically discriminable (i.e. significantly different). As in conventional analysis, every activity pattern we estimate from the data results from a combination of true effects and noise. Noise is always present and will make every pattern unique (just as in a univariate t-test there is always a small difference between the estimates of the two means to be compared, even if the null hypothesis is true). We need to determine whether the patterns associated with, say, condition A and condition B, are more different than expected under the null hypothesis of equal activity patterns in both conditions. Under the null hypothesis, any differences between the pattern estimates would be produced by noise alone.

Univariate data is usually analyzed using a t-test or analysis of variance (ANOVA). For multivariate data, the equivalent method would be a multivariate analysis of (co)variance (MANOVA). However, this method assumes that the distribution of the residuals is multivariate normal, an assumption that might not hold for fMRI data. This is one reason why most of the cited studies approach pattern analysis as a classification problem: If we can classify the experimental conditions (which elicit the representational states we are interested in) on the basis of the activity patterns better than chance, this indicates that the response pattern carries information about the experimental conditions. This approach has been referred to as “brain reading” (Cox and Savoy, 2003) or “decoding”.

1.2.1 Linear classification is the most widespread and successful pattern-information analysis in neuroimaging so far

Multivoxel patterns of activity can be viewed as points in a multidimensional space (with as many dimensions as voxels). Consider the simple case of patterns based on activity of only two voxels. Each pattern can then be thought of as a point on a plane, where the activity in each voxel determines one of the coordinates (Figure 1.2). One way to classify these patterns is to construct a line that separates the patterns belonging to condition A from the patterns belonging to condition B (solid green lines in Figure 1.2). Patterns on one side of the line will be classified as condition A, patterns on the other side will be classified as condition B. For more than two voxels, the plane becomes a higher-dimensional space and the decision line generalizes to a linear decision boundary (also called a decision hyperplane). Classifiers that use a linear decision boundary are referred to as *linear* or *hyperplane* classifiers. Linear classification is the most widespread and successful tool for pattern-information analysis in neuroimaging so far.¹ A good introductory textbook on the mathematics of pattern classification is Duda et al. (2001).

The three most widespread linear classification methods in pattern-information fMRI (Figure 1.2) are the minimum-distance classifier (e.g. Haxby et al., 2001), Fisher linear discriminant analysis (FLDA; e.g. Carlson et al., 2003) and the linear support vector machine (linear SVM; e.g. Cox and Savoy, 2003). Each of these methods places the linear decision boundary slightly differently (solid green lines in Figure 1.2).

These methods will perform optimally under different assumptions about the distribution of the response patterns. In practice, they tend to perform somewhat similarly on fMRI data and there is no strong evidence to date suggesting a general superiority of any one of them in this context (but see Ku et al., 2008; Mourao-Miranda et al., 2005). Importantly the differences concern the *sensitivity* for detection of pattern information, not the *specificity* (i.e. the false-positives rate for detecting information). Thus, any of the methods can provide a valid statistical test of pattern-information when correctly applied.

¹ Non-linear classification algorithms have also been used for pattern-information analysis (e.g. Cox and Savoy, 2003; LaConte et al., 2005). These algorithms can capture more complicated class boundaries than linear classifiers. However, non-linear classification methods are more prone to overfit the data than linear classification methods. Overfitting is a particularly severe problem in fMRI because the number of data points (condition repetitions or time points) is typically not very large in relation to the number of ROI voxels. Overfitting leads to lower generalization performance (i.e. lower accuracy on the test data set) and a decrease in power for detecting linear pattern effects (STEP 5).

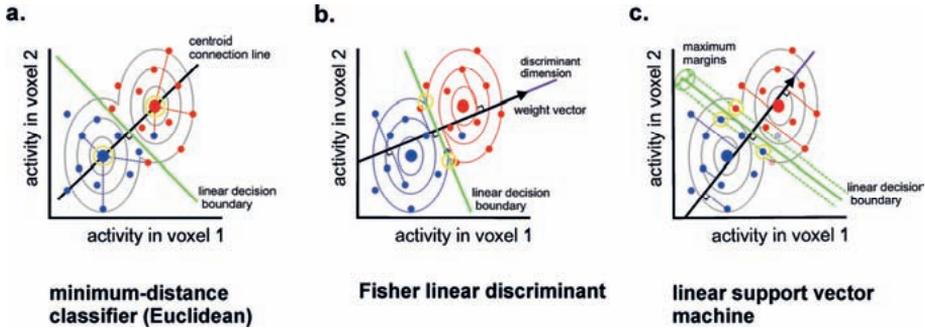


Figure 1.2 Linear classification methods all define a linear decision boundary, but the boundary is placed slightly differently. This is shown for a given set of hypothetical activity patterns. The blue dots represent activity patterns for one experimental condition (e.g. the speech sound /ra/), the red dots represent activity patterns for a second condition (e.g. the speech sound /la/). For simplicity, the displayed activity patterns are based on activity of only two voxels. Nevertheless, the classification methods generalize to higher-dimensional voxel spaces. The ellipses in the background of each panel are iso-probability-density contours describing the bivariate normal distribution of the activity patterns for each condition. The yellow circles indicate the geometrical features that define the linear decision boundary (green) for each classifier. **(a)** Minimum-distance classifier. This classifier first determines the centroids of the two multivariate distributions (large dots). Each activity pattern is then classified as the condition whose centroid it is closest to in multivariate space (using Euclidean distance here, as shown by the dotted lines). This implies a linear decision boundary (i.e. a hyperplane) orthogonal to the centroid connection line, equally dividing the distance between the two centroids. **(b)** Fisher linear discriminant analysis (FLDA). Response patterns are projected onto a linear discriminant dimension by weighting each voxel's activity in order to maximize the ratio of between-condition and within-condition variance. The voxel weights define a weight vector that points in the direction of the linear discriminant dimension. The patterns (i.e. the data points) are orthogonally projected onto the discriminant dimension and a threshold is used for classification. This implies a linear decision boundary (i.e. a hyperplane) orthogonal to the linear discriminant dimension. **(c)** Linear support vector machine (SVM). Same description as FLDA, except for the way the voxel weights are computed. The voxel weights computed by linear SVM are set to yield a linear decision boundary that maximizes the margin (i.e. the distance of the nearest data point to the decision boundary). To make this intuitive, we can imagine starting with a decision boundary that perfectly classifies the training set, then widening the margin equally on both sides while adjusting the angle and position of the decision boundary, until the margin cannot be widened anymore without including one of the training data points. The response patterns closest to the decision boundary (points in yellow circles) then define the margins and the decision boundary halfway in-between the margins. These points are therefore called "support vectors". In order to handle overlapping distributions, SVM algorithms are typically set to allow for a few misclassifications on the training set (see the two transparent points in our hypothetical example).

1.2.2 Subtle differences between linear classifiers

In this section we provide a conceptual description of the three methods to give the interested reader an intuitive sense of how the linear decision boundary is placed in each method (solid green lines in Figure 1.2).

The minimum-distance classifier assigns each activity pattern to the condition whose centroid (multivariate mean) it is closest to in multivariate space. This results in a linear decision boundary orthogonal to the centroid connection line and equally dividing the distance between the two centroids (Figure 1.2a) – assuming that the multivariate distance is simply measured as the length of a straight line connecting the two points (i.e. the Euclidean distance). Using Euclidean distance, this method performs optimally when the distributions associated with the two conditions are identical (homoscedasticity) and isotropic (i.e. they fall off in the same way in all directions of multivariate space). Alternatively, the correlation of the patterns across voxels can be used to compare patterns. A correlation-based distance can be obtained as $1-r$, where r is the correlation coefficient. Minimum-distance classification using the correlation distance is equivalent to the method used by Haxby et al. (2001). Note that using pattern correlation renders the analysis insensitive to regional-average differences (activation effects), which may be desirable. With either distance measure, the minimum-distance classifier implies a linear decision boundary.

Unlike minimum-distance classification, FLDA (Figure 1.2b) takes the covariance structure of the data into account. FLDA is equivalent to modeling each condition's distribution as a multivariate normal distribution (with a covariance estimate pooled across the two conditions) and classifying each pattern as the condition that has the greater probability density at that point in the space. As a consequence, FLDA performs optimally when the distributions associated with the two conditions actually are approximately multivariate normal² (but not necessarily isotropic) and have the same covariance structure (homoscedasticity).

Linear SVM does not assume multivariate normality. Instead it searches for a linear decision boundary that not only discriminates the two sets of points but also has the maximum margin (greatest distance to the nearest points on both sides; Figure 1.2c). The response patterns on the margins are referred to as the “support vectors”, because they “support” the margins and define the decision hyperplane. In other words, linear SVM only uses the most informative subset of data (the support vectors) for constructing the boundary. A linear SVM decision boundary will not change when data points (response patterns) far away from the boundary are moved – as long as the support vectors do not change. By con-

² Note that, in contrast to MANOVA, the specificity of FLDA is not dependent on the assumption of multivariate normality of the residuals because classification algorithms use independent data sets for training and testing. Strong violations of multivariate normality will affect sensitivity, but not specificity, so a test of pattern information is valid.

trast, an FLDA or minimum-distance-classifier decision boundary will move when any data point is shifted.

Mathematically, the linear decision boundary is defined by a vector w that points orthogonal to it in multivariate activity-pattern space and by a parameter that shifts it to the best location. We can think of each linear classifier as using a different rule for determining the vector w and the shift parameter. For a given linear decision boundary, we can use the vector w to determine which side a pattern falls on. To this end, we compute a weighted sum (also called a linear combination) of the voxel responses using the entries of the vector w as the weights, which is why w is also known as the weight vector.³ Geometrically, computing a weighted sum of voxel responses corresponds to orthogonally projecting an activity pattern (point in multivariate space) onto a linear discriminant dimension, which is a line in multivariate space. (These orthogonal projections are denoted by dashed lines in Figure 1.2b and c.) The weight vector points in the direction of the discriminant dimension, i.e. orthogonal to the decision boundary. We can apply a decision threshold to the weighted sums for all patterns so as to classify the patterns with the greatest accuracy. The threshold defines the shift of the decision boundary to the best location (Figure 1.2).

For the minimum-distance classifier, w is the difference between the centroids. For FLDA, w is the weight vector that maximizes the ratio of between-condition and within-condition variances (this constitutes an alternative but equivalent definition of FLDA to the one given above). For the linear SVM, w depends on the support vectors as determined by the training algorithm. None of these methods is superior in general. Minimum-distance classification is expected to perform better than FLDA when its assumption of isotropic distributions is actually true. FLDA is expected to perform better than linear SVM when the data are actually multivariate normal or approximately so. Actual performance will crucially depend on the amount of data available, with limited amounts of data and greater numbers of voxels favoring simpler classification methods. Minimum-distance classification is the most conceptually simple, statistically stable, and computationally efficient method. FLDA is sensitive to the covariance structure of the data, but requires more data to capitalize on this advantage. FLDA also requires slightly more computation. Compared to linear SVM, FLDA is more computa-

³ Intuitively, we would like to weight each voxel by how well its activity discriminates the two conditions. This could be achieved by using the t-values for the contrast between these two conditions (A-B) as weights. This means that a voxel responding more to condition A than B (positive t-value) will be given a positive weight, and a voxel responding more to condition B than A (negative t-value) will be given a negative weight. A voxel that responds similarly to A and B will be given a weight close to zero. The methods for voxel weighting shown in Figure 2b-c are mathematically more complex, but conceptually similar to using contrast t-values as voxel weights.

tionally efficient and arguably more straightforward, conceptually as well as mathematically. However, linear SVM handles limited data in high-dimensional spaces naturally and gracefully, whereas FLDA might require a regularized covariance estimate (Ledoit and Wolf, 2003).

1.3 Pattern-information analysis: step-by-step

In this section, we provide a step-by-step description of the methods for extracting patterns of activity from fMRI data and for analyzing these patterns. These steps are summarized in Figure 1.3.

1.3.1 STEP 1: Data splitting and preprocessing

Before analysis, the data should be split into an independent training and test set to ensure unbiased testing results. The training data set should be used for voxel selection (STEP 3) and classifier training (STEP 4). Both of these steps involve voxel weighting, either binary (voxel selection) or continuous (classifier training). Voxel weighting can bias testing results if performed on the same data and therefore it is crucial to use an independent data set for classifier testing (STEP 5). To make sure the data are independent, the two sets should be based on different scanner runs (e.g. even and odd runs) that use independent stimulus sequences. One option is to split the data into two halves. However, the training data set is generally chosen to be larger than the test set in order to obtain stable voxel weights. Efficient use of the data can be achieved by cross-validation: divide the data into a number of independent subsets (e.g. single runs in your experiment), use all but one subset as training data and use the left out subset as test data; then repeat this procedure until each subset has been used as test data once. Performance on the different subsets is combined to obtain overall classifier performance. Ideally, preprocessing should be performed separately for training and test data sets so as to avoid introducing dependencies between the data sets. Preprocessing steps are the same as in activation-based analysis (i.e. slice-scan-time correction, motion correction, trend removal). In order to preserve fine-grained pattern information, spatial smoothing of the data should be omitted or strongly reduced.

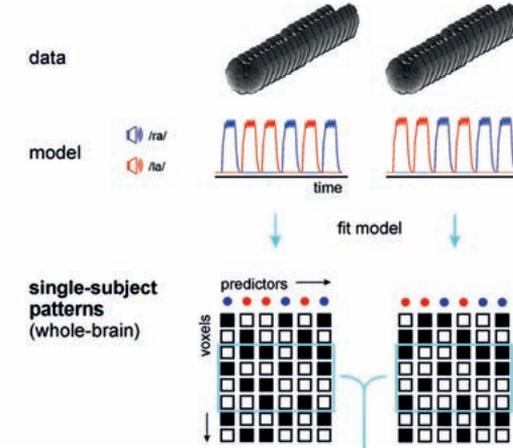
STEP 1

Data splitting
Preprocessing
(no spatial smoothing)



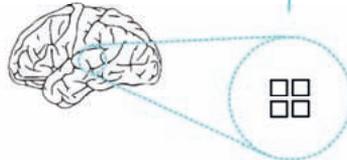
STEP 2

Estimating the
single-subject patterns



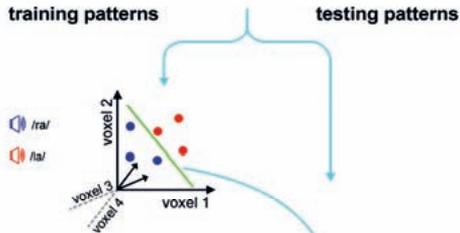
STEP 3

Selecting the voxels
based on:
* training data
* another data set
independent from test
data set (e.g. anatomy)



STEP 4

Training the classifier



use linear decision boundary
from training on test data

STEP 5

Testing the classifier

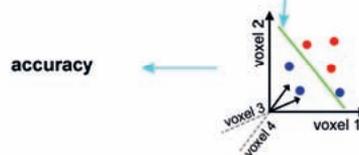


Figure 1.3 Pattern-information analysis: step by step. Schematic illustration of the five steps of pattern-information analysis as described in the text. First, data are split into a training and a test data set and preprocessed separately. Then, single-subject patterns of activity are estimated from the data using univariate analysis (GLM) at each voxel. This results in whole-brain activity patterns consisting of beta-estimates. Black boxes indicate activated voxels; white boxes indicate non-activated voxels. Note that activity levels are continuous in analysis and only stated as binary here for simplicity. There will be as many patterns as there are predictors (conditions) in the model. Pattern-estimation is done separately for the training and test data set. The third step consists of selecting voxels for pattern-information analysis. This can be done based on anatomy, function or both. For simplicity, the shown example region consists of four voxels only. Voxel selection should be based on the training data set or another data set that is independent from the test data set in order to prevent biased testing results. This also applies to STEP 4: voxel weighting should be performed on the training data set to prevent biased testing results. Voxels are weighted in order to maximize discriminability of the patterns belonging to the two conditions. The voxel weights computed in STEP 4 can then be tested on the test data set in STEP 5. If the weights capture true differences between the two conditions, good performance (classification accuracy) on the training data set will generalize to the test data set. Performance significantly better than chance indicates that the ROI contains information about the experimental conditions, i.e. the representational content of the region differs across conditions. The image for STEP 3 has been adapted with permission from Raizada et al. (2010).

1.3.2 STEP 2: Estimating the single-subject activity patterns

Previous studies have used several methods to estimate single-subject activity patterns. For block designs or slow event-related designs, where BOLD responses to different conditions do not overlap in time, it is possible to stay close to the raw data and use single-volume signal intensity values (Polyn et al., 2005) or temporally averaged normalized signal intensity values as patterns of activity (e.g. Kamitani and Tong, 2005). Single-subject patterns can also be estimated by univariate analysis at each voxel using the general linear model (GLM) (Friston et al., 1994; Friston et al., 1995ab; Worsley and Friston, 1995). This is useful, in particular, for rapid event-related designs (e.g. Kriegeskorte et al., 2007, 2008ab) because of the hemodynamic response overlap, but has also been used in combination with block designs (e.g. Haxby et al., 2001). An advantage of using the GLM is the possibility to include motion and trend predictors in the model in order to obtain better estimates. Each condition or each example belonging to a condition (if estimating the shape of the response distribution) is entered as a predictor in the model. This part of the analysis is identical to activation-based analysis and will yield a beta-value for each predictor and voxel. The beta-values for one predictor across voxels form the pattern of activity for a specific condition. Pattern estimation yields a set of training patterns and a set of test patterns. In order to preserve fine-grained subject-specific information, the patterns should not be averaged across subjects. Instead the analysis should be performed in native subject space.

1.3.3 STEP 3: Selecting the voxels

Once activity values are computed, the next step is to decide which voxels to include for pattern-information analysis. These voxels are selected using the training data set or another data set independent from the test set (e.g. anatomical data or functional data from a separate block-localizer experiment). One option would be to analyze the patterns of activity in a specific ROI. If defined by activation-based analysis, ROIs will be spatially contiguous sets of voxels, but they do not have to be. For example, to investigate object-category discrimination, the most visually responsive voxels in object-selective cortex could be selected for subsequent analysis, irrespective of whether these voxels are adjacent or not. A computationally more demanding option would be to analyze the pattern of activity across all brain voxels. This might increase informational content, but it will definitely also add substantial amounts of noise. Typically there will a decrease in performance as the number of voxels becomes very large. Possible solutions include selecting fewer voxels and transforming the original voxel space into a lower dimensional space (dimensionality reduction). Voxels can also be selected using information-based brain mapping (Kriegeskorte et al., 2006, 2007). This can be seen as the multivariate equivalent of univariate statistical parametric mapping (SPM) (Friston et al., 1995ab).

1.3.4 STEP 4: Training the classifier

To investigate whether a region's pattern of activity discriminates two conditions, we first use the training data set to determine a set of weights (one for each voxel) that linearly combines the voxel responses in such a way as to maximize the difference between the two conditions (classifier training). We described three different linear classifiers that can be used for pattern-information analysis: the minimum-distance classifier, FLDA, and linear SVM. These may differ in sensitivity, depending on factors including the brain region, experimental events, the amount of data available, and the number of voxels in the ROI. Any of the three methods can provide a valid test of pattern-information.⁴

Most classifiers can also be trained on data from multi-condition experiments (Pereira et al., 2009). However, multi-class discriminations are often approached as a combination of multiple two-class discriminations. This approach is motivated by the fact that two-class discriminations are generally of neuroscientific interest, even if an experiment contains more than two conditions. For a detailed overview on using linear classification algorithms in neuroimaging, and

⁴ If more than one method is used, all results should be reported. (Picking the significant result among different analyses would require correction for multiple comparisons.)

their mathematical descriptions, see Pereira et al. (2009). Paragraph 1.5 lists several currently available pattern-information analysis toolboxes.

1.3.5 STEP 5: Testing the classifier

The weights computed during training are set to yield optimal classification performance on the training data set. To test whether good classification performance generalizes (i.e. is not based largely on noise present in the training data set), the weights are applied to an independent test data set. Performance of the classifier on the test data set can be measured by percent correct classification (accuracy). The null hypothesis is that the classifier performs at chance level. To test whether classification accuracy is significantly better than chance, we can use a chi-square test (or a Monte-Carlo method in case of few observations). If the statistical test shows a significant result, this indicates that the region's response contains information about the experimental conditions.⁵ Another way to test the classifier is to perform a univariate t-test on the projected test patterns (Kriegeskorte et al., 2007). As described above, projection (voxel weighting) converts the activity patterns into one-dimensional values. These values can then be analyzed by a conventional univariate t-test. Similar to a classification accuracy that is significantly better than chance, a significant t-value for the difference between the two conditions would indicate that the region's response contains information about the experimental conditions.

1.4 Conclusion

Pattern-information analysis investigates the representational content of a region by analyzing the information carried by a region's pattern of activity. This information would not be detected by conventional activation-based analysis and can significantly contribute to our understanding of neural representations of mental content. In combination with high-resolution fMRI, pattern-information analysis can detect fine-grained activity-pattern information. The most popular method is linear classification, which analyzes a region's activity patterns by means of a weighted sum of the single-voxel responses, with the weights chosen to maximally discriminate different conditions. Statistical inference is performed on a data set independent of that used for ROI definition and voxel weighting so as to prevent statistical circularity. The conceptual appeal of pattern-information fMRI is that it allows us to "look into" the regions and inves-

⁵ In addition to the overall accuracy, we can examine the frequency of all four possible classifier outcomes (true/false positives, true/false negatives). This is important, in particular, when the frequencies of the two conditions are not equal.

tigate their representational content. Recent neuroscientific successes in the domain of sensation and perception suggest that higher-order cognitive functions in the domain of social and cognitive neuroscience might also benefit from the pattern-information approach.

1.5 Pattern-information analysis toolboxes

AFNI 3dsvm plug-in: <http://www.cmu.edu/laconte/3dsvm.html>

Princeton MVPA toolbox: <http://www.cs.brown.edu/mvpa/>

PyMVPA toolbox: <http://pkg-exppsy.alioth.debian.org/pymvpa/>

LIBSVM toolbox: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

Chapter 2

Face-identity change activation outside the face system: “Release from adaptation” may not always indicate neuronal selectivity

Face recognition is a complex cognitive process that requires distinguishable neuronal representations of individual faces. Previous functional magnetic resonance imaging (fMRI) studies using the “fMRI-adaptation” technique have suggested the existence of face-identity representations in face-selective regions, including the fusiform face area (FFA). Here, we present face-identity adaptation findings that are not well explained in terms of face-identity representations. We performed blood-oxygen-level-dependent (BOLD) fMRI measurements, while participants viewed familiar faces that were shown repeatedly throughout the experiment. We found decreased activation for repeated faces in face-selective regions, as expected based on previous studies. However, we found similar effects in regions that are not face-selective, including the parahippocampal place area (PPA) and early visual cortex (EVC). These effects were present for exact-image (same view and lighting) as well as different-image (different view and/or lighting) repetition, but more widespread for exact-image repetition. Given the known functional properties of PPA and EVC, it appears unlikely that they contain domain-specific face-identity representations. Alternative explanations include general attentional effects and carryover of activation from connected regions. These results remind us that fMRI stimulus-change effects can have a range of causes and do not provide conclusive evidence for a neuronal representation of the changed stimulus property.

Mur M, Ruff DA, Bodurka J, Bandettini PA, Kriegeskorte N (2010) Face-identity change activation outside the face system: “Release from adaptation” may not always indicate neuronal selectivity. *Cereb Cortex* 20, 2027-2042, doi:10.1093/cercor/bhp272.

2.1 Introduction

Face recognition is an important function of the human visual system. It requires the existence of distinct neuronal representations of individual faces (Haxby et al., 2000; Kanwisher et al., 1997; Puce et al., 1995). These representations have been investigated with the “fMRI adaptation” method (Grill-Spector and Malach, 2001). Several studies using this method (Andrews and Ewbank, 2004; Gauthier et al., 2000b; Rotshtein et al., 2005) have shown a stronger fMRI response to face-identity change than to face-identity repetition in face-selective brain regions, including the fusiform face area (FFA) (Kanwisher et al., 1997; Puce et al., 1995). FFA is defined by its stronger response to faces than to other objects, consistent with a role in detecting the presence of faces (Kanwisher et al., 1997). The fMRI-adaptation results have been interpreted as evidence for the involvement of FFA in representing face identity (see also Grill-Spector et al., 2004). Attempts to directly decode face identity from FFA activity patterns have failed so far, although they succeeded in anterior inferior temporal cortex (aIT) (Kriegeskorte et al., 2007).

Here we replicate the stronger FFA response to face-identity change than to repetition and examine whether these effects are specific to FFA and other face-selective regions or more widespread. Previous studies might have missed effects outside of face-selective cortex, because of spatially restricted analyses or lack of statistical power (but see Ng et al., 2006; Pourtois et al., 2005).

Face-identity repetition effects have been found to be influenced by face familiarity (Eger et al., 2005; George et al., 1999; Henson et al., 2002). The influence of face familiarity has also been investigated by directly comparing activation to familiar with activation to unfamiliar faces, showing modulation of activity by face familiarity in FFA (Gobbini et al., 2004; Henson et al., 2000), aIT (Gorno-Tempini et al., 1998; Sergent et al., 1992; Sugiura et al., 2001), consistent with lesion studies (Evans et al., 1995; Marotta et al., 2001), and hippocampus (Bernard et al., 2004; Eger et al., 2005; Leveroni et al., 2001). Familiar-face stimuli used in previous studies were either famous faces or faces of personal acquaintances. Identification of such a face involves both recognition of the perceptual appearance of the face and activation of associated conceptual information such as name and biographical facts (Bruce and Young, 1986; Haxby et al., 2000). Previous studies, therefore, did not dissociate the perceptual (looks) and conceptual (biographical information) components of face recognition.

The present study investigates the effects of face-identity repetition and face familiarity on activation in human inferior temporal cortex using a continuous carry-over design (Aguirre, 2007). Participants were shown faces of different levels of familiarity: “new” (never seen before), “seen” (seen previously, no fur-

ther information known), and “known” (seen previously, name and biography known). We used images of four different perceptually familiar individuals; two of them were also biographically familiar. These four familiar face identities (2 seen, 2 known) were repeated throughout the experiment to investigate the effects of consecutive face-identity repetition. The subjects’ task was to classify each face image as “new” or “familiar” (either seen or known). The task, thus, diverted attention from differences among the familiar faces. We investigated activity in face-selective as well as non-face-selective regions and also searched for effects outside these regions of interest (ROIs). For optimal stimulus control, we used renderings of textured 3D face models constructed from face photos. Non-face features were masked out and color histograms equalized to minimize low-level confounds (Figure 2.1). View and lighting were varied for each face identity. This allowed us to compare the effects of different-image repetition to those of exact-image repetition.

We expected face-identity repetition effects to be confined to face-selective regions. In addition, we expected perceptual face regions, including the occipital face area (OFA) and FFA, to be equally activated by seen and known faces, and higher-level regions, including aIT and hippocampus, to show stronger activation for known than seen faces.

2.2 Materials and Methods

2.2.1 Participants

Eight healthy male volunteers aged between 29 and 40 years (mean age = 34 years) participated in this study. All participants were right-handed and had normal or corrected to normal vision. None of the participants had a history of neurological or psychiatric disorder. Before scanning, the participants received information about the procedure of the experiment and gave their written informed consent for participating. All experiments were conducted in accordance with standards of the Institutional Review Board of the National Institutes of Mental Health, Bethesda, MD.

2.2.2 Stimuli

We took colored photographs of male faces and used FaceGen Modeller 3.1 (Singular Inversions, Vancouver, Canada) to generate 3D face models from these static frontal and profile face photographs. This software uses a morphable face model consisting of a vector space representation of the shape and texture of several 3D face model examples (Blanz and Vetter, 1999). Snapshots of the 3D face models were extracted, using combinations of two different views (-30 and

30 degrees relative to the sagittal plane) and two different lightings (perspective projection, -60 and 60 degrees relative to the sagittal plane, elevation of 30 degrees relative to the center of the head). Subsequently, the snapshots were masked with a soft-faded circular aperture and color-histogram-equalized, so that the distribution of intensity values in each color channel (RGB) was identical across stimuli (Figure 2.1b). This procedure resulted in 24-bit RGB images of 701 x 701 pixels, including the face and part of the neck (Figure 2.1a). No hair or ears were shown. All processing steps on the snapshots of the 3D face models were performed using Matlab 7.0 (The MathWorks Inc, Natick, MA). Four male identities were chosen to function as familiar faces (two seen and two known faces). New faces were generated using photographs of males other than the four chosen ones and by manipulating 3D face models in FaceGen to create new face identities (Figure 2.1c).

Stimulus presentation

Stimuli were presented using Presentation 9.81 (Neurobehavioral Systems Inc, Albany, CA), and projected onto a translucent screen positioned at the foot of the scanner (at the participants' feet), using a liquid-crystal-display projector. A mirror fixated on the head reception coil enabled participants to see the screen. Face images subtended a vertical visual angle of ~ 4 degrees.

2.2.3 Experimental design and task (face experiment)

Pre-scanning training

The participants were familiarized with the stimuli and task one or two days before scanning. The pre-scanning training consisted of one 30-minute session during which the participants were familiarized with frontal and ± 30 -degree views of the four familiar faces, two of which were accompanied by biographical information that needed to be memorized (Figure 2.1c). The biographical information was fictive, but realistic, and consisted of name, age, profession and personal background. The two descriptions were matched for information density. In order to ensure that effects were not due to the particular face images used in the different familiarity conditions, half of the participants received the biographical information with one pair of faces, and the other half with the other pair. Participants were instructed to closely examine the four familiar faces and memorize the biographical information associated with two of them. In order to equilibrate visual exposure to seen and known faces, we explicitly asked the subjects to spend an equal amount of time inspecting each of the four face images. After a learning period of 10 minutes, the participants received 18 five-option multiple choice items and a perceptual face-familiarity test (see Supplementary Material) to check whether their perceptual and conceptual knowledge

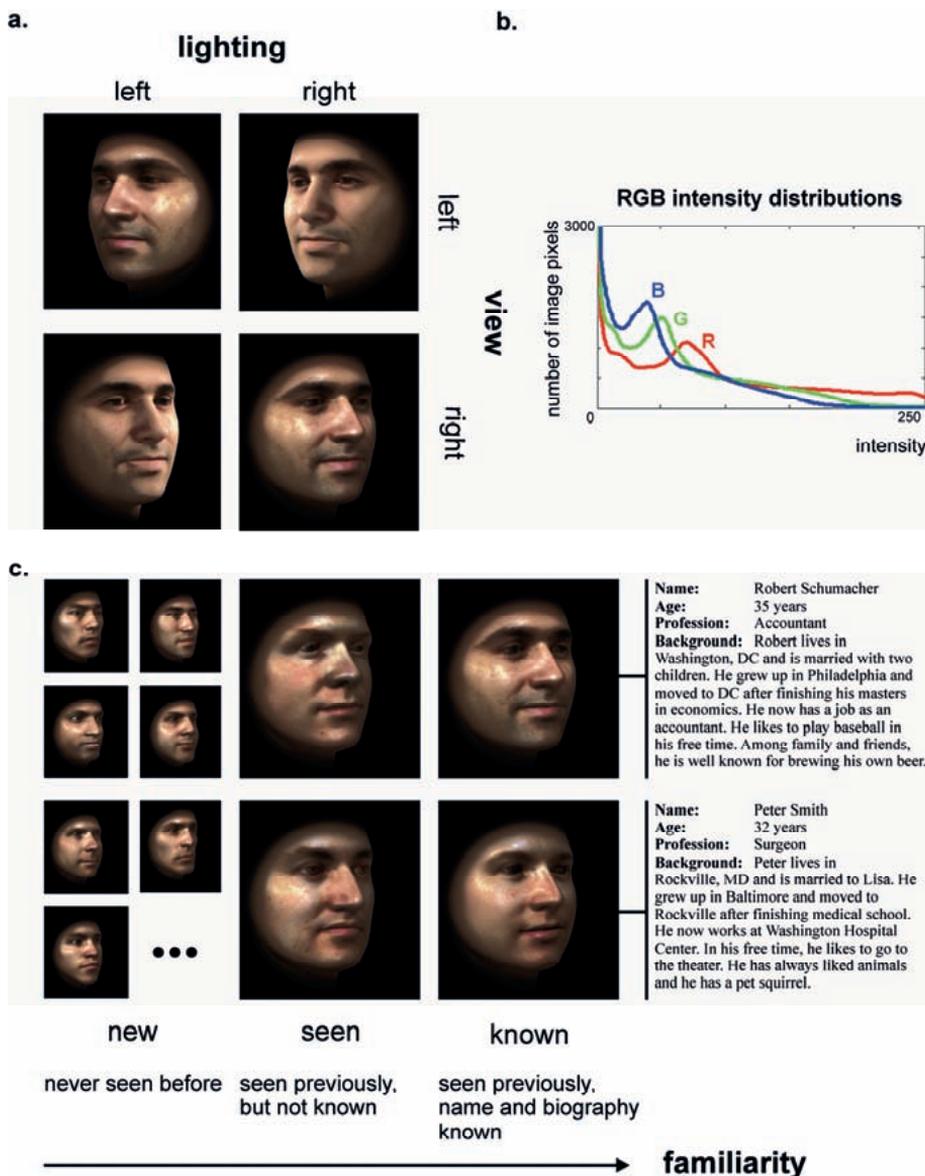


Figure 2.1 Stimuli. (a) Stimuli were male faces seen from two different views and with two different lightings. The lower left and upper right faces are counterlit (incongruent view and lighting). (b) Averaged histograms for red, green and blue (RGB) color channels that were imposed on each image. As a consequence of the color-histogram equalization, the images have the same light and spatial-signal energy. (c) Face familiarity was systematically varied, resulting in new, seen and known faces. The seven new faces that are shown are a subset of a total of 180 new faces that were used in this study. The four familiar faces are the four right-most faces.

levels were at a sufficient and stable level (accuracy > 85% per test). Participants that had not reached a sufficient level of familiarity after the first learning period examined the four familiar faces again, and were then re-tested. Errors made during the tests were reviewed and correct answers were provided. The last 10 minutes of the pre-scanning training were used to practice the task to be performed in the scanner (see *Task*). At the beginning of each subsequent scanning session, participants performed another brief learning session and face-familiarity test to refresh their memories.

Face experiment

A rapid event-related design was used, with a stimulus duration of 1 s and inter-stimulus interval (ISI) of 3 s. A single scanning session consisted of six functional runs of 126 trials each (8 min and 24 s per run). These 126 trials consisted of 48 known, 48 seen and 10 new face trials, and 20 baseline trials where no stimulus was shown. The contrasts between the different face types were our main contrasts of interest, and were assumed to be smaller than the contrast between faces and baseline. Therefore, we included more face trials than baseline trials. Stimuli were presented in pseudorandom order. Each of 16 different familiar face images (the four familiar faces * two views * two lightings) was repeated six times during one run. This resulted in 24 presentations of each face-identity per run (six repetitions * two views * two lightings). Thirty percent of these presentations were consecutive face-identity repetitions, i.e. they were directly preceded by an image of the same face-identity. These consecutive face-identity repetitions were mostly the second same face-identity in a row (first consecutive repetition), but the stimulus sequences also contained instances of more than two images of the same face-identity in a row (up to five in a row). About half of these consecutive face-identity repetitions were different-image repetitions (different view and/or lighting), the other half were exact-image repetitions (same view and lighting) (see Table S2.1). Each run contained 16 exact-image repetitions, one for each familiar face image. The sequence started and ended with four baseline trials. A new pseudorandom sequence was used for each run in the experiment. Participants were scanned two or three times, resulting in a total of up to 18 runs per participant (with at least 11 good runs per subject, see *fMRI data preprocessing*). Stimuli were presented on a black background while participants fixated a white cross that was displayed close to the bridge of the nose of each face image (see Figure S2.1).

Task

Participants were instructed to distinguish learned faces from new faces, responding with a right-thumb button press for a familiar face (either seen or known) and a left-thumb button-press for a new face.

2.2.4 Localization of OFA, FFA and PPA (functional localizer experiment)

Stimuli for the independent functional localizer experiment were grayscale photographs (252 x 252 pixels) of faces, places and objects masked with a circular aperture. Face identities shown during this experiment were different from the ones used in the face experiment. The stimuli subtended a visual angle of $\sim 6 \times 6$ degrees. Images of the different stimulus categories were presented in 30-s blocks (stimulus duration 700 ms; ISI 300 ms), intermixed with 20-s fixation blocks. Three blocks were presented for each stimulus category, resulting in a total run time of approximately 8 min. Stimuli were presented on a black background, centered with respect to a white cross that participants fixated on during the run. Participants performed a one-back repetition detection task on the images, responding with a left-thumb button press for each consecutive repetition (three to five repetitions per block).

2.2.5 Magnetic resonance imaging

Functional measurements

Blood-oxygen-level-dependent (BOLD) fMRI was performed using a 3 Tesla General Electric VH/3 MRI scanner, equipped with a custom-built 16-channel MRI digital receiver (Bodurka et al., 2004). A receive-only whole-brain 16-element surface-coil array (NOVA Medical Inc, Wilmington, MA) was used to achieve high spatial resolution for functional studies with good sensitivities to small BOLD signal changes (Bodurka et al., 2007). Twenty 2-mm axial slices (no gap) were acquired, covering the occipital and temporal lobe including the anterior pole, using single-shot full k-space gradient-recalled echo planar imaging (EPI). EPI imaging parameters were as follows: interleaved slice order, EPI matrix size = 128 x 96 pixels, voxel volume = 1.95 x 1.95 x 2 mm³, echo time (TE) = 42 ms, repetition time (TR) = 2 s. Each functional run consisted of 252 volumes (8 min and 24 s per run). The total amount of data acquired for the face experiment was equivalent to 12 h, 25 min, and 4 s of scanning (eight subjects, 11 runs per subject).

Anatomical measurements

Functional scans were superimposed on high-resolution T1-weighted whole-brain anatomical scans (voxel volume = 0.98 x 0.98 x 1.2 mm³), acquired with a fast spoiled gradient echo recalled (FSPGR) sequence.

2.2.6 Statistical analysis

Behavioral data

Reaction time data were analyzed using SPSS 15.0 (SPSS Inc, Chicago, IL). Reaction times of incorrect responses and reaction times that were more than three standard deviations from the mean were discarded. We performed a random-effects analysis of variance (ANOVA) for repeated measures including factors for face-identity repetition and familiarity. Face-identity repetition effects were tested up to the second consecutive repetition and compared for seen and known faces (reaction times for new faces could not be included since these faces were not repeated throughout the experiment). Reaction times for the third, fourth, and fifth consecutive repetition were not included in this analysis since there were only few trials for these conditions (see Table S2.1). Reaction times for the first consecutive repetition were split into a different-image and an exact-image repetition condition. An additional random-effects ANOVA for repeated measures was used to test for effects of familiarity (including new faces), view and lighting. Insignificant interaction terms were stepwise removed from the models. Post-hoc paired t-tests were performed to investigate significant main effects in more detail.

fMRI data preprocessing

MRI data preprocessing and analysis were performed using BrainVoyager QX 1.8 (Brain Innovation, Maastricht, The Netherlands). The first four data volumes of each scan were discarded to allow the fMRI signal to reach a steady state. Runs with excessive head-motion or imaging artifacts were excluded from analysis, leaving at least 11 runs per subject. To ensure that each subject-specific data set contributed equally to the results, we used exactly 11 runs per subject for analysis. For the subjects that had more than 11 runs, we randomly chose 11 runs from the whole set. Preprocessing steps performed on the functional data volumes were as follows: slice-scan-time correction, motion correction (first non-discarded volume of run as reference volume), temporal high-pass filtering with a filter of three cycles per run for the face experiment (corresponding to a cut-off frequency of .006 Hz) and two cycles per run for the functional localizer experiment (corresponding to a cut-off frequency of .004 Hz), and spatial smoothing by convolution with a Gaussian kernel of 6 mm full width at half maximum (FWHM) for the face experiment and 4 mm FWHM for the functional localizer experiment. Spatial smoothing was performed to increase sensitivity to extended activations (“activation blobs”) and improve inter-subject correspondence for group analysis (mapping). In the functional-localizer approach, intersubject correspondence is determined by means of individual ROI definitions, therefore a slightly smaller smoothing kernel (4 mm FWHM) was chosen, allowing for a more precise definition of the shape of the region in

each subject. Functional data were manually aligned to same-session high-resolution structural whole-brain scans and transformed into Talairach stereotactic space. If no same-session structural scan was available (30% of sessions), functional data were manually aligned to a structural scan from a different session for that subject. Visual comparison of activation loci before and after alignment indicated good alignment quality. All time courses were converted to percent signal change.

Multiple linear regression

We performed a fixed-effects group analysis by multiple linear regression of the time course at each voxel. Cognitive predictors were created using the Boynton hemodynamic impulse response function (Boynton et al., 1996), assuming an instantaneous rectangular neuronal response to the 1-s stimulus presentations. In order to keep the number of predictors reasonable (especially given the amount of data to be analyzed simultaneously) we constructed four slightly different models (Figure S2.1). These models investigated the effects of face-identity repetition, face familiarity, face-identity, and view and lighting. The first two models will be described below. A description of the other two models and associated results can be found in the Supplementary Methods and Results.

The first model was used to test for effects of face-identity repetition. The first consecutive repetition of a specific face identity (one of the four familiar faces) was named rep1, the second consecutive repetition of that same face identity was named rep2, and so forth, up to rep5. As described before, these repetitions could be either different-image repetitions (different view and/or lighting) or exact-image repetitions (same view and lighting). rep1 trials were most frequent and contained a relatively large proportion (0.62) of exact-image repetitions (see Table S2.1). We therefore split rep1 into two predictors to separately investigate the effects of different-image and exact-image face-identity repetition. Any familiar face stimulus that was not a consecutive face-identity repetition was named rep0 (identity change). New faces were never repeated throughout the experiment and were therefore modeled by a separate predictor. The face-identity repetition model consisted of subject-specific predictors for rep0, rep1_different, rep1_exact, rep2, rep3, rep4, rep5 and new faces, and confound-mean predictors for each subject and run (Figure S2.1a). Contrasts of interest were the following: (1) rep0 versus rep1_different, (2) rep0 versus rep1_exact, (3) rep1_different versus rep1_exact, (4) rep1_different versus rep2, and (5) rep1_exact versus rep2. The first contrast compared face-identity change to first consecutive different-image face-identity repetition, the second compared face-identity change to first consecutive exact-image face-identity repetition, the third compared different-image repetition to exact-image repetition, the fourth compared first consecutive different-image face-identity repeti-

tion to second consecutive face-identity repetition, and the fifth compared first consecutive exact-image face-identity repetition to second consecutive face-identity repetition. We did not test contrasts comparing brain responses to more than two consecutive repetitions, since there were only few trials for rep3, rep4, and rep5 (see Table S2.1). In addition, we investigated face-identity repetition effects separately for seen and known faces, and tested for an interaction between face-identity repetition and familiarity.

The second model was used to test directly for face familiarity effects and consisted of subject-specific predictors for new faces, seen faces and known faces, and confound-mean predictors for each subject and run (Figure S2.1b). The following two contrasts were performed using t-tests: (1) new versus seen faces and (2) seen versus known faces. The first contrast searched for effects of face novelty and perceptual face familiarity, while the second isolated effects of added conceptual familiarity.

Results were corrected for serial autocorrelation in the temporal domain. Maps were thresholded to control the average false-discovery rate (FDR) to be $< .05$. For both models, one contrast map was a priori chosen as reference and thresholded at $FDR < .05$. The other contrast maps were then thresholded using the t-value associated with that FDR. This created consistent thresholding across maps independent of differences across maps in the number of activated voxels. The reference contrast maps were those maps that focused on our main questions. For face-identity repetition, face-identity change versus first consecutive different-image face-identity repetition (rep0 versus rep1_different) was chosen as reference contrast map. For face familiarity, seen versus known faces was chosen as reference contrast map.

ROI definition

Six regions of interest (ROI) were defined in each hemisphere, based on (1) the block localizer experiment (OFA, FFA and PPA), (2) the contrast faces $>$ baseline performed on even runs of the face experiment (hippocampus and aIT), and (3) anatomical landmarks (early visual cortex). OFA and FFA were defined for each subject by the contrast faces $>$ objects and places, and PPA was defined by places $>$ objects and faces. Maps were thresholded using $FDR < .01$ and a cluster threshold of 200 voxels. Hippocampus and aIT were each defined at two different sizes using fixed-effect group results for faces $>$ baseline in the face experiment. Half of the data from the face experiment was used to define hippocampus and aIT, the other half was used to test for effects of interest (see *ROI analysis*). The large-sized regions were defined first, using a threshold that resulted in a contiguous set of voxels well-separated from other nearby clusters of activation (uncorrected p-values were 8.0×10^{-15} for hippocampus and 1.6×10^{-4} for aIT).

Then, in order to define the small-sized regions, the threshold was increased until the above regions were reduced by half. We only report ROI results for the large-sized hippocampus and aIT, since results for the small-sized regions were qualitatively similar. In each subject, we used high-resolution anatomical data to manually define the calcarine sulcus. Early visual cortex (EVC) was defined by centering ellipsoids (radii were 12 x 5 x 5 mm) on 11 consecutive points along the calcarine sulcus. The resulting ROI included V1 and likely portions of V2 and V3.

ROI analysis

Data from each ROI were averaged across voxels to obtain an average time course per subject. These time courses were concatenated and used for a fixed-effects group analysis. The above models for face-identity repetition and familiarity were fit to the ROI average time course using multiple linear regression. Results were corrected for serial autocorrelation in the temporal domain. Contrasts of interest identical to the ones used for mapping (see *Multiple linear regression*) were computed. To investigate whether face-identity-repetition effect sizes differed across regions, we performed paired t-tests on subject-specific contrast values (random-effects analysis) for all possible pairs of ROIs that showed significant face-identity repetition effects. For each region, subject-specific contrast values were computed by subtracting subject-specific beta-values for one condition (e.g. rep1_different) from subject-specific beta-values for the other condition (e.g. rep0). We performed region comparisons for the following two contrasts: face-identity change versus first consecutive different-image face-identity repetition (rep0 versus rep1_different), and face-identity change versus first consecutive exact-image repetition (rep0 versus rep1_exact).

Table 2.1 ROI details.

ROI	definition	hemisphere	mean TAL(x,y,z)			mean size (mm ³)
EVC	calcarine	left	-5,	-87,	-6	1817
	sulcus	right	5,	-87,	-6	1817
OFA	faces > places & objects	left	-43,	-75,	-12	1315
		right	42,	-72,	-12	1835
FFA	faces > places & objects	left	-37,	-46,	-18	1858
		right	38,	-43,	-18	2034
PPA	places > faces & objects	left	-24,	-43,	-10	3169
		right	23,	-42,	-10	3552
hippocampus	faces > baseline	left	-21,	-22,	-9	131
		right	23,	-20,	-9	142
aIT	faces > baseline	left	-26,	-6,	-27	382
		right	35,	-3,	-25	427

Mean Talairach coordinates denote the center of gravity of the ROIs. EVC was defined using anatomical landmarks; OFA, FFA and PPA were defined using independent data from a separate block

localizer experiment; and hippocampus and aIT were defined using independent data from the face experiment (even runs). EVC = early visual cortex, OFA = occipital face area, FFA = fusiform face area, PPA = parahippocampal place area, aIT = anterior inferotemporal cortex.

2.3 Results

2.3.1 Behavioral results

Subjects performed a binary classification task, in which they responded with a button press to indicate whether the presented face was “new” or “familiar”. (The familiar faces could be either “seen” or “known”, but the task did not require distinguishing between them). Accuracy across subjects was 95% or higher in all conditions.

In order to investigate the effects of face-identity repetition on reaction time, and the influence of face familiarity on these effects, we performed a 2-way ANOVA for repeated measures. Reaction times for new faces were not included in this analysis. The analysis showed a significant main effect of face-identity repetition ($F(3) = 24.537, p < .01$). Paired t-tests showed that reaction times for face-identity-change trials (rep0) were higher than for first consecutive different-image face-identity repetition trials (rep1_different) ($t(7) = 6.355, p < .01$), which in turn were higher than for first consecutive exact-image face-identity repetition trials (rep1_exact) ($t(7) = 4.732, p < .01$) (Figure 2.2). Reaction times for second consecutive face-identity repetition trials (rep2) were higher than for first consecutive exact-image face-identity repetition trials (rep1_exact) ($t(7) = 3.758, p < .01$, Figure 2.2). Note that second consecutive face-identity repetition trials were mainly different-image repetitions. Face-identity repetition effects were not significantly different for seen and known faces. A separate paired t-test showed that reaction times for new faces were significantly higher than those for identity-change trials ($t(7) = 4.905, p < .01$, Figure 2.2).

We performed an additional 3-way ANOVA for repeated measures to investigate the influence of familiarity (including new faces), view and lighting on reaction time. This analysis yielded a significant main effect of familiarity ($F(1.148, \text{Greenhouse-Geisser corrected for non-sphericity}) = 23.110, p < .01$), attributable to significantly higher reaction times for new than seen, and new than known faces ($p < .01$ for both contrasts). There was no significant difference in reaction times for seen as compared to known faces. This suggests that the seen and known faces were equally perceptually familiar, consistent with the purpose of our familiarity manipulation. We also found a significant interaction effect between view and lighting ($F(1) = 9.398, p < .05$), due to higher reaction times for faces with incongruent as compared to congruent view and lighting.

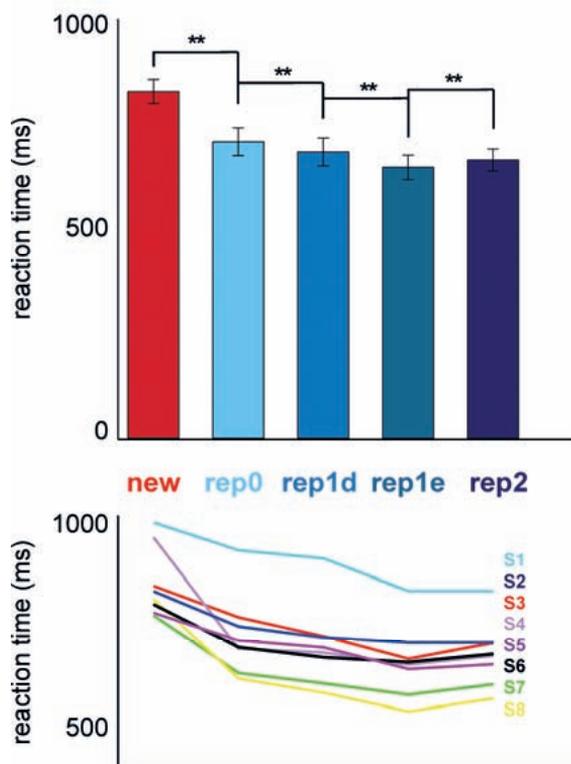


Figure 2.2 Reaction times for face-identity change (rep0) were higher than for face-identity repetition. The upper panel shows mean reaction times across subjects and associated error bars (random-effects standard error of the mean) for new faces, face-identity change trials, and face-identity consecutive repetition trials. Values for the rep3, rep4, and rep5 predictors are not shown, since these were based on only few trials. Four relevant contrasts (new versus rep0, rep0 versus rep1_different (rep1d), rep1_different versus rep1_exact (rep1e), and rep1_exact versus rep2) were tested for significance using paired t-tests. Significant contrasts are shown and denoted with ** ($p < .01$). The lower panel shows the reaction times of each individual subject for the conditions shown in the upper panel, in order to give a more detailed picture of the between subject variation.

2.3.2 fMRI results

Face-identity change versus consecutive face-identity repetition: identity change elicited more activation than repetition across early visual and posterior inferior temporal cortex

To investigate the effects of face-identity repetition, we contrasted face-identity-change trials with face-identity-repetition trials. In order to distinguish the effects of different-image face-identity repetition from the effects of exact-image repetition, we separately contrasted face-identity change (rep0) with different-image repetition (rep1_different) and with exact-image repetition (rep1_exact). Both contrasts showed a larger response for identity-change trials than identity-

repetition trials across early visual and posterior inferior temporal cortex (Figure 2.3) (thresholded using $FDR < .05$ for the contrast $rep0$ versus $rep1_different$, corresponding to a t -value of $|3.01|$). This face-identity-change effect was less widespread and more left-lateralized for the different-image contrast (Figure 2.3a) than for the exact-image contrast (Figure 2.3b). The regions that were activated more strongly for face-identity change than different-image face-identity repetition (Figure 2.3a) overlapped with anterior parts of EVC, posterior parts of OFA (little overlap), inferior medial parts of FFA, and inferior parts of PPA. A small additional cluster in anterior inferior temporal cortex (Talairach coordinates: 19, 2, -18) also responded more strongly to change than different-image repetition (Figure 2.3a). Larger, almost complete overlap was found between regions of interest (EVC, OFA, FFA, PPA) and regions showing stronger responses to face-identity change than exact-image repetition. Several small clusters in inferior frontal regions also responded more strongly to change than exact-image repetition (Figure 2.3b). Three small clusters (< 100 voxels) in left amygdala (-20, -8, -12), left middle temporal gyrus (-55, -34, -8), and left anterior middle temporal gyrus (-37, 16, -22) showed a smaller response to face-identity-change trials than to exact-image repetition trials (Figure 2.3b).

Contrasting activation to the first consecutive face-identity repetition (either $rep1_different$ or $rep1_exact$) with activation to the second consecutive face-identity repetition ($rep2$) yielded several clusters in early visual and posterior inferior temporal cortex that showed less activation to the second than to the first consecutive face-identity repetition (map thresholds identical to the threshold for the contrast $rep0$ versus $rep1_different$: $t=|3.01|$). Clusters found in the different-image contrast overlapped with clusters found in the exact-image contrast. Several additional clusters were found in the different-image contrast as compared to the exact-image contrast. These additional clusters were located in right early visual cortex and right posterior inferior temporal cortex.

ROI results were consistent with the mapping results and indicated decreased responses with face-identity repetition in EVC, OFA, FFA and PPA (Figure 2.4). All of these regions showed a decreased response to exact-image repetition as compared to face-identity change. Bilateral face-selective regions (OFA, FFA) and left EVC and PPA also showed a decreased response to different-image face-identity repetition as compared to face-identity change. A smaller response to second consecutive face-identity repetition than first consecutive face-identity repetition was found in right EVC for the exact-image repetition contrast ($rep1_exact > rep2$) (Figure 2.4) and in bilateral EVC and right OFA, FFA, and PPA for the different-image contrast ($rep1_different > rep2$) (significance not shown).

Our EVC ROI did not include cortex that represents the central visual field (foveal confluence of retinotopic areas V1/2/3), which is where we presented our stimuli. In order to investigate the effects of face-identity repetition in this region, we created two additional ROIs located at the left and right foveal confluence (FOV). These ROIs were centered at Talairach coordinates -29, -78, -11 and 25, -80, -9 (spherical ROIs with a volume of 1437 mm³ each, center Talairach coordinates taken from Dougherty et al., 2003). These ROIs as well showed a decreased response to face-identity repetition (both exact-image and different-image repetition) as compared to face-identity change (Figure S2.2). Hippocampus and aIT did not show significant face-identity repetition effects (Figure 2.4).

Different-image face-identity repetition elicited more activation than exact-image repetition in right posterior inferior temporal cortex

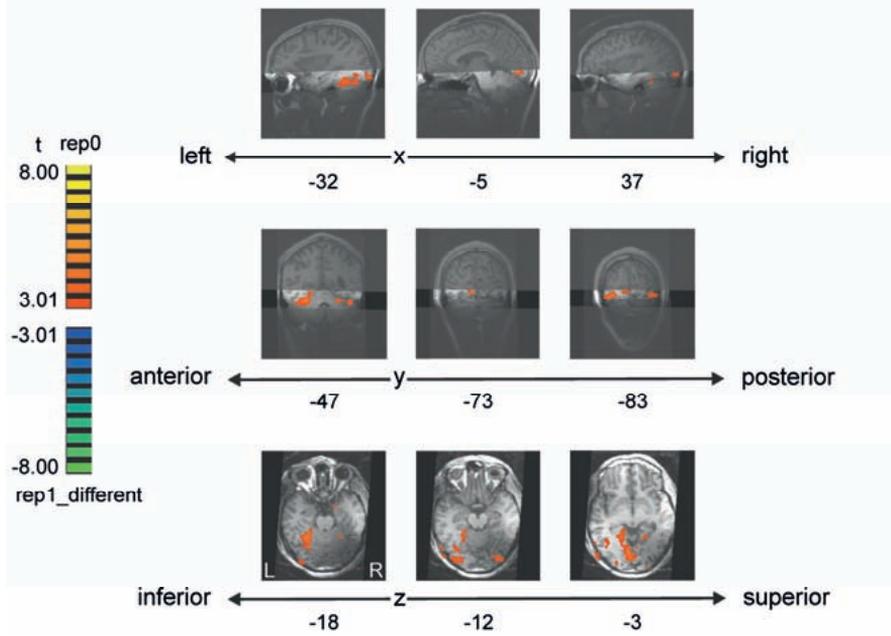
Contrasting activation to different-image face-identity repetition (rep1_different) with activation to exact-image repetition (rep1_exact) resulted in several clusters in posterior inferior temporal cortex that showed more activation for different-image than exact-image repetition (map threshold identical to the threshold for the contrast rep0 versus rep1_different: $t=|3.01|$, not shown). These clusters were mainly right-lateralized. Parts of these clusters overlapped with FFA and PPA. Exact-image repetition elicited more activation than different-image repetition in right amygdala (15, -4, -17) and left parahippocampal gyrus (-16, -26, -27) (map threshold: $t=|3.01|$). ROI results were consistent with the mapping results. Right FFA and PPA showed more activation to different-image face-identity repetition than exact-image repetition; activation to these two conditions was statistically indistinguishable in other regions (Figure 2.4).

Differences in strength of face-identity change effects across regions

We reported face-identity change effects in EVC, OFA, FFA and PPA. Previous studies have consistently reported face-identity change effects in face-selective regions (OFA, FFA), but not non-face-selective regions (EVC, PPA) (see Discussion). Could this discrepancy be explained by differences in the strength of the effect across regions? Related to that, could our widespread findings be explained by increased sensitivity due to the large amount of data we analyzed? In order to investigate these questions, we performed the following two analyses: 1) comparison of effect sizes across regions, and 2) analysis of a subset of our data (six runs per subject instead of eleven; we analyzed odd runs only).

a.

face-identity change > first consecutive different-image face-identity repetition



b.

face-identity change > first consecutive exact-image face-identity repetition

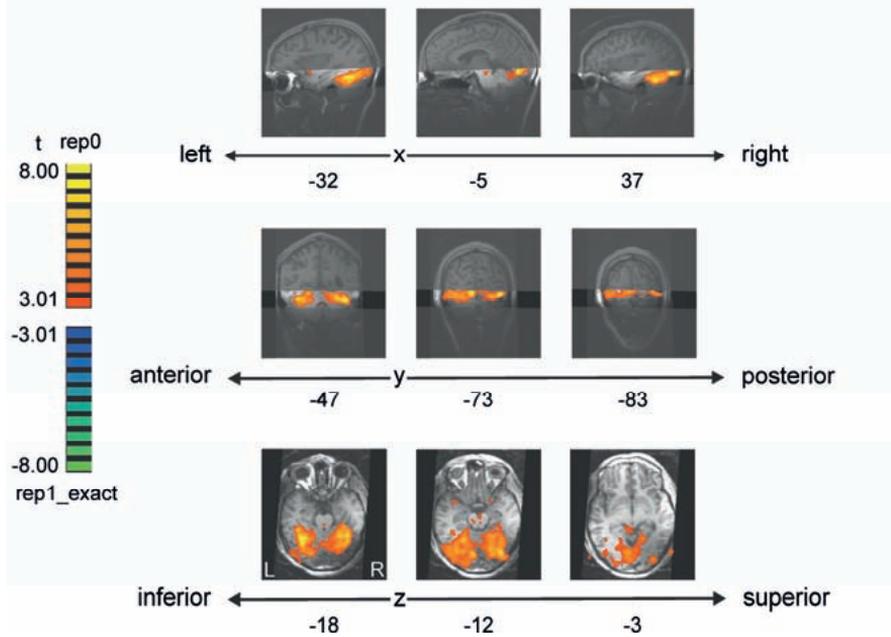


Figure 2.3 Face-identity change elicited more activation than face-identity repetition across early visual and posterior inferior temporal cortex (including regions that are not face-selective). Effects were more widespread for exact-image than different-image repetition. **(a)** Face-identity change (rep0) versus first consecutive different-image face-identity repetition (rep1_different) (FDR < .05). **(b)** Face-identity change versus first consecutive exact-image face-identity repetition (rep1_exact). In both panels, fixed-effects group results are displayed on single-subject high-resolution anatomical slices. The position of the measured slab is indicated by transparent masks overlaid on sagittal and coronal slices. Slices along different points on the x, y and z axes show stronger activation for rep0 than rep1 (orange/yellow) in early visual cortex as well as in inferior temporal regions, overlapping with OFA, FFA and PPA. More activation for rep1 than rep0 is shown in blue/green. The most superior slice along the z-axis shows activation based on only three-quarters of the data (data with very low slab position were removed).

(1) Effect-size comparison across regions. We performed effect-size comparisons for face-identity change versus first consecutive different-image face-identity repetition (rep 0 versus rep1_different, “different-image contrast”) and for face-identity change versus first consecutive exact-image repetition (rep0 versus rep1_exact, “exact-image contrast”). Different-image effect sizes were significantly smaller in right PPA than bilateral FFA ($p < .05$ for both comparisons) and left PPA ($p < .01$). Exact-image effect sizes were significantly larger in right FFA than in bilateral PPA, right EVC and right OFA ($p < .01$ for all comparisons). In addition, exact-image effect sizes were significantly smaller in bilateral PPA than in right OFA ($p < .05$ for both comparisons). These findings suggest that there are some differences in effect size across regions. In particular, face-identity change effects in (right) PPA seem to be smaller than face-identity change effects in face-selective regions. In addition, the strongest face-identity change effects are found in (right) FFA.

(2) Face-identity-change analysis on subset of data. We repeated the face-identity change analysis on a subset of our data to investigate the robustness of our effects. We used six runs per subject (odd runs only), which corresponds to about half of the data. As for the effect-size comparisons, we investigated the different-image contrast (rep0 versus rep1_different) and the exact-image contrast (rep0 versus rep1_exact). Contrast maps were thresholded using the FDR method to account for multiple testing (FDR < .05 for the contrast rep0 versus rep1_different, corresponding to a t-value of $|3.86|$). As for the full data set, face-identity change elicited more activity in early visual and posterior inferior temporal cortex. Nevertheless, the spatial extent of effects was noticeably smaller, especially for the different-image contrast (Figure S2.3). This reduction in spatial extent of the effects was less, but still noticeable, when the threshold was set to the threshold used for the full-data-set maps: $t=|3.01|$ (not shown). Reducing the amount of data affected different-image repetition effects in both face-selective and non-face-selective regions (see Figure S2.4, ROI analysis). Different-image repetition effects for OFA, FFA and PPA showed a reduction in significance or even disappeared (left PPA), while they appeared or increased in

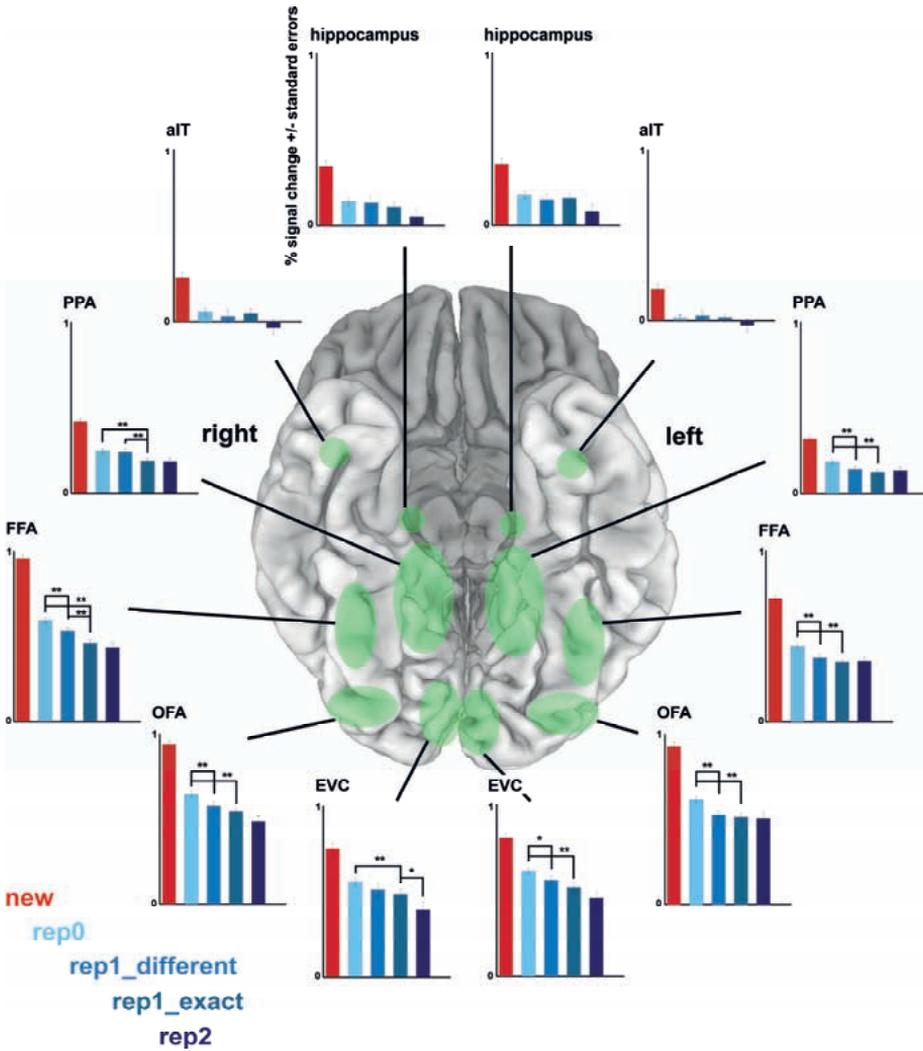


Figure 2.4 ROI analysis for face-identity repetition effects. Face-selective regions (OFA, FFA) as well regions that are not face-selective (EVC, PPA) showed more activation for face-identity change than repetition. These regions (including left EVC and PPA) showed these effects for both exact-image and different-image repetition. Approximate ROI locations are shown (in green) on a ventral view of the cortex (shown here: MNI template colin27). Graphs show beta-values and associated standard errors for the new (red), rep0 (i.e. identity change; light blue), rep1_different (blue), rep1_exact (blue grey), and rep2 (dark blue) predictors, averaged across subjects. Values for the rep3, rep4, and rep5 predictors are not shown, since these were based on only few trials. Six relevant contrasts (new versus rep0, rep0 versus rep1_different, rep0 versus rep1_exact, rep1_different versus rep1_exact, rep1_different versus rep2, and rep1_exact versus rep2) were tested for significance. The new versus rep0 contrast was significant for all tested regions ($p < .01$) (not shown). The rep1_different versus rep2 contrast was significant for bilateral EVC and right OFA, FFA, and PPA ($p < .05$) (not shown). For the other four tested contrasts, significant contrasts are shown and denoted with ** ($p < .01$) or * ($p < .05$). ROIs were defined using independent data. See Table 2.1 for abbreviations and ROI details (including ROI-defining contrasts).

significance for EVC. Exact-image repetition effects were of similar significance as for the full data set.

Effect-size comparisons between regions on the reduced data set showed that effect sizes in PPA were overall smaller than in OFA and FFA. In particular, different-image effect sizes were significantly smaller in right PPA than bilateral EVC, bilateral FFA, left OFA and left PPA ($p < .01$ for the comparison with rFFA, $p < .05$ for all other comparisons). Exact-image effect sizes were significantly smaller in bilateral PPA than bilateral OFA and right FFA ($p < .01$ for the comparisons with left OFA and right FFA, $p < .05$ for the comparison with right OFA). These effect-size-comparison results are consistent with the results of the full data set. In addition, they suggest that effect sizes in EVC are similar to effect sizes in face-selective regions.

Face-identity repetition effects were similar for seen and known faces

To test whether face familiarity would influence face-identity repetition effects, we compared face-identity repetition effects for seen and known faces. This comparison was made for each of the five contrasts that were investigated for main effects of face-identity repetition. These contrasts were: (1) rep0 versus rep1_different, (2) rep0 versus rep1_exact, (3) rep1_different versus rep1_exact, (4) rep1_different versus rep2, and (5) rep1_exact versus rep2. Mapping did not show significant differences in face-identity repetition effects between seen and known faces, except for the contrast between face-identity change (rep0) and first consecutive exact-image face-identity repetition (rep1_exact). Three small regions, located in posterior occipital cortex and cerebellum, showed the following interaction effect: for seen faces, activation for face-identity change was stronger than for exact-image repetition (rep0 > rep1_exact), while for known faces, the opposite pattern of response was found (rep0 < rep1_exact). One of these small clusters was located within our right EVC ROI. Maps were thresholded at a t-value of |3.65|, corresponding to FDR < .05 for the interaction contrast that compared the difference between rep0 and rep1_different for seen to that for known faces. ROI analysis indicated that right EVC, right aIT, and right hippocampus showed differential face-identity repetition effects for seen as compared to known faces for either one (EVC, FFA) or two (hippocampus) of the above contrasts comparing consecutive identity repetitions ($p < .05$ for each contrast). There was no clear pattern to these results: in some cases, seen faces showed a decrease in activation with repetition while known faces showed the opposite trend, and vice versa in others.

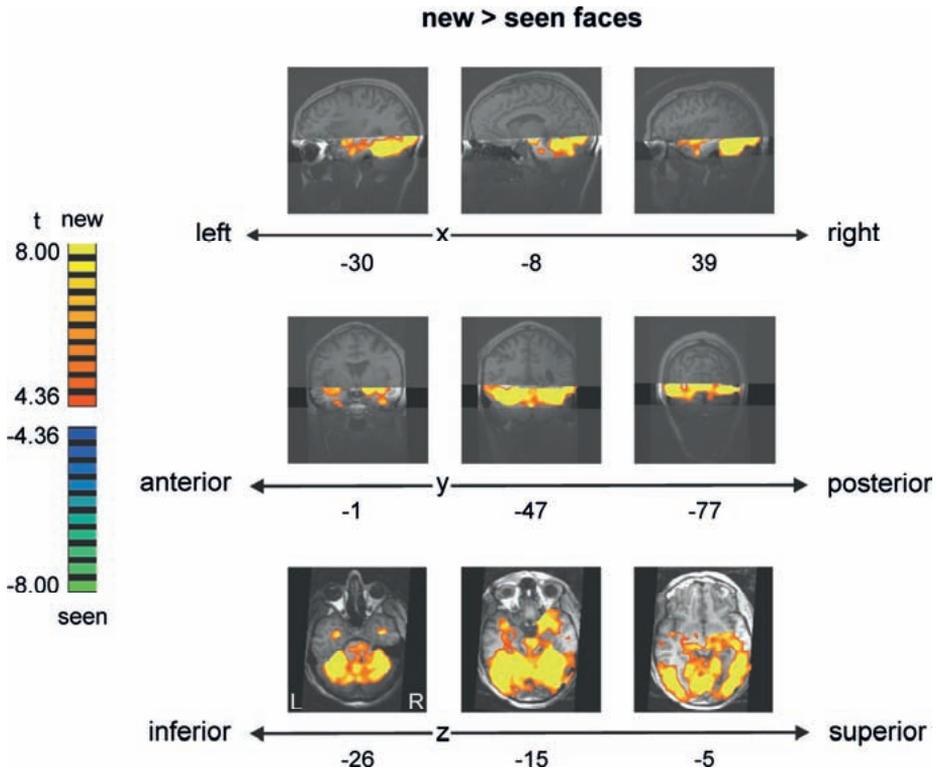


Figure 2.5 New faces elicited more activation than seen faces across early visual and inferior temporal cortex ($t=|4.36|$, associated with $FDR < .05$ for seen-known contrast, not shown). Fixed-effects group results are displayed on single-subject high-resolution anatomical slices. Position of the measured slab is indicated by transparent masks overlaid on sagittal and coronal slices. Slices along different points on the x, y and z axes show stronger activation for new than seen faces (orange/yellow) in early visual cortex as well as in (anterior) inferior temporal regions, including OFA, FFA, PPA, hippocampus and aIT. There were no regions showing more activation for seen than new faces. Note that the most superior slice along the z-axis shows activation based on only three-quarters of the data (data with very low slab position were removed).

New versus seen: new faces elicited more activation than seen faces across early visual and inferior temporal cortex

To investigate the influence of face novelty and perceptual face familiarity on face-related activation, we compared activation to new and seen faces. New faces (i.e. faces never seen before) elicited a larger response than seen faces across a large portion of occipital and inferior temporal cortex, including EVC, OFA, FFA, PPA, hippocampus and aIT (thresholded at $FDR < .05$ for the contrast seen versus known, corresponding to a t-value of $|4.36|$) (Figure 2.5). There were no regions that showed a larger response to seen than new faces. Consistent with these findings, all ROIs (see Table 2.1) showed an increased response to new as compared to seen faces (Figure 2.6).

Seen versus known: known and unknown familiar faces elicited equal activation

To localize activation associated with conceptual face information, we contrasted activation to seen (i.e. unknown familiar faces) with activation to known faces. Mapping did not yield regions activated significantly differently by seen than known faces (thresholded at $FDR < .05$ for the contrast seen versus known, corresponding to a t -value of $|4.36|$). The more powerful ROI analysis also did not reveal activation differences between seen and known faces (Figure 2.6).

2.4 Discussion

We investigated the effects of face-identity repetition and face familiarity on activation in human inferior temporal cortex. We observed a decreased BOLD response to repeated faces in OFA and FFA, as expected based on previous literature. However, these effects were not confined to face-selective regions: other regions in occipital and inferior temporal cortex, including early visual cortex (EVC) and PPA, displayed a similar effect. These face-identity repetition effects in face-selective and non-face-selective regions were present for both different-image and exact-image face-identity repetition, but were clearly reduced in spatial extent for different-image repetition. Previous studies have interpreted face-identity repetition effects in face-selective regions as an indication of the existence of specialized face-identity representations. Following this logic, our results could be taken as evidence for the presence of face-identity representations outside of face-selective regions. However, this interpretation is not plausible for early visual cortex and PPA given their known response properties. Alternative interpretations include residual attentional effects (despite our task control) and carry-over of activation from face-identity regions to other visual regions. These alternative explanations, which are discussed below, could also apply to the identity-change effects in face-selective regions, including FFA. Face-identity repetition effects were similar for seen and known faces. A direct comparison of activation to seen and known faces did not yield significant results. The infrequent new faces (never seen before), which were excluded from the identity-repetition analysis, elicited a stronger response across a large portion of occipital and inferior temporal cortex than the four familiar faces (see Supplementary Discussion for a more detailed discussion on our face-familiarity findings).

2.4.1 Face-identity repetition effects in inferior temporal cortex

Consistent with previous studies, we found a greater response to face-identity change than repetition in face-selective regions (Andrews and Ewbank, 2004;

Gauthier et al., 2000b; Pourtois et al., 2005; Winston et al., 2004, but see Epstein and Kanwisher, 1999).

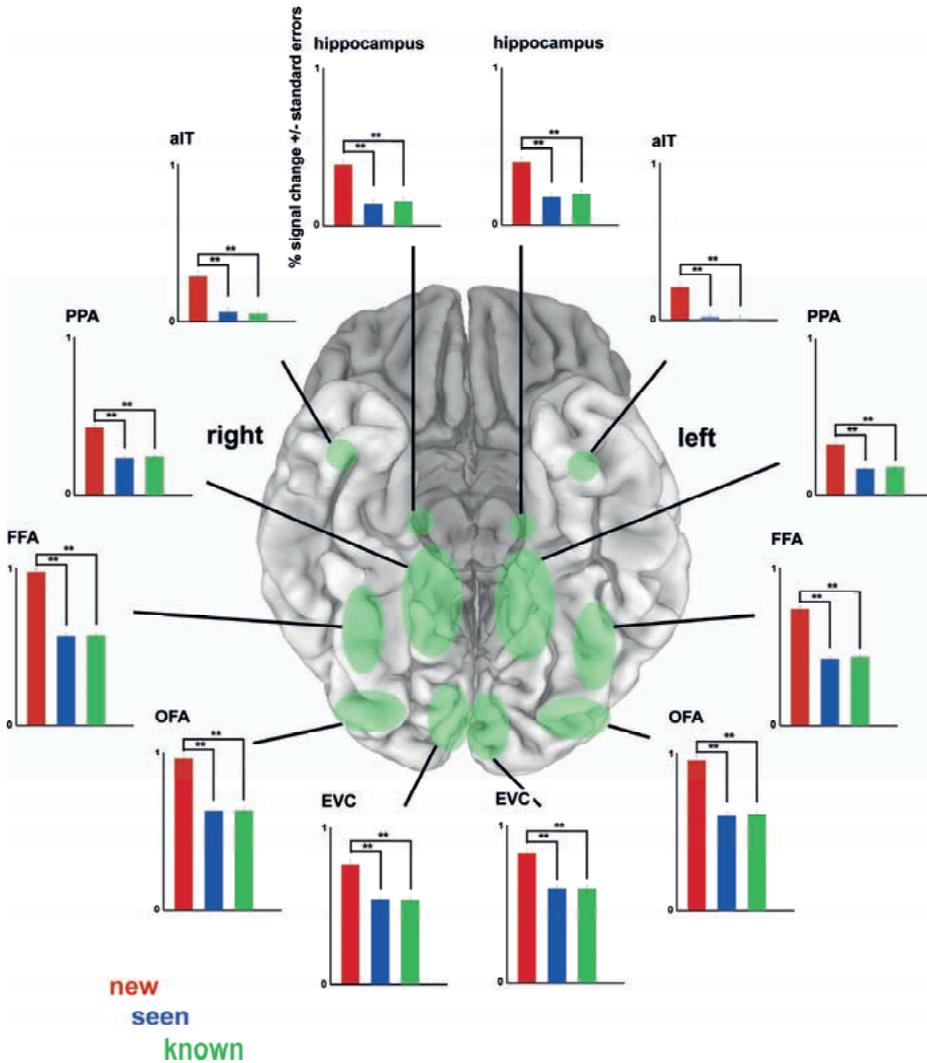


Figure 2.6 ROI analysis for face-familiarity effects. All ROIs showed more activation to new than seen and known faces. Seen and known faces elicited equal activation. Approximate ROI locations are shown (in green) on a ventral view of the cortex (shown here: MNI template colin27). Graphs show percent signal change and associated standard errors for the new (red), seen (blue) and known (green) predictors, averaged across subjects. All possible contrasts (new versus seen, new versus known, and seen versus known) were tested for significance. Significant contrasts are shown and denoted with ** ($p < .01$). ROIs were defined using independent data. See Table 2.1 for abbreviations and ROI details (including ROI-defining contrasts).

Interestingly, we observed similar face-identity repetition effects in non-face-selective regions in inferior temporal cortex (including PPA). Face-identity repetition effects outside of face-selective regions have been reported in several previous studies (Dricot et al., 2008ab; Ng et al., 2006; Pourtois et al., 2005; Sugiura et al., 2001). Widespread repetition-related response decreases have also been found using stimuli other than faces (Epstein et al., 2003) and using designs that blocked stimuli by category (Avidan et al., 2002; Chao et al., 2002). Other studies did not find face-identity repetition effects in non-face-selective inferior temporal cortex (Andrews and Ewbank, 2004; Gauthier et al., 2000b; Henson et al., 2002; Henson and Mouchlianitis, 2007; Rotshtein et al., 2005; Summerfield et al., 2008), or did not investigate activity in these regions (Eger et al., 2005; Loffler et al., 2005; Winston et al., 2004). Possible explanations of differences between studies in spatial extent of repetition effects will be discussed below.

Face identity need not be represented in face-selective regions. In line with this thought, face-identity repetition effects in inferior temporal regions that are not face-selective have often been interpreted as evidence for the existence of neuronal face-identity representations in these regions (e.g. Avidan et al., 2002; Dricot et al., 2008ab). Pattern-information analysis as well has suggested the existence of face-identity representations in a region that did not show a clear face-selective response (i.e. right anterior temporal cortex) (Kriegeskorte et al., 2007). PPA could theoretically contain small subsets of “face-identity neurons” that give rise to face-identity change effects (as suggested by Avidan et al., 2002). However, to our knowledge, there is currently no direct evidence for this possibility; it has merely been used as an interpretation of fMRI-adaptation findings. Based on its functional response properties, PPA is an unlikely candidate for representing face-identity. PPA responds strongly to scenes of the local visual environment and only weakly to faces (Epstein and Kanwisher, 1998). Consistent with this, a large proportion of parahippocampal neurons in macaque prefers eccentric stimulus positions, and only a small proportion responds to complex object images (Sato and Nakamura, 2003). Furthermore, the most prominent consequence of parahippocampal lesions is loss of the ability to navigate through spatial environments (Aguirre and D’Esposito, 1999). These response properties suggest that PPA is involved in processing spatial environments, not face-identity recognition.

2.4.2 Face-identity repetition effects in early visual cortex

We found face-identity repetition effects even in early visual cortex (including V1 and possibly portions of V2/3). Such effects have not previously been discussed to our knowledge. It appears unlikely that these effects reflect a domain-specific face-identity representation in early visual cortex. Face identity is a

high-level stimulus feature and early visual cortex is known to be sensitive to low-level stimulus properties. Activity in V1, V2, and V3 is modulated by varying low-level stimulus properties, including orientation, spatial frequency and direction of motion (e.g. Hubel and Wiesel, 1968; Levitt et al., 1994; Gegenfurtner et al., 1997). Sensitivity to low-level stimulus properties could underlie exact-image face-identity repetition effects. However, the different-image face-identity repetition effects we found are hard to explain in terms of sensitivity to low-level stimulus properties: identity change was not confounded with low-level feature change, because view and lighting changes (associated with large low-level feature changes) occurred on identity-change as well as on different-image identity-repetition trials. Nevertheless, considerable proportions of neurons in V2 and V3 have been shown to also respond to more complex stimulus features, including combinations of orientations and (moving) gratings (Anzai et al., 2007; Gegenfurtner et al., 1997; Hegdé and Van Essen, 2000). However, it appears unlikely that these response characteristics produce sensitivity to face features with invariance across view or lighting changes.

Our study is not the first to report fMRI-adaptation effects that are inconsistent with known functional properties of early visual areas (in particular V1). Boynton and Finney (2003) failed to find orientation-sensitive fMRI-adaptation effects in V1 (but see Tootell et al., 1998) despite neurophysiological evidence for sensitivity to orientation and spatial frequency in V1 (Hubel and Wiesel, 1968; Movshon and Lennie, 1979; Mueller et al., 1999), possibly attributable to the short adaptation duration that was used (Fang et al., 2005). Similar discrepancies between short-term fMRI-adaptation effects and known sensitivity to orientation were found for area V2 (Boynton and Finney, 2003; Fang et al., 2005). Another discrepancy can be found in Chao et al. (2002), who reported long-term repetition effects in early visual cortex for complex object stimuli (animals and tools). It would not be in line with known response properties of early visual cortex to interpret these findings in terms of invariant object representations. A more likely explanation of these findings would be that Chao et al. (2002) used exact-image repetitions: object changes were associated with low-level feature changes, while object repetitions were not. These reports, as well as the face-identity repetition effects in early visual cortex that we report here, indicate that fMRI-adaptation results might (1) not accurately reflect neuronal sensitivity profiles in early visual areas and (2) reflect sensitivity to stimulus properties other than the stimulus property of interest (particularly when exact-image repetitions are used).

2.4.3 Alternative explanations for stimulus-change responses

The fMRI-adaptation paradigm is based on the logic that stimulus-change fMRI effects in a specific brain region can be interpreted to indicate that the region

contains neurons that are sensitive to the changed stimulus property (for a review, see Grill-Spector et al., 2006). Sensitivity to a stimulus property is taken to indicate that the brain region represents that stimulus property (e.g. face identity). However, the findings from EVC and PPA suggest that some caution is needed when interpreting fMRI stimulus-change effects in any brain region in terms of neuronal sensitivity for the changed stimulus property. Additional support for this assertion can be found in two studies that directly investigated the relationship between neuronal selectivity as measured by classical electrophysiological methods and neuronal adaptation measured using an adaptation paradigm in higher-order visual cortex (Tolias et al., 2004; Sawamura et al., 2006; see also Krekelberg et al., 2006). Results from these studies indicated that selectivity inferred from adaptation does not consistently match directly-measured neuronal selectivity. These considerations render an interpretation of fMRI-adaptation findings in terms of local neuronal sensitivity for the changed stimulus property no more likely than other possible interpretations. We will discuss three alternative explanations for stimulus-change responses.

(1) Automatic attention. At the cognitive level of description, a plausible alternative interpretation of our effects is that a change in face identity detected by the subject results in an attentional response. Such a response could activate a wider network within the visual system. Our task drew attention away from differences among the four familiar faces. However, the attentional response to face changes could be automatic and task-independent. Under natural conditions, a new face implies the presence of a new person to be recognized, and recognition will typically be followed by more general memory access, and a host of other processes required for appropriate behavior. Attention has been shown to enhance responses to preferred stimuli in object-selective cortex (Murray and Wojciulik, 2004; O'Craven et al., 1999; Wojciulik et al., 1998) as well as early visual regions (Liu et al., 2005). In addition, attention has been shown to modulate repetition effects (Eger et al., 2004; Henson and Mouchlianitis, 2007; Murray and Wojciulik, 2004), showing decreased or abolished repetition effects for ignored as compared to attended stimuli. Task manipulations that affect the amount of attention allocated to a stimulus can also influence the strength of repetition effects (for an example, see Henson et al., 2002). Together, these results suggest that attention plays an important role in repetition-related brain responses. If a face-identity change triggered attention automatically, it could give rise to increases in activity that surpass the location where face identities are distinguished. While the strong response to the infrequent "new" faces can be accounted for as an oddball effect, our face-identity change findings among the four familiar faces cannot be explained as an oddball effect because identity-change trials were much more frequent than identity-repetition trials (70% and 30% of all presented familiar faces, respectively). An automatic atten-

tional response to face changes (even when they occur on most trials) is a more plausible explanation.

(2) *Carry-over of activation.* At the neuronal level of description, a possible cause of face-identity change responses outside face-identity regions is activation carry-over: the region distinguishing the identities and therefore exhibiting release from adaptation might activate connected regions. An example of carry-over in the visual system can be found in Tolias et al. (2004). They showed by cell recording that neurons in macaque area V4, which are not generally selective for direction of visual motion, nevertheless respond to *changes* of the direction of motion. They interpret this finding in terms of activation carry-over from area MT/V5, whose neurons are strongly selective for direction of motion. A V4 cell pooling outputs from MT/V5 cells across all directions, would not be sensitive to direction per se, but it would reflect a release from adaptation occurring in MT/V5 after a direction change. Carry-over of activation, thus, need not imply carry-over of neuronal tuning properties. In other words, carry-over could be unspecific: activation could be passed on without relaying stimulus information. Alternatively, stimulus information (e.g. face-identity information) could truly be passed on from one region to connected regions (specific carry-over). In either case, face-identity change could activate regions not primarily involved in representing face identity. Carry-over may explain our face-identity change effects in early visual regions. Feedback, which can be seen as a form of carry-over, from higher-order visual regions involved in face-identity representation could have activated early visual cortex (see also Williams et al., 2008). Such spreading (or carry-over) of activation could be functionally interpreted as an attentional effect, but carry-over could also occur in the absence of an attentional effect. Carry-over could, for example, activate a specialized network (e.g. the face network) to initiate a more comprehensive cognitive process (e.g. recognition, memory access, and response selection).

(3) *Neuronal sensitivity to stimulus changes.* Another way in which changing a specific stimulus property could elicit effects that might not reflect neuronal sensitivity to this property, is if these effects instead reflect processing of the change itself. For example, an abrupt change of stimulus position can elicit an apparent motion percept, activating the motion-sensitive human middle temporal region (hMT/V5+) (e.g. Muckli et al., 2005). By the logic of fMRI adaptation, a position-change effect in hMT/V5+ could be interpreted as indicating that the region represents the spatial location of static visual objects. However, this would not be consistent with what is known about hMT/V5+. Instead hMT/V5+ responds to visual motion, i.e. the change of spatial location. A change-detection explanation is most compelling when the change in question occurs under natural conditions. This is not the case for our study, since faces do not naturally morph from one identity to another. However, change detection is central to

visual perception. The involvement of a more general change detection mechanism cannot be ruled out. Note that neuronal adaptation provides one possible mechanism for change detection, but other mechanisms, including the Reichardt motion detector (Reichardt 1969), can serve this purpose as well.

The above face-identity-change explanations could account for neuroimaging findings associated with the behavioral face-inversion effect (Yovel and Kanwisher, 2005; Mazard et al., 2006): face-identity change might not be detected with equal reliability for upside-down faces, and therefore fail to engage general attentional or carry-over mechanisms, resulting in comparable activation for face-identity change and repetition. Similar reasoning would explain the absence of face-identity change effects for upright faces in patients with acquired prosopagnosia or developmental prosopagnosia (Schiltz et al., 2006; Williams et al., 2007a): if face-identity repetition is perceptually indistinguishable from face-identity change, then change and repetition will elicit equal activation. Interestingly, Avidan et al., (2005) reported intact face-identity change effects in congenital prosopagnosic patients. This finding seems inconsistent with the above face-identity change interpretations, however, it can be explained by an effect of *stimulus* change (face-identity change was associated with physical stimulus change while face-identity repetition was not).

2.4.4 Why did several previous studies fail to report widespread face-identity change effects?

Several previous studies reported widespread face-identity change effects consistent with our present results (Ng et al., 2006; Pourtois et al., 2005; Sugiura et al., 2001). Other studies, however, have found effects restricted to face-selective regions. This discrepancy suggests that features of the experimental design might influence the strength and spatial extent of repetition effects. We consider six different features in turn.

(1) Different-image repetition versus exact-image repetition. Most studies that reported face-identity change effects outside of face-selective regions included exact-image repetitions (e.g. Ng et al., 2006; Pourtois et al., 2005; Sugiura et al., 2001), but so did most studies that reported face-identity change effects confined to face-selective regions (e.g. Andrews and Ewbank, 2004; Gauthier et al., 2000b; Henson et al., 2002). In our study, we could directly compare the effects of different-image face-identity repetition to the effects of exact-image face-identity repetition. Face-identity change effects involving exact-image repetition were clearly more widespread than face-identity change effects involving different-image repetition (consistent with Pourtois et al., 2005). Non-face-selective regions in the right hemisphere did not respond more strongly to face-identity changes than to different-image face-identity repetitions, but did respond more

strongly to face-identity changes than to exact-image face-identity repetitions (see Vuilleumier et al., 2002 for a similar laterality effect). A likely explanation for these findings is that face-identity change trials are confounded with stimulus change for the exact-image comparison, but not for the different-image comparison. This could elicit adaptation effects in any region with sensitivity to any of the changed stimulus properties. Alternatively, this could result in a larger attentional response to face-identity change for the exact-image than the different-image comparison. Nevertheless, face-identity change effects involving different-image repetition were still quite widespread, i.e. these effects were found in face-selective regions as well as outside face-selective regions (e.g. left EVC and PPA). These results suggest that the use of exact-image repetitions contributes to widespread face-identity change effects, but cannot by itself explain the existence of face-identity change effects outside of face-selective regions.

(2) *Temporal lag between presentations.* Another factor that has been shown to influence repetition effects is the temporal lag between the first and second presentation of a stimulus. Immediate repetition (i.e. no intervening stimuli) and delayed repetition (i.e. other stimuli intervening) are associated with qualitatively different behavioral and neuronal effects (Bentin and Moscovitch, 1988; Bentin and Feled, 1990; Epstein et al., 2008). The effects of immediate repetition reported in Epstein et al. (2008) seem to be more widespread than those of delayed repetition, especially in posterior visual cortex. This finding seems consistent with our data and interpretation: immediate repetition might elicit a stronger attentional response. Repetition-lag by itself cannot account for the differences in spatial extent of repetition effects between studies: Andrews and Ewbank (2004) used immediate repetition and found effects confined to face-selective regions; Pourtois et al. (2005) used delayed repetition and found more widespread repetition effects (their Figure 2).

(3) *Repetition frequency versus change frequency.* A third factor that has been shown to modulate repetition effects is frequency of repetition. Repetition frequency influences the subject's expectations and can affect the strength of repetition effects (Summerfield et al., 2008): effects are reduced when repetitions occur with relatively low probability. This modulation of effect strength is consistent with an attentional interpretation: infrequent changes are "oddballs" and will trigger a larger attentional response. The spatial extent of repetition effects did not seem to be influenced by probability of repetition (Summerfield et al., 2008). Note that changes were frequent in our study (70% of the trials were face-identity change trials), thus the oddball explanation cannot account for our findings.

(4) *Stimulus variety.* All previous studies reporting widespread repetition effects used stimuli from one category only (Epstein et al., 2003; Ng et al., 2006; Pour-

tois et al., 2005; Sugiura et al., 2001), or blocked stimuli by category (Avidan et al., 2002; Chao et al., 2002). However, other studies with limited stimulus variety did not report widespread effects (Andrews and Ewbank, 2004; Henson et al., 2002).

(5) *Number of distinct stimuli.* Fifth, our study used a relatively small stimulus set (16 different familiar face images: 4 identities, 2 views, 2 lightings), which could have led to automatic stimulus-response binding (Dobbins et al., 2004). This could possibly have resulted in repetition effects in regions that are not primarily involved in representing face-identity. However, it is important to note that both immediate face-identity change and repetition trials can be considered delayed repetitions of the 16 specific face images. In our study, automatic stimulus-response binding would therefore apply equally to repetitions and changes.

(6) *Statistical power.* Our study had a larger amount of data than most previous studies. This might have provided us with increased power to detect widespread face-identity change effects. In order to test this possibility, we repeated our face-identity change analysis on only half of our data. This control analysis showed an overall reduction in the spatial extent of our face-identity change effects, especially for face-identity change effects involving different-image repetition. Consistent with this, ROI analysis indicated that reducing the amount of data did not significantly affect exact-image repetition effects, but did affect different-image repetition effects in both face-selective and non-face-selective regions. The strongest different-image effect reduction was seen in left PPA: the effect disappeared, resulting in an absence of the different-image effect in bilateral PPA. In contrast to the other regions, different-image repetition effects in EVC became stronger. These differences between regions suggested by our control analysis are consistent with results of effect-size comparisons between regions. These comparisons showed that face-identity change effects overall were smaller in PPA (but not EVC) than in face-selective regions. This was true for the full as well as the reduced data set and for face-identity change effects involving different-image as well as exact-image repetition. In conclusion, our different-image effects reported for left PPA might indeed be due to increased statistical power. This does not mean that there are no face-identity change effects in PPA; it does indicate that face-identity change effects in PPA are weaker than in face-selective regions and EVC.

Our findings indicate several possible causes of widespread face-identity change effects. In the literature, similar effects have sometimes, but not always, been reported. No single design feature has consistently been associated with widespread face-identity change effects. Combinations of design features might explain the discrepancies between studies in spatial extent of stimulus-change

effects. For example, frequent, immediate, exact-image repetitions of stimuli from one object category could be associated with attentional effects and elicit more widespread repetition effects. Some of these features apply to our design. However, it is important to note that our study is not special in this regard: most fMRI-adaptation studies use designs that include several of these features. Finally, even if widespread effects can be avoided by means of particular repetition designs, this would not prove that repetition-related effects indicate neuronal tuning.

2.4.5 Implications for the interpretation of FFA results

Our findings question the interpretation of face-identity change effects as conclusive evidence for the presence of neurons tuned to face identity. The alternative explanations are likely to hold for the non-face-selective regions EVC and PPA. They might also hold for FFA. Direct evidence for face-identity tuning could be provided by fMRI or cell recording of responses to single face-image presentations. In the macaque, single-unit recordings from the middle face patch, a possible homologue of the human FFA, suggested that the face-category effect is dominant, but that the region does carry some amount of face-identity information in its population response as well (Tsao et al., 2006). Using high-resolution fMRI and pattern-information analysis, we have previously attempted and failed to detect face-identity information in the FFA or its vicinity, although we did detect such information in right aIT (Kriegeskorte et al., 2007). Current evidence strongly suggests that the FFA serves a key role in face recognition. Consistent with such a role, the strongest face-identity change effects in our study were found in right FFA. However, the evidence that its role consists in distinguishing individual faces is not conclusive.

2.4.6 Conclusion

We reported widespread effects of face-identity change despite well-controlled stimuli. Effects were found in face-selective and non-face-selective regions in inferior temporal cortex and in early visual cortex. These effects were found for exact-image face-identity repetition as well as for different-image face-identity repetition, although exact-image repetition was associated with more widespread effects than different-image repetition. Face-identity-change effects found in previous fMRI-adaptation studies have commonly been interpreted to indicate the existence of face-identity representations. However, alternative interpretations, including general attentional and activation carry-over effects, are plausible as well and better account for our widespread effects. These alternative interpretations might also contribute to face-identity change effects in face-selective regions, including FFA whose fMRI activity patterns do not strongly distinguish individual faces (Kriegeskorte et al., 2007).

More generally, fMRI stimulus-change effects are widely interpreted in terms of neuronal sensitivity. This interpretation promises to reveal information represented in fine-grained patterns of activity even within a single voxel. However, our findings add to the evidence (Tolias et al., 2004; Sawamura et al., 2006) that stimulus-change effects do not provide conclusive evidence for neuronal tuning to the changed stimulus property.

2.5 Supplementary material

2.5.1 Supplementary methods

The effects of face-identity repetition and face familiarity were the main focus in the current study. However, there were several additional questions that could be addressed using the current data.

One of these questions was whether there were any regions that were activated differently by different face identities. We therefore constructed a third model to investigate these face-identity effects. It consisted of subject-specific predictors for new faces and for each of the four face identities, and confound-mean predictors for each subject and run (Figure S2.1c). We performed six pairwise contrasts, one for each combination of the four face identities. All of these contrasts searched for brain regions that discriminate different face identities based on their activation.

Another question concerned the influence of view and lighting on face-related activity along the human (ventral) visual system. To investigate the effects of changes in view (V) and lighting (L), we constructed a fourth model. This model consisted of the following subject-specific predictors: left V-left L, left V-right L, right V-left L, right V-right L and new faces, and confound-mean predictors for each subject and run (Figure S2.1d). Orthogonal contrasts were used to test for main effects of view and lighting and for their interaction. Results were corrected for serial autocorrelation in the temporal domain. Maps were thresholded at a false discovery rate (FDR) of .01. ROI analysis for face-identity and view-and-lighting effects was performed as described in the method section of the main paper. Since the contrasts that tested for face-identity effects all investigated the same conceptual question (i.e whether a specific ROI was activated differently by different face identities), the face-identity results were corrected for multiple comparisons using Bonferroni correction.

2.5.2 Supplementary results

Face-identity: weak regional activation differences between the four individual faces

In order to find out whether there were any regions discriminating individual faces based on their regional-mean response amplitude, we performed six pairwise contrasts, one for each combination of the four familiar faces. Mapping yielded significant results for only one of these comparisons ($FDR < .01$). Bilateral regions, overlapping with OFA and with the inferior and posterior parts of FFA, showed stronger activation for one face identity than for the other. The more powerful ROI analysis yielded effects in the same regions: OFA and FFA both showed significant effects for three out of the six pairwise face-identity comparisons (including the contrast that yielded significant mapping results), whereas the other ROIs did not ($p < .05$ Bonferroni-corrected).

View and lighting effects: counterlit faces elicited more activity than congruently lit faces in early visual and posterior inferior temporal cortex

We first investigated the main effects of view and lighting by comparing left with right view, and left with right lighting. A main effect of view was observed, indicated by a larger response to faces seen from a left than right view in left early visual regions and posterior medial fusiform gyrus. Regions with similar locations in the right hemisphere showed the opposite effect. This effect was qualitatively similar across the two different lightings and can be explained by the location of the fixation cross. As shown in Figure S2.1, the fixation cross was placed close to the bridge of the nose. Consequently, faces seen from a left view covered a larger part of the right visual hemifield than the left, and therefore elicited stronger responses in early visual regions in the left hemisphere, and vice versa. Comparing left with right lighting did not yield any significant results.

We then investigated the interaction between view and lighting by contrasting faces with incongruent view and lighting (counterlit faces, Figure 2.1a) with congruently lit faces. Counterlit faces elicited a larger response than congruently lit faces in early visual regions (predominantly left), bilateral lateral occipital regions and bilateral posterior medial fusiform gyrus ($FDR < .01$). Interestingly, this effect was greater in left than right early visual regions. The above regions overlapped with the lateral part of left EVC, the superior parts of OFA and the very posterior parts of FFA and PPA. These results indicate that, even though all faces in the current experiment had the same light energy, there still is an effect of view and lighting combined. A possible explanation might be that faces encountered in daily life are more often congruently lit than counterlit. Interestingly, our reaction time data showed higher reaction times to counterlit than

congruently lit faces. These findings indicate an association between behavioral and neuroimaging measures.

ROI results for view and lighting effects were consistent with the mapping results. Overall, these results suggest that face view and lighting modulate face-related activity early in the visual processing stream, but less so in higher-order visual regions, consistent with the idea that higher regions abstract from these accidental properties to some degree.

2.5.3 Supplementary tables and figures

Table S2.1 Trial counts for new and familiar faces for the full set of data.

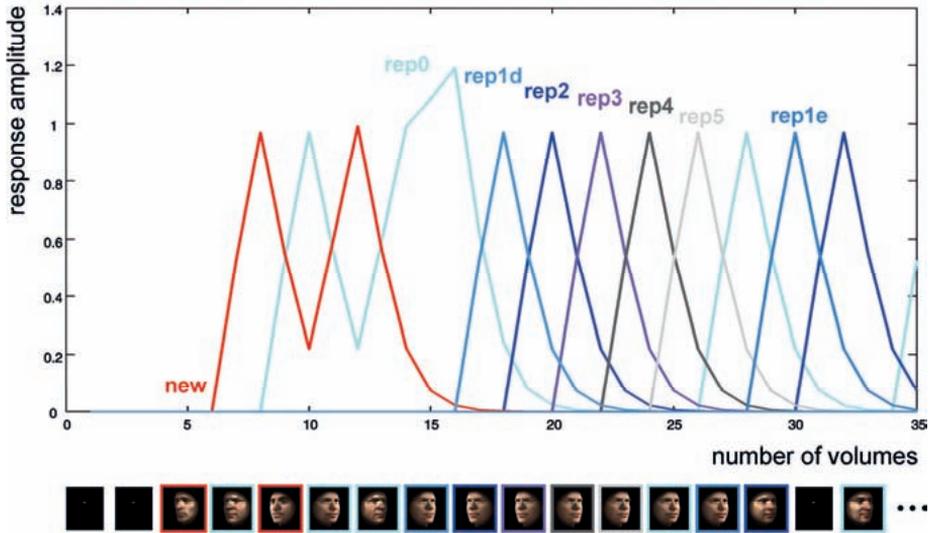
(8 subjects, 11 runs/subject)

Condition	Count	Proportion
new	880	
familiar	8448	
rep0	5929	0.70
(change)		
rep	2519	0.30
(consecutive face-identity repetition)		
rep1	2011	0.80
different	760	0.38
exact	1251	0.62
rep2	376	0.15
different	281	0.75
exact	95	0.25
rep3	90	0.04
different	39	0.43
exact	51	0.57
rep4	37	0.01
different	31	0.84
exact	6	0.16
rep5	5	0.00
different	0	0
exact	5	1.00
baseline	1760	
total	11088	

The trial counts for familiar faces are split out for the different face-identity-repetition conditions. Trial counts of subordinate categories (rep0, rep, and rep1-5) are expressed as proportion of their superordinate category.

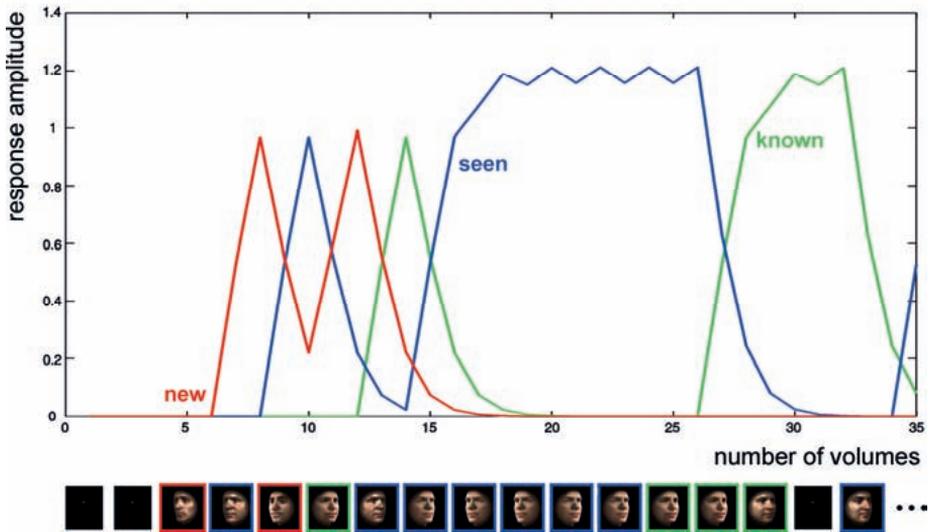
a.

consecutive face-identity repetition model



b.

face-familiarity model



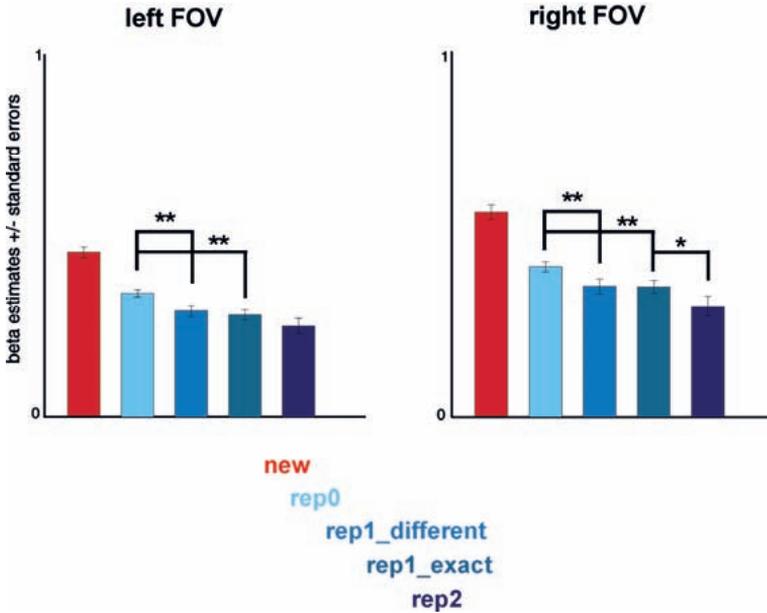
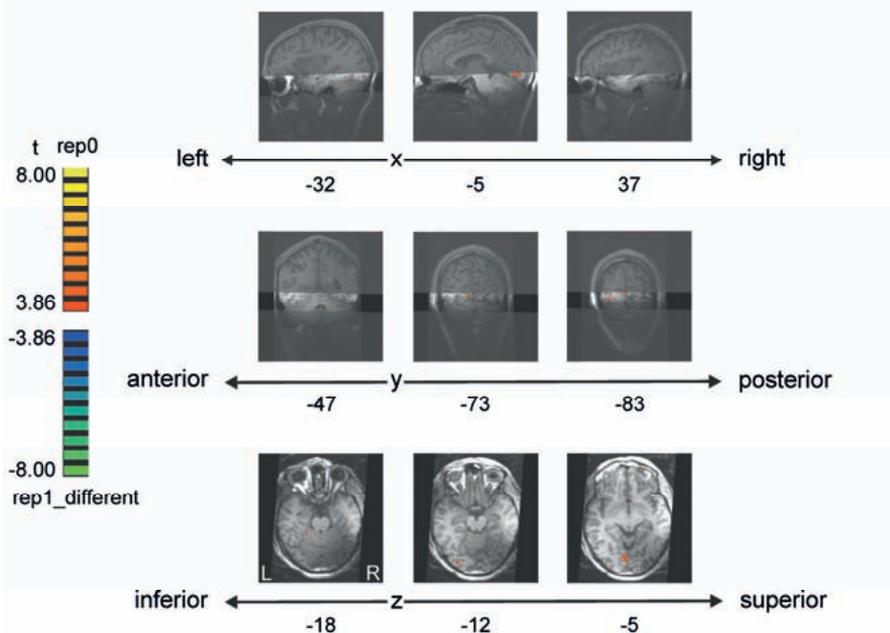


Figure S2.2 The bilateral foveal confluence (FOV) of retinotopic areas V1/2/3 also showed more activation for face-identity change than repetition. This effect was present for both exact-image and different-image face-identity repetition. Graphs show beta-values and associated standard errors for the new (red), rep0 (i.e. identity change; light blue), rep1_different (blue), rep1_exact (blue grey), and rep2 (dark blue) predictors, averaged across subjects. Values for the rep3, rep4, and rep5 predictors are not shown, since these were based on only few trials. Six relevant contrasts (new versus rep0, rep0 versus rep1_different, rep0 versus rep1_exact, rep1_different versus rep1_exact, rep1_different versus rep2, and rep1_exact versus rep2) were tested for significance. For both regions, the new versus rep0 contrast was significant ($p < .01$) and the rep1_different versus rep2 contrast was marginally significant ($p < .10$) (both not shown). For the other four tested contrasts, significant contrasts are shown and denoted with ** ($p < .01$) or * ($p < .05$). We created our left and right FOV ROIs by placing spheres with a volume of 1437 mm³ each at center Talairach coordinates taken from Dougherty et al., 2003).

a.

face-identity change > first consecutive **different-image** face-identity repetition

b.

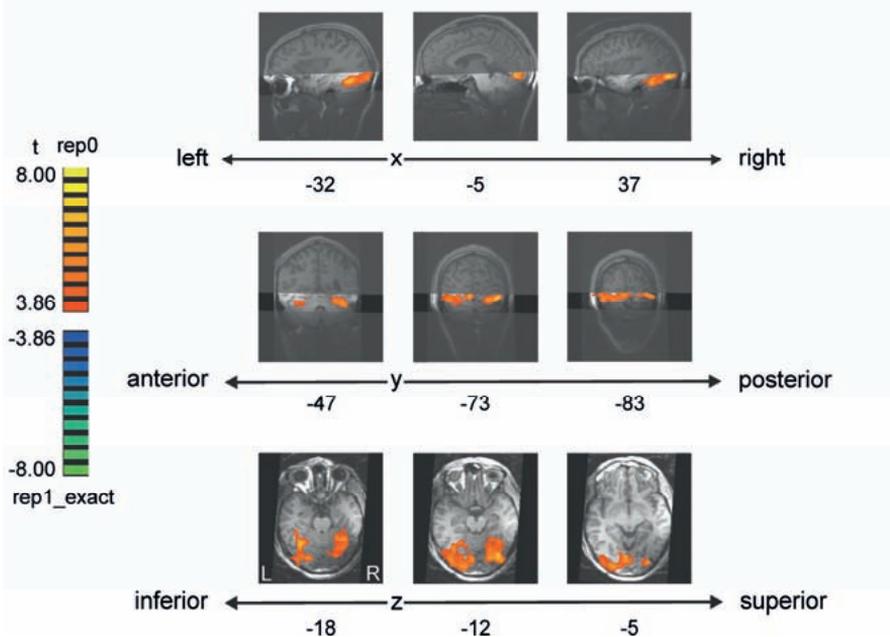
face-identity change > first consecutive **exact-image** face-identity repetition

Figure S2.3 Control analysis for face-identity repetition using half of the data (six runs per subject). As compared to the full data set results (Figure 2.3), effects are less widespread, which is clearly visible for the different-image contrast. **(a)** Face-identity change (rep0) versus first consecutive different-image face-identity repetition (rep1_different) (FDR < .05). **(b)** Face-identity change versus first consecutive exact-image face-identity repetition (rep1_exact). In both panels, fixed-effects group results are displayed on single-subject high-resolution anatomical slices. Position of the measured slab is indicated by transparent masks overlaid on sagittal and coronal slices. Slices along different points on the x, y and z axes show stronger activation for rep0 than rep1 (orange/yellow) in early visual cortex as well as in inferior temporal regions, overlapping with OFA, FFA and PPA. There were no regions showing more activation for rep1 than rep0.

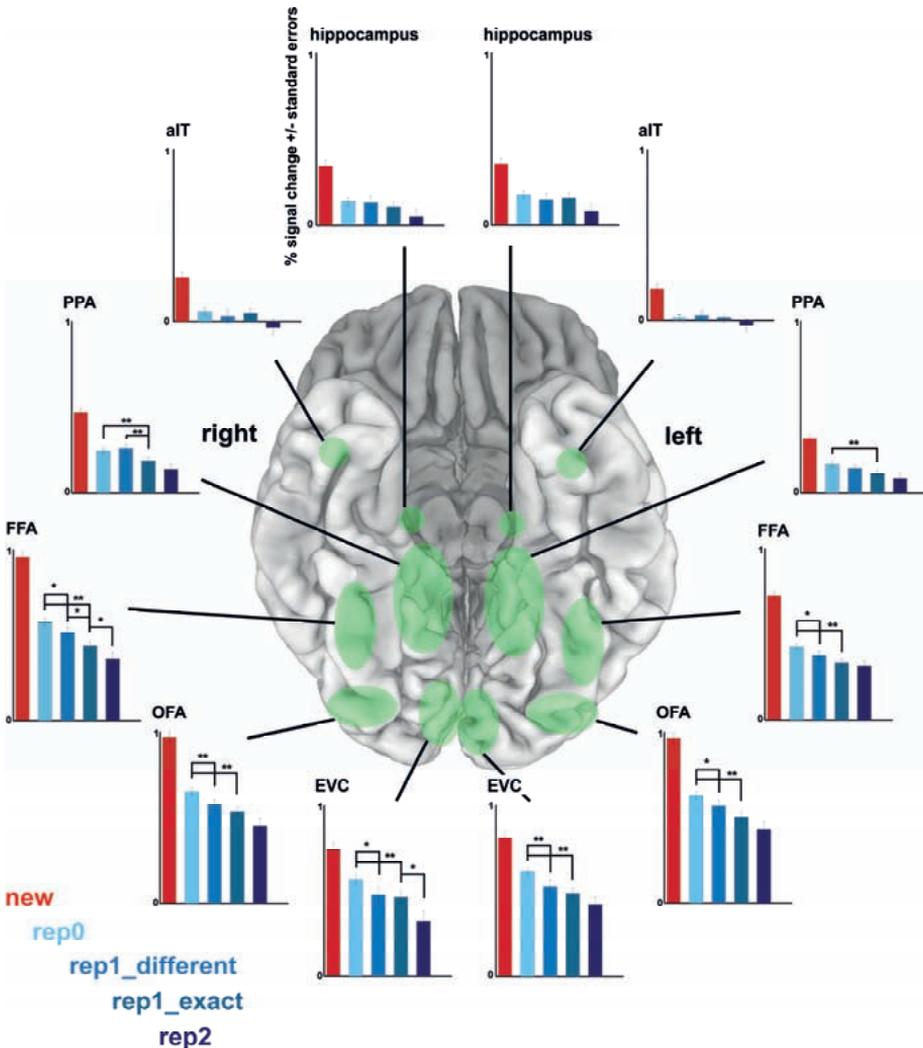


Figure S2.4 ROI control analysis for face-identity repetition using half of the data (six runs per subject). OFA and FFA, as well as EVC and PPA showed decreased activation with face-identity repetition. Most regions show weaker effects for the different-image contrast as compared to the full data set results (Figure 2.4). Approximate ROI locations are shown (in green) on a ventral view of the cortex (shown here: MNI template colin27). Graphs show beta-values and associated standard errors for the new (red), rep0 (i.e. identity change; light blue), rep1_different (blue), rep1_exact (blue grey), and rep2 (dark blue) predictors, averaged across subjects. Values for the rep3, rep4, and rep5 predictors are not shown, since these were based on only few trials. Five relevant contrasts (new versus rep0, rep0 versus rep1_different, rep0 versus rep1_exact, rep1_different versus rep1_exact, and rep1_exact versus rep2) were tested for significance. The new versus rep0 contrast was significant for all tested regions ($p < .01$) (not shown). For the other four tested contrasts, significant contrasts are shown and denoted with ** ($p < .01$) or * ($p < .05$). ROIs were defined using independent data. See Table 2.1 for abbreviations and ROI details (including ROI-defining contrasts).

2.5.4 Supplementary discussion

Widespread response increase for new as compared to seen faces

The new faces in the current experiment were not only novel stimuli; they also were targets (the subjects' task involved indicating whether the shown face was new or familiar) with a presentation frequency similar to standard oddball paradigms (~10% of the face images). Therefore, we cannot precisely determine the relative contributions of novelty, target detection, and infrequent presentation to the observed widespread increase in response. Previous literature indicates that infrequent presentation might have played a larger role than target detection or novelty (Kiehl et al., 2001; Eger et al., 2005; Leveroni et al., 2000). Any of the factors described above, or a combination of them, could have induced significant attentional effects resulting in the observed widespread response increase for new as compared to seen faces.

Why did seen and known faces elicit equal activations?

Based on previous studies, we expected to find a stronger response to known than seen faces in regions involved in the automatic activation of biographical knowledge associated with a face. A premier candidate region for showing this effect was aIT, because anterior temporal regions have been reported to be involved in representing (auto)biographical knowledge (Haxby et al., 2000; Leveroni et al., 2000; Gorno-Tempini et al., 1998; Tranel et al., 1997). In addition, atrophy of the (left) anterior and lateral temporal lobes has been associated with loss of semantic knowledge, a condition known as semantic dementia (e.g. Chan et al., 2001; McClelland and Rogers, 2003; Mummery et al., 2000). When right temporal atrophy is involved, this condition can result in prosopagnosia (Busigny et al., 2009; Czarnecki et al., 2008; see also Evans et al., 1995). Furthermore, a recent transcranial-magnetic-stimulation (TMS) study in healthy participants reported results consistent with a role of the anterior temporal

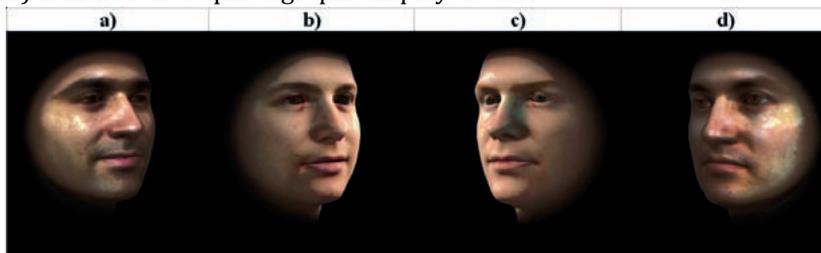
lobes in semantic processing (Pobric et al., 2007). Besides searching for “direct” effects of added biographical knowledge (by directly comparing activation to seen with activation to known faces), we also searched for “indirect” effects by investigating whether face-identity repetition effects were different for seen than known faces. However, there were no regions showing either a main or a consistent interaction effect of conceptual familiarity. Interpretations of our findings will be discussed below.

First, it could be that signal dropout in anterior temporal regions or repetition-related decreases in response (see also Gobbini et al., 2004) occluded conceptual familiarity effects. Note, however, that we acquired fMRI data at high-resolution with a 16-channel head coil. This enhances the signal in the anterior temporal lobes (Bellgowan et al., 2006). Nevertheless, it remains possible that the signal was still too weak to reveal conceptual familiarity effects in spatial-mean activity. Second, it could be that there indeed was no conceptual familiarity effect. One reason for this might be that the difference between seen and known faces was less than experienced by our participants in their daily lives. Furthermore, even though participants might have automatically activated the recently memorized biographical information (Bruce and Young, 1986; Sergent et al., 1992), this was not necessary for performing the task. Alternatively, conceptual face-information may be co-represented with the associated perceptual face information. In other words, both seen and known faces can be considered unique entities. Unique entities are exemplars that are a class of their own (e.g. famous faces), because they are associated with unique semantic and/or fine-grained physical information (Damasio et al., 1996; Tranel et al., 1997). The anterior temporal lobes have been shown to be involved in recognition of unique entities (Damasio et al., 1996; Gorno-Tempini and Price, 2001; Grabowski et al., 2001; Tranel et al., 1997). Since both seen and known faces can be considered unique entities, they may elicit indistinguishable activations in aIT. Related to this idea, it could be that face representations in aIT are perceptual in nature. Each face-identity has its own idiosyncratic perceptual features and these could form the basis of an abstract perceptual representation. Seen and known faces are both equally perceptually familiar and would therefore be associated with indistinguishable activations in aIT. However, right anterior temporal lobe atrophy is not only associated with impaired face-recognition performance, but also with impaired person-recognition from name (e.g. Busigny et al., 2009; Evans et al., 1995), suggesting that the representation may not be purely perceptual.

2.5.5 Face-familiarity test material

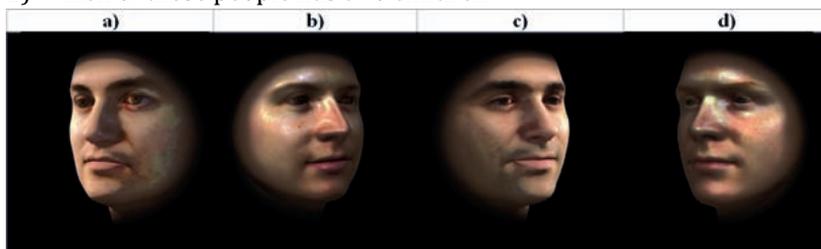
Multiple-choice example questions

1) Which of these photographs displays Peter?



e) I don't know.

2) Which of these people has two children?



e) I don't know.

3) What description fits this person best?



- a) He brews his own beer
- b) He owns a squirrel shelter
- c) He designs his own clothes
- d) He has a squirrel as pet
- e) I don't know.

Perceptual face-familiarity test

Subjects had to detect the unfamiliar faces. Shown here is the answer sheet, with circles around the unfamiliar faces.



Chapter 3

Representational similarity analysis – connecting the branches of systems neuroscience

A fundamental challenge for systems neuroscience is to quantitatively relate its three major branches of research: brain-activity measurement, behavioral measurement, and computational modeling. Using measured brain-activity patterns to evaluate computational network models is complicated by the need to define the correspondency between the units of the model and the channels of the brain-activity data, e.g. single-cell recordings or voxels from functional magnetic resonance imaging (fMRI). Similar correspondency problems complicate relating activity patterns between different modalities of brain-activity measurement (e.g. fMRI and invasive or scalp electrophysiology), and between subjects and species. In order to bridge these divides, we suggest abstracting from the activity patterns themselves and computing representational dissimilarity matrices (RDMs), which characterize the information carried by a given representation in a brain or model. Building on a rich psychological and mathematical literature on similarity analysis, we propose a new experimental and data-analytical framework called representational similarity analysis (RSA), in which multi-channel measures of neural activity are quantitatively related to each other and to computational theory and behavior by comparing RDMs. We demonstrate RSA by relating representations of visual objects as measured with fMRI in early visual cortex and the fusiform face area to computational models spanning a wide range of complexities. The RDMs are simultaneously related via second-level application of multidimensional scaling and tested using randomization and bootstrap techniques. We discuss the broad potential of RSA, including novel approaches to experimental design, and argue that these ideas, which have deep roots in psychology and neuroscience, will allow the integrated quantitative analysis of data from all three branches, thus contributing to a more unified systems neuroscience.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis – connecting the branches of systems neuroscience. *Front Syst Neurosci* 2, doi: 10.3389/neuro.06.004.2008.

3.1 Introduction

3.1.1 Relating representations in brains and models

A computational model of a single neuron (e.g. in V1) can be tested and adjusted on the basis of electrophysiological recordings of the activity of that type of neuron under a variety of circumstances (e.g. across different stimuli). This has been one successful avenue of evaluating computational models of single neurons with brain-activity data (e.g. Koch, 1999; Rieke et al., 1999; David and Galant, 2005). This single-unit fitting approach becomes intractable, however, for computational models at a larger scale of organization, which simulate comprehensive brain information processing and include populations of units with different functional properties. A major problem in relating such models to brain-activity data is the spatial correspondency problem: Which single-cell recording or fMRI voxel corresponds to which unit of the computational model? Defining a one-to-one mapping between model units and data channels will require that the functional properties of the simulated and real neurons are well characterized in advance; and finding the optimal match-up will still be challenging. To further complicate matters, a one-to-one mapping often cannot be assumed in the first place; the voxels and sensors of brain imaging, for example, reflect the activity of large numbers of neurons. Although model units as well can represent sets of neurons, we cannot in general assume a one-to-one correspondency. When a one-to-one mapping does not exist, the attempt to define such a mapping is clearly ill-motivated. Defining the correspondency more generally in terms of a linear transform would require the fitting of a weights matrix, which will often have a prohibitively large number of parameters (number of model units by number of data channels).

Similar correspondency problems arise in relating activity patterns between different modalities of brain-activity measurement. Modern techniques of multi-channel brain-activity measurement (including invasive and scalp electrophysiology, as well as fMRI) can take rich samples of neuronal pattern information. Invasive electrophysiology is the ideal modality in terms of resolution in both space (single neuron) and time (ms). However, only a very small subset of neurons can be recorded from simultaneously. Imaging techniques (fMRI and scalp electrophysiology), sample neuronal activity contiguously across large parts of the brain or across the whole brain. In imaging, however, a single channel reflects the joint activity of tens of thousands (high-resolution fMRI), or even millions of neurons (scalp electrophysiology).

If the same activity patterns are measured with two different techniques, we expect an overlap in the information sampled. However, different techniques

sample activity patterns in fundamentally different ways. Invasive electrophysiology measures the electrical activity of single cells, whereas fMRI measures the hemodynamic aspect of brain activity. Although the hemodynamic fMRI signal has been shown to reflect neuronal activity (Logothetis et al., 2001; see also Bandettini and Ungerleider, 2001), fMRI patterns are spatiotemporally displaced, smoothed, and distorted. Scalp electrophysiology combines high temporal resolution with a spatial sampling of neuronal activity that is even coarser than in fMRI.

Neuroscientific theory must abstract from the idiosyncrasies of particular empirical modalities. To this end, we need a modality-independent way of characterizing a brain region's representation. Such a characterization will also enable us to elucidate in how far different modalities provide consistent or inconsistent information. One way of characterizing the information a brain region represents is in terms of the mental states (e.g. stimulus percepts) it distinguishes (Figure 3.1). Here we suggest to relate modalities of brain-activity measurement and information-processing models by comparing activity-pattern dissimilarity matrices. Our approach obviates the need for defining explicit spatial correspondency mappings or transformations from one modality into another.

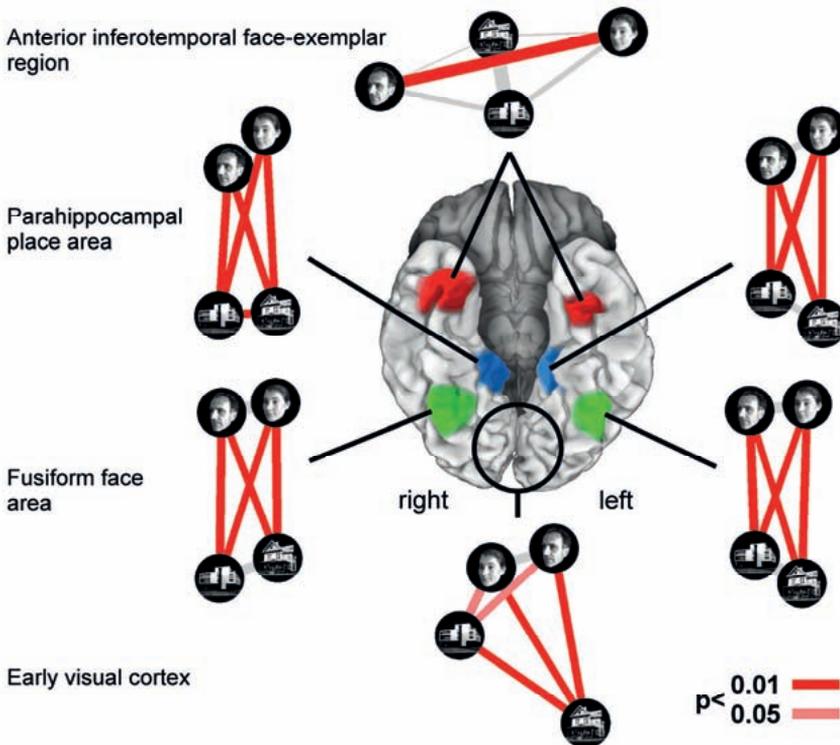


Figure 3.1 Characterizing brain regions by representational similarity structure. For each region, a similarity-graph icon shows the similarities between the activity patterns elicited by four stimulus images. Images placed close together in the icon elicited similar response patterns. Images placed far apart elicited dissimilar response patterns. The color of each connection line indicates whether the response-pattern difference was significant for the group (red: $p < 0.01$; light gray: $p \geq 0.05$, not significant). A connection line, like a rubberband, becomes thinner when stretched beyond the length that would exactly reflect the dissimilarity it represents. Connections also become thicker when compressed. Line thickness, thus, indicates the inevitable distortion of the 2D representation of the higher-dimensional similarity structure. The thickness of the connection lines is chosen such that the area of each connection (length times thickness) precisely reflects the dissimilarity measure. This novel visualization of fMRI response-pattern information combines (a) a multidimensional-scaling arrangement of activity-pattern similarity (as introduced to fMRI by Edelman et al., 1998), (b) a novel rubberband-graph depiction of inevitable distortions, and (c) the results of statistical tests of a pattern-information analysis (for details on the test, see Kriegeskorte et al., 2007). The icons show fixed-effects group analyses for regions of interest individually defined in 11 subjects. Early visual cortex was anatomically defined; all other regions were functionally defined using a data set independent of that used to compute the similarity-graph icons and statistical tests.

3.1.2 The representational dissimilarity matrix (RDM)

For a given brain region, we interpret (Dennett, 1987) the activity pattern associated with each experimental condition as a representation (e.g. a stimulus representation).⁶ By comparing the activity patterns associated with each pair of conditions (Edelman et al., 1998; Haxby et al., 2001), we obtain a representational dissimilarity matrix (RDM; Figure 3.2), which serves to characterize the representation.⁷ An RDM contains a cell for each pair of experimental conditions (Figure 3.2). Each cell contains a number reflecting the dissimilarity between the activity patterns associated with the two conditions. As a consequence, an RDM is symmetric about a diagonal of zeros. We suggest using correlation distance (1-correlation) as the dissimilarity measure, although we explore a number of measures below (Figure 3.10).

The RDM indicates the degree to which each pair of conditions is distinguished. It can thus be viewed as encapsulating the information content (in an informal sense) carried by the region. For any computational model (Figure 3.5) that can be exposed to the same experimental conditions (e.g. presented with the same

⁶ More generally, we can think of the activity pattern as the physical manifestation of the mental state induced by the experimental condition. The mental state could be the percept of an external object or something more remotely related to the external world, such as a motor program, a plan, or an emotion.

⁷ Note that similarity (a term we use here to refer to the general concept) can equally well be characterized by a similarity measure (in which greater values indicate greater similarity) or a dissimilarity measure (in which greater values indicate less similarity). We prefer the latter because of its intuitive relationship to distances in a multidimensional space.

stimuli), we can obtain an RDM for each of its processing stages in the same way as for a brain region (Figure 3.6). The RDMs serve as the signatures of regional representations in brains and models. Importantly, these signatures abstract from the spatial layout of the representations. They are indexed (horizontally and vertically) by experimental condition and can thus be directly compared between brain and model. What we are comparing, intuitively, is the represented information, not the activity patterns themselves.

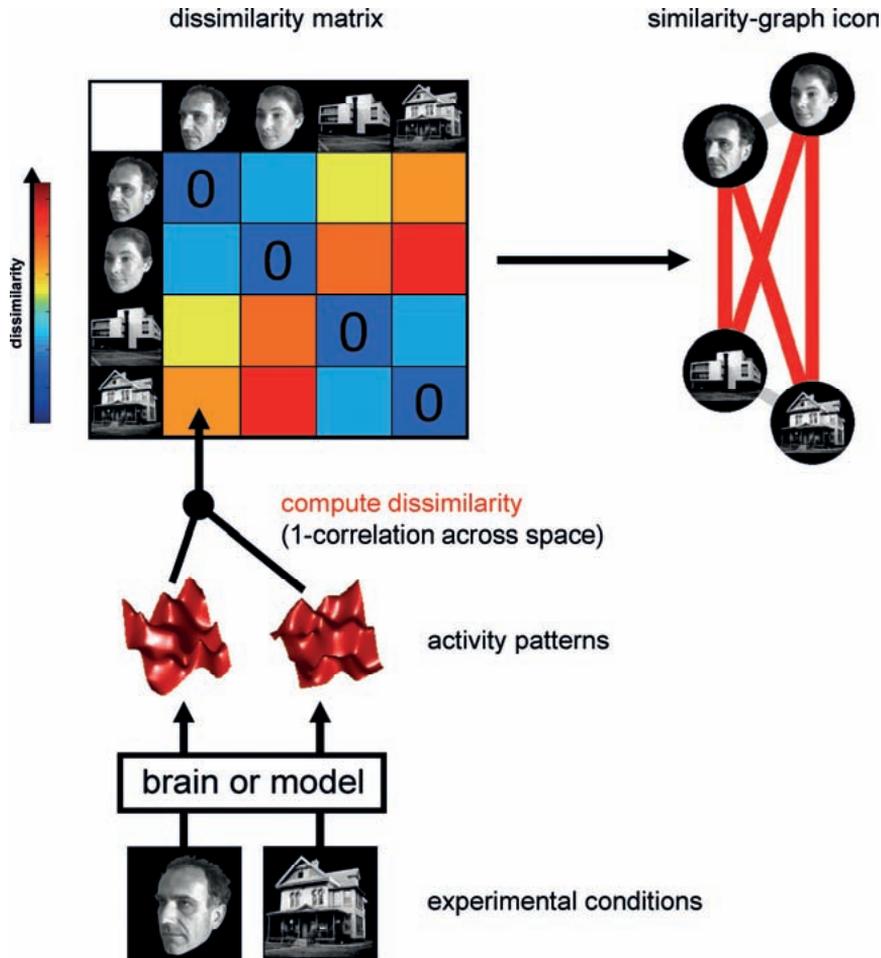


Figure 3.2 Computation of the representational dissimilarity matrix. For each pair of experimental conditions, the associated activity patterns (in a brain region or model) are compared by spatial correlation. The dissimilarity between them is measured as 1 minus the correlation (0 for perfect correlation, 1 for no correlation, 2 for perfect anticorrelation). These dissimilarities for all pairs of conditions are assembled in the RDM. Each cell of the RDM, thus, compares the response patterns elicited by two images. As a consequence, an RDM is symmetric about a diagonal of zeros. To visualize the representation for a small number of conditions, we suggest the similarity-graph icon (top right, cf. Figure 3.1).

3.1.3 Matching dissimilarity matrices: a second-order isomorphism

RDMs can be quantitatively compared just like activity patterns, e.g. using correlation distance (1-correlation) or rank-correlation distance. Because RDMs are symmetric about a diagonal of zeros, we will apply these measures using only the upper (or equivalently the lower) triangle of the matrices.

Analysis of similarity structure has a history in psychology and related fields. When exposed to a suitable sensory stimulus, our brain activity reflects many properties of the stimulus. The reflection of a stimulus property in the activity level of a neuron constitutes what has been termed a first-order isomorphism between the property and its representation in the brain. Most neuroscientific studies of brain representations have focused on the relationship between stimulus properties and brain-activity level in single cells or brain regions, i.e. on the first-order isomorphism between stimuli and their representations. One concept at the core of our approach is that of second-order isomorphism (Shepard and Chipman, 1970), i.e. the match of dissimilarity matrices.

When we encounter difficulty establishing a direct correspondence, i.e. a first-order isomorphism,⁸ in studying the relationship between stimuli and their representations, we may attempt instead to establish a correspondence between the relations among the stimuli on the one hand and the relations among their representations on the other, i.e. a second-order isomorphism. We can study the second-order isomorphism by relating the similarity structure of the objects to the similarity structure of the representations. This promises a higher-level functional perspective, which is complementary to the perspective of first-order isomorphism.

3.1.4 Related approaches in the literature

The qualitative and quantitative analysis of similarity structure has a long history in philosophy, psychology, and neuroscience. A good entry to the literature

⁸ A first-order isomorphism between object and representation can be interpreted in several ways. Naively: The representation is a replication of the object, i.e. identical with it. (Problem: A chair does not fit into the human skull.) More reasonably, we may interpret first-order isomorphism as a mere similarity of some sort. For example a retinotopic representation of an image in V1 may emit no light, be smaller and distorted, but it does bear a topological similarity to the image. More cautiously, we could maintain that first-order isomorphism requires only that the representation has properties (e.g. neuronal firing rates) that are related to properties of the objects represented (e.g. line orientation). While the naive interpretation is clearly untenable, the other interpretations are generally accepted in neuroscience. We concur with this widespread view, which motivates studies of stimulus selectivity at the level of single cells and brain regions. However, we feel that analysis of the second-order isomorphism (which can reflect a first-order isomorphism) is equally promising and offers a complementary higher-level functional perspective.

is provided by Edelman (1998), who (Edelman et al., 1998) also pioneered application of similarity analysis to fMRI activity patterns using the technique of multidimensional scaling (MDS; Torgerson, 1958; Kruskal and Wish, 1978; Shepard, 1980; Borg and Groenen, 2005). Laakso and Cottrell (2000) compared representations in hidden units of connectionist networks by correlating the dissimilarity structures of their activity patterns. They suggest that this approach could be used as a general method for comparing representations and discuss the philosophical implications. Op de Beeck et al. (2001) related the representational similarity of silhouette shapes in monkey inferior temporal cortex to physical and behavioral similarity measures for those stimuli.

At a more general level, activity-pattern similarity is related to activity-pattern information as targeted in a number of recent studies in human fMRI (Haxby et al., 2001; Strother et al., 2002; Spiridon and Kanwisher, 2002; Cox and Savoy, 2003; Carlson et al., 2003; Mitchell et al., 2004; Hanson et al., 2004; Kamitani and Tong, 2005; Haynes and Rees, 2005ab; Polyn et al., 2005; LaConte et al., 2005; Mourao-Miranda et al., 2005; Davatzikos et al., 2005; Kriegeskorte et al., 2006; Kamitani and Tong, 2006; Pessoa and Padmala, 2006; Haynes et al., 2007; Williams et al., 2007b; Serences and Boynton, 2007; Friston et al., 2008; for reviews see Haynes and Rees, 2006; Norman et al., 2006; Kriegeskorte and Bandettini, 2007a) and also in monkey electrophysiology (Hung et al., 2005; Tsao et al., 2006). Explicit similarity analyses of neuronal activity patterns have begun to be applied in human fMRI (Edelman et al., 1998; O'Toole et al., 2005; Aguirre, 2007; Drucker and Aguirre, 2009; Aguirre et al., in preparation; Kriegeskorte et al., 2008a) and monkey electrophysiology (Op de Beeck et al., 2001; Kiani et al., 2007).

3.1.5 Connecting the branches of systems neuroscience

In this paper, we argue that the theoretical concept of second-order isomorphism (Shepard and Chipman, 1970) can serve a much more general purpose than previously thought, relating not only external objects to their brain representations, but bridging the divides between the three branches of systems neuroscience: behavioral experimentation, brain-activity experimentation, and computational modeling (Figure 3.3).

We introduce an analysis framework called representational similarity analysis (RSA), which builds on a rich psychological and mathematical literature (Edelman, 1995; Edelman, 1998; Edelman and Duvdevani-Bar, 1997ab; Laakso and Cottrell, 2000; Kruskal and Wish, 1978; Shepard, 1980; Shepard and Chipman, 1970; Shepard et al., 1975; Torgerson, 1958). The core idea is to use the RDM as a signature of the representations in brain regions and computational models.

We define a specific working prototype of RSA and discuss the potential of this approach in its full breadth:

(1) Integration of computational modeling into the analysis of brain-activity data. A key advantage of RSA is that computational models of brain information processing form an integrated component of data analysis and can be directly evaluated and compared. We demonstrate how to apply multivariate analysis to a set of dissimilarity matrices from brain regions and models in order to find out (a) which model best explains the representation in each brain region and (b) to what extent representations among regions and models resemble each other. We introduce a randomization test of representational relatedness and a bootstrap technique for obtaining error bars on estimates of the goodness of fit of different models.

(2) Relating regions, subjects, species, and modalities of brain-activity measurement. We discuss how RSA can be used to quantitatively relate:

- representations in *different regions* of the same brain (“representational connectivity”),
- corresponding brain regions in *different subjects* (“intersubject information”),
- corresponding brain regions in *different species* (e.g. humans and monkeys),
- and *different modalities* of brain-activity data (e.g. cell recordings and fMRI).

(3) Relating brain and behavior. We discuss how RSA can quantitatively relate brain-activity measurements to behavioral data. This possibility has already been demonstrated in previous work (Op de Beeck et al., 2001; Kiani et al., 2007; Aguirre et al., in preparation).

(4) Addressing a broader array of neuroscientific questions with each experiment by means of condition-rich design. While RSA is applicable to conventional experimental designs, it synergizes with novel condition-rich experimental designs, where a single experiment can address a large number of neuroscientific questions. We demonstrate this with an fMRI experiment that has 96 separate conditions and discuss the broader implications.

We hope that RSA will contribute to a more integrated systems neuroscience, where different multi-channel measures of neural activity are quantitatively related to each other and to computational theory and behavior via the information-rich characterization of distributed representations provided by the RDM.

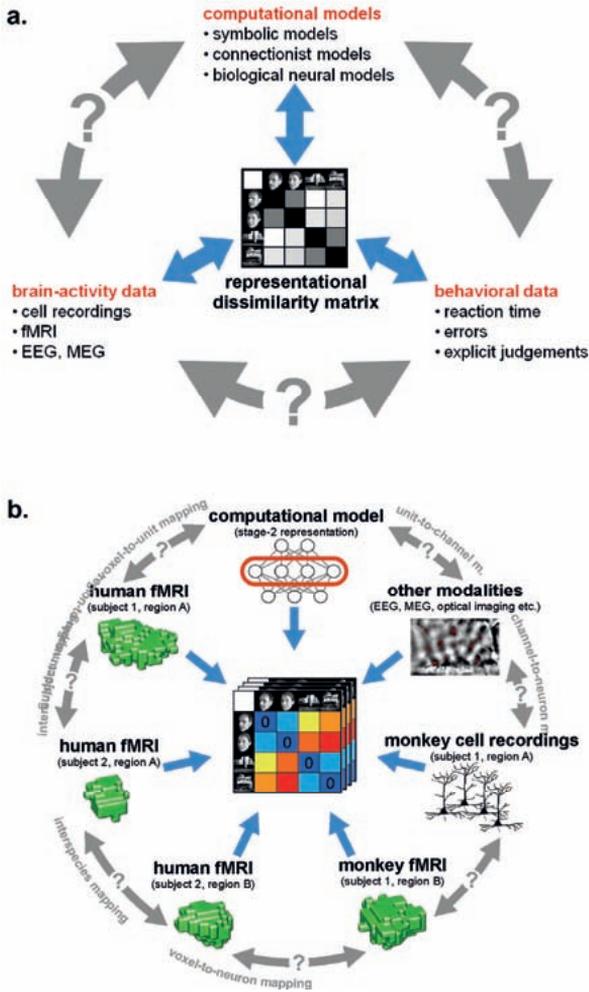


Figure 3.3 The representational dissimilarity matrix as a hub that relates different representations. **(a)** Systems neuroscience has struggled to relate its three major branches of research: behavioral experimentation, brain-activity experimentation, and computational modeling. So far these branches have interacted largely on two levels: (1) They have interacted on the level of verbal theory, i.e. by comparing conclusions drawn from separate analyses. This level is essential, but it is not quantitative. (2) They have interacted at the level characteristic functions, e.g. by comparing psychometric and neurometric functions. This form of bringing the branches in touch is equally essential and can be quantitative. However, characteristic functions typically contain only a small number of data points, so the interface is not informationally rich. Note that the RDM shown is based on only 4 conditions, yielding only $(4^2-4)/2=6$ parameters. However, since the number of parameters grows as the square of the number of conditions, the RDM can provide an informationally rich interface for relating different representations. Consider for example the 96-image experiment we discuss, where the matrix has $(96^2-96)/2=4560$ parameters. **(b)** This panel illustrates in greater detail what different representations can be related via the quantitative interface provided by the RDM. We arbitrarily chose the example of fMRI to illustrate the within-modality relationships that can be established. Note that all these relationships are difficult to establish otherwise (gray double arrows).

3.2 Representational similarity analysis – step by step

In this section we describe the core of RSA step by step. We assume that the data to be analyzed consists in a multivariate activity pattern measured for each of a set of conditions in a given brain region, whose representation is to be better understood. The data could be from single-cell or electrode-array recordings, from neuroimaging, or any other modality of brain-activity measurement. We demonstrate the analysis on an fMRI experiment, in which human subjects viewed 96 particular object images. The step-by-step description that follows describes the method. The empirical results for our example experiment are described and interpreted subsequently.

3.2.1 Step 1: Estimating the activity patterns

The first step of the analysis is the estimation of an activity pattern associated with each experimental condition. In our example, the activity patterns are spatial response patterns from early visual cortex (EVC) and from the fusiform face area (FFA). The analysis proceeds independently for each region.

Instead of spatial activity patterns we could use spatiotemporal patterns or simply temporal patterns from a single site as the input to RSA. Similarly, we could filter the measurements in some neuroscientifically meaningful way. For cell recordings, for example, we could use windowed spike counts, multi-unit activity, or local field potentials as the input. In our fMRI example, we obtain an activity estimate for each voxel and condition using massively univariate linear modeling (Figure 3.7). The design matrix used to model each voxel's response is based on the event sequence and a linear model of the hemodynamic response (Boynton et al., 1996). For each region of interest, the resulting condition-related activity patterns form the basis for computation of the representational dissimilarities.

3.2.2 Step 2: Measuring activity-pattern dissimilarity

In order to compute the RDM (Figure 3.2), we compare the activity patterns associated with each pair of conditions. A useful measure of activity-pattern dissimilarity that normalizes for both the mean level of activity and the variability of activity is correlation distance, i.e. 1 minus the linear correlation between patterns (cf. Haxby et al., 2001; Aguirre, 2007; Kiani et al., 2007). Alternative measures include the Euclidean distance (cf. Edelman et al., 1998), the Mahalanobis distance (cf. Kriegeskorte et al., 2006) and, in order to relate RSA to conventional activation-based fMRI analysis, the absolute value of the regional-average activation difference (Figure 3.10).

The dissimilarity values for all pairs of conditions are assembled in an RDM, which will have a width and height corresponding to the number of conditions and is symmetric about a diagonal of zeros (Figure 3.2). We can use MDS to visualize the similarity structure of the activity patterns. This is demonstrated in Figure 3.4, where conditions are represented by colored dots. The distances between the dots approximate the dissimilarities of the activity patterns the conditions are associated with.

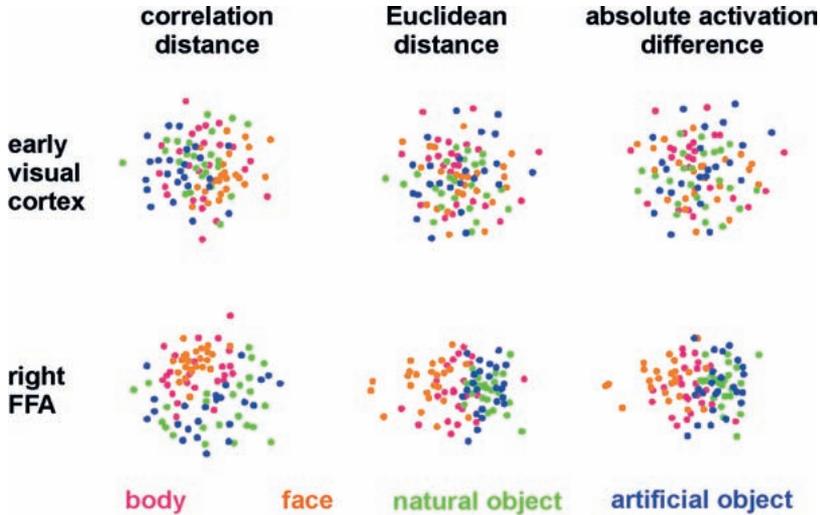


Figure 3.4 Unsupervised arrangement of 96 experimental conditions reflecting pairwise activity-pattern similarity. As in Figure 3.1, but for 96 instead of 4 conditions, the arrangements reflect the activity-pattern similarity structure. Each panel visualizes the RDM from the corresponding panel in Figure 3.10. Each condition (corresponding to the presentation of one of 96 object images) is represented by a colored dot, where the color codes for the category (legend at the bottom). In each panel, dots placed close together indicate that the two conditions were associated with similar activity patterns. Dots placed far apart indicate that the two conditions were associated with dissimilar activity patterns. The panels show results of non-metric multidimensional scaling (minimizing the loss function “stress”) for two brain regions (rows) and three activity-pattern dissimilarity measures (columns). Note that a categorical clustering of the face-image response patterns is apparent in the right FFA (bottom row), but not in early visual cortex (top row). (Note that the absolute activation differences could be represented by an arrangement along a straight line, had the dissimilarity matrices not been averaged across subjects.)

3.2.3 Step 3: Predicting representational similarity with a range of models

In this section we describe the different types of model that can be evaluated using RSA. Figure 3.5 shows the internal representations of several example models and Figure 3.6 shows the dissimilarity matrices characterizing the model representations.

Complex computational models

In order to evaluate a computational model with RSA, the model needs to simulate some aspect of the information processing occurring in the subject's brain during the experiment. The term model, thus, has a different meaning here than conventionally in statistical data analysis, where it often refers to a statistical model that does not simulate brain information processing (such as the design matrix in Figure 3.7, which was used to estimate the activity patterns). In our example, we are interested in visual object perception, so the models to be used simulate parts of the visual processing. The models are presented with the same experimental stimuli as our human subjects. Moreover, their internal representations are analyzed in the same way as the measured brain representations of our subjects.

We demonstrate RSA with three complex computational models. First, we use a model of V1 consisting in retinotopic maps of simulated simple and complex cells based on banks of Gabor filters for a range of spatial frequencies and orientations at each location (details in the Appendix). We also include a variant of this model, in which we attempted to simulate the local averaging of fMRI voxels by pooling local responses of the original V1 model (V1 model, smoothed). Second, as an example of a higher-level representation, we use a model developed in the HMAX framework (Riesenhuber and Poggio, 2002; Serre et al., 2007), which includes C2 units based on natural-image patches as filters and corresponds, approximately, to the level of representation in V4. Third, we use a computational model from computer vision, the RADON transform, whose components in the present implementation are not meant to resemble neurons in the primate visual system. However, this model could be implemented with biological neurons and has been proposed as a functional account of the representation of visual stimuli in the lateral occipital complex (Wade et al., Human Brain Mapping 2006) based on fMRI evidence. Detailed descriptions of the model representations are to be found in the *Methodological Details*.

Simple computational models

The models described above are meant to simulate brain information processing in some sense. We can additionally use simple image transformations as competing computational models. Although there may be no compelling neuroscientific motivation for such models, they can provide useful benchmarks and help us characterize the information represented in a given brain region. Here we include (1) the digital images themselves in the Lab color space (which more closely reflects human color similarity perception than the RGB color space more commonly used for image storage), (2) the luminance patterns (grayscale versions) of the images, (3) low-pass (i.e. smoothed), and (4) high-pass (i.e. edge-emphasized) versions of the luminance patterns, (5) the Lab joint histo-

grams of the images (representing the set of colors present in each image), and (6) the silhouettes of the objects, in which each figure pixel is 1 and each background pixel 0. These models as well are described in more detail in the Appendix.

Conceptual models

Model dissimilarity matrices can be obtained not only from explicit computational accounts. A theory may specify that a given brain region represents particular information and abstracts from other information without specifying how the representation is computed. In such “conceptual models”, the information processing is miraculous (i.e. unspecified) and the activity patterns unknown. However, we can still specify a hypothetical similarity structure to be tested by comparison to the similarity structures found in different brain regions.

Here we use two categorical models as examples of this model variety (Figure 3.6). The first is the animate-inanimate model, in which two object images are *identical* (dissimilarity=0) if they are either both animate or both inanimate, and *different* (dissimilarity=1) if they straddle the category boundary. The second categorical model follows the same logic for the category of faces: two object images are *identical* (dissimilarity=0) if they are either both faces or both non-faces, and *different* (dissimilarity=1) if exactly one of them is a face.

In addition, we use a “face-animal-prototype model”, which assumes that all faces elicit a prototypical response pattern (implying small dissimilarities between individual face representations) and that the same is true to a lesser degree for the more general class of animal images.

Behavior-based similarity structure

We could also use behavioral measures to define reference dissimilarity matrices. The dissimilarity values could come from explicit similarity judgments or from reaction times or confusion errors in comparison tasks (Cutzu and Edelman, 1996, 1998; Edelman et al., 1998; Op de Beeck et al., 2001; Kiani et al., 2007; Aguirre et al., in preparation; Shepard et al., 1975). Such behavioral dissimilarity matrices may reflect the representations that determine the behavioral choices, reaction times, or confusion errors. A close match between the RDM of a brain region and the behavioral dissimilarity matrix would suggest that the regional representation might play a role in determining the behavior measured.

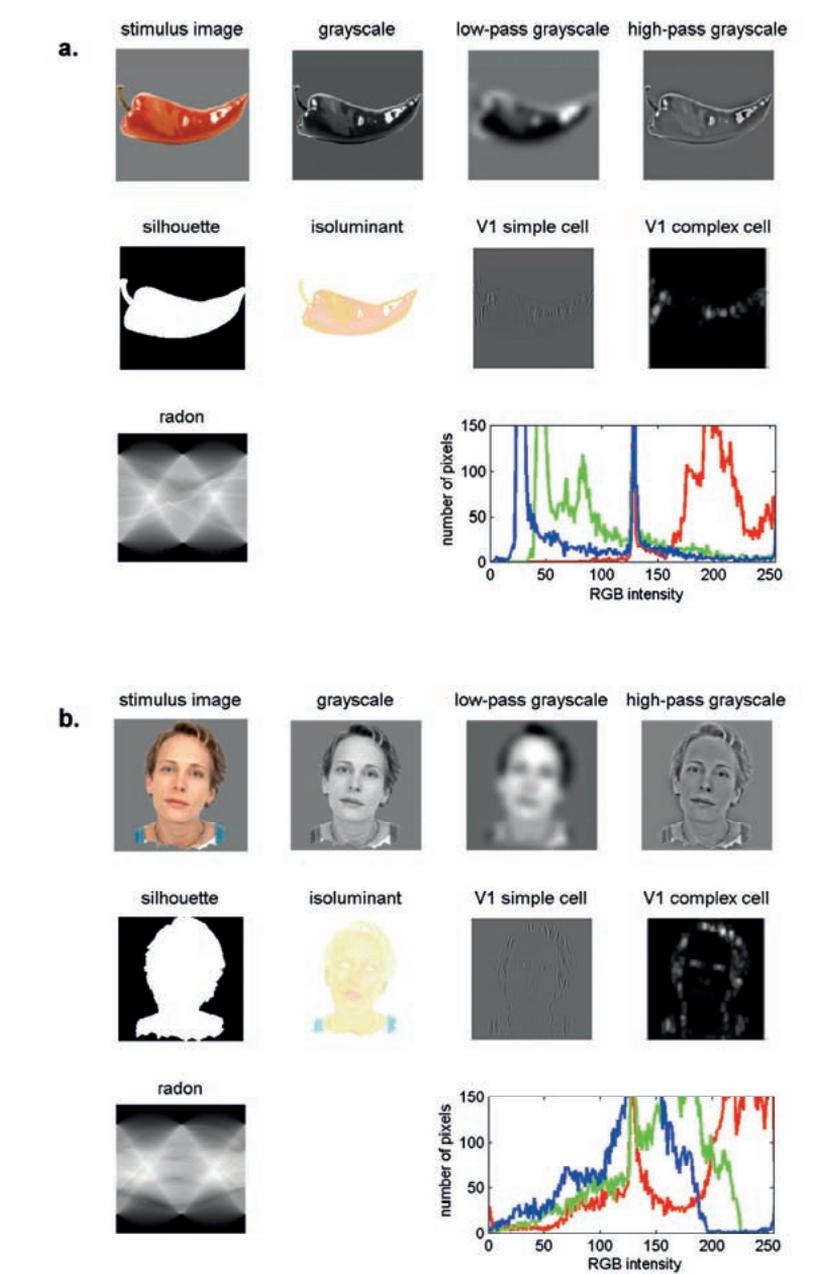


Figure 3.5 Model representations of two example images. Two example images (**a**, **b**) from the 96-image experiment and their representations in a number of computational models, including standard transformations of image processing as well as neuroscientifically motivated models. Note that each such representation defines a unique similarity structure for the 96 stimuli (as encapsulated in the RDMs of Figure 3.6).

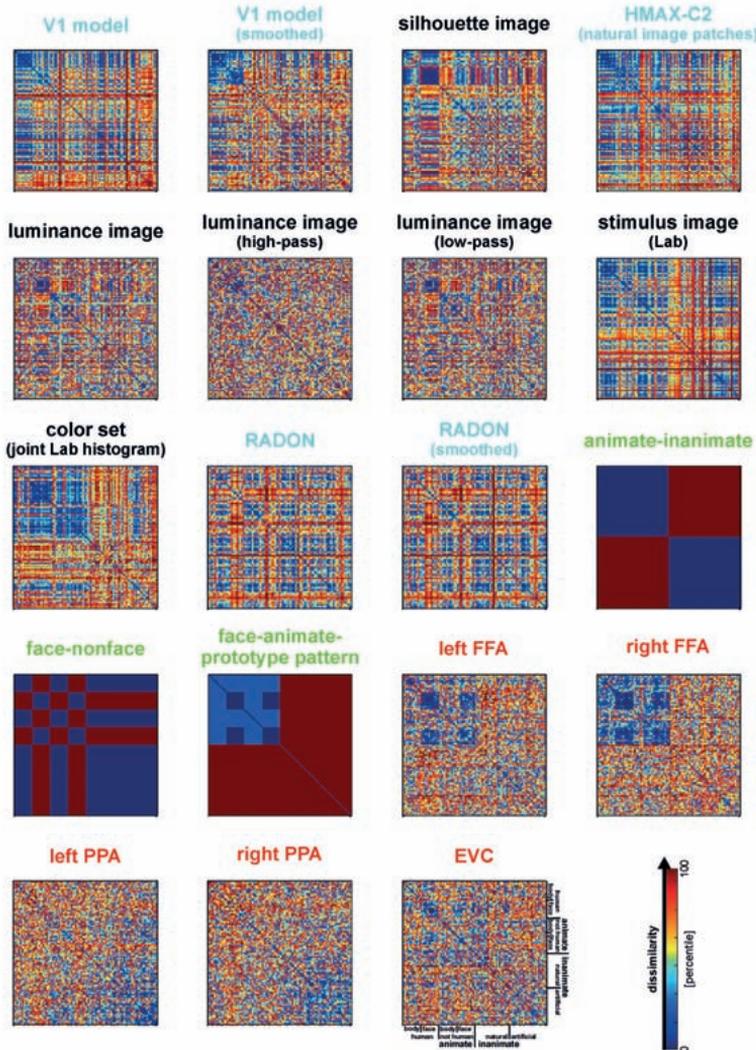


Figure 3.6 Representational dissimilarity matrices for models and brain regions. Dissimilarity matrices for model representations and regional brain representations (as introduced in Figure 3.2). The dissimilarity measure is 1-correlation (Pearson correlation across space). Note that each model yields a unique representational similarity structure that can be compared to that of each brain region (bottom five matrices). This comparison is carried out quantitatively in Figure 3.8. The text labels indicate the representation depicted with the color indicating the type: complex computational model (blue), simple computational model (black), conceptual model (green), brain representation (red).

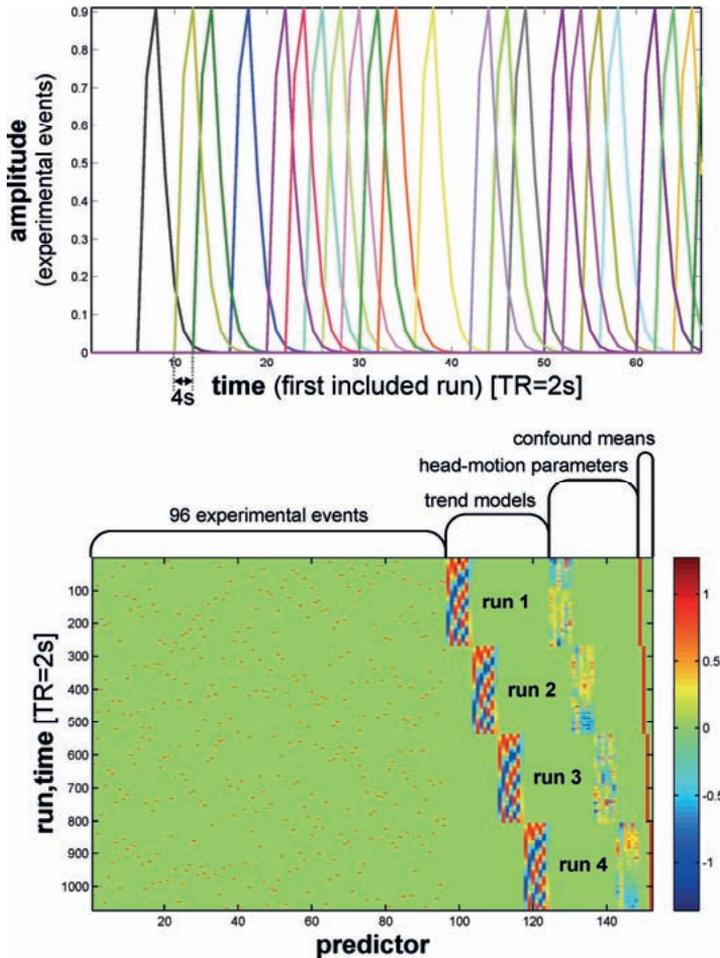


Figure 3.7 Design matrix for condition-rich ungrouped-events fMRI design. Both panels illustrate the design matrix used for the 96-image experiment, an example of a condition-rich ungrouped-events design. The top panel shows the hemodynamic predictor time courses for the experimental events occurring in the first couple of minutes of the first run. Note that events occur at 4-s trial-onset asynchrony, yielding overlapping but dissociable hemodynamic responses and a reasonable frequency of stimulus presentation. (Each of the 96 conditions occurs exactly once in each run. The condition sequence is independently randomized for each run.) The bottom panel shows the complete design matrix with predictor amplitude color coded (see colorbar on the right). In addition to the 96 predictors for the experimental conditions, the design matrix also includes components modeling slow artefactual trends and residual head-motion artefacts (after rigid-body head-motion correction), and a confound-mean predictor for each run.

3.2.4 Step 4: Comparing brain and model dissimilarity matrices

Once the dissimilarity matrices of the brain representations (Figure 3.10) and those of theoretical models (Figure 3.6) have been specified they can be visually and quantitatively compared. One way to quantify the match between two dissimilarity matrices is by means of a correlation coefficient. We use 1-correlation as a measure of the dissimilarity between RDMs (Figure 3.8). Because dissimilarity matrices are symmetrical about a diagonal of zeros, the correlation is computed over the values in the upper (or equivalently the lower) triangular region. Note that above we suggested the use of this measure for comparing activity patterns. Here we suggest using it to assess second-order dissimilarity: the dissimilarity of dissimilarity matrices.

We could use an alternative distance measure, such as the Euclidean distance, for comparing dissimilarity matrices. As for comparing activity patterns, we again prefer correlation distance, because it is invariant to differences in the mean and variability of the dissimilarities. For the models we use here, we do not wish to assume a linear match between dissimilarity matrices. We therefore use the Spearman rank correlation coefficient to compare them. In the Appendix, we present another argument for the use of rank-correlation distance (instead of the Pearson linear correlation distance or Euclidean distance) for comparing dissimilarity matrices. The argument is based on the observation that, in high-dimensional response spaces, a prominent component of the effect of activity-pattern noise on the dissimilarities can be accounted for by a monotonic transform.

Figure 3.8 shows the deviations (1-Spearman correlation) of the models from each brain region's RDM. Smaller bars indicate better fits. In order to estimate the variability of each model deviation expected if a similar experiment were to be performed with different stimuli (from the same population of stimuli), we computed each model deviation 100 times over for bootstrap resamplings of the condition set (i.e. 96 conditions chosen with replacement from the original set of 96 on each iteration).⁹ This method is attractive, (1) because it requires few assumptions, (2) because only the dissimilarity matrices are needed as input, (3) because it is computationally less intensive than modeling the noise at a

⁹ A complication of this method is that bootstrap resampling of the condition set moves zeros from the diagonal into the off-diagonal parts of the matrix whenever a condition is selected multiple times in the bootstrap resampling. The inclusion of these off-diagonal zeros leads to artefactually small model deviation estimates (because it increases the correlation between the dissimilarity values). In order to avoid underestimating the model deviations in the bootstrap simulation, these artefactual off-diagonal zeros (about 1% of the dissimilarity values here) were excluded before computing the model deviations.

lower level, and (4) because it generalizes (to the degree possible given the experimental data) from the set of conditions actually used in the experiment to the population of conditions that the actual conditions can be considered a random sample of. This bootstrap procedure would also lend itself to testing whether one model fits the data better than another model, as discussed in the Appendix.¹⁰

3.2.5 Step 5: Testing relatedness of two dissimilarity matrices by randomization

In order to decide whether two dissimilarity matrices are related, we can perform statistical inference on the RDM correlation. The classical method for testing correlations assumes independent measurements for the two variables. For dissimilarity matrices such independence cannot be assumed, because each similarity is dependent on two response patterns, each of which also codetermines the similarities of all its other pairings in the RDM. We therefore suggest testing the relatedness of dissimilarity matrices by randomizing the condition labels. We choose a random permutation of the conditions, reorder rows and columns of one of the two dissimilarity matrices to be compared according to this permutation, and compute the correlation. Repeating this step many times (e.g. 10,000 times), we obtain a distribution of correlations simulating the null hypothesis that the two dissimilarity matrices are unrelated. If the actual correlation (for consistent labeling between the two dissimilarity matrices) falls within the top $\alpha \cdot 100\%$ of the simulated null distribution of correlations, we reject the null hypothesis of unrelated dissimilarity matrices with a false-positives rate of α . The p value for each brain region's relatedness to each model is given beneath the model's bar in Figure 3.8. They are conservative estimates based on 10,000 random relabelings, so the smallest possible estimate is 10^{-4} .

¹⁰ Alternatively, we could obtain error bars and statistical tests by estimating the distribution of the model deviation estimates for repetitions of the experiment with the same stimuli and subjects or with the same stimuli and different subjects, or with different stimuli and different subjects. These approaches would provide complementary information to the condition-label bootstrap approach we have described.

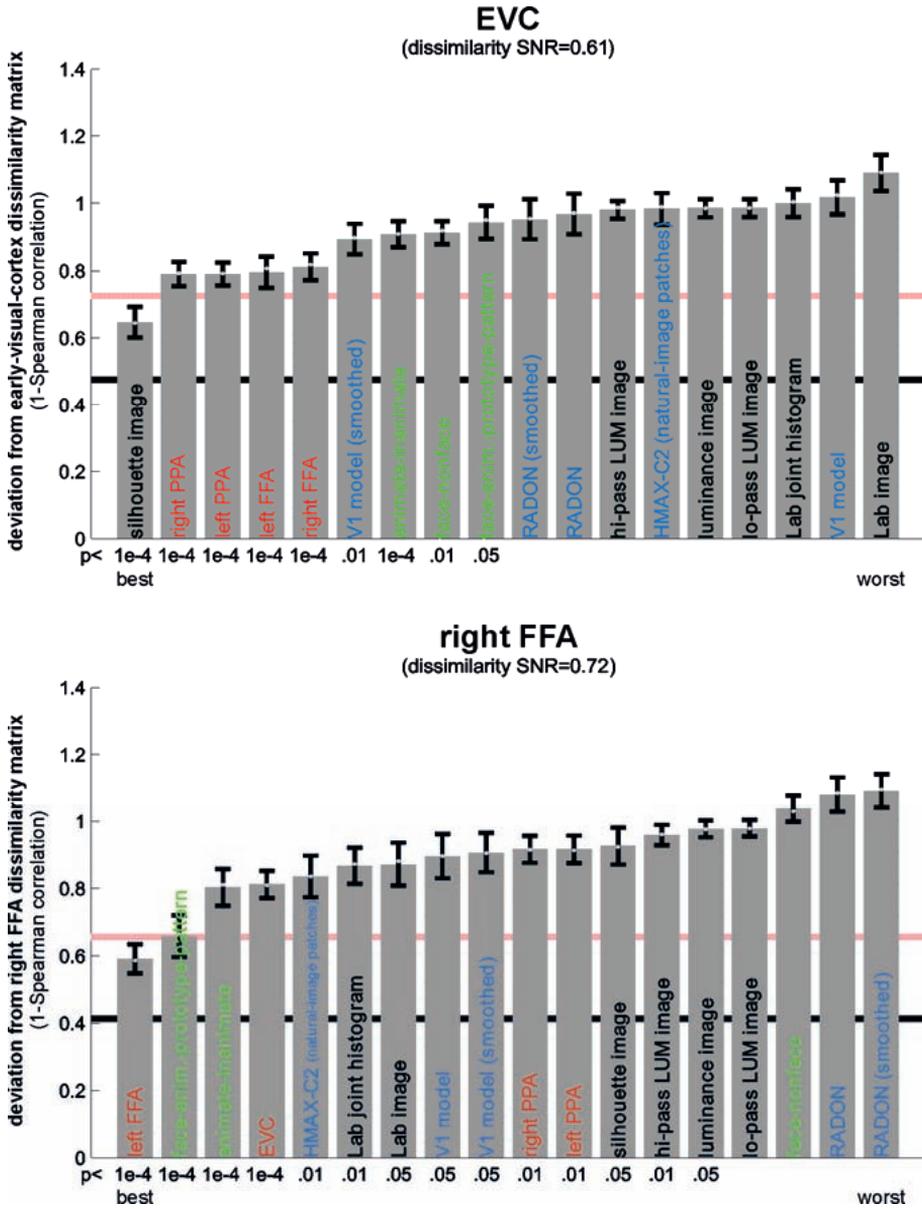


Figure 3.8 Matching models to brain regions by comparing representational dissimilarity matrices. The dissimilarity matrices characterizing the representation in early visual cortex (top) and the right FFA (bottom) are compared to dissimilarity matrices obtained from model representations and other brain regions. Each bar indicates the deviation between the RDM of the reference region (early visual cortex or the right FFA) and that of a model or other brain region. The deviation is measured as 1 minus the Spearman correlation between dissimilarity matrices (for motivation see Step 4 and Appendix). Text-label colors indicate the type of representation: complex computational model (blue), simple computational model (black), conceptual model (green), brain representation (red). Error bars indicate the standard error of the deviation estimate. (The standard error is estimated as the standard deviation of 100 deviation estimates obtained from bootstrap

resamplings of the conditions set.) The number below each bar indicates the p value for a test of relatedness between the reference matrix (early visual cortex or the right FFA) and that of the model or other region. (The test is based on 10,000 randomizations of the condition labels.) The black line indicates the noise floor, i.e. the expected deviation between the empirical reference RDM (with noise) and the underlying true RDM (without noise). The red line indicates the expected retest deviation between the empirical dissimilarity matrices that would be obtained for the reference region if the experiment were repeated with different subjects (both matrices affected by noise). Both of these reference lines as well as the dissimilarity signal-to-noise ratios (dissimilarity SNR: below the titles) are estimated from the variability of the dissimilarity estimates across subjects.

3.2.6 Step 6: Visualizing the similarity structure of representational dissimilarity matrices by MDS

MDS provides a general method for arranging entities in a low-dimensional space (e.g. the two dimensions of a figure on paper), such that their distances reflect their similarities: Similar entities will be placed together, dissimilar entities apart. In Figure 3.4 we used MDS to visualize the similarity structure of activity patterns in EVC and FFA. Here we suggest using MDS also to visualize the similarity structure of representational dissimilarity matrices. We first assemble all pairwise comparisons between activity-pattern dissimilarity matrices in a dissimilarity matrix of dissimilarity matrices (Figure 3.9a), using rank-correlation as the dissimilarity measure as suggested above. We then perform MDS on the basis of this second-order dissimilarity matrix.

This exploratory visualization technique (Figure 3.9b) simultaneously relates all RDMs (from models and brain regions) to each other. It thus summarizes the information we would get by inspecting a bar graph of RDM fits (Step 4) not just for EVC and the right FFA (as shown in Figure 3.8), but for each model and region. The conciseness of the MDS visualization comes at a cost: the distances are distorted (depending on the number of representations included) and there are no error bars or statistical indications. Nevertheless this exploratory visualization technique provides a useful overall view. It can alert us to relationships we had not considered and prompt confirmatory follow-up analysis.

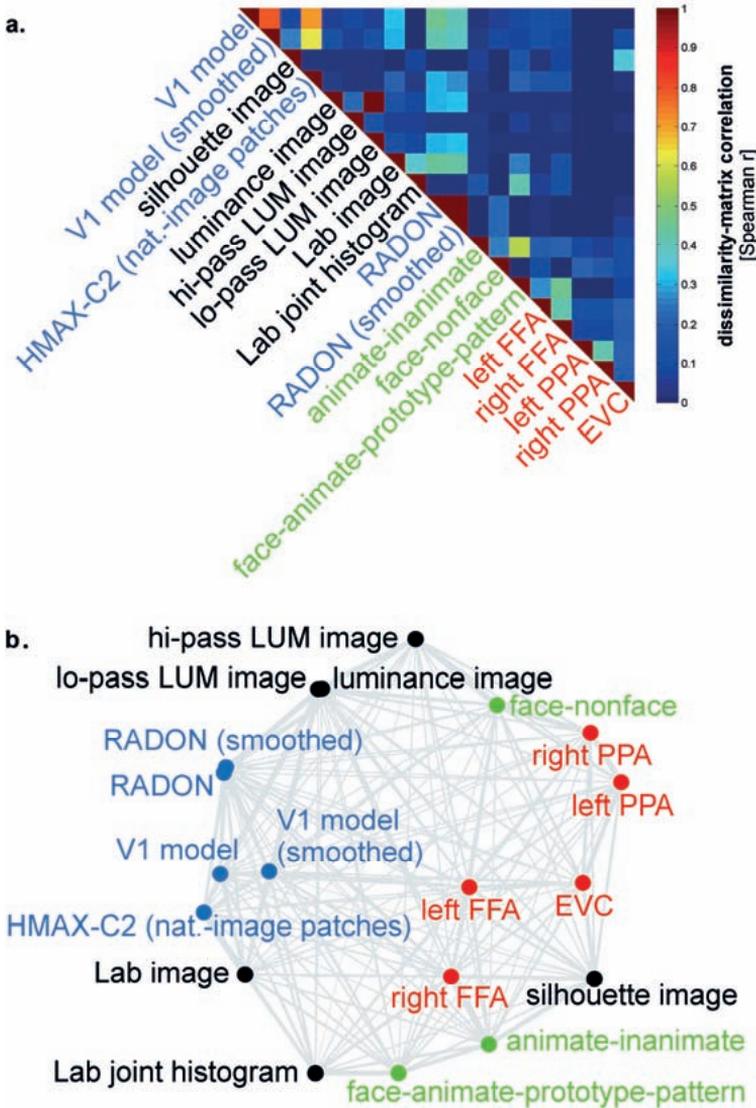


Figure 3.9 Simultaneously relating all pairs of representations. Figure 3.8 showed the relationships between two reference regions and all models and other regions. Here we simultaneously visualize the pair-relationships between all models and regions (text labels). Note that the visualization of all pair-relationships comes at a cost: statistical information is omitted here. Text-label colors indicate the type of representation: complex computational model (blue), simple computational model (black), conceptual model (green), brain representation (red). **(a)** The correlation matrix [Spearman rank correlation] of RDMs. **(b)** Multidimensional scaling arrangement (minimizing metric stress) of the representations. Note that MDS was used here to arrange not activity patterns (as in Figures 3.1 and 3.4), but dissimilarity matrices. The rubberband graph (gray connections) depicts the inevitable distortions introduced by arranging the models in 2D (see legend of Figure 3.1 for an explanation).

3.3 Empirical results and their interpretation

3.3.1 The representational dissimilarity matrices of EVC and FFA

Figure 3.10 shows the correlation-distance matrix for EVC and FFA. For the FFA, but not EVC, the matrix reflects the categorical structure of the stimuli. This structure is obvious, because the condition sequence for the dissimilarity matrices were defined by the categorical order. Note, however, that this order affects merely the visual appearance of the matrices. Reordering the conditions does not affect the results of RSA. For the FFA, the correlation-distance matrix reveals a pattern markedly different from that exhibited by the two other measures of activity-pattern dissimilarity. The absolute-activation-difference matrix shows the prominent contrast in activation level between faces and inanimate objects and less prominently between animate and inanimate objects. The correlation-distance matrix normalizes out the regional-average activation effects and reveals that the activity patterns are highly correlated among faces (human or animal) and to a lesser degree among animals. The Euclidean-distance matrix is sensitive to both the absolute activation difference and the pattern correlation. Unless indicated otherwise, subsequent analyses are based on correlation-distance matrices.

3.3.2 The similarity structure of activity patterns in EVC and FFA as revealed by MDS

Figure 3.4 visualizes the dissimilarity structure as estimated with the three measures by arranging dots that represent the 96 object images in 2D with category-color-coding, such that stimuli eliciting similar response patterns are placed close together and stimuli eliciting dissimilar response patterns are placed far apart. Such arrangements are computed by MDS. We observe some categorical clustering (for faces and, to a lesser degree, for animate objects) in FFA, but not in EVC. This is consistent with our inspection of dissimilarity matrices in Figure 3.10.

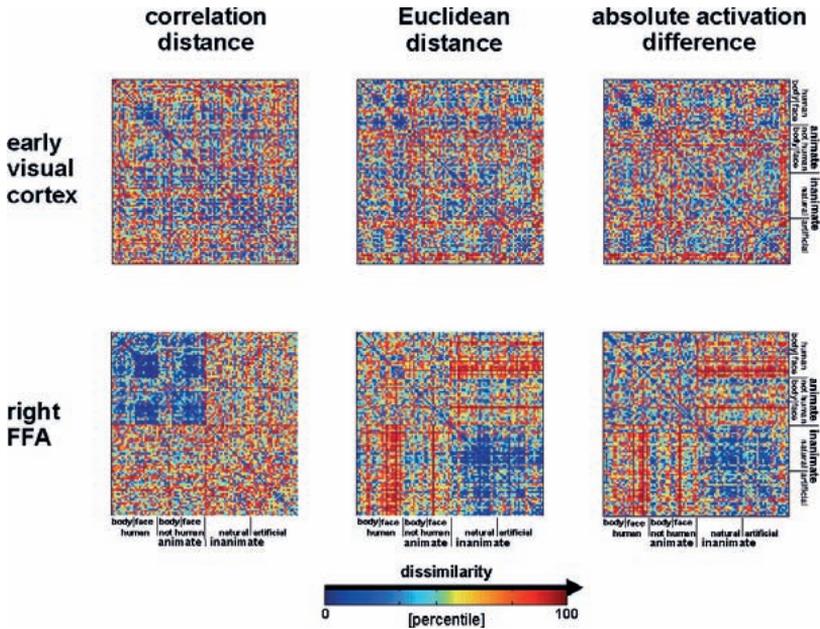


Figure 3.10 Dissimilarity matrices of activity patterns elicited in early visual cortex and FFA by viewing 96 object images. Dissimilarity matrices (as introduced in Figure 3.2) are shown for early visual cortex (top row) and right FFA (bottom row) and for three different measures of dissimilarity (columns): 1-correlation (Pearson correlation across space), the Euclidean distance between the two response patterns (in standard error units) and the absolute activation difference (i.e. the absolute value of the difference of the spatial-mean activity level). The absolute activation difference is sensitive only to the overall level of activation and has been included only because regional-average activation is conventionally targeted in fMRI analysis. Note that the correlation distance (1-correlation) normalizes for both the overall activation and the variability of activity across space. It is therefore the preferred measure for detecting distributed representations without sensitivity to the global activity level (which could be attributed e.g. to attention). The Euclidean distance combines sensitivity to pattern shape, spatial-mean activity level, and variability across space. Note that as expected using the Euclidean distance yields an RDM resembling both the one obtained with correlation distance and the one obtained with absolute activation difference. The matrices have been separately histogram-equalized (percentile units) for easier comparison. Dissimilarity matrices were averaged across 2 sessions for each of 4 subjects.

3.3.3 Model fits to EVC and FFA

Figure 3.6 shows the RDMs of the models. The first thing to note is that each matrix presents a unique pattern that characterizes the model representation. Figure 3.8 shows the deviation of each model from the empirical RDMs of EVC and FFA. We do not have the space here to fully discuss the neuroscientific implications of this analysis, but we offer some basic observations that demonstrate how RSA can help characterize regional representations:

- For EVC, note that the best-fitting model is the silhouette-image model. This is plausible because EVC is known to contain retinotopic representations of the visual input. The fMRI patterns in EVC appear to reflect primarily the shape of the retinotopic region stimulated (i.e. the shape of the figure, since the background is uniformly gray). That the simple silhouette model explains the RDM better here than the V1 model suggests that the orientation information is not as strongly reflected in the RDM. This is consistent with recent results by Kay et al. (2008), who showed that images can be identified on the basis of their fMRI responses in early visual cortex, with the major portion of the information provided by the retinotopic representation of edge energy and a smaller portion provided by the representation of edge orientation.¹¹ Early visual orientation information is likely to be attenuated in fMRI data because of its fine-scale spatial organization and pooling of columns of all orientation-preferences in each fMRI voxel.
- Among the complex computational models, the V1 model fits the EVC data best, but only the “smoothed” version, where we simulated local pooling of orientation-specific responses in fMRI voxels. Like the good fit of the silhouette model, this is consistent with the limited spatial resolution of our fMRI voxels.
- The RDMs of the fusiform face and parahippocampal place areas in either hemisphere fit the EVC matrix better than the V1 model, but not as well as the silhouette model. One explanation for this is that the conventional V1 model does not capture the full complexity of the representation in EVC. This would be plausible for two reasons: On the one hand, our EVC region contains voxels from the early visual foveal confluence, not just from V1. On the other hand, V1 itself is likely to contain a more complex representation than our Gabor-based model of simple and complex cells.
- The higher-level HMAX-C2 representation based on natural image patches, plausibly does not capture the similarity structure we find in EVC, nor do the simple image transformations.

¹¹ Note that Kay et al. (2008) used stimuli of about 20° visual angle (in contrast to the 2.9° stimuli used here) thus driving a more extended retinotopic representation, which may provide more power for detecting the subtler orientation information present in the fMRI signals. Note also that the two studies take very different approaches to activity-pattern analysis. Finally, the stimulus set always influences what aspects of a representation we are sensitive to in any neurophysiological experiment. Our stimulus set here may not afford great sensitivity to orientation information in the context of RSA: A given pair of images may be similar in orientation at one retinal location and dissimilar at another, such that the overall representational dissimilarity (across the entire extent of the image) ends up at an intermediate value for all pairs of images. Different results might be expected for grating stimuli, where some stimulus pairs are similar in orientation across the entire extent of the image, and other pairs are dissimilar in orientation everywhere (cf. Kamitani and Tong, 2005).

- For the right FFA, the best-fitting dissimilarity structure consists in the empirical dissimilarity of FFA in the opposite hemisphere. This is plausible, given the close functional relationship between the regions.
- The dissimilarities of the right FFA are best modeled by a conceptual model: the “face-animal-prototype model”. This suggests that, to a first approximation, different faces elicit a prototypical response pattern – implying small dissimilarities between individual face response patterns, consistent with Kriegeskorte et al. (2007), and that the same is true to a lesser degree for the more general class of animal images.
- Among the complex computational models, the HMAX-C2 representation based on natural image patches provides the best fit to the right FFA. This may reflect the higher-level nature of the representations in FFA.
- The right FFA resembles the EVC more closely than the V1 model, the silhouette model, or any other brain region. This could reflect feedback from FFA to EVC. Alternatively, FFA may reflect some of the more complex features of the early visual representation that are not captured by either the silhouette or the V1 model.

3.3.4 The similarity structure of representational dissimilarity matrices as revealed by MDS

Figure 3.9 simultaneously relates the RDM “signatures” of all brain regions and models to each other by means of MDS. This representation is devoid of indications of statistical significance and inevitably compromised by geometric distortions (because a higher-dimensional structure is represented in 2D). However, it provides a useful overview of *all* pairwise relationships (not just the relationships shown in Figure 3.8 of EVC and the right FFA to the other representations). Although the 2D distances do not precisely reflect the actual dissimilarities between the dissimilarity matrices, almost all observations from Figure 3.8 are also reflected in the MDS arrangement of Figure 3.9. However, the MDS arrangement provides us with a lot of additional information. As examples of the additional information, consider these observations:

- The close interhemispheric observed for the left and right FFA (Figure 3.8), also holds for the left and right parahippocampal place area.
- The smoothing applied to the V1 model and the RADON model in order to simulate pooling of responses within fMRI voxels does not appear to drastically alter the RDM of either of these models.
- The five brain regions included (red) all seem to be somewhat related in their representational similarity structure. The fact that no model appears in their midst suggests that there may be a common component to these visual representations that is not captured by any of the models.

3.4 The broad potential of representational similarity analysis

3.4.1 Relating models, brain regions, subjects, species, and behavior

Systems neuroscience has struggled to quantitatively relate its three major branches of research: behavioral experimentation, brain-activity experimentation, and computational modeling. The RDM can serve as a hub that relates representations from a variety of sources in the three branches (Figure 3.3). We can use dissimilarity matrices to compare internal representations between two models or two brain regions in the same subject (representational connectivity, see below). In addition, RSA provides a solution to the fine-grained spatial-correspondency problem encountered when relating corresponding brain regions in different subjects of an fMRI experiment. Conventionally, different subjects in an fMRI experiment are related by transforming the data into a common spatial frame of reference, such as Talairach space (Talairach and Tournoux, 1988) or cortical-surface space defined by cortex-based alignment (Fischl et al., 1999; Goebel et al., 2006; Goebel and Singer, 1999). However, these available common spaces do not have sufficient precision to relate high-resolution fMRI voxels. Establishing spatial correspondency is not merely a technical challenge. It is a fundamental empirical question to what spatial precision intersubject correspondency even exists in different functional areas (Kriegeskorte and Bandettini, 2007a). RSA offers an attractive way of abstracting from the spatial layout and even from the linear basis of the representation, allowing us to relate fine-grained activity patterns between subjects. Even different species and modalities of brain-activity data (e.g. single-cell recording and fMRI; Kriegeskorte et al., 2008a) can be meaningfully related with RSA.

3.4.2 Advanced types of representational similarity analysis

Similarity searchlight: Finding brain regions matching a model

RSA also allows us to localize a brain region whose intrinsic representation resembles that of a specified model. For this purpose we can move a spherical or cortex-patch searchlight (Kriegeskorte et al., 2006) throughout the measured volume to select, at each location, a local contiguous set of voxels, for which RSA is performed. The results, for each model, form a continuous statistical brain map reflecting how well that model fits in each local neighborhood.

Representational connectivity analysis

In order to assess to what extent two brain regions in the same subject represent the same information, we can compare the two regions' condition- or time-

point-based dissimilarity matrices (Kriegeskorte et al., 2008a). The latter approach can be applied to either the raw data or residuals of the linear modeling of stimulus-related effects. Using the residuals will focus the analysis on the internal representational dynamics of the system including stochastic innovations. In analogy to functional connectivity analysis, we refer to this approach as “representational connectivity analysis”. It can be combined with the searchlight approach (Kriegeskorte et al., 2006) in order to find a set of regions representationally connected to a given region.

Fitting parameters of computational models

The computational models we present as examples here are fixed models in that they do not have any parameters fitted on the basis of the data. It will be interesting to extend our approach to the fitting of model parameters on the basis of an empirical RDM. For example, a network model could be trained (supervised learning) to fit a given RDM. In order to avoid circular (i.e. self-fulfilling) inference, a separate set of conditions (e.g. different experimental stimuli) will then be needed to assess the fit of the computational model to the experimental data.

Composite modeling of a brain region’s representational dissimilarity matrix

In our demonstration here, we have treated the models as separate accounts of the data to be evaluated independently. A complementary approach is to model the RDM of a brain region by combining several models. To this end, one could combine units from the internal representations of several models (as we have done for simple and complex V1-model units) and compute the overall representational dissimilarity. One could then fit parameters, including the number of units from each model to include in the representation, so as to best account for an empirical RDM. A simpler approach is to directly model an empirical RDM as a combination of model dissimilarity matrices. If we use Euclidean distance to compare activity patterns and assume that the different models account for orthogonal components of the activity patterns (e.g. separate sets of units), then we can account for the squared empirical Euclidean distance matrix as a linear combination of the squared model Euclidean distance matrices. (Note that this does not require the dissimilarity patterns of the models to be orthogonal; the linear model would use the dissimilarity variance uniquely explained by each model to disambiguate the explanation of shared dissimilarity variance.) A more generally applicable approach would be to explain the empirical RDM as a weighted sum of monotonically transformed model dissimilarity matrices, where a separate monotonic transform is estimated for each model simultaneously with the weights.

Weighted representational readout analysis

So far we have thought of a region's representation as characterized by a single RDM. Alternatively, we can consider the representation as a high-dimensional structure that is viewed from different perspectives by the regions that read it out. If readout consists in multiple linear weightings of the representational units, then it amounts to a linear projection that can be likened to the transformation of a 3D structure to a 2D "view" of it. In this spirit, we can reverse the logic of the previous paragraph and see to what extent we can read out a particular dissimilarity structure from the representation by weighting the units before computing the RDM. Again, using the squared Euclidean distance yields a simple relationship: Each unit (e.g. a voxel or a neuron) yields a separate RDM. The overall squared Euclidean distance matrix is the sum of the single-unit squared Euclidean distance matrices. Now we can "account for" each model's dissimilarity pattern as a linear combination of the single-unit dissimilarity matrices. This avenue can be construed as a generalization of linear discriminant analysis from a single contrast to a complex pattern of contrast predictions. It is interesting because of its neuroscientific motivation in terms of readout by other brain regions. As in linear discriminant analysis and classification in general, independent test data will be needed to confirm any relationships suggested by such a fit.

3.4.3 Core concepts for experimental design

What experimental designs lend themselves to RSA? A distinguishing feature of RSA is its potential to simultaneously exploit the spatial and temporal richness of multi-channel brain-activity data. Although RSA can be applied to a wide range of conventional experimental designs, there may be little conceptual motivation for it in the context of certain experiments, e.g. a low-resolution block-design fMRI experiment that targets regional activation and averages across very different processes (e.g. perception of different stimuli within a given category). The benefits of RSA will be greatest for condition-rich experimental designs targeting activity-pattern information with high-resolution measurement. In this section we describe novel types of experimental design that are feasible with RSA and optimally exploit its potential.

Condition-rich design

RSA is particularly useful in conjunction with condition-rich designs. One example of such a design is the 96-object-image experiment we presented to demonstrate the approach. We refer to a design as condition-rich if the number of effective experimental conditions (that is brain states to be discerned) is large. Condition-rich designs approach the limit of the temporal complexity of the signal measured in order to amply sample the space of all possible conditions.

Within the classical approach of massively univariate activation-based analysis (Worsley et al., 1992; Friston et al., 1994; Friston et al., 1995ab; Worsley and Friston, 1995), one way of enriching design has been to parameterize the conditions. The result is a larger number of conditions that might not singly yield stable estimates, but the correlation between condition parameters and brain activity – combining evidence across conditions – can be stably estimated. Such designs also lend themselves to RSA: The model dissimilarity matrices can be computed from the condition parameters. However, RSA is not limited to designs whose conditions sample a predefined parameter space in a regular way. In RSA, the parametric statistical models describing activity variation across time are replaced by computational models exposed to the same experimental conditions. Regular parameterization may help focus the experiment on particular hypotheses, but RSA also accommodates less restricted designs such as the 96-object-image design we use as an example here.

Ungrouped-events design

In the classical block-design approach to fMRI experimentation, an experimental block corresponding to one of the conditions typically includes a variety of brain states (e.g. corresponding to percepts of a variety of stimuli from the same category) that are to be averaged across. While differences between block-average activation can be very sensitively detected with this method, the average results will be ambiguous with respect to single-trial processing (Bedny et al., 2007; Kriegeskorte et al., 2007). Equally importantly, the temporal capacity of the fMRI signal to discern a large number of separate brain states is largely wasted. In event-related designs (Buckner, 1998), stimuli can appear in complex temporal sequences allowing for a wider range of experimental tasks. However, the experimental events are usually still grouped in condition sets and the variety of events forming a single condition is averaged across in the analysis (e.g. by modeling each condition by a single predictor). The sequence of experimental events is often designed to maximize estimation efficiency for the condition contrasts of interest. In that case the design itself will imply a grouping of the experimental events.

We propose to avoid any predefined grouping of experimental events (ungrouped-events design). Each experimental event (e.g. each stimulus) is treated as a separate condition (Figure 3.7; Kriegeskorte et al., 2007; Aguirre, 2007; Kriegeskorte et al., 2008a). The 4-image experiment is an example of an ungrouped-events design. The 96-image experiment is an example of an ungrouped events design, which is also condition-rich. One approach is to have events occur in a random sequence implying no grouping. In order to include a reasonable number of events, but still be able to discern the activity patterns

they are associated with, we use a design with temporally overlapping but still separable single-trial hemodynamic responses here. Our example employs a design with a trial-onset asynchrony (TOA) of 4s (Figure 3.7). The effects of varying the TOA are explored in Figure 3.11. A more detailed discussion of optimal event sequences for condition-rich designs (including ungrouped-events designs) is to be found in the Appendix (Section *Optimal condition-rich fMRI design*).

For estimation of a given contrast of interest, a condition-rich ungrouped-events design with a random sequence will be less efficient than a block-design or a sequence-optimized rapid event-related design. In our view, however, the statistical cost is more than offset by the ability to group the events into arbitrary sets and, more generally, to study the rich space they populate and its relationship to the brain-activity patterns they are associated with. RSA provides an attractive method for exploring this rich empirical information and testing particular hypotheses.

Unique-events design and time-continuous experimentation

An ungrouped-events design does not group different experimental events into a condition set, but it may contain repetitions of identical experimental events. An extreme type of ungrouped-events design would be a unique-events design, in which no experimental event is ever repeated. RSA can handle unique-events designs just like any other design. This is an important property, because unique-events designs take the complexity of the conditions set to the limit of the temporal capacity of the measured signal. In addition, there are neuroscientific domains, where exact repetition of an experimental event is a questionable concept. Strictly speaking each experimental event in any experiment – and in fact any experienced event at all – permanently changes the brain. In many studies, we may choose a design that minimizes such effects so that we can neglect them in the analysis. For studies of plasticity, however, it may be attractive to track changes to the system along with its activity dynamics. RSA in conjunction with a suitably plastic computational model could address this challenge.

We can go one step further and abolish the notion of discrete experimental events in favor of that of time-continuous experimentation (e.g. Hasson et al., 2004). For time-continuous designs, we can treat each acquired volume as a separate condition and directly compute the RDM from the data. For each region of interest, the resulting RDM will then have a width and height corresponding to the number of time points. We refer to such a dissimilarity matrix as a time² dissimilarity matrix. For fMRI data, the time² dissimilarity matrix will reflect the temporal characteristics of the hemodynamic response. Time-continuous RSA is attractive for studies of time-continuous perception of stimuli, including com-

plex natural stimuli such as movies (Hasson et al., 2004), and, more generally, for studies of time-continuous interactions, such as playing computer games or interacting with a virtual-reality environment (Schneider et al., HBM 2007). Note that time-continuous experimentation allows for greater ecological validity (i.e. the subject's experimental experience can be made more similar to experiences in natural environments). However, time-continuous experimentation can also utilize stimuli and interactions that are unnatural and designed to address a particular hypothesis – trading ecological validity for experimental control.

3.4.4 Data-driven and hypothesis-driven representational similarity analysis

RSA lends itself to a broad spectrum of analyses from data-driven (where results richly reflect the data) to hypothesis-driven (where results are strongly constrained by theoretical assumptions and the data serve to test predefined hypotheses). On the former end of the spectrum, the RDM itself richly reflects a given region's representation. A multidimensional-scaling arrangement of the conditions set in two dimensions (Figures 3.1 and 3.4) provides a data-driven, exploratory visualization that can allow us to discover natural groupings within the representational space (Edelman et al., 1998). But RSA becomes distinctly hypothesis-driven when we test whether a predefined model fits a brain region's representation (Figure 3.8). One hallmark of hypothesis-driven analysis is complexity reduction. When we test a model fit by comparing two dissimilarity matrices, the voxel-by-time data matrix is reduced to a single fit parameter or the result of a statistical test.

The RDM at the front end of RSA certainly is a more data-driven representation than a scalar measure of model fit. But how rich is it exactly? That depends on the number of conditions. Usually computing the RDM will reduce the amount of data. Consider a single-subject experiment with 96 conditions (as in our example here). Let's assume we are analyzing a region of 100 voxels and the experiment has 500 time points. The data matrix has $100 \times 500 = 50,000$ numbers. The RDM (symmetrical about a diagonal of zeros) has $(96^2 - 96) / 2 = 4560$ parameters. Computing the RDM, thus, constitutes a complexity reduction. If we consider the time² dissimilarity matrix, on the other hand, we have expanded the data matrix into a 500 by 500 matrix with $(500^2 - 500) / 2 = 124,750$ parameters.

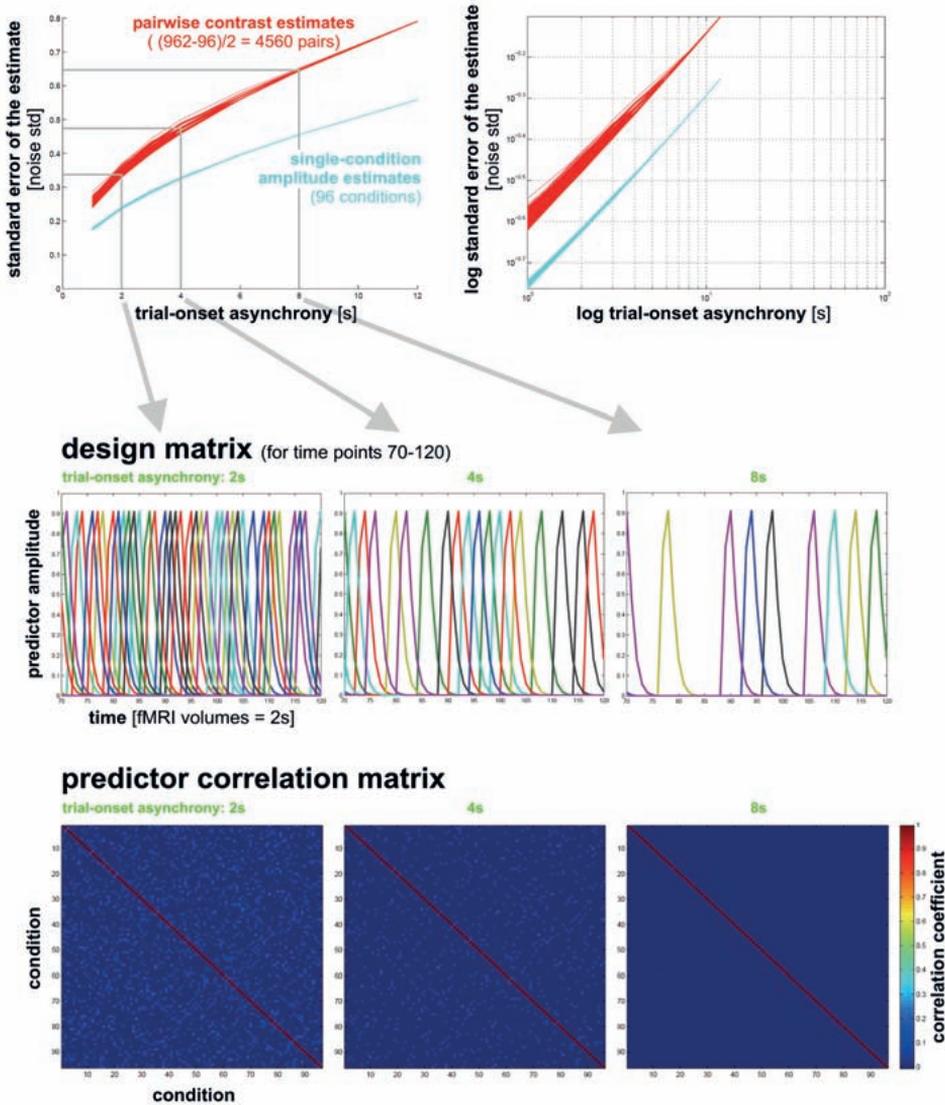


Figure 3.11 Design efficiency as a function of trial-onset asynchrony for a 96-condition fMRI design. This figure shows simulation results exploring how statistical efficiency depends on the trial-onset asynchrony (TOA) under linear-systems assumptions for a 96-condition design with one hemodynamic-response predictor per condition and a random sequence of experimental events (including 25% null events for baseline estimation). We assume that about 50 minutes of fMRI data are to be collected in a single subject. The simulation suggests a simple conclusion: The more closely the trials are spaced in time, the higher the efficiency will be (top panels) for single-conditions (cyan) and pairwise condition contrasts (red). Doubling the number of trials packed into the same 50-minute period, then, would improve efficiency about as much as performing the whole experiment twice: decreasing the standard errors of the estimates roughly by a factor of $\sqrt{2}$. In other words, the standard errors are proportional to $\sqrt{\text{TOA}}$. (Why doesn't the greater response overlap decrease efficiency? For an intuitive understanding, consider that although the greater response overlap for shorter TOAs corre-

lates predictors, the greater number of event repetitions decorrelates them.) Importantly, however, the straightforward relationship suggested by the simulation rests on the assumption of a linear neuronal and hemodynamic response system. In reality, the effects of closely spaced events may interact at the neuronal level and the hemodynamic responses may also not behave linearly (e.g. three 16-ms stimuli at a TOA of 32-ms are unlikely to elicit a hemodynamic response that is three times higher than that to a single such stimulus). The choice of TOA therefore requires an informed guess regarding the short-TOA nonlinearity for the particular experimental events used. For the 96-image experiment, we chose a TOA of 4s. Details on the simulation and an intuitive explanation for the result are given in the Appendix (Section *Optimal condition-rich fMRI design*), along with further discussion of design choices including the TOA.

Meaningful statistical summaries

In order to learn from the massive amounts of brain-activity data we can acquire today with techniques including fMRI as well as scalp and invasive multi-channel electrophysiological techniques and voltage-sensitive dye imaging, we need meaningful statistical summaries that relate a complex data set to systems-level theory. First, statistical summaries are needed to reduce the complexity of the effects and relate them to theory. Second, statistical summaries combine the evidence of many noisy measurements, thus helping us separate effects from noise. The most obvious and widespread method of summarizing data is averaging. While potentially powerful, averaging applied too early in the analysis can remove the effects of greatest neuroscientific interest. In fMRI, for example, data are often locally averaged (i.e. smoothed) prior to mapping analysis. This removes fine-grained spatial-pattern effects that reflect each functional region's intrinsic representation (Kriegeskorte and Bandettini, 2007a; Kriegeskorte et al., 2006). Similarly in the temporal dimension, grouped-events designs (including block designs) average across very different experimental events, rendering results ambiguous with regard to single-trial processing (Bedny et al., 2007; Kriegeskorte et al., 2007).

Late combination of evidence

A central theme of RSA is late combination of evidence: In order to better exploit the complexity of the data toward neuroscientific insights, spatial as well as temporal averaging (across sets of different experimental events) is omitted. This does not mean that the analysis involves less combination of evidence for reduction of complexity. Instead the combination of the evidence occurs later on, in ways that are conceptually better motivated. Evidence is combined in RSA, for example, when (1) the patterns of activity within an extended region of interest are summarized in an RDM, when (2) dissimilarity matrices for a given functional region are averaged across subjects, and when (3) the complex structure of the resulting group-average RDM is compared to model dissimilarity matrices (summarizing the region's function by its goodness of fit to several models or by the index of the best-fitting model). Combining evidence requires theoretical assumptions. If we take a step back to look at the empirical cycle as a

whole, we can motivate late combination of evidence in terms of late commitment to theoretical assumptions.

Late commitment: Using theoretical assumptions to constrain analysis, not design

In the first step of the empirical cycle, we strive to minimize the theoretical assumptions built into the experimental design. This approach is motivated by the observation that designs, e.g. of fMRI experiments, can be made much more versatile (allowing us to address more neuroscientific questions) at moderate costs in terms of statistical efficiency (for addressing a given question). A general design that can address a hundred questions appears more useful than a restricted design that addresses a single question with slightly greater efficiency.

Statistical power is afforded by combining the evidence – usually by averaging. When we decide on a grouping of experimental events (e.g. for a block design), we commit to a particular way of combining the evidence and thus give up versatility. Ungrouped-events designs allow us to combine the evidence in many different ways *during analysis*. First, this approach allows for exploratory analyses, which can (a) test basic assumptions of a field, (b) usefully direct our attention to larger phenomena (in terms of explained variance) and (c) lead to unexpected discoveries. Second, ungrouped-events designs allow a broad set of theoretically constrained analyses to be performed on the same data. And third, as a consequence, such designs allow us to combine data across studies and research groups in order to address a particular question with a power otherwise unattainable. In the Appendix, we assess this third point, the potential of data sharing within subfields of neuroscientific inquiry, in detail.

3.5 Discussion

3.5.1 To what extent does measured pattern information reflect neuronal representations?

A fundamental question in systems neuroscience is to what extent brain-activity patterns measured with different techniques reflect neuronal pattern information. RSA characterizes pattern information in terms of pattern similarity and, thus, provides one attractive avenue for addressing this issue. We will focus our discussion here on blood-oxygen-level-dependent fMRI (Ogawa et al., 1990; Bandettini et al., 1992; Kwong et al., 1992; Ogawa et al., 1992), but similar arguments hold for other modalities.

What pattern information will be shared between fMRI and neuronal activity is difficult to predict, because fMRI voxels sample neuronal activity through a complex spatiotemporal transform: the hemodynamics. If voxels reflected simply the spatiotemporally local average of neuronal activity, then any neuronal pattern differences in the attenuated high spatial and temporal frequency bands would be reduced or eliminated in the fMRI similarity structures. However, fMRI voxel sampling is likely to be more complex than local averaging and may have sensitivity to neuronal pattern information in unexpectedly high spatial (and possibly temporal) frequencies (consider Kamitani and Tong, 2005; Haynes and Rees, 2005a). The unexpected sensitivity of fMRI is encouraging, but also suggests a more complex transform from neuronal to fMRI patterns, making it more difficult to predict what aspects of neuronal information exactly are reflected in fMRI patterns.

We used RSA to relate neuronal patterns recorded in monkey IT (Kiani et al., 2007) to fMRI patterns elicited by the same set of 92 object images (the set also used in our example here) in human IT (Kriegeskorte et al., 2008a). Despite the confounding species difference, results show a surprising match between the two dissimilarity matrices (linear correlation = 0.49, $p < 0.0001$). This indicates not only that monkey and human IT represent similar object-image information, but also that this information is similarly reflected in single-cell recordings and high-resolution fMRI, when analyzed with massively multivariate information-based techniques. The convergence of fMRI and neuronal recordings had not previously been addressed at the level of pattern information and our results are encouraging. Ultimately, however, assessing to what extent pattern information is shared between neuronal activity and fMRI will require simultaneous measurement in both modalities, just as for local activity (Logothetis et al., 2001; Shmuel et al., HBM 2007).

It appears likely that high-resolution fMRI (Cheng et al., 2001; Hyde et al., 2001; Duong et al., 2001; Yacoub et al., 2003; Harel et al., 2006; Kriegeskorte and Bandettini, 2007a) and cell recording will turn out to convey overlapping but non-identical components of the underlying neuronal pattern information. While fMRI is limited by hemodynamic signal confluence yielding an ambiguous combination signal at each voxel, invasive electrophysiological techniques are limited by selective subsampling of neuronal responses. It will be interesting to see if fMRI provides us with merely a subset of the information recorded by implanted multi-electrode arrays or if it can also give us neuronal pattern information missing in a given array recording. RSA appears attractive for relating modalities and also for use in each modality, no matter what their relationship turns out to be.

3.5.2 Relation between RSA and other tools of pattern-information analysis

Multivariate techniques of pattern-information analysis have recently gained momentum in fMRI and electrophysiology (see list of citations in the Introduction). RSA shares a key feature with the cited pattern-information approaches: it is motivated by the theoretical concept of distributed representation and targets activity-pattern information, combining evidence across space and time. However, RSA differs from the cited pattern-information approaches in that it considers how the activity-pattern dissimilarity matrix relates to dissimilarity matrices predicted by theoretical models, i.e. a second-order isomorphism. The cited pattern-information approaches, in contrast, attempt to demonstrate that each condition is associated with a distinct activity pattern, i.e. a first-order isomorphism. RSA can be thought of as a particular variant of pattern-information analysis, which need not involve decoding or classification of internal representations. But at the same time RSA can be construed as a generalization of pattern-information analysis, where many pattern-contrast predictions are tested together. A test of the discriminability of the activity patterns associated with two conditions is handled as a special case, using a binary model dissimilarity matrix.¹²

An important feature of RSA is the goal of understanding and quantitatively explaining the empirical RDM. This entails a healthy focus on the major variance-explaining components in the data. In classifier-based pattern-information analysis, by contrast, we typically focus on a particular dimension defined by the sets of experimental conditions we set out to discriminate. Classifier-based pattern-information analysis, therefore, typically has a stronger theoretical bias than RSA. However, we are free to trade off variance for bias by means of testing constrained model spaces. For example, instead of asking, which of a range of models best explains the FFA representational dissimilarity (RSA), we could ask simply if animacy can be decoded from the FFA response patterns (pattern-information analysis). Or we could address the same smaller question with RSA by asking if the animate-inanimate matrix explains any dissimilarity variance.

The simple implementation of RSA that we describe here is less sophisticated than classification approaches in how it accounts for structured noise and

¹² We would enter a 1 in the model dissimilarity matrix when the hypothesis predicts distinct activity patterns and a 0 otherwise. In order to obtain enough dissimilarity values for correlation of dissimilarities, we might need to use multiple activity-pattern estimates obtained for replications of each condition. Consider the simplest case of a two-condition experiment. The lower triangle of the dissimilarity matrix would contain a single cell, rendering dissimilarity correlation impossible. However, we could split the data to get two independent activity-pattern estimates per condition, or we could use each trial as a separate estimate.

nonlinear representational geometries. This may suggest the use of more complex dissimilarity measures. However, estimating nonlinear relationships requires substantial amounts of data. One strength of RSA is its ability to deal with and integrate information about a large number of conditions. For condition-rich experiments, the amount of data per condition pair will be small and techniques accounting for more complex geometries will likely need to combine information across many conditions in order to provide stable estimates.

3.5.3 RSA and information-theoretic quantification

Considering the RDM is motivated by the idea that it encapsulates, in an intuitive sense, the pattern information a region conveys about the experimental condition. It is natural to ask for formal information-theoretic quantification. It would be desirable to obtain pattern-information estimates (i.e. mutual information between experimental condition and spatiotemporal activity pattern) that do not depend on assumptions about the code (as pattern classification, or “decoding”, approaches do). To this end, we could estimate a multivariate pattern distribution for each condition and compute the plug-in estimate of mutual information. But estimates of high-dimensional distributions from small numbers of data points (as are available in fMRI in relation to the number of voxels) are highly susceptible to noise, unless constrained by strong assumptions. The difficulty will grow with the number of conditions. Modern approaches to estimation of mutual information circumvent explicit multivariate distribution estimates and take a graph-theoretical approach (Kraskov et al., 2004). In our hands, however, grasping for such generality in fMRI analysis has been associated with prohibitive penalties in terms of estimate stability. Nevertheless information-theoretic quantification is an important direction for further exploration.

3.6 Appendix

3.6.1 Optimal condition-rich fMRI design

How should the sequence of events be designed for a condition-rich fMRI experiment like the 96-image experiment? The field has developed sophisticated methods for designing experimental event sequences to optimize statistical efficiency (e.g. Wager and Nichols, 2003). These methods are general and apply to condition-rich designs as a special case. However, the large number of conditions has some consequences that merit consideration. Our goal here is the estimation of (a) a response amplitude for each condition and (b) a response-amplitude contrast for each pair of conditions. We assume a linear hemodynamic response model (Boynton et al., 1996) to obtain a design matrix for the

experiment (Figure 3.7). Optimizing the event sequence so as to maximize the stability of these estimates will have two main consequences: (1) Events belonging to the same condition (identical events in ungrouped-events design) will become clustered in time. (This improves estimate stability because temporally overlapping hemodynamic responses to successive trials will add up so as to increase the sum of squares, i.e. the predictor energy.) (2) Events will be sequenced so as to approximately orthogonalize the hemodynamic response predictors for all pairs of conditions. (This improves estimate stability because it reduces mutual dependency for pairs of condition estimates, thus disambiguating the joint least-squares estimate.) We will argue that in the context of condition-rich design, (1) temporal clustering may be undesirable, (2) random sequences may yield sufficiently low predictor correlation, and (3) shorter trial-onset asynchrony (TOA) yields greater power, as long as linearity of the responses holds.

(1) Temporal clustering may be undesirable. For a condition-rich design, temporal clustering of conditions may not be desirable. Consider our 96-condition example. We will assume the realistic scenario that, within a single experimental session, we acquire about 50 minutes of fMRI data for the main experiment. At a TOA of 2s (about the minimum if we are to avoid nonlinearities of the hemodynamic response), we can only repeat each condition about 12 times per session. Temporal clustering of such few repetitions over a 50-minute experiment is undesirable: it would entail that a given condition occurs only a handful of times (with two or more consecutive repetitions) over the course of the entire session, rendering temporal confounds (e.g. subject fatigue) a serious concern. This problem will be even more pronounced at longer TOAs (such as the 4-s TOA used in our experiment), because there will be even fewer repetitions. We prefer to distribute the repetitions of each condition equally across the experiment. In our experiment here, we repeated each condition exactly once in each run, which has the added benefit that failed runs do not create imbalances in the amount of data available for each condition.

(2) Random sequences may yield sufficiently low predictor correlation. Sequence optimization can serve to orthogonalize predictors. How large a benefit does this promise? Figure 3.11 explores design-efficiency for 96-condition designs as a function of TOA. We used an unoptimized random sequence for each run (with 25% null events interspersed at random), concatenating such sequences to fill the 50-minute experimental session. The predictor correlation matrices for these unoptimized random sequences suggest that predictor correlation is already low. For short TOAs (e.g. 2s), there is some room for improvement. For slightly longer TOAs (e.g. 4s as used in the experiment here), predictor correlation depends mainly on the immediate temporal neighbors of each condition (because the hemodynamic response overlap is negligible for trials that have an

intervening trial between them). In the 4-s TOA case, each condition is repeated 6 times in the 50-minute experiment. Using random sequences, most conditions have no repeated temporal neighbors, about a third of the conditions have one repeated temporal neighbor. This is reflected in the predictor correlation matrix, which shows homogeneously low correlations (below .1) across pairs of conditions. Sequence optimization might bring the design slightly closer to the ideal of predictor orthogonality, but efficiency gains will be very small, because there is little room for improvement. Practical considerations add to the argument in favor of using random sequences: We may have to deal with failed runs. Moreover, during analysis we may want to divide the data into subsets of runs (e.g. odd runs as training set, even runs as test set). Sequence optimization should ideally anticipate these eventualities, thus complicating the process. In sum, event-sequence optimization should be considered in designing a condition-rich fMRI experiment. However, in certain scenarios, such as the present example, the benefits may be negligible.

(3) Shorter trial-onset asynchrony yields greater power, as long as linearity holds. What is the optimal TOA for a condition-rich fMRI experiment? Figure 3.11 explores how statistical efficiency depends on TOA for a 96-condition design using a random event sequence (including 25% null events). The simulation suggests a simple conclusion: The more closely the trials are spaced in time, the higher the efficiency will be (Figure 3.11, top panels) for single-condition amplitude estimates (cyan) and pairwise amplitude contrasts (red) – assuming linearity of the responses. The choice of TOA therefore requires an informed guess: it should be the shortest TOA, for which linearity holds for the particular experimental events used.

We now describe the simulation in detail and explain why the linear-systems assumption does not predict a greater cost of response overlap in time. We, again, assume that about 50 minutes of fMRI data are to be collected in a single subject in a given session. If the TOA is 8s, we can repeat each of the 96 conditions 3 times (with 25% null events in each run). If the TOA is 4s, we can repeat each of the 96 conditions 6 times, but now the hemodynamic responses clearly overlap. The greater number of repetitions in the measurement time increases design efficiency. However, hemodynamic-response overlap renders predictors nonorthogonal, which decreases design efficiency. Predictor nonorthogonality is reflected in the predictor correlation matrices (Figure 3.11, bottom row) for a TOA of 2s (left), 4s (middle), and 8s (right). We use the standard general linear model framework to predict the standard errors of the estimates of (a) response amplitudes for single conditions (cyan in top row) and (b) contrasts between pairs of conditions (red in top row). The standard error estimate is $\sqrt{\text{var}(\text{residuals}) \cdot c^T (X^T X)^{-1} c}$, where c is the contrast of interest and X the design matrix. We plot $\sqrt{c^T (X^T X)^{-1} c}$, which can be interpreted as the stan-

standard error in noise standard deviation units ($\sqrt{\text{var}(\text{residuals})}$). The standard error is plotted as a function of the TOA (inversely related to the number of repetitions, as each simulation assumes the same overall measurement duration of 50 min). The simulation suggests that the loss due to hemodynamic overlap is negligible. This is because shorter TOAs also allow more repetitions: A given condition will overlap more, but also be repeated more (and overlap with different other conditions on each repetition). As a result, doubling the number of trials roughly divides the standard error by $\sqrt{2}$, as expected for no overlap. For shorter TOAs, however, the standard error of pairwise contrast estimates varies more across contrasts, because some pairs of conditions overlap more than others (for long TOAs, there is no overlap for any pair of conditions). The simulation is based on the assumption of a linear system, which will break down for short TOAs, because responses to successive trials will interact. Such interactions may occur as part of the hemodynamics and as part of the neuro-cognitive processes occurring in the experiment. For each experiment, thus, we need to judge how closely we think we can space the trials and still rely on the linear-systems assumption for analysis.

Our example here employs a design with a TOA of 4s (Figure 3.7). Because this is faster than a slow event-related design (i.e. a design with nonoverlapping hemodynamic responses to successive events, $\text{TOA} \geq 12\text{s}$), but slower than most rapid event-related designs (which have overlapping hemodynamic responses, $\text{TOA} < 4\text{s}$), we refer to it as a *quick event-related* fMRI design. If linearity holds, a more rapid design, e.g. using a TOA of 2s should yield greater statistical efficiency. Estimating single-trial responses would be compromised for a 2-s TOA, but this may not be considered a drawback.

Should trials be temporally jittered on a grid finer than the minimal TOA? Temporal jittering is important when the goal is the estimation of the shape of the hemodynamic response (e.g. using a finite-impulse-response model). Here our goal is the estimation of response amplitudes and pairwise amplitude contrasts under the assumption of a shape for the hemodynamic response (Boynton et al., 1996). Fine-scale temporal jittering does not in general improve estimate stability in this context.

3.6.2 RSA and data sharing within subfields of neuroscientific inquiry

Data sharing has great potential in many fields including the different disciplines of neuroscience. For human fMRI, the National fMRI Data Center (Van Horn et al., 2005) has pioneered the central facilitation of data sharing. A problem to be overcome is the complexity of individual experiments to be described and understood by other researchers. The fact that experiments are often designed to test particular hypotheses reduces the versatility of the data. The sci-

entist reanalyzing a given data set may find that particular details of the design are detrimental to answering the question to be addressed by the reanalysis. This can render reanalysis less attractive than performing a new experiment designed specifically for the hypothesis at hand.

The approach suggested here of keeping design more general with respect to the hypotheses to be addressed enhances the potential for data sharing. In order to overcome the disconnect impeding data sharing today, greater generality of design needs to be compounded by data-sharing efforts specialized to specific subfields. This promises collaborative synergies previously difficult to imagine. For example, it may allow us to test a given novel hypothesis instantly using a large amount of data acquired by multiple groups over a number of years. Within subfields, experimental designs are often similar in many of their generic features. This is certainly the case for subfields of the field of visual perception. Consider object-vision fMRI, where the only essential differences between a large number of experiments concern the images presented and their grouping. (There are certainly studies with unique designs or task manipulations. However, a sizable subset could be assimilated to a generic approach.) Essential similarities of design are also evident within subfields of the fields of auditory perception, memory research, higher cognition, and motor control.

Within object-vision fMRI, it would be useful to collect stimulus images along with the response patterns they elicit in individual subjects. The collection of experimental data in this format of stimulus pattern and response pattern should be combined with the collection of computational models (e.g. in Matlab) capable of processing arbitrary stimulus images. We envision a phase of informal data and model sharing (during which formats will be negotiated) to culminate in the development of a web-based collaboration portal for object-vision fMRI (and perhaps other modalities). The object-vision fMRI portal would allow downloading of data sets and computational models as well as online testing of computational models and theoretical hypotheses. As a result, separate populations of theoretical and experimental neuroscientists could relate their contributions via an information-rich quantitative interface. On the one hand, this will enable individuals to specialize in either theoretical or experimental work, while keeping the other aspect an integral part of their quantitative analyses. On the other hand, it will empower researchers interested in both computational theory and experimental work to take their transdisciplinary approach to another level.

3.6.3 Additional statistical tests for RSA

There is an extended literature on finding lower-dimensional representations on the basis of dissimilarity or distance matrices. Popular techniques include

MDS (Torgerson, 1958; Kruskal and Wish, 1978; Shepard, 1980) and clustering algorithms (e.g. Johnson, 1967; von Luxburg, 2007), as well as nonlinear manifold-learning techniques such as isomap (Tenenbaum et al., 2000) and locally linear embedding (Roweis and Saul, 2000). However, we are not aware of a literature on statistical testing of the relationships between two or more dissimilarity matrices. Analysis of RDM relationships is an interesting special case of multivariate analysis, where the space typically has a very large number of the dimensions (4560 in our 96-condition example), and those dimensions are related in a particular way – as each corresponds to a pair of conditions. We will briefly discuss some basic statistical tests for dissimilarity matrices that have yet to be developed (or found in the literature in case they exist).

Difference between two dissimilarity matrices

We have proposed a randomization procedure for testing the *relatedness* of two dissimilarity matrices (Step 5). A separate statistical question is whether two dissimilarity matrices are different. Why is this a separate question? First, a failure to find a significant relatedness does not imply that there is no relation; the noise in the data may just obscure the effect. Second, multivariate entities such as dissimilarity matrices can be at once related and distinct – just like two people (e.g. brothers) can be related without being identical. In order to test the difference between two dissimilarity matrices, we need to estimate the distribution of the measure of fit (e.g. correlation between the matrices) under the null hypothesis that the two dissimilarity matrices are identical. The measure of fit will vary due to measurement noise affecting one or both dissimilarity matrices.

Difference in fit of two model dissimilarity matrices to a brain-data dissimilarity matrix

We may wish to assess whether one model RDM fits the data RDM for a given brain region better than another one. The previously discussed tests do not have direct implications for this one. Consider, for example, a case in which both models are significantly related to and significantly different from the data RDM. One of them may still fit the data significantly better (given measurement noise) than the other. Figure 3.8 shows two bar graphs of RDM model fits (to early visual cortex and FFA). The standard-error bars are estimated as the standard deviation of the fit parameter obtained for bootstrap resamplings of the conditions set. Bootstrap resampling could also be used for a formal test of the difference between two models in fitting a data RDM.

Inference from experimental sample of conditions to the population of conditions

Statistical inference in neuroscience usually generalizes within subjects (i.e. to potential replications of the experiment with the same subjects) or across sub-

jects (i.e. to the population the subjects were randomly selected from). Both of these forms of inference can be performed in RSA, but the methods have yet to be developed. In addition, condition-rich design promises the possibility of performing statistical inference to generalize from the experimental conditions actually used in the experiment to the population of experimental conditions the actual conditions were randomly selected from. Bootstrap resampling of the conditions set (as used to compute the standard-error bars in Figure 3.8) is one method of estimating the distribution of RDM fits for random sets of experimental conditions. Formal statistical inference to the conditions population is an exciting topic for further research.

3.6.4 A motivation for the use of rank-correlation distance in comparing representational dissimilarity matrices

Given the nature of the computational and conceptual models and the noise affecting the brain dissimilarity matrices, we cannot in general rely on a direct match of the dissimilarity magnitudes between models and regions. The Euclidean distance therefore does not appear appropriate for comparing dissimilarity matrices, unless the matrices are first normalized in some way. Normalization could consist in a rank-transform of each RDM (i.e. replacing each value by its rank in the context of all the other values in the matrix). This yields a uniform distribution of dissimilarity values, which conserves the order. Alternatively, we could impose a Gaussian distribution of dissimilarities, again preserving the order.¹³

Instead of normalizing each RDM before computing Euclidean distances, we could choose a distance measure that implies a normalization, for example correlation distance, i.e. $1 - r$. If we expect the true relationship between the dissimilarity values in two matrices to be linear, we can use the Pearson linear correlation coefficient to compute r . Whenever one of the matrices is of merely ordinal scale or a nonlinear monotonic relationship between the dissimilarities is plausible, a rank correlation coefficient is more appropriate.

Another line of argument suggests using rank correlation to compare brain and model dissimilarities, even when a linear relationship between the true dissimilarities is expected. The argument is based on the effect of activity-pattern noise on a brain region's RDM. We assume (1) that the activity-patterns are high-

¹³ Gaussianization may be a useful transformation before averaging dissimilarity matrices (e.g. across sessions or subjects). Because the resulting distances between dissimilarity matrices are not limited in range (as is the case for correlation distance or any rank-transformed distance), the distribution of noise displacements in RDM space may be closer to isotropic, rendering the average a more meaningful measure of central tendency.

dimensional (hundreds or thousands of values in each activity pattern), and (2) that the activity pattern noise is additive, independent of the activity patterns, and isotropic. The high dimensionality of the activity-pattern space has a desirable consequence (a blessing of dimensionality, if you will): The displacement of each true activity pattern by an additive noise pattern is likely to be (1) approximately orthogonal to each of the activity-pattern differences and to each other noise displacement, and (2) of approximately constant Euclidean length. The approximate orthogonality results from the fact that there are so many directions in a high-dimensional space and most of them are approximately orthogonal to any given direction. The approximately constant length results from the fact that the variability of the displacements' Euclidean lengths (relative to their mean length) becomes smaller and smaller as dimensionality increases. We can, thus, think of the activity-pattern noise as affecting the Euclidean distance matrix (condition-by-condition) approximately as follows: $d_i' = \sqrt{d_i^2 + 2c^2}$, where d_i are the true distances, d_i' the approximate distance estimates from noisy data, i the condition-pair index, and c the norm of the noise displacement affecting each activity-pattern estimate. The activity-pattern noise, thus, places the squared Euclidean distances on a pedestal. As a result, the Euclidean distance matrix is nonlinearly, but monotonically transformed. The transform is monotonic because none of the three operations (squaring, adding c , and taking the square root) changes the order of the values. The most prominent features of the transform are that the values are scaled down (smaller variance of dissimilarities across the matrix) and shifted up (greater mean dissimilarity).

In practice, the effect of the activity-pattern noise on the RDM will not precisely conform to this prediction, (1) because activity-pattern dimensionality is finite, and therefore the noise displacements of the activity patterns will not be of exactly constant length or exactly orthogonal to the true pattern differences, (2) because the assumptions about the noise may not hold, and (3) because we may use a distance other than Euclidean distance (e.g. correlation distance) for the activity-pattern dissimilarity matrix. Nevertheless, this relationship may hold approximately. The expected prominent shift up of all values in the RDM and its nonlinear and approximately monotonic transform suggest using a rank-correlation distance (e.g. 1 - Spearman rank correlation) for comparing representational similarity matrices.

3.7 Methodological details

3.7.1 FMRI experiments

The results shown here to demonstrate RSA have not been presented before. However, the experiments have been previously described and analyzed to address different questions in Kriegeskorte et al. (2007; 4-image experiment) and Kriegeskorte et al. (2008; 96-image experiment), where further experimental details can be found.

Ungrouped-events designs and tasks

4-image experiment. We performed an ungrouped-events design using 4 object photos as stimuli. The particular stimuli are shown in Figure 3.1. Subjects were familiarized with the four images before the experiment and instructed to continually fixate a central cross, which was always visible, and to perform an anomaly-detection task during the experiment. On 12% of the trials of each experimental run, subtle variations of the four images were presented. In each anomalous version, the global shape of the object as well as several details were slightly distorted. Subjects were asked to press a button placed underneath their right index finger on a regular trial and a button underneath their left index finger when they detected an anomalous image. The task served to motivate subjects to attend to each image presentation even after many repetitions and allowed us to monitor attentive viewing. We used a rapid event-related design with a basic trial duration of 3 s (minimal trial-onset asynchrony), corresponding to two functional volumes of TR=1.5 s. The event sequence was optimized for estimation of the contrasts between the responses to the four original images by a method based on a genetic algorithm (Wager and Nichols, 2003). Each image was presented for 400 ms. In each run, there were 63 presentations of each of the four original images, 33 presentations of anomalous versions of the images and nine null trials, on which the image presentation was omitted and the fixation cross remained visible. The total number of 3-s time slots was, thus, $4 \times 63 + 33 + 9 = 294$, and the duration of the run including two empty time slots at the end was $(294 + 2) \times 3\text{s} = 14.8\text{min}$.

96-image experiment. We performed an ungrouped-events design using 96 object photos as stimuli. The stimuli were chosen from the set used in Kiani et al. (2007), so as to include human and animal bodies (including faces) as well as natural and artificial objects. Stimuli were run-unique with each image presented exactly once in each run. The stimuli were presented at a width of 2.9° visual angle for a duration of 300 ms at a minimal trial-onset asynchrony of 4 s (Figure 3.7). For estimation of baseline activity, the sequence also included null events (25% of trials) with no stimulus presented. Stimuli were presented in

random order (no sequence optimization) on a constantly visible uniform gray background while subjects fixated a white fixation cross. Subjects performed a color-discrimination task: During stimulus presentation the fixation cross turned either green or blue and the subject responded with a right-thumb button press for blue and a left-thumb button press for green. We used a different random event sequence on each of up to 18 runs (spread over up to three fMRI sessions) per subject. The fixation-cross changes to blue or green were chosen according to an independent random sequence. Stimuli were centered with respect to the fixation cross.

fMRI measurements

4-image experiment. We acquired 15 transversal functional slices with a Siemens Magnetom Trio scanner (3 Tesla) using a single-shot gradient-echo echo-planar-imaging (EPI) sequence and a standard birdcage headcoil. The imaged volume consisted in a 3-cm thick temporal-occipital slab including early visual regions as well as the entire ventral visual stream. The pulse-sequence parameters were as follows: in-plane resolution: 2×2 mm², slice thickness: 2 mm (no gap), slice acquisition order: interleaved, field of view (FoV): 256×256 mm², acquisition matrix: 128×128 , time to repeat (TR): 1.5 s, time to echo (TE): 32 ms, flip angle (FA): 75 deg. A functional run lasted 14.8 min. Each subject underwent a single imaging session including two functional runs and a high-resolution T1-weighted anatomical magnetization prepared rapid gradient echo (MPRAGE) scan lasting 9.8 min (192 slices, slice thickness: 1 mm, TR: 2.3 s, TE: 3.93, FA: 8 deg, FoV: 256×256 mm², matrix: 256×256). The experiments were performed at the Donders Centre for Cognitive Neuroimaging (Nijmegen, The Netherlands).

96-image experiment. Blood-oxygen-level-dependent measurements were performed at high spatial resolution using a 3T GE HDx MRI scanner. For signal reception, we used a receive-only whole-brain surface-coil array (16 elements, NOVA Medical Inc., Wilmington, MA). Twenty-five 2-mm axial slices (no gap) were acquired, covering the occipital and temporal lobe, using single-shot interleaved gradient-recalled EPI. Imaging parameters were as follows: EPI matrix size: 128×96 , voxel size: $1.95 \times 1.95 \times 2$ mm³, TE: 30 ms, TR: 2 s. Each functional run consisted of 272 volumes (9 min and 4 s per run). Four subjects were scanned in two separate sessions each, resulting in 11 to 14 runs per subject, yielding a total of 49 runs (equivalent to 7 h, 24 min, and 16 s of fMRI data). As an anatomical reference, we acquired high-resolution T1-weighted whole-brain anatomical scans with an MPRAGE sequence. Imaging parameters were as follows: matrix size: 256×256 , voxel size: $0.86 \times 0.86 \times 1.2$ mm³, 124 slices.

Data preprocessing

The fMRI data sets were subjected to slice-scan-time adjustment and head-motion correction (in this order) using the BrainVoyagerQX software package (R. Goebel, Maastricht, The Netherlands). (1) Slice-scan-time adjustment was performed by resampling the time courses with linear interpolation such that all voxels in a given volume represent the signal at the same point in time. (2) Small head movements were automatically detected and corrected by utilizing the anatomical contrast present in functional MR images. The Levenberg-Marquardt algorithm was used to determine translation and rotation parameters (6 parameters) that minimize the sum of squares of the voxelwise intensity differences between each volume and the first volume of the first run of each session. Each volume was then resampled using trilinear interpolation in 3-D space so as to align it with the first volume of the first run of the session. All further analysis was conducted in Matlab. The cortical surface reconstruction in Figure 3.1 was performed with the AFNI-SUMA software package (R. Cox and Z. Saad, Bethesda, MD, USA).

Extracting condition responses by univariate linear modelling

We concatenated the runs within a session along the temporal dimension. For each voxel, we performed a single univariate linear model fit to extract an activity-amplitude estimate for each of the 96 stimuli. The model (Figure 3.7) included a hemodynamic-response predictor for each of the 96 stimuli. Since each stimulus occurred once in each run, each of the 96 predictors had one hemodynamic response per run and extended across all within-session runs included. The predictor time courses were computed using a linear model of the hemodynamic response (Boynton et al., 1996) and assuming an instant-onset rectangular neural response during each condition of visual stimulation. For each run, the design matrix included these stimulus predictors along with six head-motion-parameter time courses, a linear-trend predictor, a 6-predictor Fourier basis for nonlinear trends (sines and cosines of up to 3 cycles per run) and a confound-mean predictor. Trends were, thus, modeled by a separate set of predictors for each run. The trend predictors for a particular run had zero entries for all other runs along time. For head-motion models and confound means as well, separate predictors accounted for each run (Figure 3.7). Because of the large amount of data concatenated along the temporal dimension for each session, the model fitting was performed in spatial chunks. For each of the 96 stimuli, we converted the activity-amplitude (beta) estimate map into a t map. The resulting 96 t maps were used for RSA.

Definition of regions of interest

All regions of interest (ROIs) were defined on the basis of independent experimental data. In the 4-image experiment (Kriegeskorte et al., 2007), we used a subset of the main-experimental data to define the fusiform face area (FFA; Kanwisher et al., 1997) by means of the contrast faces minus buildings. In the 96-image experiment (Kriegeskorte et al., 2008a), we defined FFA by means of a separate block-design experiment including blocks with faces, places and objects (see below for details on the localizer experiment). The FFA was defined by the contrast faces minus objects. The resulting t contrast map was thresholded so as to define FFA at a range of sizes (for details, see Kriegeskorte et al., 2008a). To define early visual cortex, we selected the most visually responsive voxels within a manually defined anatomical mask selecting an extended cortical region around the calcarine sulcus. Visual responsiveness was assessed using the t map for the average response to the 96 images as assessed for one third of the runs within each session. The remaining runs were used to perform RSA on the ROI. (Since visual responsiveness is orthogonal to the effects of interest here, the data splitting may not be crucial for the present analyses. However, we prefer to consistently use separate data sets for defining ROIs, because it allows us to define ROIs by analyses related to the analyses performed on the ROIs. Using the same data in this context would render the ROI analysis circular.)

Localizer block-design experiment. Along with the 96-image experiment, we performed a functional localizer experiment using the same fMRI sequence as for the 96-image main experiment. Subjects viewed grayscale photos of faces, places, and objects presented in category blocks. Each block lasted 30 s (SOA: 1 s; stimulus duration: 700 ms), alternating with 20-s fixation blocks. Three blocks were presented for each stimulus category (face, place, object), resulting in a total run duration of 7 min and 50 s. Stimuli were presented on a constantly visible uniform black background while subjects fixated a white fixation cross. Subjects continually fixated a central cross and performed a one-back repetition-detection task on the images, responding with a left-thumb button press for each consecutive repetition (3 to 5 repetitions per block). Each stimulus was only presented once, except for the immediate repetitions to be detected in the one-back task. Stimuli were centered with respect to the fixation cross.

Subject-group statistics

In order to combine information across subjects we simply average the dissimilarity matrices computed for each subject separately. Note that this allows the representational patterns to be unique in each subject, while requiring consistency across subjects at the level of the similarity structure. As an alternative to averaging across subjects, one could compute the RDM on the union of ROI voxel sets across subjects (group-brain method). These two alternatives are similar

but not equivalent for correlation distance. Computing a separate RDM for each subject will be required if generalization to the population is to rely on a random-effects analysis. A fixed-effects analysis will afford greater statistical sensitivity. However, generalization to the population will then depend on the assumption that the brain function under study has a neuronal mechanism consistent across the population. This assumption may be reasonable for basic visual functions shared even across species.

3.7.2 Model representations of the stimuli

We processed our stimuli to obtain their representations in a number of low-level models. We then analyzed these model representations in the same way as the brain-activity data. Each image was converted to a representational vector as described below for each model. As for the brain-activity data, each representational vector was then compared to each other representational vector by means of $1-r$ as the dissimilarity measure (where r is the Pearson linear correlation).

Color image (CIELAB)

The RGB color images (175×175 pixels) were converted to the CIELAB color space, which approximates a linear representation of human perceptual color space. Each CIELAB image was then converted to a pixel vector (175×175×3 numbers).

Luminance image

The RGB color images (175×175 pixels) were converted to luminance images. Each luminance image was then converted to a pixel vector (175×175 numbers). We additionally used smoothed versions of these images (low-passed), which were computed by convolving the images with a Gaussian kernel of 11.75 pixels (0.2° visual angle) full width at half maximum. We also used high-passed versions of the images, which were the complements of the low-passed versions (original image minus low-passed version).

Binary silhouette image

The RGB color images (175×175 pixels) were converted to binary silhouette images, in which all background pixels had the value 0 and all figure pixels had the value 1. Each binary silhouette image was then converted to a pixel vector (175×175 binary numbers).

CIELAB joint histogram (6×6×6 bins)

The RGB color images (175×175 pixels) were converted to the CIELAB color space. The three CIELAB dimensions (L, a, b), were then divided into 6 bins of equal width. The joint CIELAB histogram was computed by counting the number of figure pixels (gray background left out) falling into each of the 6×6×6 bins. The joint histogram was converted to a vector (6×6×6 numbers).

V1 model

The luminance images (175×175 pixels, 2.9° visual angle) were given as input to a population of modeled V1 simple and complex cells (Lampl et al., 2004; Riesenhuber and Poggio, 2002; Kiani et al., 2007). The receptive fields (RFs) of simple cells were simulated by Gabor filters of 4 different orientations (0°, 90°, -45° and 45°) and 12 sizes (7-29 pixels). Cell RFs were distributed over the stimulus image at 0.017° intervals in a cartesian grid (for each image pixel there was a simple and a complex cell of each selectivity that had its RF centered on that pixel). Negative values in outputs were rectified to zero. The RFs of complex cells were modeled by the MAX operation performed on outputs of neighboring simple cells with similar orientation selectivity. The MAX operation consists in selecting the strongest (maximum) input to determine the output. This renders the output of a complex cell invariant to the precise location of the stimulus feature that drives it. Simple cells were divided into four groups based on their RF size (7-9 pixels, 11-15 pixels, 17-21 pixels, 23-29 pixels) and each complex cell pooled responses of neighboring simple cells in one of these groups. The spatial range of pooling varied across the four groups (4×4, 6×6, 9×9, and 12×12 pixels for the four groups, respectively). This yielded 4 (orientation selectivities) × 12 (RF sizes) = 48 simple-cell maps and 4 (orientation selectivities) × 4 (sets of simple-cell RF sizes pooled) = 16 complex-cell maps of 175×175 pixels. All maps of simple and complex cell outputs were vectorized and concatenated to obtain a representational vector for each stimulus image.

HMAX-C2 model based on natural image fragments

This model representation developed by Serre et al. (2005) builds on the complex-cell outputs of the V1 model described above (implemented by the same group). The C2 features used in the analysis may be comparable to those found in primate V4 and posterior IT. The model has four sequential stages: S1-C1-S2-C2. The first two stages correspond to the simple and complex cells described above, respectively. Stages S2 and C2 use the same pooling mechanisms as stages S1 and C1, respectively. Each unit in stage S2 locally pools information from the C1 stage by a linear filter and behaves as a radial basis function, responding most strongly to a particular prototype input pattern. The prototypes correspond to random fragments extracted from a set of natural images (stimuli

independent of those used in the present study). S2 outputs are locally pooled by C2 units utilizing the MAX operation for a degree of position and scale tolerance. A detailed description of the model (including the parameter settings and map sizes we used here) can be found in Serre et al. (2005). The model, including the natural image fragments, was downloaded from the author's website in January 2007 (for the current version, see <http://cbcl.mit.edu/software-datasets/standardmodel/index.html>).

Radon transform

As an example of a model inspired by image processing, we included the Radon transform, which has been proposed as a functional account of the representation of visual stimuli in the lateral occipital complex (Wade et al., Human Brain Mapping 2006). The Radon transform of a two-dimensional image is a matrix, each column of which corresponds to a set of integrals of the image intensities along parallel lines of a given angle. We used the Matlab function `radon` to compute the Radon transform for each luminance image.

Chapter 4

Single-image activation of category-selective regions in human inferior temporal cortex

Human inferior temporal (hIT) cortex contains regions that respond preferentially to particular object categories, including faces and places. These regions are defined by their greater category-average activation to the preferred relative to the non-preferred category. Since these regions are defined by activation averaged over category exemplars, it is unclear to what extent category selectivity holds for individual objects. Here, we use a single-image approach to investigate (1) whether category selectivity holds in general or is violated by particular single images, and (2) whether there are within-category activation differences. We measured single-image activation of the fusiform face area (FFA) and the parahippocampal place area (PPA) during perception of 96 object images from a wide range of categories, including faces and places, using functional magnetic resonance imaging (fMRI). We address our questions using a signal-detection approach. We found no evidence of any images outside the preferred category eliciting a stronger response than any images inside the preferred category for either FFA or PPA. Regional-average activation might thus perfectly reflect category membership of individual exemplars. However, within each category, individual images elicited different levels of activation, suggesting a graded rather than a pure step-function response profile. We relate our regional-average findings to single-image pattern-information studies in humans and to single-image results reported in the monkey literature.

Mur M, Ruff D, Bodurka J, De Weerd P, Bandettini P, Kriegeskorte N. Single-image activation of category-selective regions in human inferior temporal cortex. In revision, *J Neurosci*.

4.1 Introduction

Human inferior temporal (hIT) cortex has been shown to contain category-selective regions that respond more strongly to object images of one specific category than to images belonging to other categories. The two most well-known category-selective regions are the fusiform face area (FFA), which responds selectively to faces (Kanwisher et al., 1997; Puce et al., 1995), and the parahippocampal place area (PPA), which responds selectively to places (Epstein and Kanwisher, 1998). The category-selectivity of these regions has been shown for a wide range of stimuli (e.g. Downing et al., 2006; Kanwisher et al., 1999). However, previous studies grouped stimuli into predefined natural categories, and therefore only assessed category-average activation. To investigate responses to individual stimuli, stimuli would need to be ungrouped so that each stimulus is treated as a separate condition (i.e. single-image design). Despite common use of single-image designs in monkey electrophysiology (e.g. Vogels, 1999; Földiák et al., 2004; Tsao et al., 2006; Kiani et al., 2007) and occasional use of item-specific designs in human studies in other domains (e.g. Bedny et al. 2007), single-image responses in human visual cortex have not been investigated until recently.

We measured single-image fMRI activity elicited by 96 stimuli from a wide range of object categories without assuming any predefined grouping in design or analysis. In Kriegeskorte et al. (2008), we analyzed these data for multi-voxel pattern effects. We found that single-image activity patterns in hIT (including the lateral occipital complex (Malach et al., 1995), FFA and PPA) reflect natural categories: when activity patterns are grouped by their similarity, patterns elicited by the same category fall into the same cluster. Here, we analyze these data for activation effects. In contrast to Kriegeskorte et al. (2008), we focus on category-selective regions (rather than hIT as a whole) and on regional-average activation (rather than pattern information), thus relating the single-image approach to the earlier literature on category selectivity in human visual cortex. This enables us to investigate (1) whether category selectivity holds in general or is violated by particular single images (i.e. are there images from the non-preferred category that elicit greater activation than any of the images from the preferred category), and (2) whether there are within-category activation differences (i.e. do some preferred-category images activate a region more strongly than others).

We measure single-image category-selectivity using a signal-detection approach and address our questions by testing for replicability of single-image activation differences. We focus our analysis on single-image activation of category-selective regions FFA and PPA. Based on neurophysiological studies in monkeys (Tsao et al., 2006; Földiák et al., 2004), we expect single-image activation of FFA

to show strong face selectivity. Nevertheless, this does not exclude the possibility that some non-faces might drive FFA more strongly than some faces (see Kiani et al., 2007). Previous studies in monkeys provide little information on place-selectivity at the single-image level, but do suggest that we might find within-category activation differences in FFA and PPA for images of their preferred category (Bell et al., 2009).

4.2 Materials and methods

4.2.1 Subjects

Four healthy human volunteers participated in the fMRI experiment (mean age = 35 years; two females). Subjects were right-handed and had normal or corrected-to-normal vision. Before scanning, the subjects received information about the procedure of the experiment and gave their written informed consent for participating. The experiment was conducted in accordance with the Institutional Review Board of the National Institutes of Mental Health, Bethesda, MD.

4.2.2 Stimuli

We used 96 colored photos of isolated objects spanning a wide range of categories, including faces and places (subset of stimuli from Kiani et al., 2007). The 96 object photos are displayed in Figure 6.1.

4.2.3 Experimental design and task

Ranking experiment

Stimuli were presented using a rapid event-related design (stimulus duration: 300 ms, interstimulus interval: 3700 ms) while subjects performed a fixation-cross-color detection task. Stimuli were displayed at fixation on a uniform gray background at a width of 2.9° visual angle. Each of the 96 object images was presented once per run in random order. Each run included 40 randomly interleaved baseline trials where no stimulus was shown. Subjects participated in two sessions of 6 nine-minute runs each. The sessions were acquired on separate days.

Localizer experiment

Subjects participated in an independent block-design experiment that was designed to localize regions of interest (ROIs) for the ranking analysis. The block-localizer experiment used the same fMRI sequence as the ranking experiment

and a separate set of stimuli. Stimuli were grayscale photos of faces, objects, and places, displayed at a width of 5.7° of visual angle, centered with respect to a fixation cross. The photos were presented in 30-s category blocks (stimulus duration: 700 ms, interstimulus interval: 300 ms), intermixed with 20-s fixation blocks, for a total run time of approximately eight minutes. Subjects performed a one-back repetition-detection task on the images.

4.2.4 Functional magnetic resonance imaging (fMRI) measurements

Blood-oxygen-level-dependent (BOLD) fMRI measurements were performed at high spatial resolution (voxel volume: $1.95 \times 1.95 \times 2 \text{ mm}^3$), using a 3 Tesla General Electric HDx MRI scanner, and a custom-made 16-channel head coil (Nova Medical Inc.). Single-shot gradient-recalled Echo Planar Imaging with Sensitivity Encoding (matrix size: 128×96 , TR: 2s, TE: 30ms, 272 volumes per run) was used to acquire 25 axial slices that covered inferior temporal and early visual cortex bilaterally.

4.2.5 Data analysis

fMRI data preprocessing

fMRI data preprocessing was performed using BrainVoyager QX 1.8 (Brain Innovation). The first three data volumes of each run were discarded to allow the fMRI signal to reach a steady state. All functional runs were subjected to slice-scan-time correction and 3D motion correction. In addition, the localizer runs were high-pass filtered in the temporal domain with a filter of two cycles per run (corresponding to a cut-off frequency of 0.004 Hz) and spatially smoothed by convolution of a Gaussian kernel of 4 mm full width at half maximum. Data were converted to percent signal change. Analyses were performed in native subject space (i.e. no Talairach transformation).

Definition of regions of interest

All ROIs were defined based on the independent block-localizer experiment and restricted to a cortex mask manually drawn on each subject's fMRI slices. Both the left and right fusiform face area (FFA) were defined as a contiguous cluster consisting of the most face-selective voxels in that hemisphere. These clusters were defined at five sizes, ranging from 10 to 300 voxels. Face-selectivity was assessed by the contrast faces minus places and objects. The parahippocampal place area (PPA) was defined in an identical way but then using the contrast places minus faces and objects. For control analyses, we defined the following two regions. Human inferior temporal cortex (hIT) was defined by selecting the most visually responsive voxels within the inferior temporal portion of the bi-

lateral cortex mask. hIT was defined at five sizes as well, ranging from 20 to 600 voxels. Visual responsiveness was assessed by the contrast visual stimulation (face, object, place) minus baseline. To ensure that hIT results would not be driven by face-selective or place-selective voxels, FFA and PPA were excluded from selection. For this purpose, FFA and PPA were defined at 150 and 200 voxels in each hemisphere, respectively. To define early visual cortex (EVC), we selected the most visually responsive voxels, as for hIT, but within a manually defined anatomical region around the calcarine sulcus within the bilateral cortex mask. EVC was defined at the same five sizes as hIT.

Estimation of single-image activation

Single-image BOLD fMRI activation was estimated by univariate linear modeling. We concatenated the runs within a session along the temporal dimension. For each ROI, data were extracted and averaged across space. We then performed a single univariate linear model fit for each ROI to obtain a response-amplitude estimate for each of the 96 stimuli. The model included a hemodynamic-response predictor for each of the 96 stimuli. Since each stimulus occurred once in each run, each of the 96 predictors had one hemodynamic response per run and extended across all within-session runs. The predictor time courses were computed using a linear model of the hemodynamic response (Boynton et al., 1996) and assuming an instant-onset rectangular neuronal response during each condition of visual stimulation. For each run, the design matrix included these stimulus-response predictors along with six head-motion-parameter time courses, a linear-trend predictor, a six-predictor Fourier basis for nonlinear trends (sines and cosines of up to three cycles per run) and a confound-mean predictor. The resulting response-amplitude (beta) estimates, one for each of the 96 stimuli, were used for the ranking analyses.

Assessment of category selectivity for single images

In order to investigate the category-selectivity of single-image responses, the 96 object images were ranked by their beta estimates, i.e. by the activation they elicited in each ROI. To quantify how well activation discriminated faces from non-faces and places from non-places, we computed receiver operating characteristic (ROC) curves and associated areas under the curves (AUCs) for each ROI. The AUC represents the probability that a randomly chosen face (or place) is ranked before a randomly chosen non-face (or non-place) based on the activation elicited by these two images. Taking faces as an example, an AUC of .5 would indicate chance accuracy at discriminating faces from non-faces. An AUC of 1 would indicate perfect discrimination, i.e. each face was ranked before each non-face. To determine whether discrimination performance was significantly above chance, we used a one-sided label-randomization test on the AUC (10,000 randomizations). p-values were corrected for multiple comparisons using Bon-

ferroni correction based on the number of ROI sizes tested per region. For group analysis, we averaged the beta estimates across sessions and subjects and performed the AUC test on these subject-average beta estimates.

We expect category-selective regions to discriminate preferred from non-preferred images significantly above chance. However, taking FFA as an example, even if each face elicited greater regional-average activation than any non-face, we still expect the AUC to be smaller than 1 because of the noise in the data. We therefore need a separate test for violation of category-consistent ranking. If there were indeed non-faces that consistently activate FFA more strongly than faces, these inverted pairs (i.e. non-preferred image ranked before preferred image) should replicate. We use the proportion of replicated inverted pairs (PRIP) from one session to the next as our test statistic. We computed the PRIP for each subject by dividing the number of inverted pairs that replicated from session one to session two by the total number of inverted pairs in session one. A PRIP of 1 would indicate that all inverted pairs replicated from one session to the next (perfect replicability). A PRIP of .5 would indicate that half of the inverted pairs replicated from one session to the next (chance level). A PRIP of 0 would indicate that none of the inverted pairs replicated from one session to the next (zero replicability). In other words, this would indicate that all inverted pairs reverted to “category-preference pairs” (i.e. preferred image ranked before non-preferred image). To determine whether the replicability of inverted pairs differed significantly from chance, we used a two-sided label-randomization test on the PRIP (10,000 randomizations). p-values were corrected for multiple comparisons using Bonferroni correction based on the number of ROI sizes tested per region. For group analysis, we averaged the subject-specific PRIPs and label-randomization null distributions and then performed the PRIP test using these subject-average values. Note that this allows the particular image pairs inverted to differ across subjects.

Investigation of within-category activation profiles

Would images of its preferred category all activate a category-selective region equally strongly or would some of them activate the region more strongly than others? To address this question, we tested whether within-category ranking order replicated across sessions. If all images of one specific category would activate a region equally strongly (i.e. flat within-category activation profile), we would expect their ranking order to be random and therefore not replicable across sessions. If, however, some images of a specific category would consistently activate the region more strongly than other images of the same category (i.e. graded within-category activation profile), we would expect the ranking order of these images to replicate across sessions. We assessed replicability of within-category ranking by computing Spearman’s rank correlation coefficient

(Spearman's r) between beta estimates for one specific category of images in session one, and beta estimates for the same subset of images in session two. We performed a one-sided test to determine whether Spearman's r was significantly larger than zero, i.e. whether replicability of within-category ranking was significantly higher than expected by chance. p -values were corrected for multiple comparisons using Bonferroni correction based on the number of ROI sizes tested per region. For group analysis, we first combined single-subject data for each session separately and then performed the across-session replicability test on the combined data. We used two approaches for combining the single-subject data. The first approach consisted in concatenating the session-specific within-category beta estimates across subjects, the second in averaging them across subjects. The concatenation approach is sensitive to replicable within-category ranking across sessions even if ranking order would differ across subjects. The averaging approach is sensitive to replicable within-category ranking that is consistent across subjects.

4.3 Results

4.3.1 Strong category selectivity for single images

To visualize the degree of category selectivity for single images, we ranked the 96 object images by the activation they elicited in each ROI (Figure 4.1 and 4.2). Visual inspection of the ranking results indicates that category-selective regions FFA and PPA show a clear preference for images of their preferred category: activation of PPA ranks (almost) all places before all non-places and activation of FFA ranks most faces before most non-faces (Figure 4.1). Control regions hIT and early visual cortex (EVC) do not show a clear category preference at first inspection (Figure 4.2).

To quantify these results, we computed ROCs and AUCs for each ROI. Consistent with visual inspection, single-image activation of FFA showed very good discrimination of faces from non-faces, with right FFA (AUC = .94) showing better discrimination performance than left FFA (AUC = .82). Single-image activation of PPA showed (near) perfect discrimination of places from non-places (AUC = 1). A condition-label randomization test on the AUCs indicated that discrimination performance of FFA and PPA for their preferred category was significantly above chance ($p < .001$ for each region, Figure 4.1). Discrimination performance of hIT was not significantly different from chance for either category. EVC showed above-chance accuracy for places (AUC = .74, $p < .05$, Figure 4.2) but not for faces. The above-chance accuracy for places suggests that place images differ to some extent from non-place images in terms of their lower-level visual properties (see also Rajimehr et al., 2011).

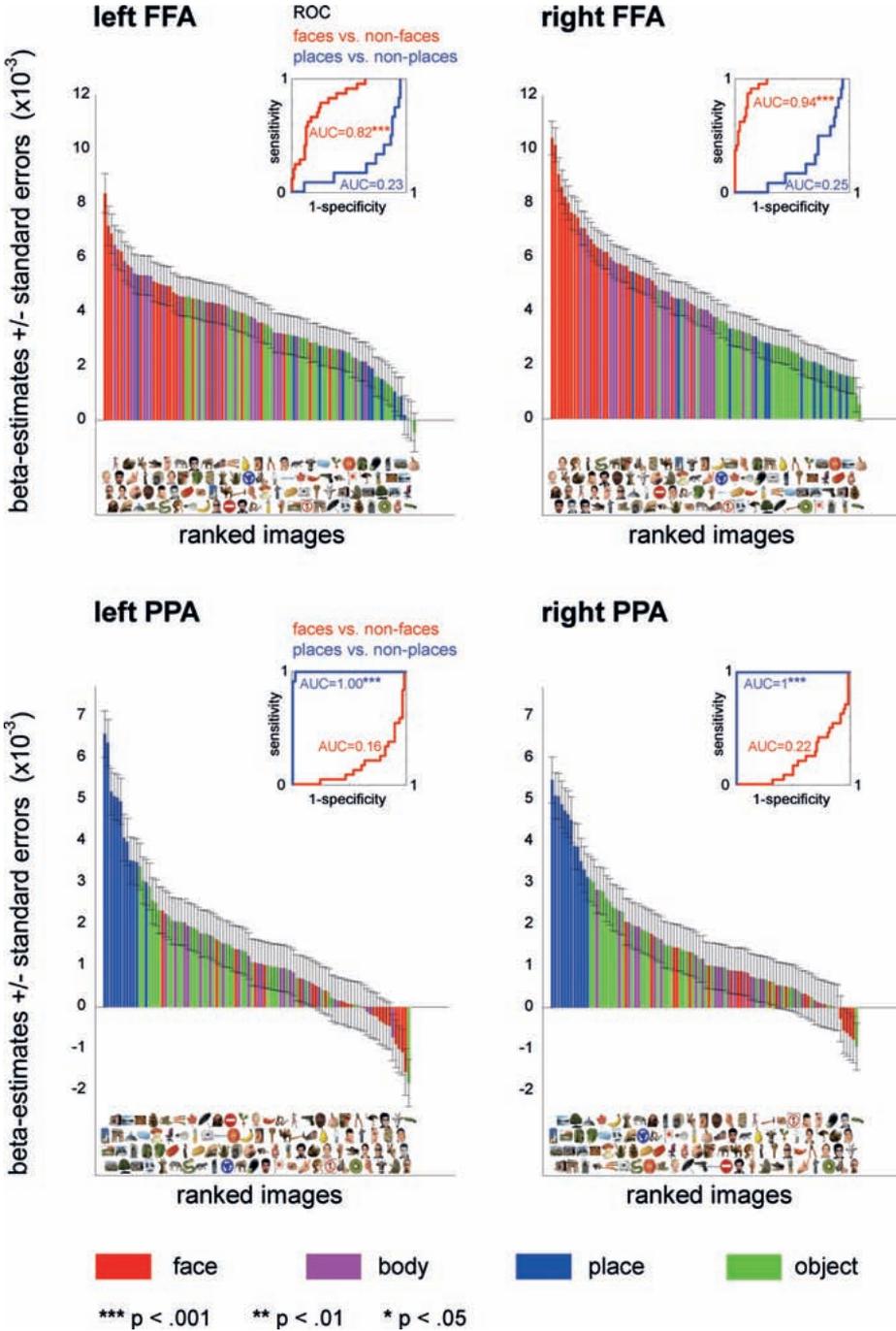


Figure 4.1 Single-image activation of category-selective regions discriminates preferred from non-preferred images with high accuracy. The graphs show the 96 object images ranked by the activation they elicited in each ROI. Each bar represents activation to one of the 96 object images in proportion signal change, averaged across four subjects. Each image is placed exactly below the bar that reflects its activation, so that the images are ordered from left to right (i.e. only the x-coordinate is meaningful). The left-most image activated the region most strongly, the right-most image activated the region most weakly. The bars are color-coded for category to give an overall impression of category-selectivity without having to inspect all single images. The insets show ROC curves and associated AUCs, indicating accuracy for discriminating faces from non-faces (red) and places from non-places (blue). We used a one-sided label- randomization test to determine whether accuracy was significantly above chance ($H_0: AUC = .5$). Since we tested discrimination accuracy at five different ROI sizes for each region, we corrected p-values for multiple (five) comparisons using Bonferroni correction. Error bars represent standard error of the beta estimates, averaged across four subjects. FFA and PPA were each defined at 128 voxels in each hemisphere, based on an independent block-localizer experiment.

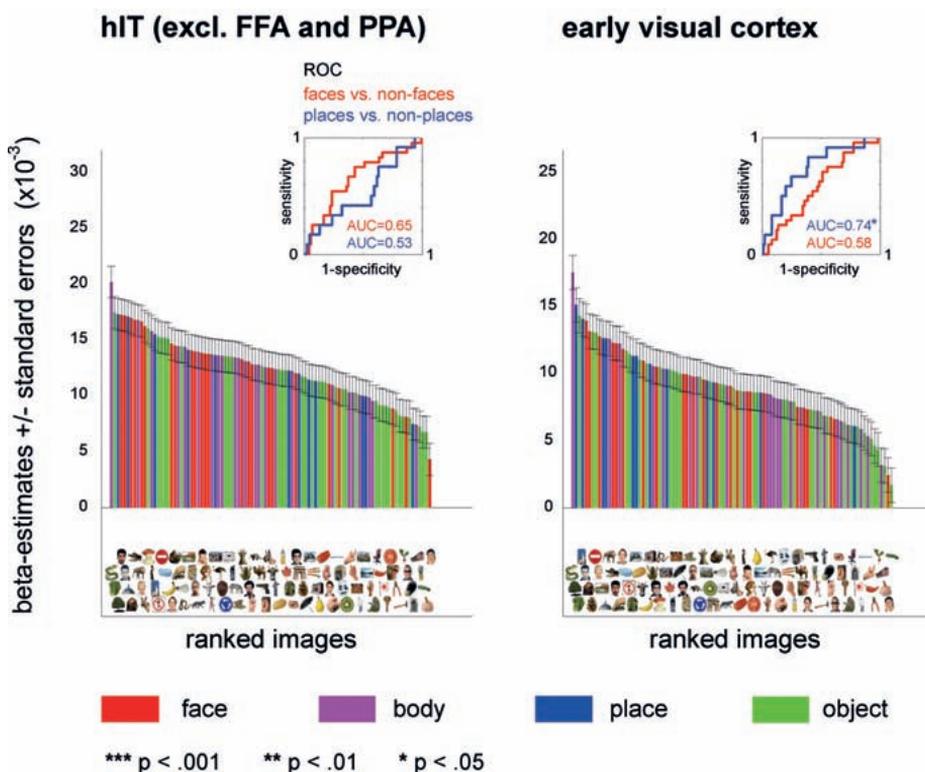


Figure 4.2 Single-image activation of hIT and early visual cortex does not show a strong category preference. As in Figure 4.1, images are ranked by the activation they elicited in each ROI, and insets show discrimination accuracy. Statistical tests as described in Figure 4.1. hIT and EVC were defined bilaterally at 256 voxels each, based on visual responsiveness during an independent block-localizer experiment.

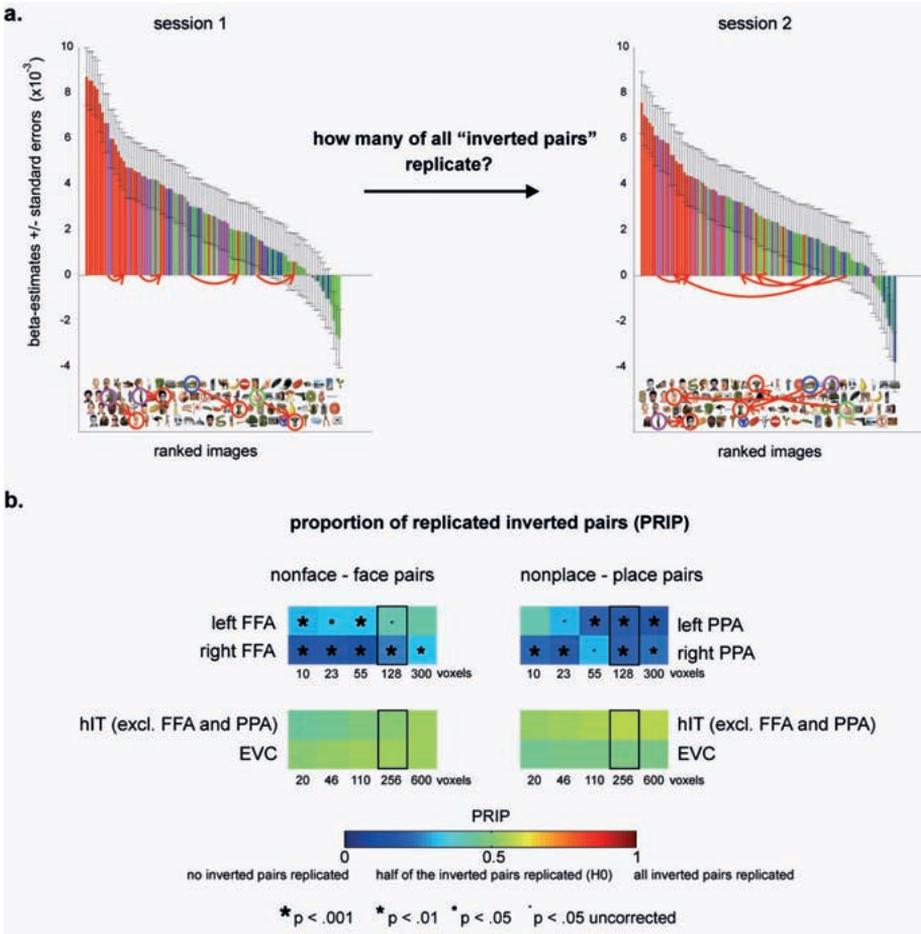


Figure 4.3 No evidence for violations of category-consistent activation ranking. If there were some non-preferred images that would consistently activate category-selective regions more strongly than some preferred images, we would expect the high ranking position of these non-preferred images to replicate across sessions. We investigated this by computing the proportion of replicated inverted pairs (PRIP). **(a)** Computation of the PRIP. The graphs display single-image activation of right FFA, defined at 128 voxels in one specific subject, shown separately for each session. Colored circles connected by red arrows highlight four inverted pairs (i.e. non-preferred image ranked before preferred image) in session one. Only one of those four inverted pairs replicates in session two, the other three switch to category-preference pairs (i.e. preferred image ranked before non-preferred image). If these four example pairs were the only inverted pairs in session one, the PRIP would be .25. Color coding is the same as in Figure 4.1. **(b)** Group analysis of the PRIP for category-selective regions FFA and PPA and control regions hIT and EVC. The PRIP was averaged across subjects, allowing for different particular image pairs to be inverted in each subject. Since inverted pairs are defined based on the notion of category preference, the analysis was based on nonface-face pairs for FFA and nonplace-place pairs for PPA. hIT and EVC do not have a strong category preference and were tested for both types of pairs, serving as a control for category-selective regions. We used a two-sided label-randomization test to determine whether the PRIP differed significantly from 0.5, the level we expect under the null hypothesis that there are no inverted pairs (i.e. all apparent inversions are caused by noise in the data). A PRIP significantly above

0.5 indicates replicable inversions of activation rank for particular images. A PRIP significantly below 0.5 indicates that inverted pairs tend to switch to category-preference pairs from one session to the next. p-values were corrected for multiple comparisons as described in Figure 4.1, unless otherwise stated. Black boxes highlight the ROI sizes that results were displayed at in Figure 4.1 (FFA and PPA) and 4.2 (hIT and EVC).

Figure 4.1 indicates that, despite the clear preference of FFA and PPA for images of their preferred category, some non-preferred images appear before some preferred images in this descriptive analysis. This can be seen most clearly for FFA: some non-face images activated FFA more strongly than some face images. To test whether high-ranked non-preferred images consistently activated the category-selective regions more strongly than lower-ranked preferred images, we computed the proportion of replicated inverted pairs (PRIP) (Figure 4.3 and *Materials and Methods*). The PRIP gives an indication of the rate at which inverted pairs (i.e. non-preferred image ranked before preferred image) replicate from one session to the next.

Results show that the PRIP for both FFA and PPA was significantly below 0.5 for nearly all ROI sizes (Figure 4.3b), indicating that inverted pairs replicated at a rate that is less than expected by chance. This indicates that inverted pairs had a significant tendency to switch back to category-preference pairs from one session to the next. In other words, we found no evidence for violations of category-consistent activation ranking. This is consistent with the idea that FFA will prefer any face over any non-face (in terms of its regional-average activation), and that PPA will similarly prefer any place over any non-place. Control regions hIT and EVC do not have a strong category preference. However, for completeness, we performed the same analysis for these regions and found that their PRIP values were not significantly different from chance (Figure 4.3b).

4.3.2 Category selectivity across ROI sizes

We performed our analyses of single-image ranking and pair inversion for five different ROI sizes, ranging from 10 to 300 voxels for unilateral FFA and PPA and from 20 to 600 voxels for bilateral hIT and EVC. Figure 4.1 displays category-discrimination accuracy (AUC) based on the ranking results for one of the larger ROI sizes (128 voxels), chosen to match most closely to previously reported volume of right FFA (Kanwisher et al., 1997). Discrimination performance for the other ROI sizes can be found in Table 4.1. This table shows that single-image activation of both left and right FFA discriminated faces from non-faces with high accuracy at all ROI sizes. Accuracy was highest for the smallest ROI size (10 most face-selective voxels) and decreased with increasing ROI size. The effect of ROI size was more pronounced for left than right FFA, resulting in a considerable difference in accuracy between left and right FFA for the two larg-

est ROI sizes. The bottom panel of Table 4.1 shows that single-image activation of both left and right PPA discriminated places from non-places with (near) perfect accuracy at all ROI sizes. Accuracy of right PPA was not influenced by ROI size; accuracy of left PPA was a bit lower for the two smallest ROI sizes. It should be noted that hIT showed above-chance accuracy for discriminating faces from non-faces at small ROI sizes ($AUC = .72$, $p < .01$), which can be attributed to the inclusion of some weakly face-selective voxels in a subset of the subjects, and that EVC showed above-chance accuracy for discriminating places from non-places at all ROI sizes ($.71 < AUC < .74$, $p < .05$). With respect to pair inversions, Figure 4.3b indicates that right FFA and PPA showed PRIP effects for their preferred category at all ROI sizes. Left FFA and PPA showed PRIP effects at most ROI sizes, with stronger effects at smaller ROI sizes for FFA and larger ROI sizes for PPA.

In sum, these findings indicate that the strong single-image preference for faces over non-faces in FFA and places over non-places in PPA can be found at all ROI sizes. Nevertheless, ROI size does affect measured category selectivity, especially for (left) FFA. Strongest category selectivity is found at smaller ROI sizes for FFA and at larger ROI sizes for left PPA. The clear decrease in discrimination accuracy for left FFA with increasing ROI size might simply reflect the previously reported finding that left FFA contains fewer strongly face-selective voxels than right FFA (Kanwisher et al., 1997).

Table 4.1 Single-image category-discrimination accuracy (AUC) for faces and places.

faces vs. non-faces

ROI size (voxels)	10 (20)	23 (46)	55 (110)	128 (256)	300 (600)
left FFA	0.96***	0.96***	0.94***	0.82***	0.75***
right FFA	0.98***	0.99***	0.99***	0.94***	0.91***
left PPA	0.25	0.22	0.18	0.16	0.16
right PPA	0.27	0.28	0.22	0.22	0.21
hIT	0.72**	0.72**	0.72**	0.65	0.58
EVC	0.62	0.63	0.62	0.58	0.56

places vs. non-places

ROI size (voxels)	10 (20)	23 (46)	55 (110)	128 (256)	300 (600)
left FFA	0.27	0.20	0.18	0.23	0.32
right FFA	0.22	0.24	0.25	0.25	0.22
left PPA	0.97***	0.99***	1.00***	1.00***	1.00***
right PPA	1.00***	1***	1.00***	1***	1***
hIT	0.56	0.5	0.5	0.53	0.54
EVC	0.71*	0.71*	0.72*	0.74*	0.73*

*** $p < .001$ ** $p < .01$ * $p < .05$

p-values were computed using a right-sided label-randomization test and were corrected for multiple (five) comparisons using Bonferroni correction. Voxel numbers in between brackets describe ROI sizes for bilateral hIT and EVC.

4.3.3 Within-category activation differences

Figure 4.1 suggests that, within the preferred category, some images activated category-selective regions more strongly than others. We tested this hypothesis by examining the replicability of within-category ranking (Figure 4.4), which we estimated by rank correlating within-category beta estimates across sessions (Figure 4.4a). If the ranking order for a category-subset of images (e.g. faces) would be replicable across sessions, this would indicate that some of these images indeed consistently activated the region more strongly than others. Group results are shown in Figure 4.4b. The top panel shows replicability of within-face ranking; the bottom panel shows replicability of within-place ranking. The left column shows results for data concatenated across subjects; the right column for data averaged across subjects. The concatenation approach is sensitive to replicable ranking irrespective of differences in particular ranking order among subjects, while the averaging approach is sensitive to replicable ranking that is consistent among subjects (see *Materials and Methods*).

The top panel of Figure 4.4b shows that both left and right FFA showed replicable ranking for faces, especially at small ROI sizes. This indicates that some faces consistently activated FFA more strongly than others. The significant results for the averaging approach further suggest that within-face activation profiles were similar across the four subjects. This conclusion was supported by visual inspection of single-subject within-face ranking orders and by inter-subject correlation analyses. Effects were somewhat stronger in left than right FFA. Control regions hIT and EVC showed replicable within-face ranking as well, but only for the concatenation approach. This suggests that ranking order was not consistent across subjects. In addition, effects in EVC were small and were only found for small ROI sizes.

The bottom panel of Figure 4.4b shows that right, but not left, PPA shows replicable ranking for places at most ROI sizes. This indicates that some places consistently activated right PPA more strongly than others. The significant results for the averaging approach suggest that within-place activation profiles were similar across the four subjects. This conclusion was supported by visual inspection of single-subject within-place ranking orders and by inter-subject correlation analyses. Right FFA and control regions hIT and EVC showed replicable within-place ranking as well. Right FFA showed effects at smaller ROI sizes; hIT at larger ROI sizes. Effects in right FFA disappeared when using the averaging approach while effects in the other two regions remained present (hIT) or even increased in strength (EVC). These findings suggest that within-place activation profiles were similar across the four subjects for hIT and EVC, but not for right FFA. This conclusion was supported by visual inspection of single-subject within-place ranking orders and by inter-subject correlation analyses.

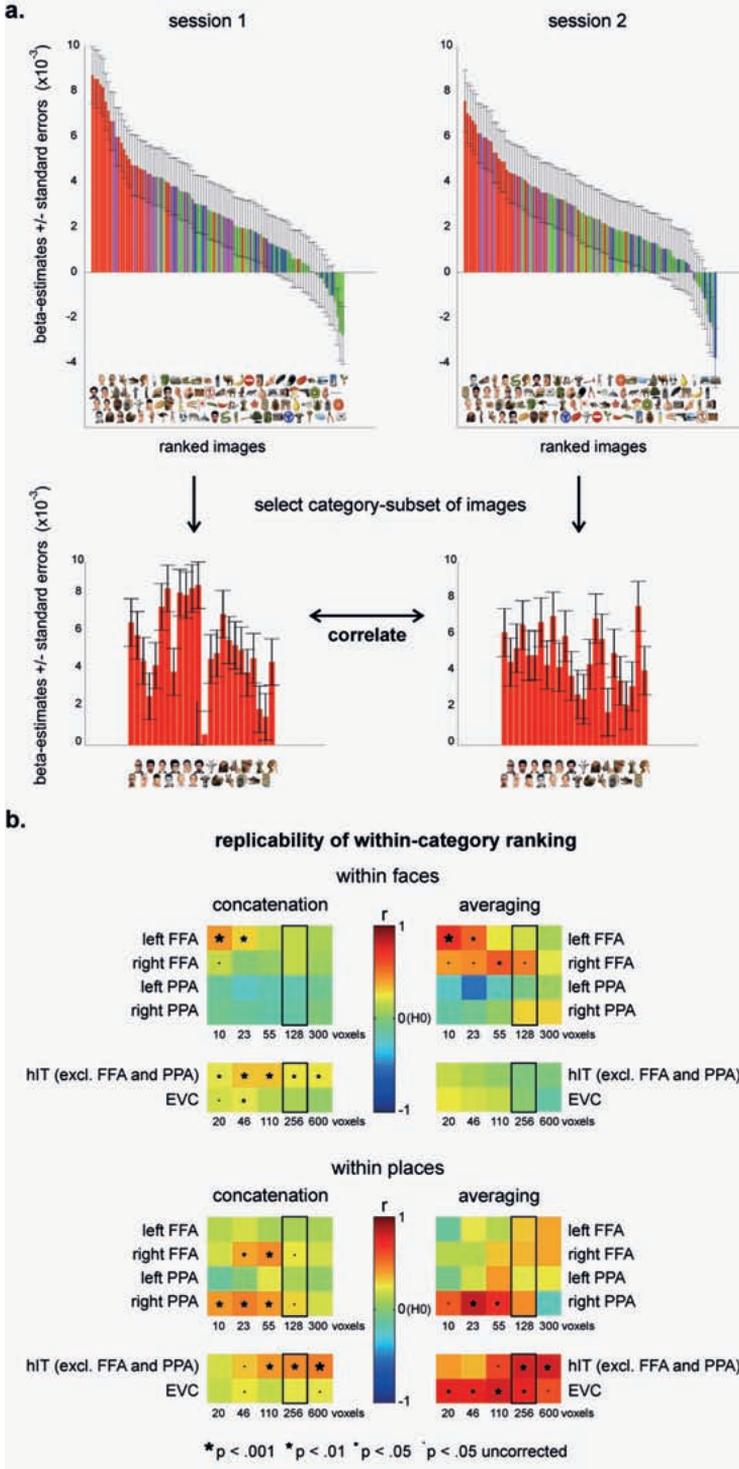


Figure 4.4 Category-selective regions show graded activation profiles for images of their preferred category. (a) If some images consistently activated a region more strongly than other images of the same category (i.e. graded within-category activation profile), the within-category ranking order should replicate across sessions. We computed the replicability of within-category ranking by selecting the same category-subset of images in both sessions, and correlating their beta estimates using Spearman's r . This procedure is illustrated for within-face ranking in right FFA defined at 128 voxels in one specific subject. Color coding is the same as in Figure 4.1. (b) Group analysis of replicability of within-category ranking for category-selective regions FFA and PPA and for control regions hIT and EVC. Analysis was performed for the image subsets of faces (top) and places (bottom), either using the concatenation approach (left) or the averaging approach (right) for combining single-subject data (see *Materials and Methods*). We performed a standard one-sided test on Spearman's r to determine whether replicability of within-category ranking was significantly higher than expected by chance ($H_0: r = 0$). p -values were corrected for multiple comparisons as described in Figure 4.1, unless otherwise stated. Black boxes highlight the ROI sizes that results were displayed at in Figure 4.1 (FFA and PPA) and 4.2 (hIT and EVC).

These findings confirm that category-selective regions are activated more strongly by some images of their preferred category than by others, i.e. they show a graded activation profile for images of their preferred category. These effects are not confined to category-selective regions: hIT and EVC show graded within-category activation profiles as well, especially for places.

4.4 Discussion

4.4.1 Single-image activation of category-selective regions might perfectly reflect category membership

We measured single-image activation of human category-selective regions to 96 object images from a wide range of categories, and investigated whether category selectivity holds in general or is violated by particular single images. We found that single-image activation of category-selective regions FFA and PPA discriminates preferred from non-preferred stimuli with near perfect accuracy at almost all tested ROI sizes. Furthermore, we did not find evidence for violations of category-consistent ranking by particular single images. Together, these findings suggest that activation of category-selective regions might perfectly reflect category membership of object exemplars.

This conclusion is consistent with several single-image studies in monkeys that showed strong face-selectivity in the macaque middle and anterior superior temporal sulcus (STS) (Tsao et al., 2006; Földiák et al., 2004). These studies reported cells that responded almost exclusively to faces. It should be noted that many of the recorded cells in the middle macaque face patch, a suggested homologue of FFA located in the STS (Tsao et al., 2003; 2006), also responded significantly to several non-face images (Tsao et al., 2006). These non-face images

shared lower-level visual properties with face images (i.e. round shape). However, at the population level (i.e. when responses were averaged across the population of visually-responsive cells in the middle face patch), the responses elicited by these non-face images were clearly weaker than those elicited by any of the face images (Tsao et al., 2006). Kiani et al. (2007) reported a similar finding: they measured responses of face-selective cells in macaque inferior temporal (IT) cortex, and reported imperfect face selectivity at the single-cell level but close-to-perfect face selectivity when responses were averaged over a small population of face-selective IT cells. These findings are consistent with the idea that category membership of natural objects is encoded at the population level (Vogels, 1999; Kiani et al., 2007).

In sum, our results indicate that category selectivity of FFA and PPA, conventionally defined using category-average activation, holds for single images. Our results strengthen the suggestion of a homology between the human FFA and the macaque middle face patch and predict strong single-image selectivity for places in a recently reported place-selective region in the macaque (Bell et al., 2009).

4.4.2 Exemplar activation differences: gradation, not simple step function

Previous category-average studies left open whether category-selective regions simply act as a binary classifier or whether they show graded responses to individual category members (i.e. exemplars). In the first scenario, the activation of the region would follow a step function, i.e. there would be only two possible levels of activation: high for exemplars of the preferred category and low for exemplars of non-preferred categories. In the second scenario, the activation of the region would show gradation, i.e. some category members would activate the region more strongly than others. Our results support the second scenario: FFA and PPA showed a graded activation profile for exemplars of their preferred category. This finding suggests that we need more than a simple step function to explain single-image activation of category-selective regions.

Our findings are consistent with a recent monkey fMRI study that reported activation differences in face- and place-selective regions in IT between visually dissimilar exemplars of the preferred category (Bell et al., 2009). They are also in line with an earlier monkey electrophysiology study that reported a population of tree-selective cells in IT whose mean response differed across tree exemplars (Vogels, 1999). Other reports on gradation of activation focused on differences between non-preferred categories (Downing et al., 2006; Kiani et al., 2007), and did not investigate differences between exemplars.

There are several possible interpretations of the within-category activation differences reported here. First, it could be that activation differences between exemplars reflect differences in low-level visual features. Consistent with this idea, we found within-category activation differences in early visual cortex, especially for places. Our image set consisted of photographs of natural objects which were not controlled for low-level visual properties in order to preserve ecological validity. Low-level effects could be reduced by using a more controlled stimulus set. Second, activation differences between exemplars might reflect differences between the underlying distributed patterns of activity that are thought to represent them (Young and Yamane, 1992; Edelman et al., 1998; Tsao et al., 2006; Kiani et al., 2007; Eger et al., 2008; Kriegeskorte et al., 2008a). Exemplar information carried by distributed activity patterns might get lost by pooling (Kriegeskorte et al., 2006; Kriegeskorte et al., 2007; Eger et al., 2008), but could also to some extent be reflected in regional-average activation. The underlying distributed activity patterns are not confined to category-selective regions (Haxby et al., 2001; Kriegeskorte et al., 2008a), which could explain the exemplar activation effects in hIT. Third, our within-category activation differences could be interpreted as attentional effects. Attention enhances responses to stimuli in object-selective cortex (Wojciulik et al., 1998; O'Craven et al., 1999) and early visual regions (Liu et al., 2005). Stimuli might differ in the extent to which they trigger attention. For example, high-valence stimuli (e.g. angry face) might trigger more attention than low-valence stimuli (e.g. neutral face), resulting in activation differences among stimuli (Breiter et al., 1996; Lane et al., 1999; Palermo and Rhodes, 2007).

In any case, our findings suggest that a single-image approach explains additional variance as compared to a category-average approach.

4.4.3 Single-image designs for studying regional-average activation and pattern information

The classical fMRI category-block-design studies (e.g. Kanwisher et al., 1997; Epstein and Kanwisher, 1998) averaged across stimuli within predefined categories and across response channels (i.e. voxels within contiguous regions). Haxby et al. (2001) studied pattern information, but still averaged patterns within predefined categories. Kriegeskorte et al. (2008) studied pattern information of single-image response patterns, enabling data-driven discovery of category structure (see also Edelman et al., 1998). The present study constitutes a missing link in the sense that it considers single-image responses, but in terms of regional-average activation levels.

Our results, together with previous single-image pattern-information reports (Edelman et al., 1998; Kriegeskorte et al., 2007, 2008; Aguirre, 2007; Eger et al.,

2008; Haushofer et al., 2008), demonstrate the feasibility of single-image designs for fMRI. Single-image designs reduce experimenter bias because they do not assume any grouping of the stimuli in design or analysis. They enable exemplar-based analyses and empirical discovery of (new) categories represented in high-level visual cortex and can thus be used to address key questions that cannot be addressed with classical fMRI designs.

4.4.4 In what sense is the representation categorical? And in what sense is it not categorical?

The current study combines a single-image approach with the analysis of regional-average activation. Our results extend previous findings on categorical object representations in IT by showing (1) that regional-average activation of category-selective regions might perfectly reflect category membership of individual objects and (2) that we need more than a simple step function to explain single-image activation of category-selective regions. In what sense do these results, combined with previous findings in both humans and monkeys, support the idea of a categorical representation in high-level visual cortex?

The object representation in IT might perfectly reflect category membership of individual objects but does not seem to be categorical in the sense of a pure step-wise response function. This has been demonstrated both at the level of single-cell responses in the monkey (Vogels, 1999; Kiani et al., 2007; Tsao et al., 2006) and at the level of regional-average activation in the human (current study). These studies suggest gradation of responses instead of a pure step-wise response function and raise questions on the strongest claims of univariate selectivity in IT. Lateral prefrontal cortex, which receives input from IT, seems a more likely candidate for the neuronal representation of step-wise category boundaries (Freedman et al., 2001). However, the object representation in IT *is* categorical in the sense of potentially perfect rank-ordering by category (current study) and in the sense of categorical clustering of activity patterns (Kiani et al., 2007; Kriegeskorte et al., 2008a). Activity patterns also showed within-category variation that matched between man and monkey (Kriegeskorte et al., 2008a), consistent with previous reports of pattern-information differences between exemplars of the same category (Tsao et al., 2006; Kriegeskorte et al., 2007; Eger et al., 2008).

To conclude, the object representation in IT might be the result of striking a balance between maximizing both the between- and the within-category response variation. The optimal solution would enable representation of both object category (largest component of variance) and object identity. Such a solution might be implemented by feature-selectivity at the columnar level (Tanaka, 1996) that is tuned to those object features that are most informative for dis-

criminating categories as well as exemplars (Sigala and Logothetis, 2002; Ullman et al., 2002; Lerner et al., 2008), untangling category and exemplar distinctions in multivariate space (DiCarlo and Cox, 2007). Despite the absence of pure step-function responses, categoricity remains an important concept for understanding the IT representation at the functional level. The underlying neuronal mechanisms remain to be elucidated.

Chapter 5

Matching categorical object representations in inferior temporal cortex of man and monkey

Inferior temporal (IT) object representations have been intensively studied in monkeys and humans, but representations of the same particular objects have never been compared between the species. Moreover, IT's role in categorization is not well understood. Here, we presented monkeys and humans with the same images of real-world objects and measured the IT response pattern elicited by each image. In order to relate the representations between the species and to computational models, we compare response pattern dissimilarity matrices. IT response patterns form category clusters, which match between man and monkey. The clusters correspond to animate and inanimate objects; within the animate objects, faces and bodies form subclusters. Within each category, IT distinguishes individual exemplars, and the within-category exemplar similarities also match between the species. Our findings suggest that primate IT across species may host a common code, which combines a categorical and a continuous representation of objects.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126-1141.

5.1 Introduction

Do monkeys and humans see the world similarly? Do monkeys categorize objects as humans do? What main distinctions between objects define their cortical representation in each species? The comparison between monkey and human brains is important from an evolutionary perspective. High-level visual object representations are of particular interest in this context, because they are at the interface between perception and cognition and have been extensively studied in each species. Moreover, the monkey brain provides the major model system for understanding primate and, in particular, human brain function. Understanding the species relationship is therefore a challenge central not only to comparative neuroscience, but to systems neuroscience in general.

Great progress has been made by comparing monkey and human brains with functional magnetic resonance imaging (fMRI). Previous studies have used classical activation mapping in both species and cortical-surface-based alignment to define a spatial correspondency mapping between the species. Within the visual system, this approach has revealed a coarse-scale regional homologies for early visual areas and object-sensitive inferior temporal (IT) cortex (Van Essen et al., 2001; Tsao et al., 2003; Tootell et al., 2003; Denys et al., 2004; Orban et al., 2004; Van Essen and Dierker, 2007).

These studies employed classical activation mapping, in which activity patterns elicited by different particular stimuli within the same class (e.g. an object category) are averaged. Moreover, in order to increase statistical power and relate individuals and species, spatial smoothing is typically applied to the data. As a result, this classical approach reveals regions involved in the processing of particular stimulus classes. It does not reveal how those regions represent particular stimuli. In order to address the questions posed above, however, we need to understand how particular real-world object images are represented in fine-grained activity patterns within each region, and how their representations are related between the species. Here we take a first step in that direction, by studying IT response patterns elicited by the same 92 object images in monkeys and humans.

One way to relate the IT representations between the species would be to compare the activity patterns on the basis of a spatial correspondency mapping between monkey and human IT. However, this approach is bound to fail at some level of spatial detail even within a species: Every individual primate brain is unique by nature and nurture. A neuron-to-neuron functional correspondency cannot exist. (For proof, consider that different individuals have different numbers of neurons.) However, even if a fine-grained representation is unique in each individual, like a fingerprint, the region containing the representation may

be homologous, like a finger – serving the same function in both species. For example, the region may serve the function of representing a particular kind of object information. In this study, relating the species is additionally complicated by the fact that activity was measured with single-cell recording in the monkeys and fMRI in the humans. For these reasons, we do not attempt to define a spatial correspondency mapping between monkey and human IT. Instead we compare each response pattern elicited by a stimulus to each other response pattern in the same individual animal, so as to obtain a “representational dissimilarity matrix” (RDM) for each species. An RDM shows which distinctions between stimuli are emphasized and which are deemphasized in the representation, thus encapsulating, in an intuitive sense, the information content of the representation. Since RDMs are indexed horizontally and vertically by the stimuli, they can be directly compared between the species.

Our approach has the following key features: (1) The same particular images of real-world objects are presented to both species, while measuring brain activity in IT (with electrode recording in monkeys and high-resolution fMRI in humans). (2) Stimuli are presented in random sequences; neither the experimental design nor the analysis is biased by any predefined grouping. (3) Each stimulus is treated as a separate condition, for which a response pattern is estimated without spatial smoothing or averaging (Kriegeskorte et al., 2007; Eger et al., 2008; Kay et al.; 2008). (4) The analysis targets the information in distributed response patterns (Haxby et al., 2001; Cox and Savoy, 2003; Carlson et al., 2003; Kamitani and Tong, 2005; Kriegeskorte et al., 2006). (5) We introduce a framework for “representational similarity analysis”, in which RDMs are visualized and quantitatively compared to relate the single-stimulus representations between the species and to computational models.

Population representations of the same stimuli have not previously been compared between monkey and human. However, our approach is deeply rooted in the similarity analyses of mathematical psychology (Shepard and Chipman, 1970). An introduction is provided by Edelman (1998), who pioneered the application of similarity analysis to fMRI activity patterns (Edelman et al., 1998) using the technique of multidimensional scaling (Torgerson, 1958; Kruskal and Wish, 1978; Shepard, 1980). Several studies have applied similarity analyses to brain activity patterns and computational models (Laakso and Cottrell, 2000; Op de Beeck et al., 2001; Haxby et al., 2001; Hanson et al., 2004; Kayaert et al., 2005; O’Toole et al., 2005; Aguirre, 2007; Lehky and Sereno, 2007; Kiani et al., 2007; Kay et al., 2008).

Beyond the species comparison, our approach allows us to address the question of categoricity. IT is thought to contain a population code of features for the representation of natural images of objects (e.g. Desimone et al., 1984; Tanaka,

1996; Grill-Spector et al., 2001; Haxby et al., 2001). Does IT simply represent the visual appearance of objects? Or are the IT features designed to distinguish categories defined independent of the visual appearance of their members? Whether IT is optimized for the discrimination of object categories is unresolved. Human neuroimaging has investigated category-average responses for predefined conventional object categories (Puce et al., 1995; Martin et al., 1996; Kanwisher et al., 1997; Aguirre et al., 1998; Epstein and Kanwisher, 1998; Haxby et al., 2001; Downing et al., 2001; Cox and Savoy, 2003; Carlson et al., 2003; Downing et al., 2006, but see Edelman et al., 1998). This approach requires the assumption of a particular category structure and therefore cannot address whether the representation is inherently categorical. Monkey studies have reported IT responses that are correlated with categories (Vogels, 1999; Sigala and Logothetis, 2002; Baker et al., 2002; Tsao et al., 2003; Freedman et al., 2003; Kiani et al., 2005; Hung et al., 2005; Tsao et al., 2006; Afraz et al., 2006). However, more clearly categorical responses have been found in other regions (Kreiman et al., 2000; Freedman et al., 2001; Quiroga et al., 2005; Freedman and Assad, 2006), suggesting that IT has a lesser role in categorization (Freedman et al., 2003). A brief summary of the previous evidence on IT categoricity is given in the Supplementary Material.

Kiani et al. (2007) investigated monkey-IT response patterns elicited by over 1000 images of real-world objects to address whether IT is inherently categorical. The present study uses the same monkey data and a subset of the stimuli to compare the species. Cluster analysis of the monkey data revealed a detailed hierarchy of natural categories inherent to the monkey-IT representation. Will human IT show a similar categorical structure? Our approach allows us to address the question of categoricity without the bias of predefined categories. Independent of the result, this provides a crucial piece of evidence for current theory. The question of the inherent category structure of IT is of particular interest with respect to the species comparison, because the prevalent categorical distinctions might be expected to differ between species.

Our goal is to investigate, to what extent monkey and human IT represent the same object information. In particular, we ask: (a) Do human-IT response patterns form category clusters as reported for monkey IT (Kiani et al., 2007)? If so, what is the categorical structure and does it match between species? (b) Is within-category exemplar information present in IT? If so, is this continuous information consistent between the species? (c) How is the representation of the objects transformed between early visual cortex and IT? (d) What computational models can account for the IT representation?

5.2 Results

We presented the same 92 images of isolated real-world objects (Figure S5.1) to monkeys and humans while measuring IT response patterns with single-cell recording and high-resolution fMRI, respectively. Two monkeys were presented with the 92 images in rapid succession (stimulus duration: 105 ms, interstimulus interval: 0 ms) as part of a larger set while they performed a fixation task. Neuronal activity was recorded extracellularly with tungsten electrodes, one cell at a time. The cells were located in anterior IT cortex, in the right hemisphere in monkey 1 and in the left in monkey 2. The analyses are based on all cells that could be isolated and for which sufficient data was available across the stimuli. This yielded a total of 674 neurons for both monkeys combined. For each stimulus, each neuron's response amplitude was estimated as the average spike rate within a 140-ms window starting 71 ms after stimulus onset (for details on this experiment, see Kiani et al., 2007).

Four humans were presented with the same images (stimulus duration: 300 ms, interstimulus interval: 3700 ms) while they performed a fixation task in a rapid event-related fMRI experiment. Each stimulus was presented once in each run in random order and repeated across runs within a given session. The amplitudes of the overlapping single-image responses were estimated by fitting a linear model. The task required discrimination of fixation-cross color changes occurring during image presentation. We measured brain activity with high-resolution blood-oxygen-level-dependent fMRI (3-Tesla, voxels: $1.95 \times 1.95 \times 2$ mm³, SENSE acquisition; Prüssmann, 2004; Kriegeskorte and Bandettini, 2007a; Bodurka et al., 2007) within a 5-cm-thick slab including all of inferior temporal and early visual cortex bilaterally. Voxels within an anatomically defined IT-cortex mask were selected according to their visual responsiveness to the images in an independent set of experimental runs.

5.2.1 Representational dissimilarity matrices: The same categorical structure may be inherent to IT in both species

What stimulus distinctions are emphasized by IT in each species? Figure 5.1 shows the RDMs for monkey and human IT. Each cell of a given RDM compares the response patterns elicited by two stimuli. The dissimilarity between two response patterns is measured by correlation distance, i.e. $1-r$ (Pearson correlation), where the correlation is computed across the population of neurons or voxels (Haxby et al., 2001; Kiani et al., 2007). An RDM is symmetric about a diagonal of zeros here, because we use a single set of response-pattern estimates. The RDMs allow us to compare the representations between the species, although there may not be a precise correspondency of the representational features between monkey IT and human IT, and although we used radically differ-

ent measurement modalities (single-cell recordings and fMRI) in the two species. Our approach of representational similarity analysis requires comparisons only between response patterns within the same individual animal, obviating the need for a monkey-to-human correspondency mapping within IT.

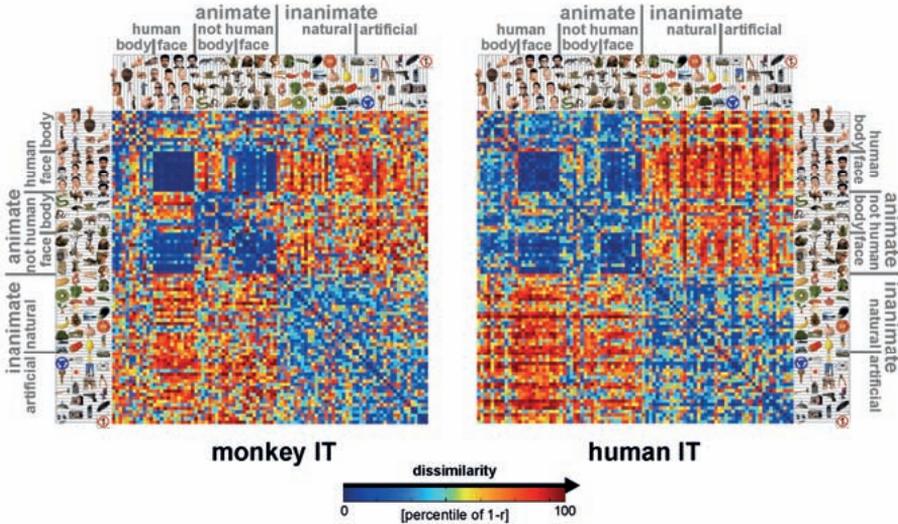


Figure 5.1 Representational dissimilarity matrices for monkey and human IT. For each pair of stimuli, each RDM (monkey, human) color codes the dissimilarity of the two response patterns elicited by the stimuli in IT. The dissimilarity measure is $1-r$ (Pearson correlation across space). The color code reflects percentiles (see colorbar) computed separately for each RDM (for $1-r$ values and their histograms, see Figure 5.3a). The two RDMs are the product of completely separate experiments and analysis pipelines (data not selected to match). Human data is from 316 bilateral inferior temporal voxels ($1.95 \times 1.95 \times 2 \text{ mm}^3$) with the greatest visual-object response in an independent data set. For control analyses using different definitions of the IT region of interest (size, laterality, exclusion of category-sensitive regions), see Figures S5.9-S5.11. RDMs were averaged across 2 sessions for each of 4 subjects. Monkey data is from 674 IT single cells isolated in two monkeys (left IT in one monkey, right in the other; Kiani et al., 2007).

Several important results (to be quantified in subsequent analyses) are apparent by visual inspection of the RDMs (Figure 5.1). First, there is a striking match between the RDMs of monkey and human IT. Two stimuli tend to be dissimilar in the human-IT representation to the extent that they are dissimilar in the monkey-IT representation, and vice versa. This is unexpected because the behaviorally relevant stimulus distinctions might be very different between the species. Moreover, single-cell recording and fMRI sample brain activity in fundamentally different ways, and it is not well understood to what extent they similarly reflect distributed representations. Second, the dissimilarity tends to be large when one of the depicted objects is animate and the other inanimate and smaller when the objects are either both animate or both inanimate. Third, dissimilarities are particularly small between faces (including human and ani-

mal faces). These observations suggest that the IT representation reflects conventional category boundaries in the same way in both species and that there may be a hierarchical structure inherent to the representation. The categorical structure of the matching dissimilarity patterns raises the question, whether the fine-scale patterns of dissimilarities within the categories also match between the species. Alternatively, the categorical structure may fully account for the apparent match. These questions cannot be decided by visual inspection and are addressed by quantitative analysis of the RDMs in the subsequent figures.

It is important to note that the two RDMs (Figure 5.1), which form the basis of the subsequent interspecies analyses (Figures 5.2-5.4), are the product of completely independent experiments and analysis pipelines. In particular, voxels and cells were not selected to maximize the match in any way. The human RDMs are averages of RDMs computed separately for each of 2 fMRI sessions in each of 4 subjects. Note that averaging RDMs for corresponding functional regions is a useful way of combining the data across subjects. As for the species comparison, a precise intra-regional spatial correspondency mapping between human subjects is not required. In total, 7 h, 24 min and 16 s of fMRI data were used for these analyses. For Figures 5.1-5.6, 316 inferior temporal voxels ($1.95 \times 1.95 \times 2 \text{ mm}^3$) with the greatest visual response were selected bilaterally in each subject and session.

5.2.2 Multidimensional scaling: Category determines global grouping when stimuli are arranged by representational similarity

Figure 5.2a shows unsupervised arrangements of the stimuli reflecting the response-pattern similarity for monkey and human IT (multidimensional scaling, criterion: metric stress; Torgerson, 1958; Shepard, 1980; Edelman et al., 1998). In each arrangement, stimuli placed close together elicited similar response patterns; stimuli placed far apart elicited dissimilar response patterns. The arrangement is unsupervised in that it does not presuppose a categorical structure. For ease of visual comparison, the two arrangements have been scaled to equal size (matching the areas of their convex hulls) and rigidly aligned (Procrustes alignment). Stimulus arrangements computed by multidimensional scaling are data-driven and serve an important exploratory function: they can reveal the properties that dominate the representation of our stimuli in the population code without any prior hypotheses. For IT in both species, the global grouping reflects the categorical distinctions between animates and inanimates, and between faces and bodies among the animates. This suggests that category is the dominant factor determining the IT response pattern in both species: If any other stimulus property were more important, it would dominate the stimulus arrangement.

Note also that neighboring stimuli within a category often differ markedly in both shape and color. The arrangements are very similar between the species. Both are characterized by a clean separation of animate and inanimate objects. Furthermore, body parts and faces occupy separate regions among the animate objects. Note that faces appear to form a particularly tight cluster in the IT response-pattern space of both species. For the human face images this could reflect their similarity in shape and color. However, the face cluster also includes animal faces. Although human and animal faces may be somewhat separated within the face cluster (see also Figure 5.4), very visually dissimilar animal faces appear to group together. These and other hypotheses inspired by exploring the stimulus arrangement will need to be tested in separate experiments.

5.2.3 Interspecies dissimilarity correlation (1): Most single stimuli are consistently represented in both species

Inspecting the stimulus arrangements for monkey and human separately, reveals their overall similarity. However, from the arrangements alone is not easy to see to what extent particular stimuli within a category appear in different “neighborhoods” of the representational space in the two species. The “fiber-flow” visualization of Figure 5.2b reproduces both stimulus arrangements and relates them by “fibers” linking dots that represent the same stimulus. This makes it easier to see how stimuli within the same category match up between species. Most fibers flow in a roughly straight line (i.e. without much displacement) from the monkey to the human representation. This reflects the within-category match of the representations, which is analyzed and tested for significance in Figure 5.3.

In order to reveal the species differences, we chose the thickness of the fibers in Figure 5.2b to reflect the extent to which each stimulus is inconsistently represented in monkey and human IT. For each stimulus i , its place in the high-dimensional monkey-IT response-pattern space is characterized by the vector m_i of its dissimilarities to the other 91 stimuli. Its place in the human-IT representation is characterized analogously by dissimilarity vector h_i . The interspecies correlation r_i (Pearson) between m_i and h_i reflects the consistency of placement of the stimulus in the representations of both species (see Figure S5.2 for details). For each stimulus i , the thickness of its fiber in Figure 5.2b is proportional to $(1-r_i)^2$, thus emphasizing the most inconsistently represented stimuli. The prevalence of thin fibers (which tend to be straight) reflects the overall interspecies consistency.

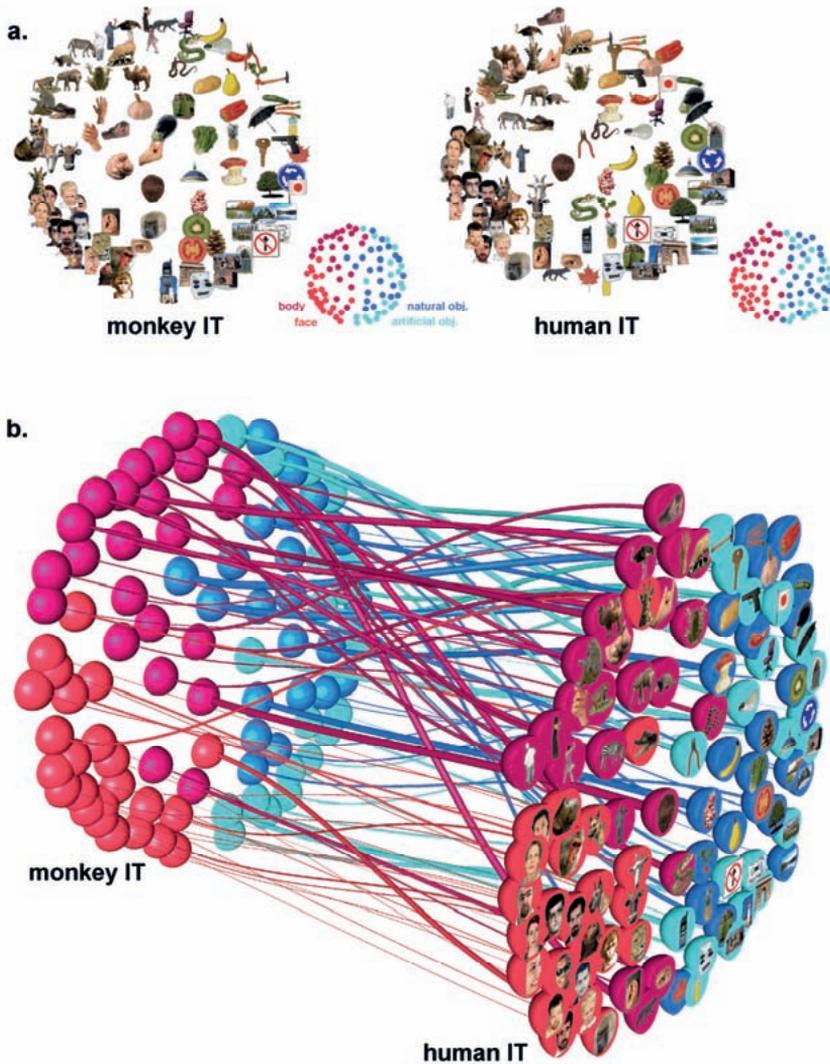


Figure 5.2 Stimulus arrangements reflecting IT response-pattern similarity in monkey and human and the interspecies relationship. (a) The experimental stimuli have been arranged such that their pairwise distances approximately reflect response-pattern similarity (multidimensional scaling, dissimilarity: $1 - \text{Pearson } r$, criterion: metric stress). In each arrangement, images placed close together elicited similar response patterns. Images placed far apart elicited dissimilar response patterns. The arrangement is unsupervised: it does not presuppose any categorical structure. The two arrangements have been scaled to match the areas of their convex hulls and rigidly aligned for easier comparison (Procrustes alignment). The correlations between the high-dimensional response-pattern dissimilarities ($1-r$) and the 2-dimensional Euclidean distances in the figure are 0.67 (Pearson) and 0.69 (Spearman) for monkey IT and 0.78 (Pearson) and 0.78 (Spearman) for human IT. (b) Fiber-flow visualization emphasizing the interspecies differences. This visualization combines all the information from (a) and links each pair of dots representing a stimulus in monkey and human IT by a “fiber”. The thickness of each fiber reflects to what extent the corresponding stimulus is inconsistently represented in monkey and human IT. The interspe-

cies consistency r_i of stimulus i is defined as the Pearson correlation between vectors of its 91 dissimilarities to the other stimuli in monkey and human IT. The thickness of the fiber for stimulus i is proportional to $(1 - r_i)^2$, thus emphasizing the most inconsistently represented stimuli. The analysis of single-stimulus interspecies consistency is pursued further in Figures S5.2 and S5.3.

The single-stimulus interspecies dissimilarity correlations r_i are further visualized and statistically analyzed in Figures S5.2 and S5.3. Results show significant interspecies consistency for about two thirds of the single stimuli ($p < 0.05$, corrected for multiple tests). Furthermore human faces exhibit significantly higher interspecies correlations than the stimulus set as a whole and several stimuli (including images of animate and inanimate objects) exhibit significantly lower interspecies correlations. The two stimuli with the lowest interspecies correlations (eggplant, back-view of human head) were the only two stimuli described as ambiguous by human subjects during debriefing (Figure S5.1). This is consistent with the idea that the IT representation reflects not only the visual appearance, but also the conceptual interpretation of a stimulus.

5.2.4 Interspecies dissimilarity correlation (2): IT emphasizes the same stimulus distinctions in both species within and between categories

The RDMs of Figure 5.1 suggest similar representations in monkey and human IT. Figure 5.3a quantifies this impression. The scatterplot of the monkey-IT dissimilarities (horizontal axis) and the human-IT dissimilarities (vertical axis) across pairs of stimuli reveals a substantial correlation ($r = 0.49$, $p < 0.0001$ estimated by means of 10,000 randomizations of the stimulus labels).

Does the matching categorical structure fully explain the interspecies correlation of dissimilarities? Figure 5.3a (colored subsets) shows that the correlation is substantial also within animates ($r = 0.51$, $p < 0.0001$) and, to a lesser extent, within inanimates ($r = 0.20$, $p < 0.0001$), as well as across pairs of stimuli crossing the animate-inanimate boundary ($r = 0.19$, $p < 0.0001$). The monkey-to-human correlation is also present (Figure 5.3b) within images of humans ($r = 0.66$, $p < 0.0001$), within images of nonhuman animals ($r = 0.35$, $p < 0.0001$), within images of faces (including human and animal faces, $r = 0.31$, $p < 0.0058$), within images of bodies (including human and animal bodies, $r = 0.31$, $p < 0.0001$; not shown in Figure 5.3), within images of human bodies ($r = 0.53$, $p < 0.0001$; not shown in Figure 5.3), within images of natural objects ($r = 0.19$, $p < 0.0039$), and within images of artificial objects ($r = 0.23$, $p < 0.0139$). These within-category dissimilarity correlations between the species indicate that the continuous variation of response patterns within each category cluster is not noise, but distinguishes exemplars within each category in a way that is consistent between monkey and human.

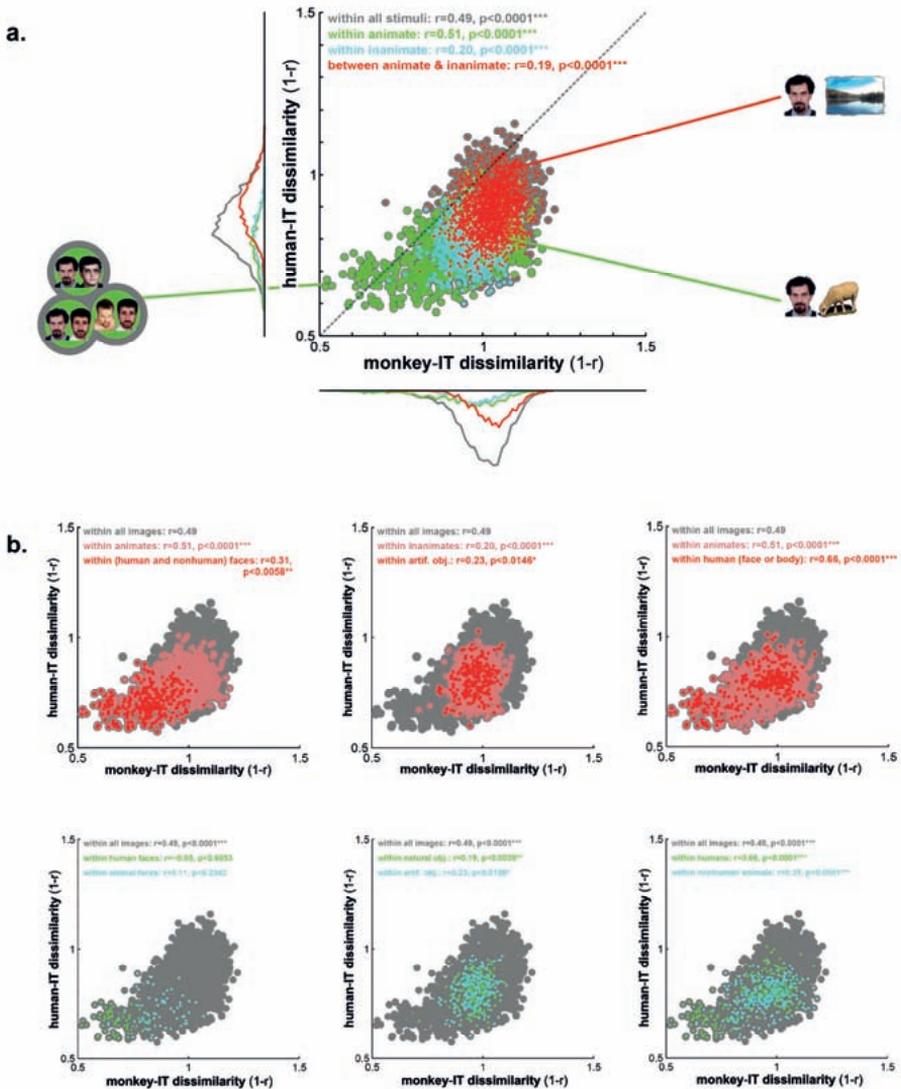


Figure 5.3 Correlation of representational dissimilarities between monkey and human IT. (a)

For each pair of stimuli, a dot is placed according to the IT response-pattern dissimilarity in monkey (horizontal axis) and human (vertical axis). As before, the dissimilarity between the two response patterns elicited by each stimulus pair is measured as 1-r (Pearson correlation). Dot colors correspond to all pairs of stimuli (gray), pairs within the animate objects (green), pairs within the inanimate objects (cyan), and pairs crossing the animate-inanimate boundary (red). Marginal histograms are shown for the three subsets of pairs using the same color code. For detailed exploratory analysis of the species differences, Figures S5.13 and S5.14 show the stimulus pairs corresponding to the dots for the three apical regions of the scatterplot. **(b)** The same analysis as in (a), but for within-category correlations between human- and monkey-IT object dissimilarities. Colored dots correspond to all pairs of stimuli (gray), or pairs within stimulus-category subsets (colors). In the top row, each panel shows the whole set (gray), a subset (pink), and a subset nested within that subset (red), as indicated in the colored legend of each panel. In the bottom row, each

panel shows the whole set (gray) and two disjoint subsets (green and cyan), as indicated in the colored legend of each panel. In both (a) and (b), each panel's color legend (top inset) also states the correlations (r , Pearson) between monkey- and human-IT dissimilarities and their significance ($p < 0.05$ indicated by *, $p < 0.01$ by **, $p < 0.001$ by ***). The dissimilarity correlations are tested by randomization of the stimulus labels. This test correctly handles the dependency structure within each RDM. All p values < 0.0001 are stated as $p < 0.0001$ because the randomization test terminates after 10,000 iterations.

We did not find a significant monkey-to-human dissimilarity correlation either within images of human faces ($r = -0.05$, $p < 0.605$) or within images of animal faces ($r = 0.12$, $p < 0.234$; Figure 5.3b, bottom left). We also did not find a significant correlation within images of nonhuman bodies ($r = 0.12$, $p < 0.21$; not shown in Figure 5.3). The negative findings all occurred for small subsets of images (12 images, 66 pairs), for which we have reduced power. Note, however, that the effect sizes (r) were also smaller for the insignificant correlations than for all significant correlations. That the correlation is significant within human bodies, but not within nonhuman bodies, could reflect the fact that the human-body images included whole bodies as well as body parts, whereas the nonhuman body images were all of whole bodies and 9 of the 12 images were of four-limbed animals (Figure S5.1), which may constitute a separate subordinate category (Kiani et al., 2007). Regarding the absence of a significant interspecies dissimilarity correlation within human faces and within nonhuman faces, one interpretation of particular interest is that human and monkey differ in how they represent individual human faces as well as individual nonhuman faces. For example, within each species the representation of images of its own members may have a special status.

5.2.5 Species-specific face analysis: IT might better distinguish conspecific faces in each species

We observed greater dissimilarities in the human representation of human faces than in the monkey representation of human faces (Figure S5.8). Figure S5.4 explores the possibility of a species-specific face representation. We selectively analyzed the representation of monkey, ape, and human faces in monkey and human IT. The dissimilarities among human faces are significantly larger in human IT than in monkey IT ($p = 0.009$). The dissimilarities among monkey-and-ape faces are larger in monkey IT than in human IT in our data, but the effect is not significant ($p = 0.12$). The difference between the two effects is significant ($p = 0.02$). This analysis provides an interesting lead for future studies designed to address species-specific face representation (for details, see Figure S5.4).

5.2.6 Hierarchical clustering: A nested categorical structure matching between species is inherent to IT

The stimulus arrangements of Figure 5.2 suggest that the categories correspond to contiguous regions in IT response-pattern space. However, it is not apparent from Figure 5.2, whether the response patterns form clusters corresponding to the categories. Contiguous category regions in response-pattern space could exist for a unimodal response-pattern distribution. Category clusters would imply separate modes in response-pattern space, separated by boundaries of lower probability density. We therefore ask whether the category boundaries can be determined without knowledge of the category labels.

Figure 5.4 shows hierarchical cluster trees computed for the IT response patterns in monkey and human. Unlike the unsupervised stimulus arrangements (Figure 5.2), hierarchical cluster analysis (Johnson, 1967) assumes the existence of some categorical structure, but it does not assume any particular grouping into categories. We find very similar cluster trees for both species. The top-level distinction is that between animate and inanimate objects. Faces and body parts form subclusters within the animate objects. Note that the clustering conforms closely, though not perfectly, to these human-conventional categories. The deviating placements could be a consequence of inaccurate response-pattern estimation: Because of the large number of conditions (92) in these experiments, our response-pattern estimates are noisier than they would be for a small number of conditions based on the same amount of data.

5.2.7 Results similar between hemispheres and robust to exclusion of category-sensitive regions and to varying the number of voxels

The results we describe for bilateral IT are similar when IT is restricted to either cortical hemisphere (for details, see Figure S5.9). Results are also robust to changes of the number of voxels selected. The categorical structure is present already at 100 voxels and decays only when thousands of voxels are selected (for details, see Figure S5.10). Finally, the categorical structure appears unaffected when the fusiform face area (FFA) (Kanwisher et al., 1997) and the parahippocampal place area (PPA) (Epstein and Kanwisher, 1998) are bilaterally excluded from the selected voxel set (for details, see Figure S5.11). After exclusion of FFA and PPA, the region of interest has most of its voxels in the lateral occipital complex, but also includes more anterior object-sensitive voxels within IT.

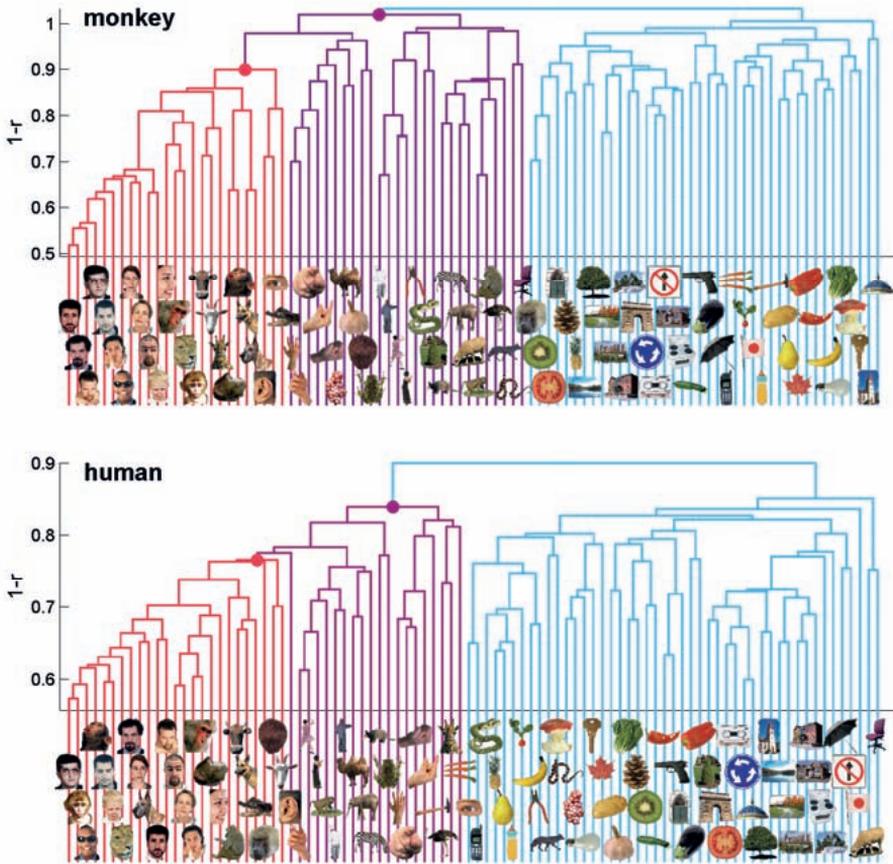


Figure 5.4 Hierarchical clustering of IT response patterns. In order to assess whether IT response patterns form clusters corresponding to natural categories, we performed hierarchical cluster analysis for human (top) and monkey (bottom). This analysis proceeds from single-image clusters (bottom of each panel) and successively combines the two clusters closest to each other in terms of the average response-pattern dissimilarity, so as to form a hierarchy of clusters (tree structure in each panel). The vertical height of each horizontal link indicates the average response-pattern dissimilarity (the clustering criterion) between the stimuli of the two linked subclusters (dissimilarity: $1-r$). The cluster trees for monkey and human are the result of completely independent experiments and analysis pipelines. This data-driven technique reveals natural-category clusters that are consistent between monkey and human. For easier comparison, we colored subcluster trees (faces: red, bodies: magenta, inanimate objects: light blue). Early visual cortex (Figures 5.5, 5.6, S5.5) and low-level computational models (Figures S5.6, S5.7) did not reveal such category clusters.

5.2.8 Category-boundary effect: weak in early visual cortex and strong in IT

The human fMRI data allowed us to compare the representations between early visual cortex and IT (Figures 5.5, 5.6, S5.5). Visual inspection of the RDMs (Fig-

ure 5.5) suggests a categorical representation in IT, but not in early visual cortex. The multidimensional scaling arrangements and hierarchical cluster trees also do not support that early visual cortex contains an inherently categorical representation (Figure S5.5). Note, however, that a subset of the human faces appears to be associated with somewhat lower dissimilarities in early visual cortex (Figure 5.5). This could be caused by similarities in shape and color among these stimuli.

Although the early visual representation does not exhibit a categorical structure as observed for IT, the top-level animate-inanimate distinction might be reflected in the early visual responses in more subtle ways. To test this possibility for the top-level animate-inanimate distinction, we analyze the category-boundary effect, which we define as the difference between the mean dissimilarity for between-category pairs (i.e. one is animate, the other inanimate) and the mean dissimilarity for within category pairs (i.e. both are animate or both are inanimate). As in Figure 5.1, the dissimilarities are correlation distances between spatial response patterns, converted to percentiles for each RDM separately (for histograms of the original correlation distances, see Figure 5.6). The analysis indicates that the category-boundary effect is strong in IT and weak, but present, in early visual cortex. The category-boundary effect estimates in percentile points are 36% ($p < 0.001$) for IT (defined at 316 voxels as before), and 7% ($0.01 < p < 0.05$), 5% ($0.01 < p < 0.05$), and 2% (not significant) for early visual cortex, defined at 224, 1057, and 5000 voxels, respectively. We further investigated the category distinction by a linear decoding analysis (Figure S5.12), which suggested linear separability of animates and inanimates in IT, but not early visual cortex.

The p values for the category-boundary effect were computed by bootstrap resampling of the stimulus set, thus simulating the distributions of mean dissimilarities expected if the experiment were to be repeated with different stimuli from the same categories and with the same subjects. We also tested the category-boundary effect by a randomization test, in which the stimulus labels are randomly permuted, thus simulating the null hypothesis of no difference between the response patterns elicited by the stimuli, but not generalizing to different stimuli from the same categories. This test yields the same pattern of significant results as the bootstrap test ($p < 0.001$ for IT and early visual cortex defined at 224 or 1057 voxels, $p \geq 0.05$ for human early visual cortex defined at 5000 voxels). In addition, both bootstrap and randomization tests show a significantly larger category-boundary effect in IT than in early visual cortex ($p < 0.0001$ for each of the three sizes of the early visual region of interest).

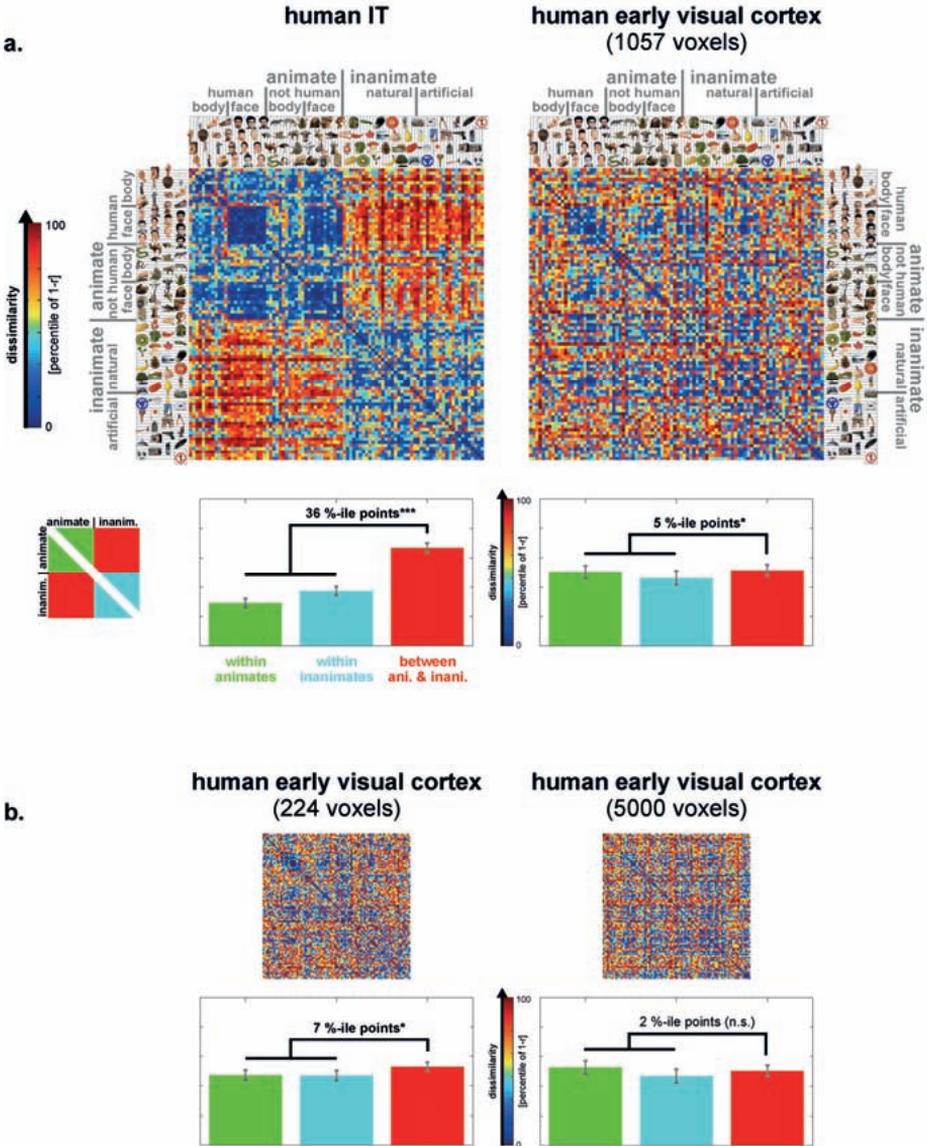


Figure 5.5 Early visual cortex and IT in the human: representational dissimilarity matrices and category-boundary effects. (a) RDMs for human IT (top left, same as in Figure 5.1) and human early visual cortex (top right). As in Figure 5.1, the color code reflects percentiles (see colorbar) computed separately for each RDM (for 1-r values and their histograms, see Figure 5.6). The bar graph below each RDM shows the average dissimilarity (percentile of 1-r) within the animates (green bars), within the inanimates (cyan bars), and for pairs crossing the category boundary (red bars). Error bars indicate the standard error of the mean estimated by bootstrap resampling of the stimulus set. We define the category-boundary effect as the difference (in percentile points) between the mean dissimilarity for between-animate-and-inanimate pairs and the mean dissimilarity for within-animate and within-inanimate pairs. The zeros on the diagonal are excluded in computing these means. The category-boundary effect sizes are given above the bars in each panel with signifi-

cant effects marked by stars ($p \geq 0.05$ indicated by n.s. for “not significant”, $p < 0.05$ indicated by *, $p < 0.01$ by **, $p < 0.001$ by ***). The p values are from a bootstrap test; a randomization test yields the same pattern of significant effects (see Results). Here, as in Figures 5.1-5.4, human IT has been defined at 316 voxels (for IT at 100-10,000 voxels, see Figure S5.10) and human early visual cortex at 1057 voxels. **(b)** The same analyses for smaller and larger definitions of human early visual cortex (224 and 5000 voxels, respectively).

5.2.9 Representational connectivity analysis: early visual cortex and IT share visual-similarity information

We have compared monkey and human IT by correlating representational dissimilarities across pairs of stimuli. The same approach can serve to characterize the relationship between the representations in two brain regions of a given species. In analogy to the concept of functional connectivity, we refer to the correlation of representational dissimilarities between two brain regions as their “representational connectivity”. Like functional connectivity, representational connectivity does not imply an anatomical connection or a directed influence. Unlike functional connectivity, representational connectivity is a multivariate, nonlinear, and design-dependent connectivity measure.

Visual comparison of the RDMs for early visual cortex and IT does not suggest a strong correlation, because the categorical structure dominating IT appears absent in early visual cortex. However, the representational-connectivity scatterplot (Figure 5.6) reveals a substantial correlation of the dissimilarities ($r = 0.38$, $p < 0.0001$). Pairs of stimuli eliciting more dissimilar response patterns in early visual cortex also tend to elicit more dissimilar response patterns in IT. The analysis for between- and within-category subsets of pairs, reveals what drives this effect (see diagram in Figure 5.6b): First, the between-category distribution (red) is shifted relative to the within-category distributions, but only along the IT axis, not along the early-visual axis. This is evident in the marginal dissimilarity histograms framing the scatterplot (Figure 5.6) and reflects the category-boundary effect, which is strong in IT and weak in early visual cortex (Figure 5.5). Second, in addition to its category-boundary effect, IT reflects the dissimilarity structure of the early visual representation for within- as well as between-category pairs (diagonally elongated distributions, $p < 0.0001$ for all correlations, tested by randomization of the stimulus labels). These results do not depend on the size of the early-visual region of interest (Figure 5.6c). Early visual response patterns are likely to reflect shape similarity in this experiment, because all stimuli were presented at the same retinal location (fovea) and size (2.9° visual angle). Shape similarity as reflected in the early visual representation (see also Kay et al., 2008) may therefore carry over to the IT representation, even if IT is more tolerant to changes of position and size.

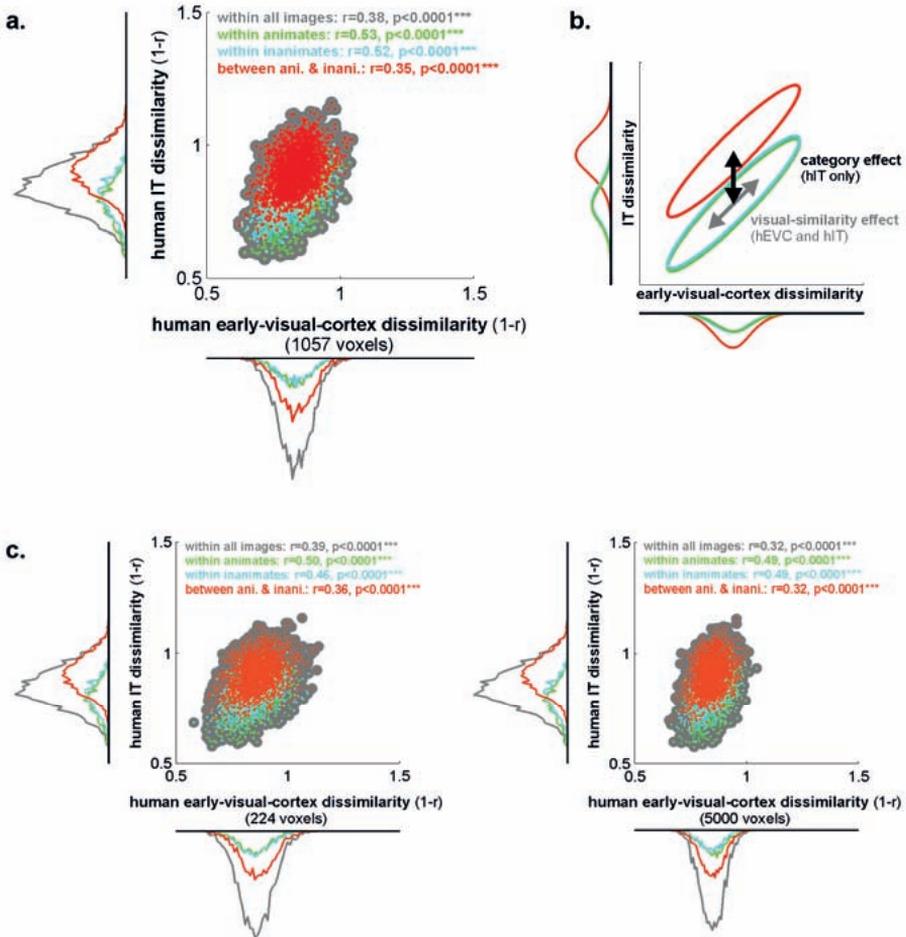


Figure 5.6 Representational connectivity between early visual cortex and IT in the human. (a) For each pair of stimuli, we plot a dot with horizontal position reflecting early visual response-pattern dissimilarity and vertical position reflecting IT response-pattern dissimilarity. Scatterplots and correlation analyses (insets) show that pairs of stimuli eliciting more dissimilar response patterns in early visual cortex also tend to elicit more dissimilar response patterns in IT. This suggests that visual similarity as reflected in the early visual representation carries over into the IT representation. However, IT additionally exhibits a strong category-boundary effect: when a stimulus pair crosses the animate-inanimate boundary (red) the two response patterns tend to be more dissimilar than when both stimuli are from the same category (green, cyan). The category-boundary effect is evident in the marginal dissimilarity histograms framing the scatterplot (for statistical analysis, see Figure 5.5). (b) In this conceptual diagram, the distributions from the scatterplots are depicted as ellipsoids (iso-probability-density contours) with the same color code. The visual-similarity effect is shared between early visual and IT representations (each distribution diagonally elongated), whereas the category-boundary effect is only present in IT (red distribution vertically, but not horizontally shifted with respect to the within-category distributions). (c) The same analyses for smaller and larger definitions of human early visual cortex (224 and 5000 voxels, respectively) show that the findings above do not depend on the size of the early visual region of interest.

5.2.10 Computational modeling: A range of low- and intermediate-level representations cannot account for the categorical structure observed in IT

Can low-level feature similarity account for our results? We compared the IT representation to several low-level model representations. The low-level models included: the color images themselves (in CIELAB color space), simple processed versions of the images (low-resolution color image, grayscale image, low-resolution grayscale image, spatial low- and high-pass-filtered grayscale image, binary silhouette image), CIELAB joint color histograms, and a computational model of V1 (including simple and complex cells). We also tested an intermediate-level computational model corresponding approximately to the level of V4 and posterior IT, the HMAX-C2 representation based on natural image patches. (For details on these models, see Supplementary Material.) The RDMs, multidimensional scaling arrangements, and hierarchical cluster trees for these models (Figures S5.6, S5.7) suggest that none of them can account for the category clustering we observed in IT cortex.

5.3 Discussion

5.3.1 Matching information in monkey and human IT

IT is thought to contain a high-level representation of visual objects at the interface between perception and cognition. Our results show that monkey and human IT emphasize very similar distinctions among objects. To answer the questions posed at the end of the Introduction: (a) IT response patterns elicited by object images appear to cluster according to the same categorical structure in monkey and human. (b) Within each category, primate IT appears to represent more fine-grained object information. This information as well is remarkably consistent across species and may reflect subordinate categorical distinctions as well as a high-level form of visual similarity (Op de Beeck et al., 2001; Eger et al., 2008). (c) Categoricality appears to arise in IT; it is largely absent in human early visual cortex. (d) A range of low- and intermediate-level computational models did not reproduce the categorical structure observed for human and monkey IT.

5.3.2 A hierarchical category structure inherent to IT

The categorical structure inherent to IT in both species appears hierarchical: Animate and inanimate objects form the two major clusters; faces and bodies form subclusters within the animate cluster. The hierarchy observed for the present set of 92 stimuli has only two levels. However, the previous study by

Kiani et al. (2007), using over 1000 stimuli, reported a hierarchy for monkey IT, which is consistent with our findings here, but extends into finer distinctions. This raises the question, whether finer categorical distinctions are also present in the human and, if so, if they match between the species.

5.3.3 Relationship between category-sensitive regions and IT pattern information

Human neuropsychology has described category-specific deficits resulting from temporal brain damage and suggested a special status for the living/nonliving distinction (Martin, 2007; Capitani et al., 2003; Humphreys and Forde, 2001; Martin et al., 1996). Our results support the view that this distinction has a special status. We note that fruit and vegetables fall into the inanimate category in the IT cluster structure we observed (Figure 5.4). Our findings are also consistent with the idea that IT contains specialized features or processing mechanisms for faces and bodies (Puce et al., 1995; Kanwisher et al., 1997; Downing et al., 2001; but see also Gauthier et al., 2000a).

Can the previous findings on human-IT regions sensitive to these categories explain the IT response-pattern clustering we report? Let us assume that the FFA responds with a similar overall activation to each individual face (Kriegeskorte et al., 2007). Including the FFA in the region of interest will then render IT response patterns to faces more similar, thus contributing to their clustering. More generally, a sufficiently category-sensitive feature set will exhibit categorical clustering of the response patterns. However, the existence of the category-sensitive regions does not predict (a) that the category effects will dominate the representation such that response patterns form category clusters that are separable without prior knowledge of the categories (alternatively, the category-sensitive component could be too weak – in relation to the total response-pattern variance – to form clusters), (b) that the cluster structure will be hierarchical, or (c) what categorical distinction is at the top of the hierarchy (explaining most response-pattern variance). Moreover, our human results hardly changed when FFA and PPA were excluded (Figure S5.11), leaving the lateral occipital complex as the main focus within the IT region of interest. (Voxels were selected by their average response to objects versus fixation using independent data.) In the monkeys, the IT recordings did not target category-sensitive regions (Tsao et al., 2003); nevertheless the population exhibited a complex categorical clustering of its response patterns (Figure 5.4), and for each pair of categories, discrimination by the cell population was robust to exclusion of cells responding maximally to either category (Figure 10 of Kiani et al., 2007).

If FFA and PPA can be excluded without a qualitative change to the representational dissimilarity structure (Figure S5.11), the remaining portion of human IT must have similarly category-sensitive features. One interpretation is that the prominent category regions are just particularly conspicuous concentrations of related features within a larger category-sensitive feature map (Haxby et al., 2001). Such large, consistently localized foci of features may only exist for a few categories (Downing et al., 2006). Their number is limited by the available space in the brain. Beyond discovering those regions, our larger goal should be to understand the representation as a whole, including the contribution of its less prevalent – or perhaps just more scattered – features. After all, IT categoricity is inherently a population phenomenon: Step-function-like categorical responses as reported for cells in the medial temporal lobe (Kreiman et al., 2000) and prefrontal cortex (Freedman et al., 2001) are not typically observed in either single IT cells (Vogels, 1999; Freedman et al., 2003; Kiani et al., 2007; but see Tsao et al., 2006) or category-sensitive fMRI responses (Haxby et al., 2001). Categorical clustering of response patterns indicates that the categorical distinctions explain a lot of variance across the population. It does not imply that any single cell exhibits a step-function-like response.

5.3.4 Explaining the IT representational similarity structure

Can low-level features explain the IT representational similarity structure? The categorical cluster structure observed in IT was absent in the fMRI response patterns in human early visual cortex (Figure 5.5, S5.5) and also in several low-level model representations of the images (luminance pattern, color pattern, color histogram, silhouette pattern, V1 model representation; Figures S5.6, S5.7). The possibility that our findings can be explained by low-level features can never be formally excluded, because the space of models to be tested is infinite. However, our results suggest that the categorical clustering in IT does not reflect only low-level features. Can more complex natural-image features explain the IT representational similarity structure? Categorical clustering was not evident in the intermediate-complexity HMAX-C2 model based on natural image fragments (Figure S5.7; Serre et al., 2005). In addition, a high-level representation composed of shape-tuned units adapted to real-world object images in the HMAX framework has previously been shown not to exhibit categorical clustering (Kiani et al., 2007). Our interpretation of the current evidence is that evolution and development leave primate IT with features optimized not only for representing natural images (as the features of the models described above), but also for discriminating between object categories. This suggests that an IT model should acquire category-discriminating features by supervised learning (Ullman, 2007). A recent study suggests that human IT responds preferentially to such category-discriminating features (Lerner et al., 2008).

Does IT categoricity arise from feedforward or feedback processing? Our tasks (in both species) minimize the top-down component by withdrawing attention from briefly presented stimuli. Although this does not abolish local recurrent processing, it minimizes feedback from higher regions, suggesting that IT categoricity is not a product of top-down influences. One interpretation is that IT categoricity arises from feedforward connectivity. Rapid feedforward animate-inanimate discrimination would explain reports that humans can perform animal detection at latencies allowing for limited recurrent processing (Thorpe et al., 1996; Kirchner and Thorpe, 2006). Serre et al. (2007) proposed a feedforward model of rapid categorization (see also Riesenhuber and Poggio, 2002), which summarizes a wealth of neuroscientific findings. Their architecture may be able to account for our findings. However, these authors associate the category-discrimination stage with prefrontal cortex. Our results suggest that features at the stage of IT already are optimized for category discrimination.

Beyond visual features optimized for categorization, could IT represent more complex semantic information? So far we have considered the features a means to the end of categorization. Instead we could argue, more generally, that the features serve to infer nonvisual properties from the visual input. The features, then, are the end, and category clusters may arise as a consequence of the feature set. It has been suggested, for example, that IT represents action-related properties (Mahon et al., 2007). This perspective relates our findings to the literature on semantic representations (Tyler and Moss, 2001; McClelland and Rogers, 2003; Patterson et al., 2007). In order to test semantic-feature hypotheses along with computational models, we could predict the IT representational similarity from semantic property descriptions of the stimuli.

To find a model that reproduces the empirical representational similarity structure of IT (Figure 5.1) would constitute a substantial theoretical advance. The reader is invited to join us in testing additional models by exposing them to our stimuli and comparing the RDMs of the model representations to our empirical RDMs from IT.¹⁴ We will provide both stimuli and RDMs of monkey and human IT upon request.

5.3.5 Representational similarity analysis

Studying a brain region's pairwise response-pattern dissimilarities for a sizable set of stimuli reveals what distinctions are emphasized and what distinctions are abstracted from by the representation. Representational similarity analysis

¹⁴ If the models have parameters fitted, so as best to predict the empirical RDMs, independent stimulus sets will be needed for fitting and testing.

allows us to make comparisons between brain regions (Figure 5.6), between species (Figure 5.3), between measurement modalities (Figure 5.3, confounded with the species-effect here), and between biological brains and computational models (Figures S5.6, S5.7). An RDM usefully combines the evidence across the patterns of response within a functional region (thus allowing us to see the forest), but it requires no averaging of activity across space, time, or stimuli (thus honoring the trees). The RDM has a very intricate structure ($(n^2-n)/2$ dissimilarities, where n is the number of stimuli), thus providing a rich characterization of the representation.

In order to understand a population code, representational similarity analysis must be complemented with a wide range of methods. For example, we need to quantify the pairwise stimulus information, address how the representation can be read out (e.g. is a given distinction explicit in the sense of linear decodability?; Figure S5.12), how it relates to other brain representations (Figure 5.6) and behavior, and how the activity patterns are organized in space and time.

5.3.6 Implications for the relationship between fMRI and single-cell data

A single voxel in blood-oxygen-level-dependent fMRI reflects the activity of tens of thousands of neurons (Logothetis et al., 2001). We therefore expect to find somewhat different stimulus information in hemodynamic and neuronal response patterns. fMRI patterns may contain more information about fine-grained neuronal activity patterns than voxel size would suggest (Kamitani and Tong, 2005). But to what extent neuronal pattern information is reflected in fMRI pattern information is not well understood, because a voxel's signal does not provide us simply with the average activity within its boundaries, but rather reflects the complex spatiotemporal transform of the hemodynamic response. The close match we report here between the RDMs from single-cell recording and fMRI provides some hope that data from these two modalities, for all their differences, may somewhat consistently reveal neuronal representations when subjected to massively multivariate analyses of activity-pattern information (Kriegeskorte and Bandettini, 2007a).

5.3.7 A common code in primate IT

Taken together, our results suggest that evolution and individual development leave primate IT with representational features that emphasize behaviorally important categorical distinctions. The major distinctions, animate-inanimate and face-body, are so basic that their conservation across species appears plausible. However, the IT representation is not purely categorical. Within category clusters, object exemplars are represented in a continuous object space, which may reflect a form of visual similarity. The categorical and continuous aspects of

the representation are both consistent between man and monkey, suggesting that a code common across species may characterize primate IT.

5.4 Experimental Procedures

This section describes the experimental designs and brain-activity measurements in monkey and human. The monkey experiments have previously been described in detail (Kiani et al., 2007), so we only give a brief summary here. Detailed descriptions of the statistical analysis and human localizer experiments are in the Supplementary Material.

5.4.1 Stimuli presented to humans and monkeys

The stimuli presented to monkeys and humans were 92 color photographs (175×175 pixels) of isolated real-world objects on a gray background (Figure S5.1). The objects included natural and artificial inanimate objects as well as faces and bodies of humans and nonhuman animals. No predefined stimulus grouping was implied in either the experimental design or the core analyses for either species.

5.4.2 Monkey experiments

Experimental design and task

Two alert monkeys were presented with the 92 images in rapid succession (stimulus duration: 105 ms, interstimulus interval: 0 ms) as part of a larger set of over 1000 similar images while they performed a fixation task. Fixation was monitored with an infra-red eye-tracking system. Stimuli were presented in a pseudorandom order. The stimulus sequence started after the monkey maintained fixation for 300 ms. Stimuli spanned a visual angle of about 7°. Each stimulus lasted for 105 ms and was followed by another stimulus without intervening interstimulus interval.

Brain-activity measurements

Neuronal activity was recorded extracellularly with tungsten electrodes, one cell at a time. The cells were located in anterior IT cortex (anterior 13-20 mm, distributed over the ventral bank of the superior temporal sulcus and the ventral convexity up to the medial bank of the anterior middle temporal sulcus), in the right hemisphere in monkey 1 and in the left in monkey 2. On average, the stimulus set was repeated 9 ± 2 (median, 10) times for each recording site. A different random stimulus sequence was used on each repetition for each re-

coding site in order to avoid consistent interactions between successively presented stimuli.

5.4.3 Human experiments

Experimental design and task

We presented the 92 images to subjects in a “quick” event-related fMRI experiment, which balances the need for separable hemodynamic responses (suggesting a slow event-related design) and the need for presenting many stimuli in the limited time-span of the fMRI experiment (suggesting a rapid event-related design). The experiment included four additional images, which were excluded from the interspecies analyses because of insufficient monkey data (Figure S5.1). Stimuli spanned a visual angle of 2.9° and were presented foveally for a duration of 300 ms on a constantly visible uniform gray background. Stimuli were centered with respect to a fixation cross superimposed to them.

Each stimulus was presented exactly once in each run. The sequence also included 40 null trials with no stimulus presented (4 of them at the beginning, 4 of them at the end, and 32 randomly interspersed in the sequence). The trial-onset asynchrony was 4 s; the stimulus-onset asynchrony was either 4 s or a multiple of that duration when null trials occurred in the sequence. The trials (including 96 stimulus presentations and 32 interspersed null trials) occurred in random order (no sequence optimization). We used a different random sequence on each of up to 14 runs (spread over two fMRI sessions) per subject. A run lasted 9 min and 4 s ($4+96+32+4=136$ trials, each 4 s long).

Subjects continually fixated a fixation cross superimposed to the stimuli and performed a color-discrimination task: During stimulus presentation the fixation cross turned from white to either green or blue and the subject responded with a right-thumb button press for blue and a left-thumb button press for green. The fixation-cross changes to blue or green were chosen according to an independent random sequence.

Brain-activity measurements

Blood-oxygen-level-dependent fMRI was performed at high spatial resolution using a 3T GE HDx MRI scanner. For signal reception, we used a receive-only whole-brain surface-coil array (16 elements, NOVA Medical Inc., Wilmington, MA). Twenty-five 2-mm axial slices (no gap) were acquired, covering the occipital and temporal lobe, using single-shot interleaved gradient-recalled Echo Planar Imaging (EPI) with a sensitivity-encoding sequence (SENSE, acceleration factor: 2, Prüssmann 2004). Imaging parameters were as follows: EPI matrix

size: 128×96 , voxel size: $1.95 \times 1.95 \times 2$ mm³, echo time (TE): 30 ms, repetition time (TR): 2 s. Each functional run consisted of 272 volumes (9 min and 4s per run). Four subjects were scanned in two separate sessions each, resulting in 11 to 14 runs per subject, yielding a total of 49 runs (equivalent to 7 h, 24 min, and 16 s of fMRI data). As an anatomical reference, we acquired high-resolution T1-weighted whole-brain anatomical scans with a Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence. Imaging parameters were as follows: matrix size: 256×256 , voxel size: $0.86 \times 0.86 \times 1.2$ mm³, 124 slices.

5.5 Supplementary material

5.5.1 Supplementary figures

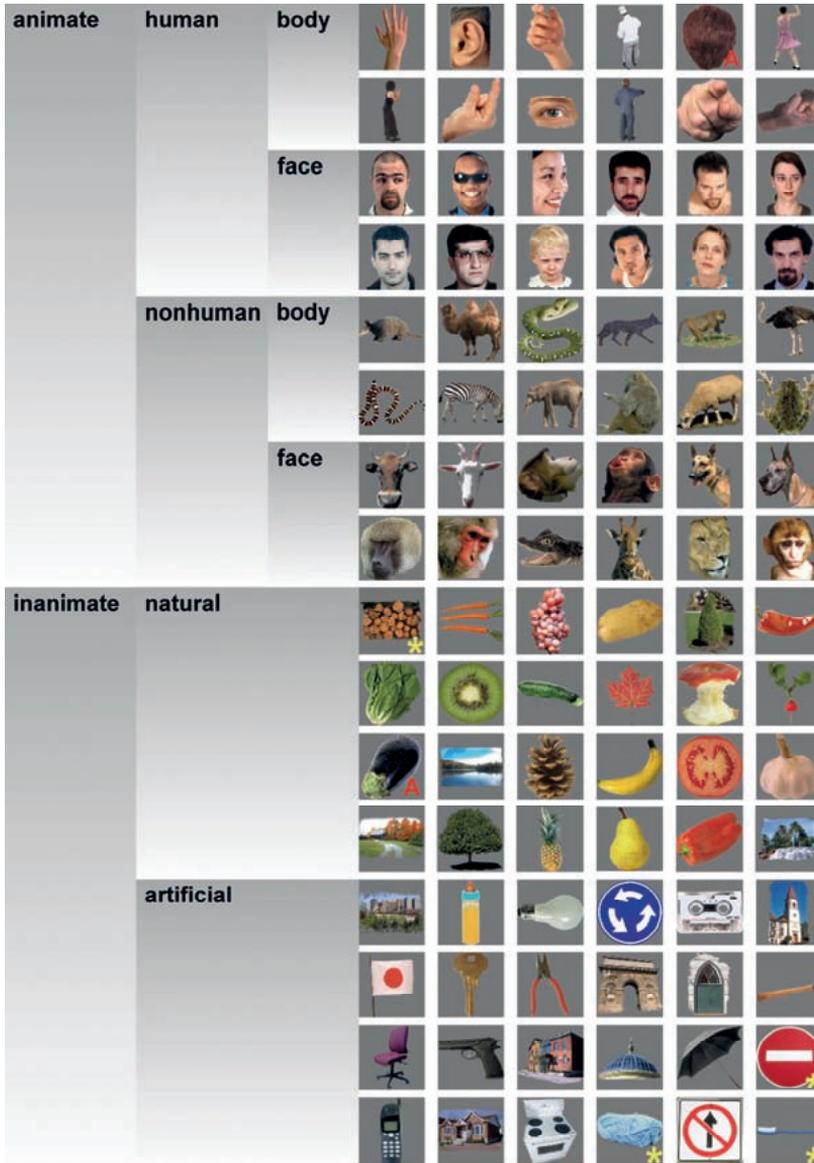


Figure S5.1 Stimuli. The object images presented to monkeys and humans. The four images marked by yellow stars were excluded from the analysis because of insufficient data in the monkey experiments. Responses to the remaining 92 form the basis of all analyses. Several of our human subjects described two of the stimuli as ambiguous during debriefing. These two stimuli (egg plant, back of a human head) are marked by a red “A”.

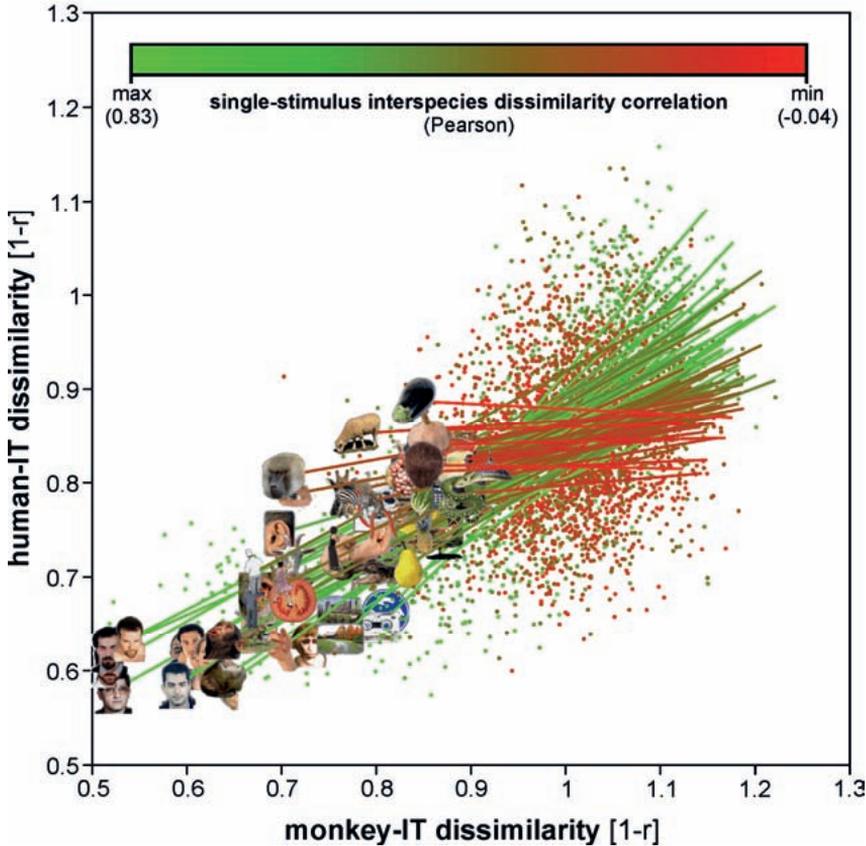


Figure S5.2 Interspecies correlation of IT object dissimilarities related to single stimuli. This figure addresses the question to what extent each of the stimuli is similarly represented in both species. A stimulus is considered “similarly represented” if its pattern of representational dissimilarities to the other 91 stimuli is correlated between human and monkey IT. For each stimulus, we consider its row (or, equivalently, its column) in the representational dissimilarity matrix in each species (Figure 5.1) and plot the monkey-IT dissimilarities against the human-IT dissimilarities (analogously to Figure 5.3). In addition, we plot a straight line for each stimulus, which is obtained as the least-squares fit to the 91 human-IT dissimilarities (vertical axis) of that stimulus and extends horizontally along the range of the corresponding 91 monkey-IT dissimilarities. The stimulus itself is plotted on the left end of the line. In order to highlight the stimuli most inconsistently represented in monkey and human, the scatterplots, fit lines, and stimuli are overplotted in the order of their interspecies representational-dissimilarity correlation, starting from the most highly correlated (green scatterplot and fit line, thin fibers in Figure 5.2b) and progressing to the least interspecies-correlated stimulus (red scatterplot and fit line, thick fibers in Figure 5.2b). The scatterplots and fit lines for intermediate stimuli are plotted in intermediate colors ranging from green to red, which linearly reflect the interspecies correlation (see colorbar).

and using the Bonferroni method to control the family-wise false-positives rate. Using either method, about 61 of the 92 stimuli, including all faces, exhibit significant interspecies correlation. **(b)** Statistical analysis addressing the question, whether stimuli vary in terms of interspecies correlation. The analysis in (a) highlights some stimuli and not others as significantly consistently represented in IT of both species. However, this does not mean that there are significant differences between single-stimulus interspecies correlations. A mixture of significant and insignificant interspecies correlations as obtained in (a) could result from an interspecies correlation constant across all stimuli in conjunction with noise. We therefore tested the null hypothesis that all single-stimulus interspecies correlations are equal. We simulated the null distribution of equal interspecies correlation across all stimuli by shuffling interspecies pairs of dissimilarities across stimuli (without replacement). This conserves the overall interspecies correlation and yields single-stimulus interspecies correlations that differ only because of the noise and limited data points (91 interspecies dissimilarity pairs for each stimulus). Each of 1000 null simulations yielded an interspecies correlation for each stimulus (1000 histograms shown in gray). We define a lower interspecies correlation threshold (dashed black line on the left), such that lower interspecies correlations occur for even a single stimulus in only 2.5% of the 1000 null simulations (i.e. by thresholding the randomization distributions of minima). Analogously, we define an upper threshold (dashed black line on the right), such that higher interspecies correlations occur for even a single stimulus in only 2.5% of the 1000 null simulations (i.e. by thresholding the randomization distributions of maxima). These two thresholds limit the family-wise false-positives rate at $p < 0.05$. Results show that human faces exhibit significantly higher interspecies correlations than the stimulus set as a whole and several stimuli (including images of animate and inanimate objects) exhibit significantly lower interspecies correlations. The two stimuli with the lowest interspecies correlation (eggplant, back-view of human head) were the only two stimuli described as ambiguous by human subjects during debriefing (Figure S5.1). Their significantly low interspecies correlation is consistent with the idea that the IT representation reflects not only the visual appearance, but also the conceptual interpretation of a stimulus.

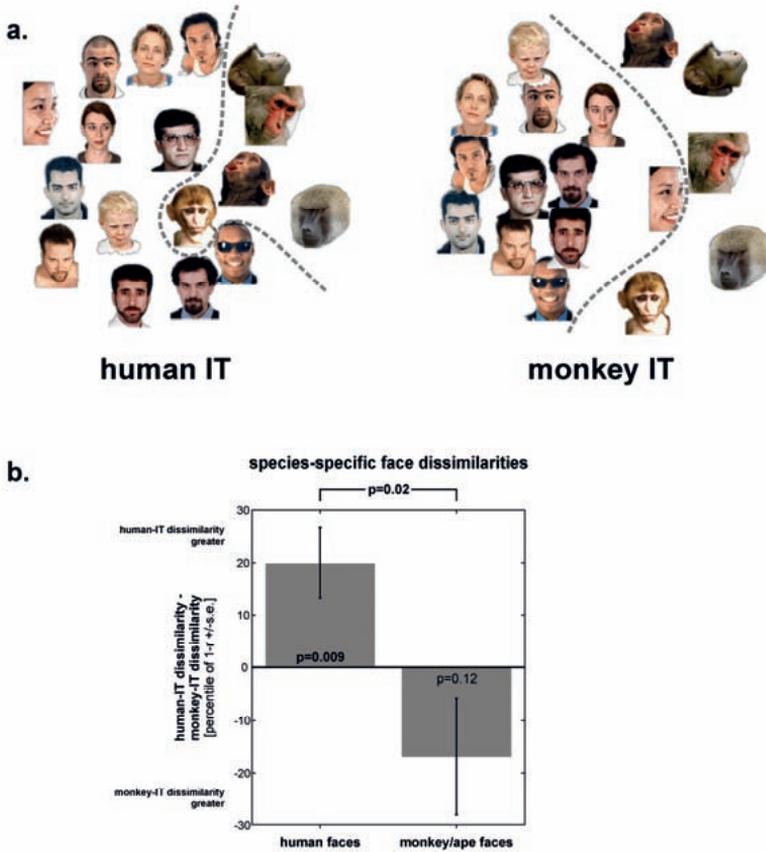


Figure S5.4 Species-specific face representation. Here we selectively analyzed the representation of monkey, ape, and human faces in monkey and human IT. **(a)** The face stimuli have been arranged such that their pairwise distances approximately reflect response-pattern similarity. The arrangement was computed by multidimensional scaling with the same settings as in Figure 5.2 and elsewhere in this paper (dissimilarity: $1 - \text{Pearson } r$, criterion: metric stress, arrangements scaled to match the areas of their convex hulls and rigidly aligned for easier comparison with the Procrustes method). A line (dashed gray) separating the monkey/ape faces from the human faces has been manually added. Visual inspection suggests that human IT may better discriminate the human faces than the monkey faces and that the converse may hold for monkey IT. **(b)** Statistical analysis comparing human- and monkey-IT mean dissimilarities for human faces (left) and for monkey/ape faces (right). The left bar shows that dissimilarities among human faces are significantly larger in human IT than in monkey IT ($p=0.009$). The right bar shows that dissimilarities among monkey/ape faces are larger in monkey IT than in human IT in our data, although the effect is not significant ($p=0.12$). The difference between the two effects is significant ($p=0.02$). Because the dissimilarities are not independent or normal, the statistical tests and error bars (indicate ± 1 standard error) are based on bootstrap resampling of the stimulus set. Note that our stimulus set is not well-suited for comparing the representation of human and monkey/ape faces, because faces were a small subset of our stimuli and because the monkey/ape faces were few and varied in species, pose, and view more than the human faces. The comparison in (b) of the representations of a given set of stimuli (either human faces or monkey faces) between human and monkey IT nevertheless provides an interesting lead for future studies designed to address this question.

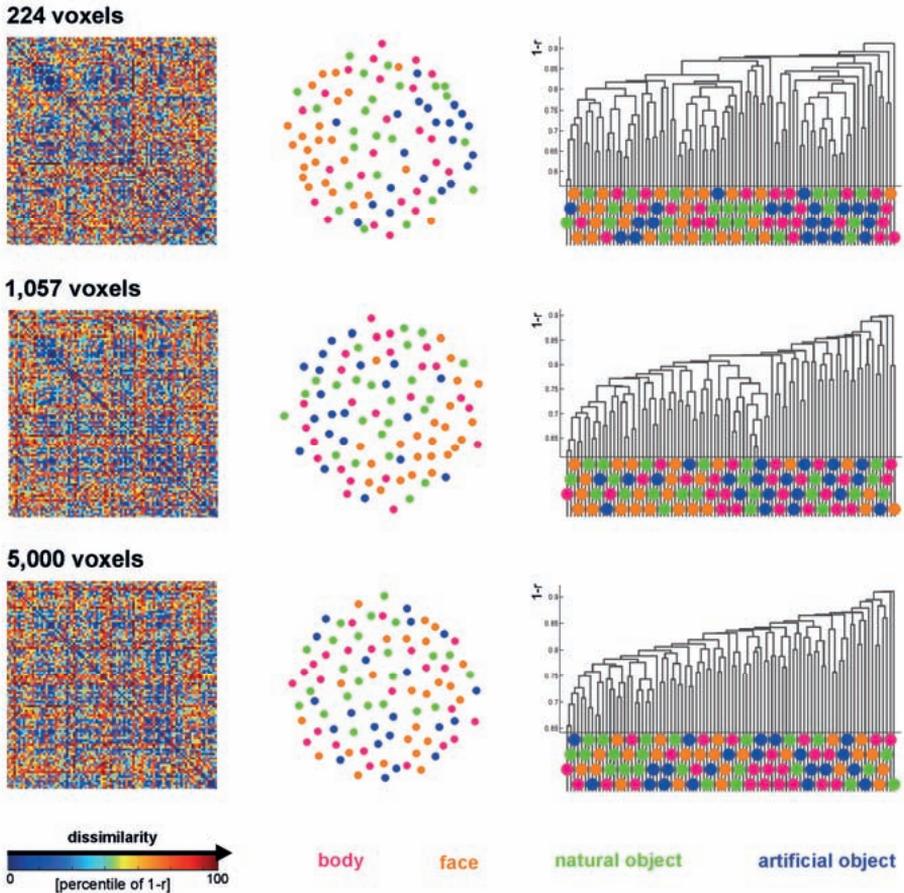


Figure S5.5 Representation in human early visual cortex defined at 224-5000 voxels. Human early visual cortex shows no evidence of categorical clustering in the fMRI data. This result is independent of the number of voxels included in the region of interest (rows). Early visual cortex was defined by selecting the most visually responsive voxels within a manually drawn anatomical mask in each subject. As for human IT, independent data were used for voxel selection. Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for IT (Figures 5.1, 5.2, 5.4, respectively).

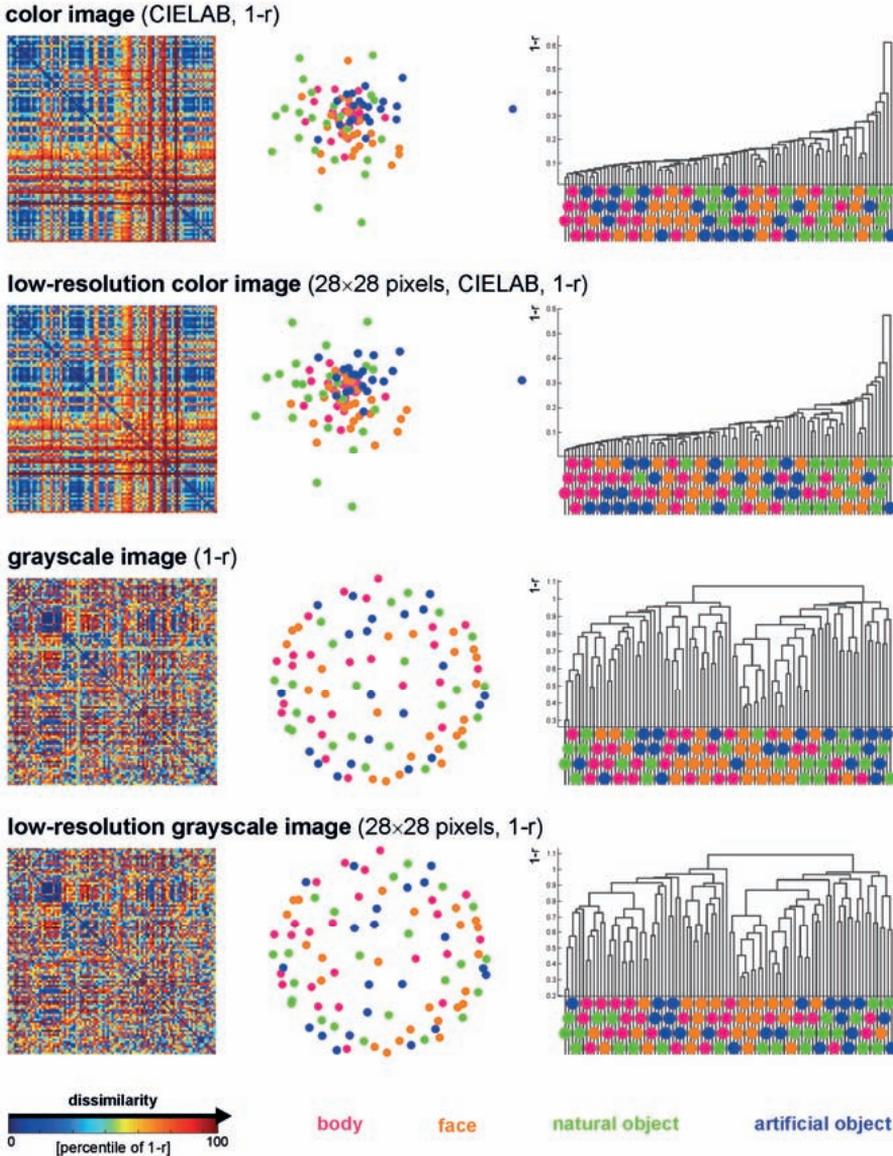


Figure S5.6 Model representations (1). We processed our stimuli to obtain their representations in a number of low-level models (rows, continued in Figure S5.7). We analyzed these model representations in the same way as the brain-activity data from early visual cortex and IT. None of the models could account for the categorical clustering found in monkey and human IT. The models are described in the section *Model representations* in the Supplementary Material. Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for IT (Figures 5.1, 5.2, 5.4, respectively).

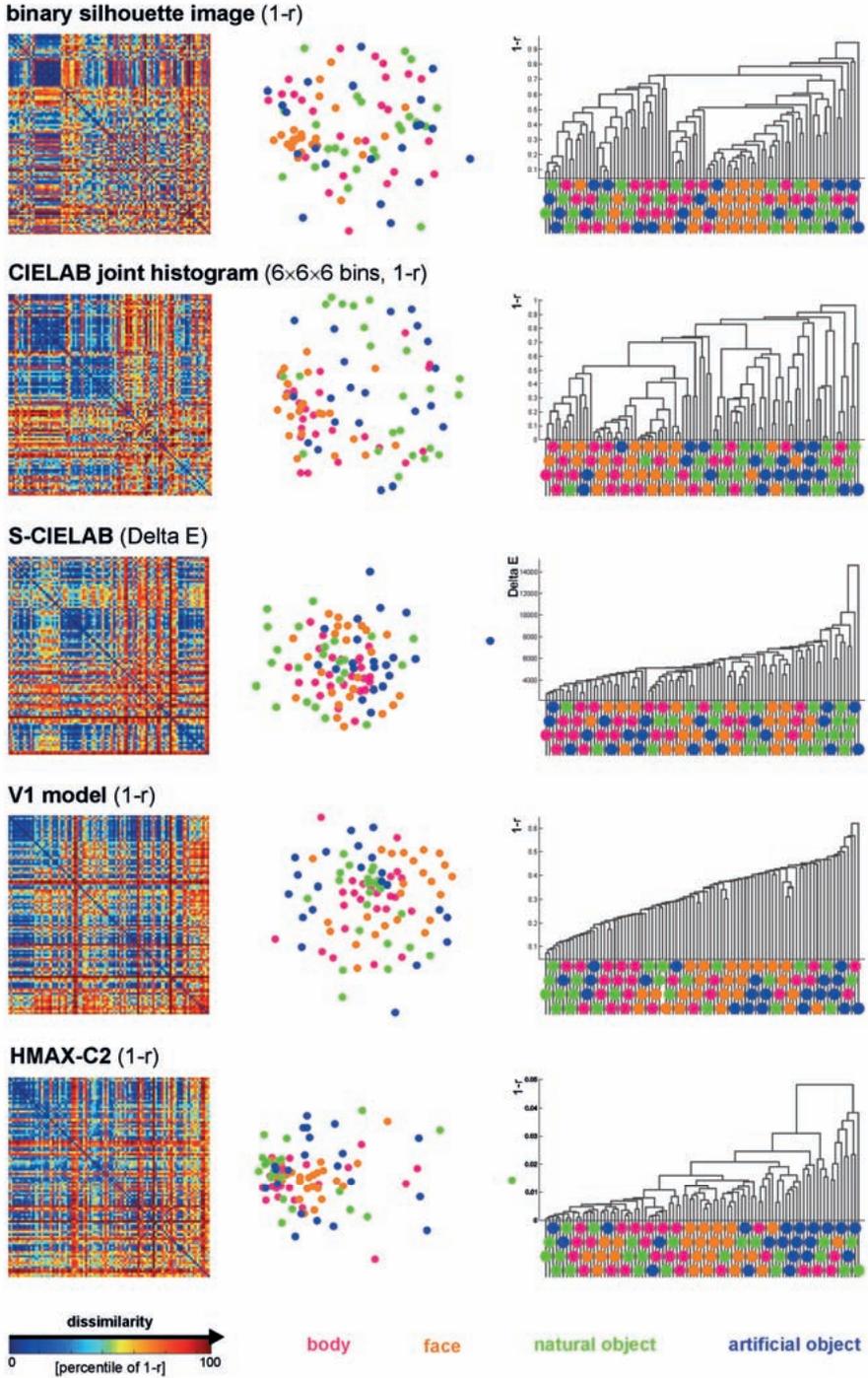


Figure S5.7 Model representations (2). Continuation of Figure S5.6. See legend of Figure S5.6.

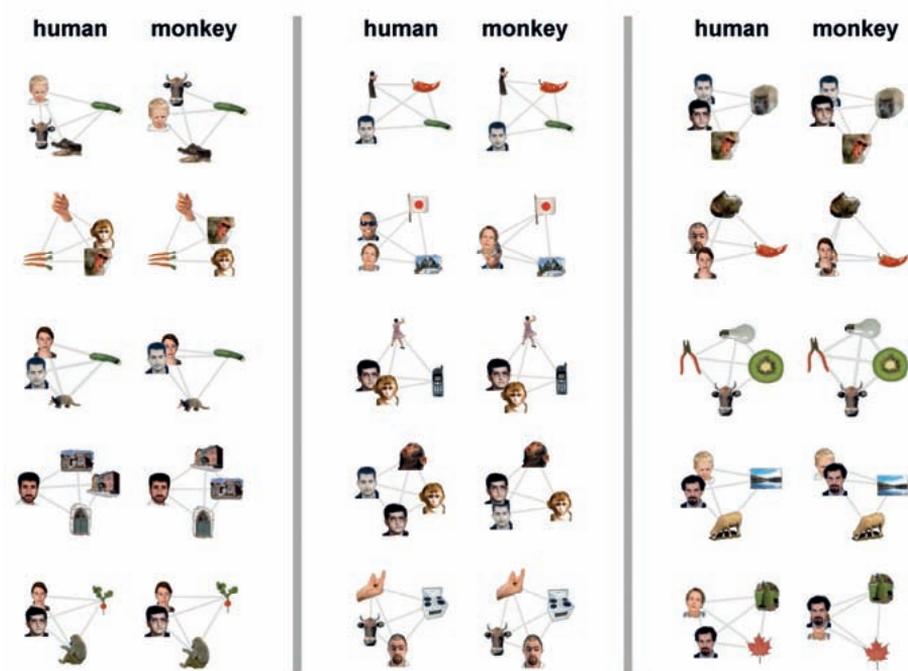


Figure S5.8 Unsupervised stimulus-quartet arrangements for monkey and human IT. For 15 representative stimulus quartets, this figure depicts the representation in human and monkey IT in terms of unsupervised arrangements reflecting response-pattern similarity. As in Figure 5.2, images placed close together elicited similar response patterns; images placed far apart elicited dissimilar response patterns. There is no special significance to the choice of 4 as the number of stimuli. However, considering quartets allows us to appreciate the underlying dissimilarity relationships at a glance. Moreover, the inevitable distortion of the original dissimilarities in the 2-dimensional arrangement is very small when only 4 stimuli are considered at a time. The arrangements were computed using multidimensional scaling with the same settings as for Figure 5.2 (dissimilarity: 1-Pearson correlation, criterion: metric stress). For each stimulus quartet, the two arrangements (human, monkey) have been rigidly aligned for easier comparison (Procrustes alignment) and scaled to the same approximate size. We introduce “rubberband graphs” (gray lines connecting the stimuli) to depict the residual distortion: the gray lines behave like rubberbands, thinning when stretched beyond the length they are to represent and thickening when compressed. More precisely, the actual dissimilarity equals the area of the rubberband connection (dissimilarity = line thickness \times line length).

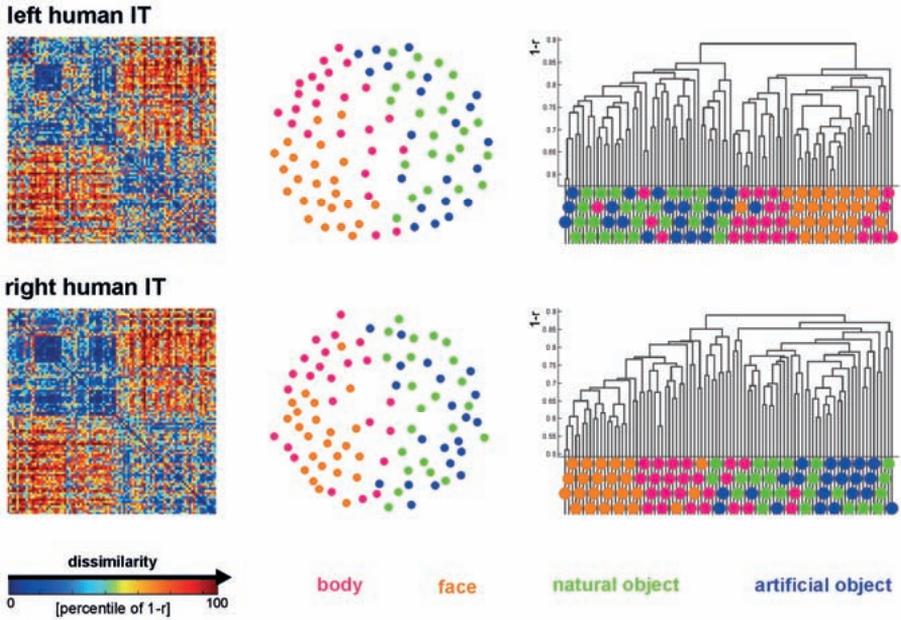
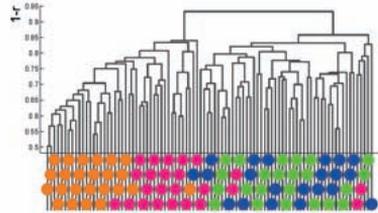
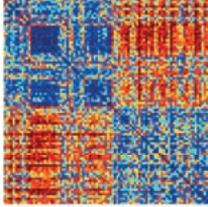
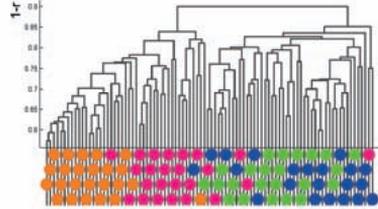
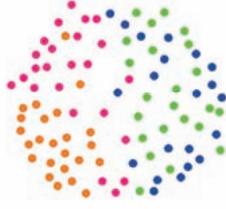
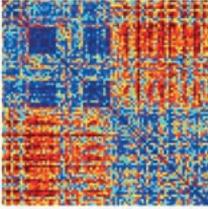


Figure S5.9 Representation in left and right human IT. The categorical clustering in human IT is only weakly dependent on the cortical hemisphere (left human IT in top row, right human IT in bottom row). Here we selected 266 voxels according to their visual responsiveness (independent data) within each hemisphere's manually defined IT mask. Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for Figures 5.1, 5.2, 5.4.

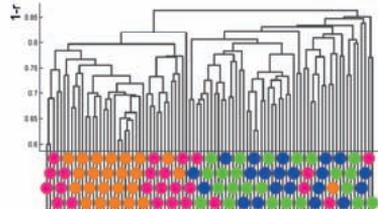
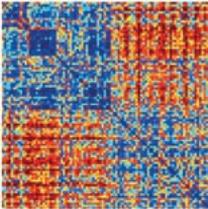
100 voxels



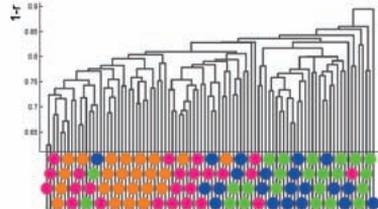
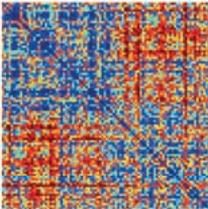
316 voxels



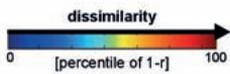
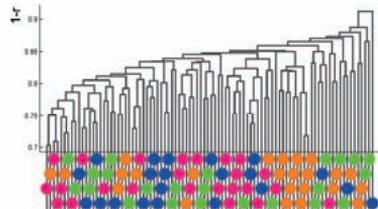
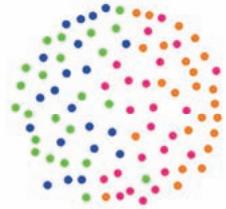
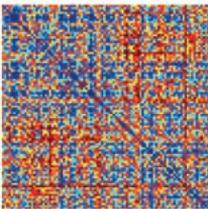
1,000 voxels



3,162 voxels



10,000 voxels



body

face

natural object

artificial object

Figure S5.10 Representation in human IT defined at 100-10,000 voxels. The similarity structure and categorical clustering characteristic of human IT is only weakly dependent on the number of voxels selected for inclusion in the region of interest. Voxels were selected according to their visual responsiveness as assessed with independent data. The human-IT region shown in the second row (316 voxels) is that used for Figures 5.1-5.6. When thousands of voxels are included in the region of interest, the categorical structure becomes less distinct. Nevertheless multidimensional scaling still separates animate and inanimate objects at 10,000 voxels (bottom row, middle panel). Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for Figures 5.1, 5.2, 5.4.

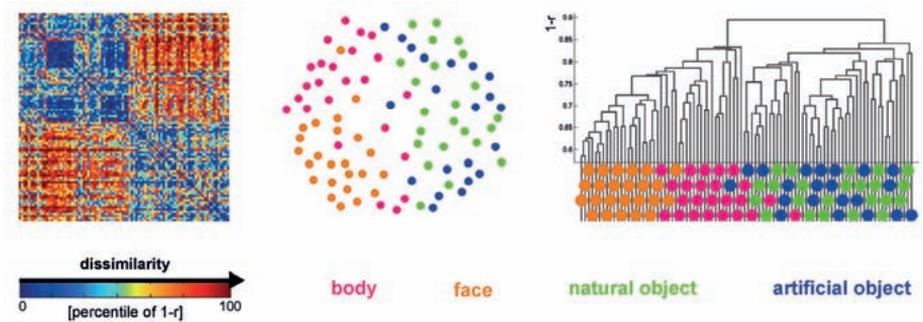


Figure S5.11 Representation in human IT without FFA and PPA. Excluding FFA and PPA bilaterally from the voxels selected to define human IT did not qualitatively change the similarity structure or categorical clustering. FFA and PPA were defined in each hemisphere at 1141 mm³ (150 voxels) and 1521 mm³ (200 voxels), respectively, by means of an independent block-design localizer experiment. Human IT was defined bilaterally at 316 voxels as for Figures 5.1-5.4, but FFA and PPA were first excluded from the cortex mask in both hemispheres. The dissimilarity matrix (left), multidimensional scaling arrangement (middle), and hierarchical clustering tree (right) were computed with the same parameters as for Figures 5.1, 5.2, 5.4.

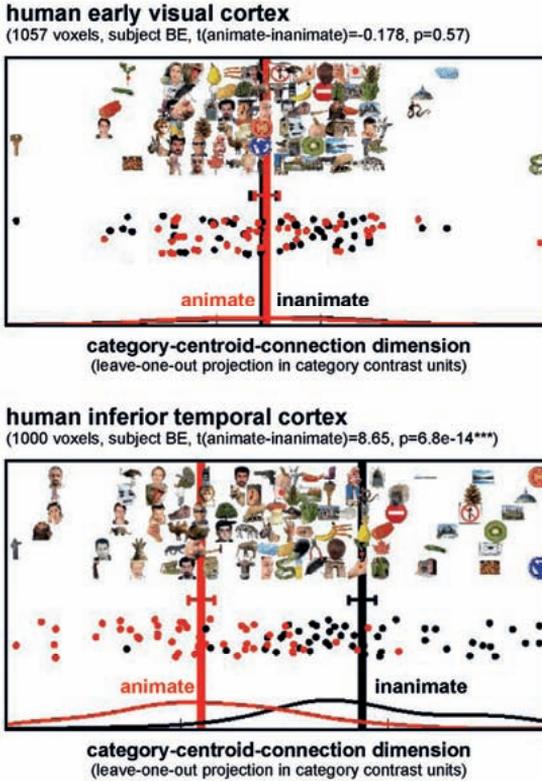


Figure S5.12 Response patterns in human IT, but not human early visual cortex, allow linear discrimination between animates and inanimates. Single-subject data show highly significant linear discriminability of animates and inanimates in human IT ($t(\text{animate-inanimate})=8.65$, $p=6.8e-14$), whereas linear discriminability of these categories is not evident in human early visual cortex ($t(\text{animate-inanimate})=-0.178$, $p=0.57$). The single-image response patterns in early visual cortex (top) and IT (bottom) have been projected onto a dimension defined to discriminate animates from inanimates. The dimension used is the category-centroid-connection dimension (equivalent to the Fisher linear discriminant computed with the assumption of isotropic, homoscedastic noise). To avoid circularity, the discriminant is computed using a leave-one-out procedure: In order to determine the location of a given single-image response pattern on the discriminant dimension, the other 95 single-image response patterns are used to compute the category centroids defining the discriminant (the thin tick marks indicate the centroid locations defining the discriminant; left for animates, right for inanimates). Note that this approach uses not only independent response estimates, but also different images (the other 95) for defining the discriminant. The thick vertical lines indicate the category means (red for animate, black for inanimate) on the discriminant dimension. Error bars indicate the ± 1 standard error of the mean. The 96 stimulus images (Figure S5.1) have been located along the discriminant (upper portion of each panel) with vertical scattering to allow a larger image size. Colored dots (red for animate, black for inanimate) show the two category distributions (middle portion of each panel), again with vertical random scattering. Probability-density estimates (kernel-smoother method) are shown in the lower portion of each panel (red for animate, black for inanimate).

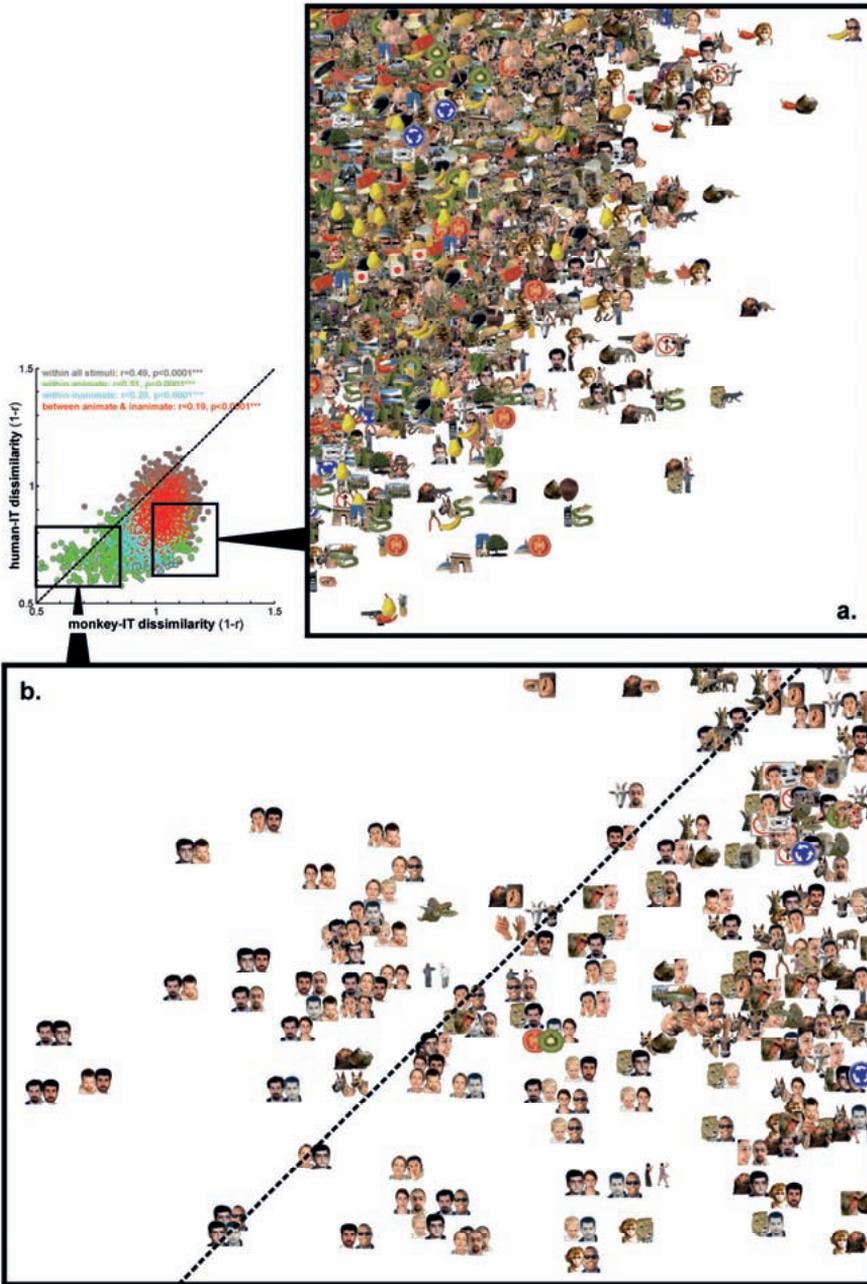


Figure S5.13 Scatterplot of stimulus-pairs relating monkey- and human-IT representations. The scatterplot in Figure 5.3a, containing a dot for each stimulus pair, relates monkey- and human-IT dissimilarities. Here we zoom in on two regions of Figure 5.3a, in order to make space for plotting each pair of stimuli, whose dissimilarity in monkey and human IT determines the horizontal and vertical location of the corresponding dot in the scatterplot. For each pair, the two stimuli are placed side by side, centered on the location of the dot that represents the pair in Figure 5.3a. **(a)**

This region contains the stimulus pairs that are furthest from the line of identity. They have the greatest difference between monkey- and human-IT dissimilarity ($1 - \text{Pearson } r$) with the monkey dissimilarity greater than the human dissimilarity. (Note that the corresponding region with greater human- than monkey-IT dissimilarities is unpopulated.) An attractive interpretation would be that these pairs are more similar to humans than to monkeys. Note, however, that the relationship between human and monkey representational dissimilarities may not be linear. Plausible interpretations suggest themselves for many of the pairs, but remain speculative. **(b)** This region contains the stimulus pairs eliciting the most similar activity patterns in both monkey and human IT. This region is dominated by pairs of human faces.

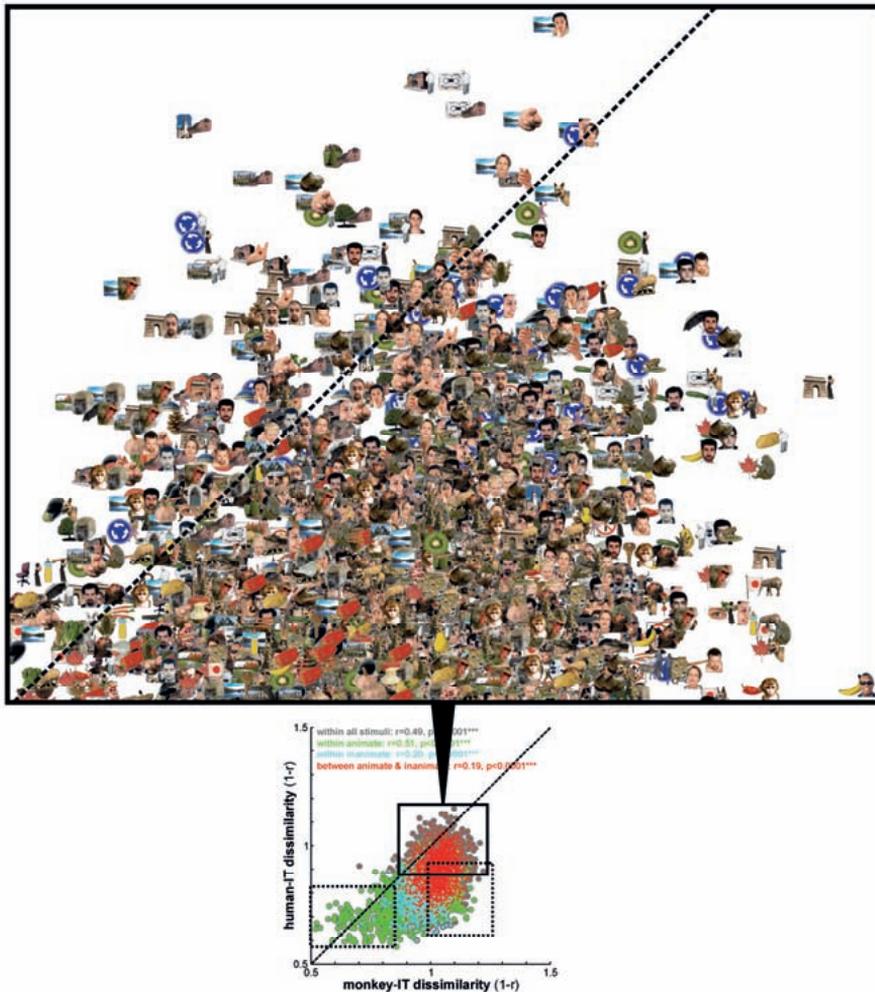


Figure S5.14 Scatterplot of stimulus-pairs relating monkey- and human-IT representations (continued). This figure follows the same logic as the previous one, but zooms in on a third region of Figure 5.3a. This region contains the stimulus pairs eliciting the most dissimilar activity patterns in both monkey and human IT. This region is dominated by pairs of stimuli crossing the animate-inanimate category boundary. (Dotted rectangles indicate the two regions zoomed in on in the previous Figure.)

5.5.2 Summary of the previous evidence on IT categoricity

Ever since neuropsychology described object-category-related deficits following brain damage (Humphreys and Forde, 2001; Capitani et al., 2003; Martin, 2007), it has been generally accepted that there is some relationship between conventional categories and human-IT representations. Human neuroimaging has investigated category-average responses, showing that IT contains focal regions whose activation is correlated with conventional categories (Puce et al., 1995; Martin et al., 1996; Kanwisher et al., 1997; Aguirre et al., 1998; Epstein and Kanwisher, 1998; Downing et al., 2001; Downing et al., 2006) and that the category can be read out from the IT response pattern with a linear classifier (Haxby et al., 2001; Cox and Savoy, 2003; Carlson et al., 2003). However, these studies investigated responses averaged across many different stimuli within predefined categories. This approach requires the assumption of a particular category structure and therefore cannot address whether the representation is inherently categorical. IT features might respond to natural image fragments that happen to be correlated with categories, without being optimized to distinguish categories.

Monkey studies, as well, have reported IT responses correlated with natural categories (Vogels, 1999; Tsao et al., 2003; Kiani et al., 2005; Hung et al., 2005; Tsao et al., 2006; Afraz et al., 2006) and novel, experimentally defined, categories (Sigala and Logothetis, 2002; Baker et al., 2002; Freedman et al., 2003). However, step-function-like categorical responses as reported for cells in medial temporal (Kreiman et al., 2000; Quiroga et al., 2005), prefrontal (Freedman et al., 2001), and parietal regions (Freedman and Assad, 2006) are not typically observed in either single IT cells (Vogels, 1999; Freedman et al., 2003; Kiani et al., 2007; but see Tsao et al., 2006) or category-sensitive fMRI responses (Haxby et al., 2001), suggesting that IT may have a lesser role in categorization (Freedman et al., 2003). Consistent with this perspective, one influential computational model of primate object recognition (Riesenhuber and Poggio, 2002; Serre et al., 2007) employs categorization training (i.e. supervised learning) to optimize the stage thought to correspond to prefrontal cortex. The model's IT stage is not optimized for categorization.

5.5.3 Representational similarity analysis

Estimation of single-image response patterns in the monkeys

The analyses are based on all cells that could be isolated and for which sufficient data was available across the stimuli. This yielded a total of 674 neurons for both monkeys combined (322 in monkey 1 and 352 in monkey 2). For each stimulus, each neuron's response amplitude was estimated as the average spike

rate within a 140-ms window starting 71 ms after stimulus onset (for details, see Kiani et al., 2007).

Estimation of single-image response patterns in the humans

Single-image response patterns were estimated by univariate linear modeling. We concatenated the runs within a session along the temporal dimension. For each voxel, we performed a single univariate linear model fit to obtain a response-amplitude estimate for each of the 96 stimuli. The model included a hemodynamic-response predictor for each of the 96 stimuli. Since each stimulus occurred once in each run, each of the 96 predictors had one hemodynamic response per run and extended across all within-session runs included (i.e. all runs except those used for region-of-interest definition). The predictor time courses were computed using a linear model of the hemodynamic response (Boynton et al., 1996) and assuming an instant-onset rectangular neural response during each condition of visual stimulation. For each run, the design matrix included these stimulus-response predictors along with six head-motion-parameter time courses, a linear-trend predictor, a 6-predictor Fourier basis for nonlinear trends (sines and cosines of up to 3 cycles per run) and a confound-mean predictor. Trends were, thus, modeled by a separate set of predictors for each run. The trend predictors for a particular run had zero entries for all other runs along time. For head-motion models and confound means as well, separate predictors accounted for each run. For each stimulus, we converted the response-amplitude (beta) estimate map into a t map. The resulting t maps (one for each stimulus) were used for representational similarity analysis.

Computation of representational dissimilarity matrices

For each pair of stimuli, the dissimilarity between the associated response patterns is measured as 1 minus the Pearson linear correlation across cells or voxels within a region of interest (0 for perfect correlation, 1 for no correlation, 2 for perfect anticorrelation). The resulting dissimilarities for all pairs of object images are assembled in a representational dissimilarity matrix (RDM; Figure 5.1). Each cell of the RDM, thus, compares the response patterns elicited by two stimuli. As a consequence, an RDM is symmetric about a diagonal of zeros.¹⁵

¹⁵ Alternatively, two separate data sets can be used, such that the vertical dimension of the RDM indexes pattern estimates from data set 1 and the horizontal dimension indexes pattern estimates from data set 2. The diagonal then contains dissimilarity estimates for replications of the same condition. Moreover, for each pair of conditions, two entries symmetrical about the diagonal of the RDM contain separate estimates of the pattern dissimilarity. We use the single-pattern-set approach here, because it provides more data along time for each fMRI pattern estimate. The overlapping hemodynamic responses to the stimuli are more precisely estimated when more runs are available not only by a factor of \sqrt{n} , where n is the number of runs, but by a larger factor, because longer

Testing relatedness of two representational dissimilarity matrices by randomization of condition labels

We use the Pearson correlation coefficient r to assess the relatedness of two RDMs (e.g. Figures 5.3, S5.2). The correlation is restricted to the upper (or equivalently the lower) triangle of each RDM. In order to decide whether two RDMs are related, we perform statistical inference on the RDM correlation. The classical method for testing correlations assumes independent pairs of measurements for the two variables. For RDMs such independence cannot be assumed, because each dissimilarity is dependent on two response patterns, each of which also codetermines the dissimilarities of all its other pairings in the RDM. We therefore test the relatedness of RDMs by randomization (e.g. Nichols and Holmes 2002). In particular, we use randomization of the condition labels to rearrange the rows and columns of an RDM. We choose a random permutation of the conditions (i.e. of the 92 stimuli), reorder rows and columns of one of the two RDMs to be compared according to this permutation, and compute the correlation. By repeating this step 10,000 times, we obtain a distribution of correlations simulating the null hypothesis that the two RDMs are unrelated. If the actual correlation (i.e. the one for consistent labeling between the two RDMs) falls within the top 5% of the simulated null distribution of correlations, we reject the null hypothesis of unrelated RDMs. More generally, we estimate the p value as the percent rank/100 of the actual correlation in the randomization distribution. The percent rank is conservatively estimated, such that $p < 0.0001$ indicates that the actual correlation was higher than any of the 10,000 correlations obtained after randomization of the condition labels.

The condition-label randomization test is justified by the random assignment of conditions (i.e. stimuli) to experimental trials. Under the null hypothesis of no relation between the RDMs, the conditions are exchangeable, i.e. the true labeling and each random relabeling of one RDM yield correlations drawn from the same distribution (i.e. from the null distribution). More generally, condition-label randomization can be used to test various RDM statistics against the null hypothesis that the condition labels are interchangeable (in the sense of not affecting the true test statistic). Note that the central results of this paper all rely on the condition-label randomization test (Figures 5.3, 5.5, 5.6, S5.3a), but condition-bootstrap resampling has been used to test additional hypotheses (Figures 5.5, S5.4).

random sequences more closely approximate the ideal of uncorrelated hemodynamic-response predictors.

Testing statistics of representational dissimilarity matrices by bootstrap resampling of the set of conditions

Not all hypotheses about RDMs can be tested by randomization. For example, the randomization test cannot be used to assess whether the mean of an RDM is greater than some constant, because the mean will be the same for each relabeling. Moreover, by condition-label randomization we test the null hypothesis that the condition labels are interchangeable. This may not be the desired null hypothesis. A less rigorous, but more versatile approach is bootstrap resampling (Efron and Tibshirani, 1993), which we apply here to the set of experimental conditions (i.e. the stimuli), in order to simulate a distribution of RDMs. Like the randomization test, the bootstrap test is appropriate for RDMs in that it does not rely on either distributional assumptions or the assumption of independence of the dissimilarity estimates. As mentioned above, tests assuming independent data may not be valid for RDMs, because dissimilarities within an RDM have a complex dependency structure. Like the condition-label randomization described above, the bootstrap resampling here operates at the level of the experimental conditions and, thus, simulates the dependency structure of RDMs.

The condition bootstrap test proceeds by resampling the set of conditions with replacement: If there are n_c condition labels, we draw n_c times from the whole set, to obtain a set of n_c labels. The bootstrap set of condition labels may include multiple instances of some labels and exclude others altogether. We construct a bootstrap RDM by resampling the original RDM according to the bootstrap sample of condition labels. We then compute the statistic of interest (e.g. the mean of all dissimilarities). We repeat this process many times (e.g. 10,000 times) to obtain a bootstrap distribution for the statistic.

The bootstrap resampling can be stratified in order to compare sets of conditions. For example, in order to test whether between-category dissimilarities are greater than within-category dissimilarities (Figure 5.5), we separately bootstrap resampled the animates set and the inanimates set, recomputing mean B of between-category dissimilarities and the mean W of within-category dissimilarities to obtain the test statistic B minus W.

The bootstrap distribution of the statistic allows us to obtain error bars on arbitrary RDM statistics: the standard deviation of the bootstrap distribution is the standard error of the estimate of the statistic. In addition, we can define a 95% confidence interval by excluding 5% of the extreme values in the distribution (either on one side for a one-sided test or on both sides symmetrically for a two-sided test). If the value assigned to the test statistic by the null hypothesis falls outside the confidence interval, the null hypothesis is rejected. This is a valid test, because the confidence interval will include the null value with 95% prob-

ability given that the null hypothesis is true and assuming that the bootstrap resampling is an accurate simulation. The latter assumption is questionable, therefore the bootstrap procedures we describe here should be considered rough, approximate methods of inference.

A further complication of this method is that the bootstrap resampling of the set of condition labels moves zeros from the diagonal into the off-diagonal parts of the RDM whenever a condition is selected multiple times in the bootstrap resampling. (For 96 conditions, this is a small proportion of the entries: on the order of 1%.) In order to prevent these zeros from biasing the statistic, we exclude them before computing the statistic. For the purpose of condition bootstrapping, it may be preferable to use two data sets for computing the RDM (as suggested above), such that each diagonal value reflects a dissimilarity between two replications of the same response pattern.

The rationale for bootstrap resampling of a set of experimental conditions is to simulate the distribution of the statistic of interest that we expect to obtain for repetitions of the experiment performed with the same subjects but with different experimental conditions (e.g. stimuli) drawn from the same population of possible conditions that could have been used for the experiment (e.g. stimuli from the same categories). An interesting feature of this approach is its potential to generalize from the set of conditions actually used in the experiment to the hypothetical population of conditions, of which the actually chosen conditions can be considered a random sample. Note, however, that a sufficient number of conditions is required for this and the accuracy of the simulation is not guaranteed. For caveats and advanced bootstrap methods, see Hesterberg (2007).

5.5.4 Human localizer experiments and definition of regions of interest

Definition of regions of interest

All regions of interest (ROIs) were defined on the basis of independent experimental data and restricted to a cortex mask manually drawn on each subject's fMRI slices. Human IT was defined by selecting a variable number of voxels (316 voxels in Figures 5.1-5.6; 100-10,000 voxels in Figure S5.11) within the inferior temporal portion of the bilateral cortex mask according to their visual responsiveness. Visual responsiveness was assessed using the t map for the average response to the 96 object images. The t map was computed on the basis of one third of the runs of the main experiment within each session. The remaining runs were used to perform all further analyses. To define early visual cortex, we selected the most visually responsive voxels, as for IT, but within a manually defined anatomical region around the calcarine sulcus within the cortex mask (Figures 5.5, 5.6, S5.5). For control analyses (Figure S5.11), we defined the FFA

(Kanwisher et al., 1997) using the contrast faces minus objects, and the PPA (Epstein and Kanwisher, 1998) using the contrast places minus objects in analyzing the separate localizer block-design experiment described below.

Localizer block-design experiment

We performed a functional localizer experiment using the same fMRI sequence as for the main experiment and a separate set of stimuli. Subjects viewed grayscale photos of faces, places, and objects (spanning a visual angle of about 5.7°) in category blocks. Each block lasted 30 s (stimulus-onset asynchrony: 1 s; stimulus duration: 700 ms), alternating with 20-s fixation blocks. Three blocks were presented for each stimulus category (face, place, object), resulting in a total run duration of 7 min and 50 s. Stimuli were presented on a constantly visible uniform black background. Subjects continually fixated a central white cross and performed a one-back repetition-detection task on the images, responding with a left-thumb button press for each consecutive repetition (3 to 5 repetitions per block). Each stimulus was only presented once, except for the immediate repetitions to be detected in the one-back task. Stimuli were centered with respect to the fixation cross.

5.5.5 Model representations

We processed our stimuli to obtain their representations in a number of low-level models. We then analyzed these model representations (Figures S5.6, S5.7) in the same way as the brain-activity data from early visual cortex and IT (Figures 5.1, 5.2, 5.4, S5.5, S5.9-S5.11). Each image was converted to a representational vector as described below for each model. As for the brain-activity data, each representational vector was then compared to each other representational vector by means of $1-r$ as the dissimilarity measure (where r is the Pearson linear correlation; only for the S-CIELAB representation, the conventional Delta E measure was used instead of $1-r$) to obtain a representational dissimilarity matrix, on which further analyses (Figures S5.6, S5.7) were based.

Color image (CIELAB)

The RGB color images (175×175 pixels) were converted to the CIELAB color space, which approximates a linear representation of human perceptual color space. Each CIELAB image was then converted to a pixel vector (175×175×3 numbers).

Low-resolution color image (28×28 pixels, CIELAB)

The RGB color images (175×175 pixels) were downsampled to 28×28 pixels (with bicubic interpolation) and subsequently converted to the CIELAB color

space. Each 28×28 CIELAB image was then converted to a pixel vector ($28 \times 28 \times 3$ numbers).

Grayscale image

The RGB color images (175×175 pixels) were converted to grayscale. Each grayscale image was then converted to a pixel vector (175×175 numbers).

Low-resolution grayscale image (28×28 pixels)

The RGB color images (175×175 pixels) were converted to grayscale and subsequently downsampled to 28×28 pixels (with bicubic interpolation). Each grayscale image was then converted to a pixel vector (28×28 numbers).

Binary silhouette image

The RGB color images (175×175 pixels) were converted to binary silhouette images, in which all background pixels had the value 0 and all figure pixels had the value 1. Each binary silhouette image was then converted to a pixel vector (175×175 binary numbers).

CIELAB joint histogram (6×6×6 bins)

The RGB color images (175×175 pixels) were converted to the CIELAB color space. The three CIELAB dimensions (L, a, b), were then divided into 6 bins of equal width. The joint CIELAB histogram was computed by counting the number of figure pixels (gray background left out) falling into each of the $6 \times 6 \times 6$ bins. The joint histogram was converted to a vector ($6 \times 6 \times 6$ numbers).

S-CIELAB (Delta E)

The RGB color images (175×175 pixels, 2.9° visual angle) were compared (each to each other, to obtain a representational dissimilarity matrix) by means of the S-CIELAB Delta-E dissimilarity measure (Zhang and Wandell, 1997), which models the perceptual similarity of color images and, unlike CIELAB Delta E, accounts for pattern-color sensitivity results (Poirson and Wandell, 1993) by separating the image into components corresponding to different spatial-frequency bands.

V1 model

The RGB color images (175×175 pixels, 2.9° visual angle) were converted to grayscale and given as input to population of modeled V1 simple and complex cells (Lampl et al., 2004; Riesenhuber and Poggio, 2002; Kiani et al., 2007). The receptive fields (RFs) of simple cells were simulated by Gabor filters of 4 different orientations (0° , 90° , -45° and 45°) and 12 sizes (7-29 pixels). Cell RFs were

distributed over the stimulus image at 0.017° intervals in a cartesian grid (for each image pixel there was a simple and a complex cell of each selectivity that had its RF centered on that pixel). Negative values in outputs were rectified to zero. The RFs of complex cells were modeled by the MAX operation performed on outputs of neighboring simple cells with similar orientation selectivity. The MAX operation consists in selecting the strongest (maximum) input to determine the output. This renders the output of a complex cell invariant to the precise location of the stimulus feature that drives it. Simple cells were divided into four groups based on their RF size (7-9 pixels, 11-15 pixels, 17-21 pixels, 23-29 pixels) and each complex cell pooled responses of neighboring simple cells in one of these groups. The spatial range of pooling varied across the four groups (4×4 , 6×6 , 9×9 , and 12×12 pixels for the four groups, respectively). This yielded 4 (orientation selectivities) $\times 12$ (RF sizes) = 48 simple-cell maps and 4 (orientation selectivities) $\times 4$ (sets of simple-cell RF sizes pooled) = 16 complex-cell maps of 175×175 pixels. All maps of simple and complex cell outputs were vectorized and concatenated to obtain a representational vector for each stimulus image.

HMAX-C2 model based on natural image fragments

This model representation developed by Serre et al. (2005) builds on the complex-cell outputs of the V1 model described above (implemented by the same group). The C2 features used in the analysis (Figure S5.7) may be comparable to those found in primate V4 and posterior IT. The model has four sequential stages: S1-C1-S2-C2. The first two stages correspond to the simple and complex cells described above, respectively. Stages S2 and C2 use the same pooling mechanisms as stages S1 and C1, respectively. Each unit in stage S2 locally pools information from the C1 stage by a linear filter and behaves as a radial basis function, responding most strongly to a particular prototype input pattern. The prototypes correspond to random fragments extracted from a set of natural images (stimuli independent of those used in the present study). S2 outputs are locally pooled by C2 units utilizing the MAX operation for a degree of position and scale tolerance. A detailed description of the model (including the parameter settings and map sizes we used here) can be found in Serre et al. (2005). The model, including the natural image fragments, was downloaded from the author's website in January 2007 (for the current version, see <http://cbcl.mit.edu/software-datasets/standardmodel/index.html>).

Additional model representations

In addition to the models described above and analyzed for our stimulus set in Figures S5.6 and S7, we tried a number of additional models (not shown). These included (1) low-passed and high-passed grayscale representations, (2) a version of the V1 model described above, in which we averaged all simple and

complex cell responses representing the same retinal location (averaging also across orientation selectivities and RF sizes) in order to mimic the effect of downsampling by population averaging within fMRI voxels, and (3) higher-level shape-tuned units created within the HMAX model framework (Riesenhuber and Poggio, 2002) as described in Kiani et al. (2007). None of these model representations exhibited categorical clustering.

Chapter 6

Human object-similarity judgments reflect and transcend primate-IT categorical object representations

Little is known about the relationship between object representations in high-level visual cortex and perceived object similarity. Using representational similarity analysis, we show that objects that are perceived as similar tend to elicit similar activity patterns in inferior temporal cortex (IT). Human similarity judgments cluster the objects by category and reflect several categorical distinctions that characterize the IT representation in both man and monkey, including the basic distinction between animate and inanimate objects. Human similarity judgments and IT object representations also show a match within category clusters. However, the similarity judgments show stronger categorical clustering and a slightly different category hierarchy, which additionally emphasizes the human/non-human distinction among animate stimuli and the natural/artificial distinction among inanimate stimuli. Our findings suggest that object-similarity judgments reflect the IT object representation, but also transcend the IT representational stage, in terms of a stronger categorical component and the introduction of species-dependent (human/non-human) and evolutionarily recent (natural/artificial) distinctions. These additional distinctions may reflect a prefrontal contribution allowing more flexible categorical distinctions.

Mur M, Meys M, Bodurka J, Goebel R, Bandettini P, Kriegeskorte N. Human object-similarity judgments reflect and transcend primate IT categorical object representations. In revision, *Front Psychology*.

6.1 Introduction

Perceived objects and their relationships are thought to be represented in conceptual spaces at the cognitive level (Gärdenfors, 2000). A conceptual space can be seen as analogous to the spatial environment that we live in: both are geometrical structures in which the location of an object is determined by its values on a set of dimensions. The difference lies in the dimensions that define the space: for our spatial environment, the location of an object can be specified by three spatial coordinates (x , y , and z dimensions); for a conceptual space, the dimensions could be any object property, including its perceived color, shape, or semantic category. The location of an object in the conceptual space is interpreted as the mental representation of that object. Distances between object representations inform us about their relationships: the higher the perceived similarity between two objects, the closer their representations will be in space.

Object representations are thought to be implemented in the brain by means of distributed activity patterns (McClelland and Rogers, 2003; Haxby et al., 2001). In line with this idea, distributed activity patterns in human inferior temporal cortex (hIT) - a large patch of object-selective cortex located in the ventral visual stream - contain information about category membership of visual objects (Cox and Savoy, 2003; Haxby et al., 2001). Moreover, when real-world objects are grouped based on similarity of the activity patterns that they elicit in IT, this results in category clusters corresponding to well-known object categories, including animate and inanimate objects and, within the animates, faces and bodies (Kiani et al., 2007; Kriegeskorte et al., 2008a). The emergence of these category clusters from human and monkey IT data indicates that object representations in primate IT are inherently categorical. Furthermore, the fact that major category clusters (e.g. animates) contained smaller clusters (e.g. faces and bodies) indicates that the category clusters are organized in a hierarchical fashion.

The presence of clusters that correspond to well-known object categories is consistent with a link to human perception (see also Edelman et al., 1998). Recent studies have suggested a relationship between perceived similarities and activity-pattern similarities in primate object-selective cortex for abstract visual shapes (Haushofer et al., 2008; Kayaert et al., 2005; Op de Beeck et al., 2008). However, these studies have not investigated the human perceptual similarity representation of real-world object images and its relation to the inherently categorical IT representation. Do human object-similarity judgments reflect the IT object space, including its hierarchy of category clusters?

To address this question, we measured the perceptual and hIT similarity representation of 96 colored photos of isolated objects, spanning a wide range of

object categories, including faces and bodies (subset of the stimuli used in Kiani et al., 2007) (Figure 6.1). The hIT similarity representation was based on single-image fMRI activity patterns across IT in four human observers and has been described previously (Kriegeskorte et al., 2008a). The perceptual similarity representation was based on object-similarity judgments from sixteen different human subjects, acquired with a novel multi-arrangement (MA) method. The MA method enables efficient and subject-tailored measurement of perceived similarity for large sets of objects.

The human object-similarity judgments acquired with the MA method were related to the hIT activity patterns using representational similarity analysis (RSA), which enables quantitative comparison of data from different branches of systems neuroscience (Kriegeskorte et al., 2008b). The core concept of RSA is the representational dissimilarity matrix (RDM) (Figures 6.2, 6.3a), which serves to capture the similarity representation of a specific set of experimental conditions (e.g. our 96 object images). The similarity representation is captured by all possible pairwise dissimilarities between the 96 object images, which are stored in the RDM. The perceptual and hIT similarity representations were related to each other by correlating the perceptual with the hIT dissimilarities. A significant correlation would indicate that the two representations match better than expected by chance.

Within the same framework, we additionally related the human similarity judgments to computational models of varying complexity, and to brain-activity measurements from visual regions other than IT, including early visual cortex. The tested models included simple models that were based on specific image features (e.g. color) and more complex computational models simulating brain information processing of visual input (e.g. HMAX model, Serre et al., 2005). Results of these additional analyses were used as a reference frame for evaluating the match between similarity judgments and hIT activity patterns.

6.2 Materials and methods

6.2.1 fMRI experiment

Subjects

Four healthy human volunteers participated in the fMRI experiment (mean age = 35 years; two females). Subjects were right-handed and had normal or corrected-to-normal vision. Before scanning, the subjects received information about the procedure of the experiment and gave their written informed consent

for participating. The experiment was conducted in accordance with the Institutional Review Board of the National Institutes of Mental Health, Bethesda, MD.

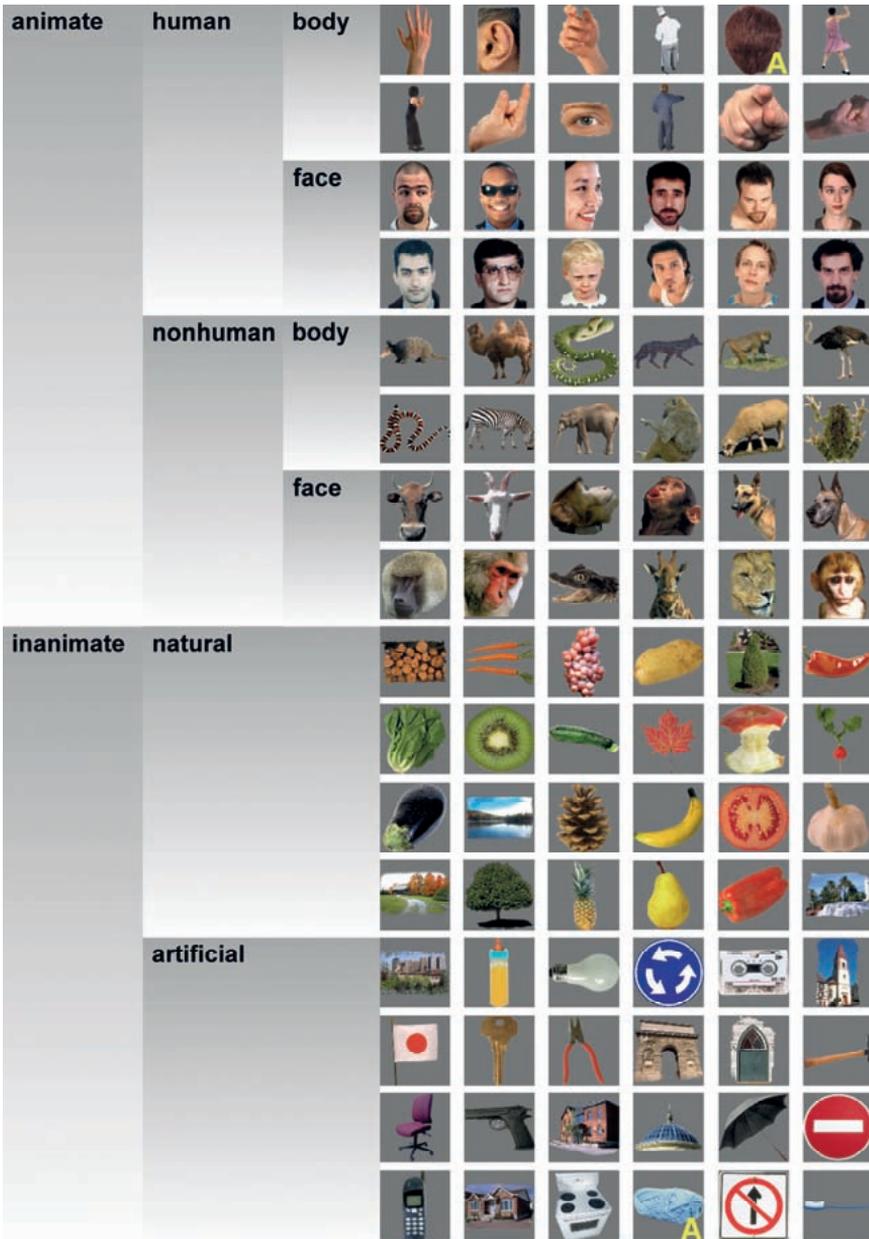


Figure 6.1 Stimuli. This figure shows the object images that we presented to our subjects. Two stimuli were described as ambiguous by several of our subjects during debriefing. These stimuli (back of a human head, knitting wool) are marked with a yellow “A”.

Experimental design and task

Stimuli were presented using a rapid event-related design (stimulus duration, 300 ms; interstimulus interval, 3700 ms) while subjects performed a fixation-cross-color detection task. Stimuli were displayed on a uniform gray background at a width of 2.9° visual angle. Each of the 96 object images was presented once per run. Subjects participated in two sessions of six nine-minute runs each. In addition, subjects participated in a separate block-localizer experiment. Stimuli (grayscale photos of faces, objects, and places) were presented in 30-s category blocks (stimulus duration, 700 ms; interstimulus interval 300 ms). Subjects performed a one-back repetition-detection task on the images.

Functional magnetic resonance imaging

Blood-oxygen-level-dependent fMRI measurements were performed at high resolution (voxel volume: $1.95 \times 1.95 \times 2$ mm³), using a 3 Tesla General Electric HDx MRI scanner, and a custom-made 16-channel head coil (Nova Medical Inc.). We acquired 25 axial slices that covered inferior temporal (IT) and early visual cortex bilaterally (single-shot, gradient-recalled Echo Planar Imaging; matrix size: 128 x 96, TR: 2s, TE: 30ms, 272 volumes per run, SENSE acquisition).

Estimation of single-image activity patterns

fMRI data were preprocessed in BrainVoyager QX (Brain Innovation) using slice-scan-time correction and head-motion correction. All further analyses were conducted in Matlab (The MathWorks Inc.). Single-image activity patterns were estimated for each session by voxel-wise univariate linear modelling (using all runs except those used for region-of-interest definition). The model included a hemodynamic-response predictor for each of the 96 stimuli along with run-specific motion, trend and confound-mean predictors. For each stimulus, we converted the response-amplitude (beta) estimate map into a t map.

Definition of regions of interest

All regions of interest (ROIs) were defined on the basis of independent experimental data and restricted to a cortex mask manually drawn on each subject's fMRI slices. Human IT was defined by selecting the 316 most visually responsive voxels within the inferior temporal portion of the cortex mask. Visual responsiveness was assessed using the t map for the average response to the 96 object images. The t map was computed on the basis of one third of the runs of the main experiment within each session. To define early visual cortex (EVC), we selected the 1057 most visually responsive voxels, as for IT, but within a manually defined anatomical region around the calcarine sulcus within the cortex mask. The fusiform face area (FFA) (Kanwisher et al., 1997) and parahippocam-

pal place area (PPA) (Epstein and Kanwisher, 1998) were defined based on the separate block-localizer experiment. The FFA was defined by the contrast faces minus objects and places; the PPA was defined by the contrast places minus objects and faces. Each of the four resulting unilateral regions contained 128 voxels.

Computation of the representational dissimilarity matrix (RDM)

For each ROI, we extracted a multivoxel pattern of activity (t map) for each of the 96 stimuli. For each pair of stimuli, activity-pattern dissimilarity was measured as 1 minus the Pearson linear correlation across voxels within the ROI (0 for perfect correlation, 1 for no correlation, 2 for perfect anticorrelation). The resulting 4560 pairwise dissimilarity estimates were stored in an RDM. RDMs were computed for each subject and session separately and then combined into a group RDM by averaging. The group RDM was used for comparison to the similarity judgements.

6.2.2 Object-similarity judgments

Subjects

Sixteen healthy human volunteers (mean age = 28 years; twelve females) participated in the multi-arrangement experiment. Subjects had normal or corrected-to-normal vision; thirteen of them were right-handed. Before participating, the subjects received information about the procedure of the experiment and gave their written informed consent for participating. The experiment was conducted in accordance with the faculty ethics committee.

Multi-arrangement method

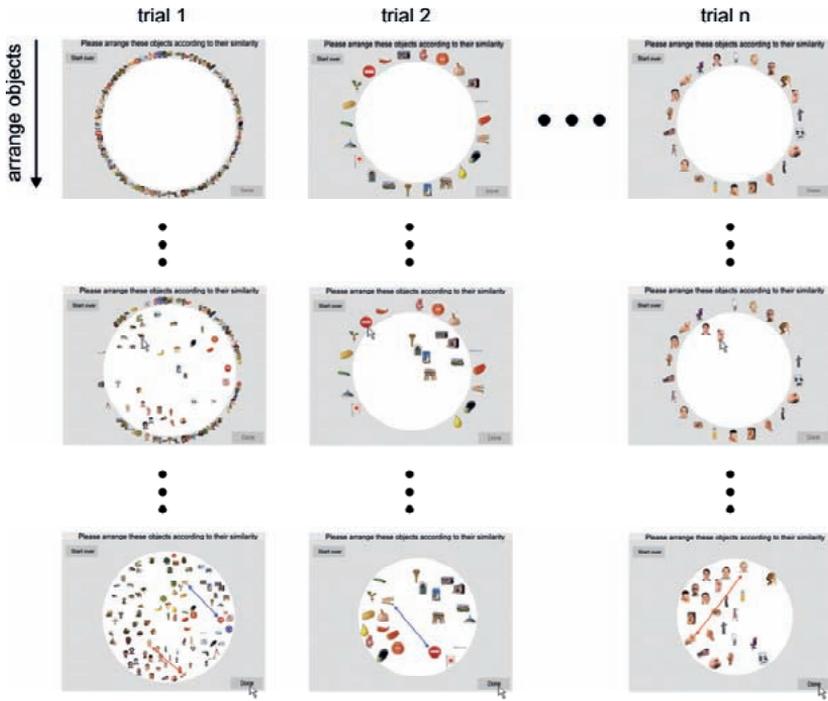
Perceived object similarity is conventionally measured using pairwise similarity ratings (e.g. Cortese and Dyre, 1996; Cooke et al., 2007). Given the large number of object-pair similarities to be measured in our study (96 objects, 4560 possible pairs), acquiring pairwise similarity ratings, or any other measure that considers each possible pair of objects separately, would be practically difficult. Data acquisition would require many hours and multiple sessions. Moreover, subjects might change their implicit criteria when rating pairwise similarities one-by-one over different sessions. The multi-arrangement (MA) method solves these problems by allowing subjects to communicate multiple object-pair similarities at once (Figure 6.2). In the MA method, subjects communicate perceived object similarity by arranging multiple object images in 2D on a computer screen by mouse drag-and-drop. The use of spatial arrangement as a measure of perceived similarity has been proposed before (Goldstone, 1994; Risvik et al., 1994). Our MA method extends this earlier work by introducing adaptive selection of object

subsets during measurement, in order to efficiently and optimally estimate perceived similarity for each individual subject. Using our MA method, the acquisition of the 4560 pairwise similarities only required one hour per subject.

The method can be summarized as follows. Each trial consists of multiple (> 2) objects that have to be arranged in a circular “arena” such that inter-object distances reflect perceived similarity (similar objects are placed close together, dissimilar objects are placed further apart). This approach enables time-efficient measurement of perceived object similarity because moving one object changes multiple object-pair similarities at once. Single-trial estimates of perceived object dissimilarity are computed as Euclidean distances between the objects (after normalization of object positions by the diameter of the arena). On the first trial, subjects arrange all objects. On subsequent trials, they arrange subsets of objects. To optimize the object subsets to be presented on subsequent trials, we assume that the arrangements are affected by isotropic placement noise in 2D. The dissimilarity signal-to-noise ratio of the estimates then depends on how closely the objects are placed together in the arena: if two objects are placed close together (smaller dissimilarity signal), the dissimilarity estimate will have a smaller signal-to-noise ratio than when they are placed further apart. After each trial, the object subset for the next trial is constructed adaptively so as to provide more evidence for the object pairs whose current combined estimates are expected to have the greatest error, thus aiming to minimize the maximum error of the final dissimilarity estimates. For example, the object pair placed closest together on the first trial will be sampled again on the next trial so as to increase the evidence for estimating the dissimilarity of these two objects. The use of multiple trials also enables the subjects to communicate similarity relationships that would require more than two dimensions to accurately reflect the perceptual similarity structure of the entire object set. The duration of the MA acquisition can either be fixed (e.g. one hour as in our experiment) or contingent upon the quality of the estimated dissimilarities (e.g. ensuring that the maximum error margin across all pairs is below a certain threshold). The MA method was implemented in Matlab (The MathWorks Inc.).

We instructed our subjects to “Please arrange these objects according to their similarity”, such that similar objects were placed close together and dissimilar objects were placed further apart. The instruction intentionally did not specify which object properties to focus on, as this would have biased our perspective on the mental representation of the objects. In other words, the general instruction enabled us to investigate which properties subjects would spontaneously use when judging object similarity for a large set of real-world object images. After performing the experiment, subjects were asked to report which object features they had used for object arrangement.

multi-object arrangements



↓
"inverse"
multidimensional
scaling

representational dissimilarity matrix

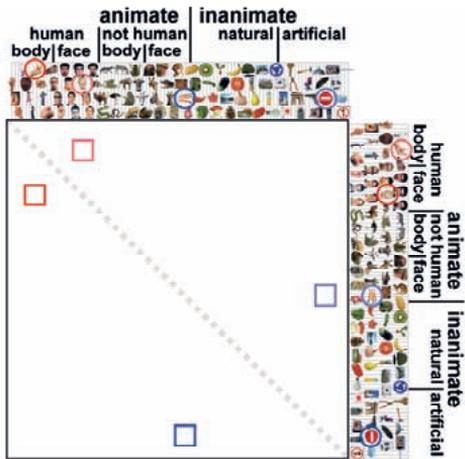


Figure 6.2 Multi-arrangement method. Object-similarity judgments were acquired using a novel multi-arrangement (MA) method, which allows efficient and subject-tailored acquisition of perceived similarity for large sets of objects. Subjects were asked to arrange the objects according to their similarity, using mouse drag-and-drop on a computer display. Perceived similarity was communicated by adjusting the distances between the objects: objects perceived as similar were placed close together; objects perceived as dissimilar were placed further apart. The upper panel of the figure shows screenshots taken at different moments during the acquisition of the similarity judgments for one subject. Columns correspond to trials and rows show object arrangements over time, running from the start (first row) to the end of each trial (final arrangement, last row). The first trial contained all object images; subsequent trials contained subsets of images that were adaptively selected to optimally estimate the perceptual dissimilarities for each subject. The black dots represent not-shown arrangements during a trial (small dots) and not-shown trials (large dots). Once acquisition of the object-similarity judgments was completed, inter-object distances of the final trial arrangements were combined over trials by rescaling and averaging to yield a single dissimilarity estimate for each object pair. Conceptually, this step can be seen as “inverse” multidimensional scaling, since it combines several lower-dimensional (2D) similarity representations into one higher-dimensional similarity representation, which can then be “stored” in an RDM. This process is shown for two example objects pairs: a boy’s face and a hand (red), and carrots and a stop sign (blue). Their single-trial dissimilarity estimates (arrows) are combined into a single dissimilarity estimate, which is placed at the corresponding entry of the RDM (lower panel). Mirror-symmetric entries are indicated by lighter colors.

Computation of the representational dissimilarity matrix (RDM)

For each subject, the dissimilarities acquired for a given stimulus pair (on-screen distance between the arranged stimuli) were averaged across trials. Rescaling of each trial’s dissimilarities was required before averaging, because subjects were instructed to use the entire arena for each arrangement, making only the relations between distances on a single trial, but not the absolute on-screen distances meaningful. For example, a given dissimilarity between two objects tended to correspond to a greater on-screen distance when the two objects appeared in a smaller subset on a given trial. The single-trial RDMs were therefore iteratively rescaled so as to align them to the overall average (minimizing the sum of squared deviations) until convergence. The result of this procedure was the trial-average RDM. We computed a trial-average RDM for each subject and then averaged the RDMs across subjects to obtain the group RDM.

6.2.3 Relating the similarity representations

In order to relate the hIT and perceptual similarity representations in terms of categorical structure, we visualized and explored them in multiple ways (Figure 6.3). Figure 6.3 displays not only the RDMs, but also the associated multidimensional scaling (MDS) plots (Torgerson, 1958; Shepard, 1980) and hierarchical cluster trees (Shepard, 1980) (see also Figure 6.4). The MDS plots display the multi-dimensional similarity representations in 2D: the closer the objects, the more similar their activity patterns or perceived similarity. The hierarchical

cluster trees explore which object-clusters emerge from the data when objects are grouped based on activity-pattern or perceived similarity.

In order to relate the hIT and perceptual similarity representations quantitatively, we correlated the 4560 pairwise hIT dissimilarities with the 4560 pairwise perceptual dissimilarities. In other words, we correlated the corresponding entries from the hIT and perceptual RDMs. We used the Spearman rank correlation coefficient to assess the relatedness of the RDMs since we expected a monotonic, but not necessary linear, relationship between hIT and perceptual dissimilarities. The correlation was restricted to the lower triangle of each RDM, which contained all possible (4560) pairwise dissimilarities.

The classical method for testing correlations assumes independent pairs of measurements for the two to-be-correlated variables. Such independence cannot be assumed for RDMs, because each dissimilarity is dependent on two stimuli (or their associated activity patterns), each of which also codetermines the dissimilarities of all its other pairings in the RDM. We therefore tested the relatedness of the group-average RDMs by randomization of the condition labels (Figures 6.5a, 6.6a, 6.7b, 6.8). This condition-label randomization test was implemented as follows. We chose a random permutation of the conditions (i.e. of the 96 stimuli), reordered rows and columns of one of the two RDMs to be compared according to this permutation, and computed the correlation. By repeating this step 10,000 times, we obtained a distribution of correlations simulating the null hypothesis that the two RDMs are unrelated. If the actual correlation (i.e. the one for consistent labeling between the two RDMs) falls within the top 5% of the simulated null distribution of correlations, we reject the null hypothesis of unrelated RDMs. The p value is estimated as 1 minus the proportion rank of the actual correlation in the randomization distribution. The proportion rank is conservatively estimated, such that $p < 0.0001$ indicates that the actual correlation is higher than any of the 10,000 correlations obtained after randomization of the condition labels.

We also tested the relatedness of the RDMs in a random-effects analysis across subjects (Figures 6.5b, 6.6b). This analysis enables generalization of the results to the population and does not require independence of the dissimilarity estimates in an RDM. We first computed single-subject Spearman rank correlation coefficients by correlating each single-subject similarity-judgment RDM with the subject-average hIT RDM. We then transformed these correlation coefficients using the Fisher z transform and performed a standard one-sample t test on the resulting values. The t test was used to determine whether the average of these sixteen Fisher-z-transformed correlation coefficients was larger than zero.

6.2.4 Model representations of the stimuli

We processed our stimuli to obtain their representations in a number of simple and complex computational models. A description of the model representations can be found in Chapter 3 (3.7.2). The model RDMs were compared to the similarity-judgment RDM (Figure 6.7).

6.3 Results

6.3.1 Do the hIT and perceptual similarity representations match in terms of categorical structure?

The hIT and perceptual similarity representations are displayed in Figure 6.3. Both the RDMs (Figure 6.3a) and MDS plots (Figure 6.3b) show that the similarity representation based on human object-similarity judgments is inherently categorical and reflects the top-level animate/inanimate distinction of the hIT representation. In addition, both representations show a tight cluster of human faces. Compared to the hIT representation, the perceptual similarity representation shows more and tighter (sub)clusters, suggesting stronger categoricity (Figures 6.3a,b). Furthermore, an exploratory cluster analysis indicated that, in addition to the animate/inanimate and face/body divisions that are present in both representations, the perceptual representation shows a natural/artificial distinction among the inanimate objects and a prominent human/nonhuman distinction among the animate objects, which is prioritized over the face/body distinction (Figure 6.3c, 4). An inferential analysis confirmed these results, showing that both the perceptual and the hIT representation were highly correlated with a simple two-category model representing the top-level animate/inanimate distinction ($r_{\text{perceptual}}=0.59$, $r_{\text{hIT}}=0.60$, $p<0.0001$ for both correlations as determined by a condition-label randomization test). A four-category model (human/nonhuman/natural/artificial) also correlated significantly with both representations ($p<0.0001$), but its correlation with the perceptual representation was clearly higher than that with the hIT representation ($r_{\text{perceptual}}=0.70$, $r_{\text{hIT}}=0.36$), underscoring the difference between the two representations with respect to the hierarchy of category subclusters.

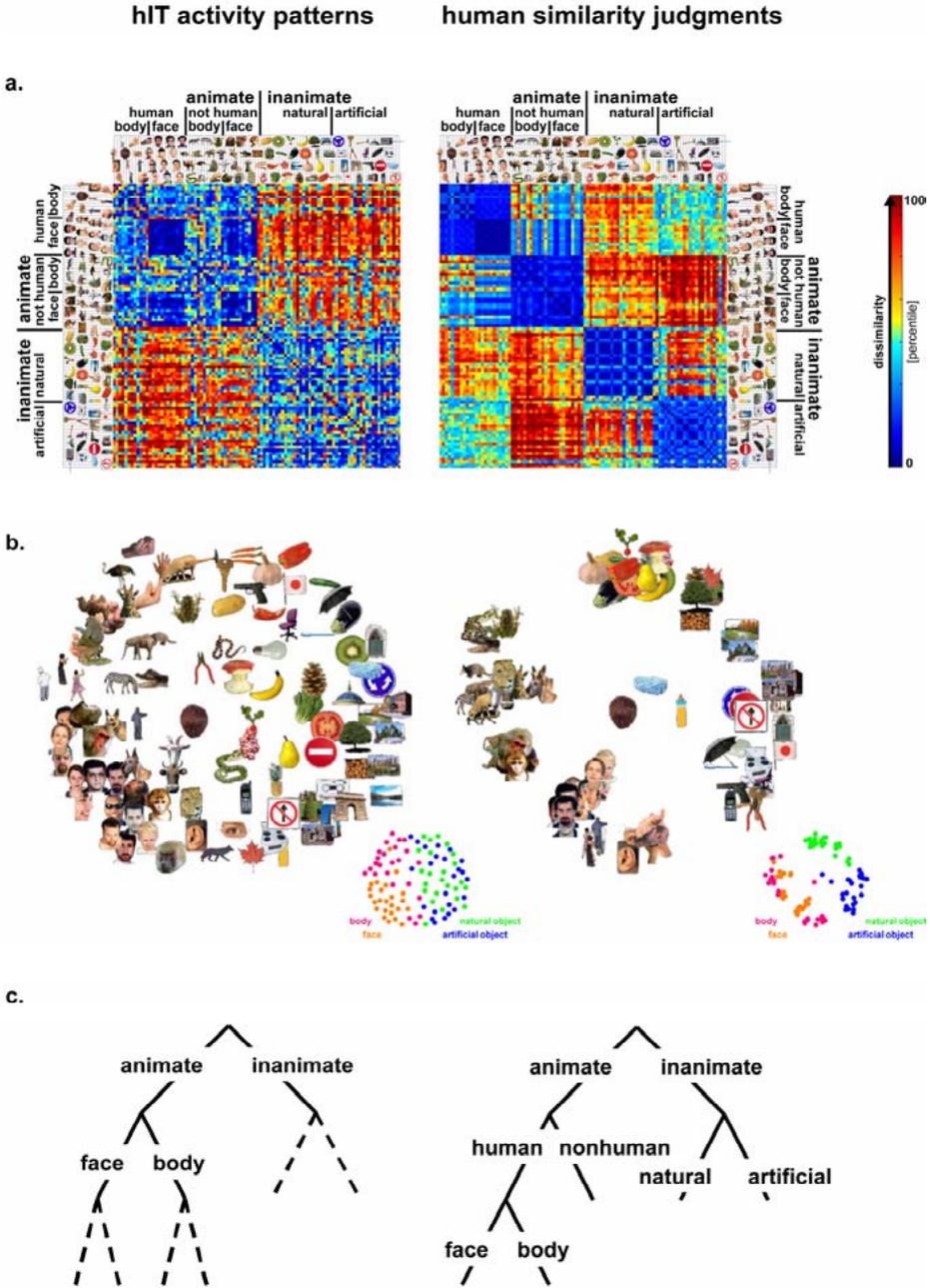


Figure 6.3 Human IT activity patterns and human similarity judgments both show an inherently categorical representation of real-world object images with an animate/inanimate top-level distinction. At the same time, the perceptual similarity representation shows additional categorical distinctions and stronger clustering than the hIT similarity representation. **(a)** Representational dissimilarity matrices (RDMs) based on hIT activity patterns and human similarity judgments. Each RDM is based on data from multiple subjects (4 and 16, respectively), averaged at the

level of the dissimilarities. Each entry of a matrix represents hIT activity-pattern dissimilarity ($1-r$, where r is Pearson correlation coefficient; 316 most visually responsive bilateral hIT voxels defined using independent data) or perceptual dissimilarity (distance as measured by the multi-arrangement method) for a pair of objects. The matrices were separately histogram-equalized for easier comparison. The color code reflects dissimilarity percentiles (see color bar). **(b)** Multidimensional scaling (MDS; criterion: metric stress) was used to visualize the hIT and perceptual similarity representations of the 96 real-world object images. Distances between images reflect the dissimilarities that are shown in the RDMs in panel a: images that elicited similar activity patterns or that were judged as similar are placed close together; images that elicited dissimilar activity patterns or were judged as dissimilar are placed further apart. **(c)** Schematic of the hierarchical clustering results (Figure 6.4) showing a common top-level distinction but different sub-level hierarchical structure for the perceptual as compared to the hIT representation.

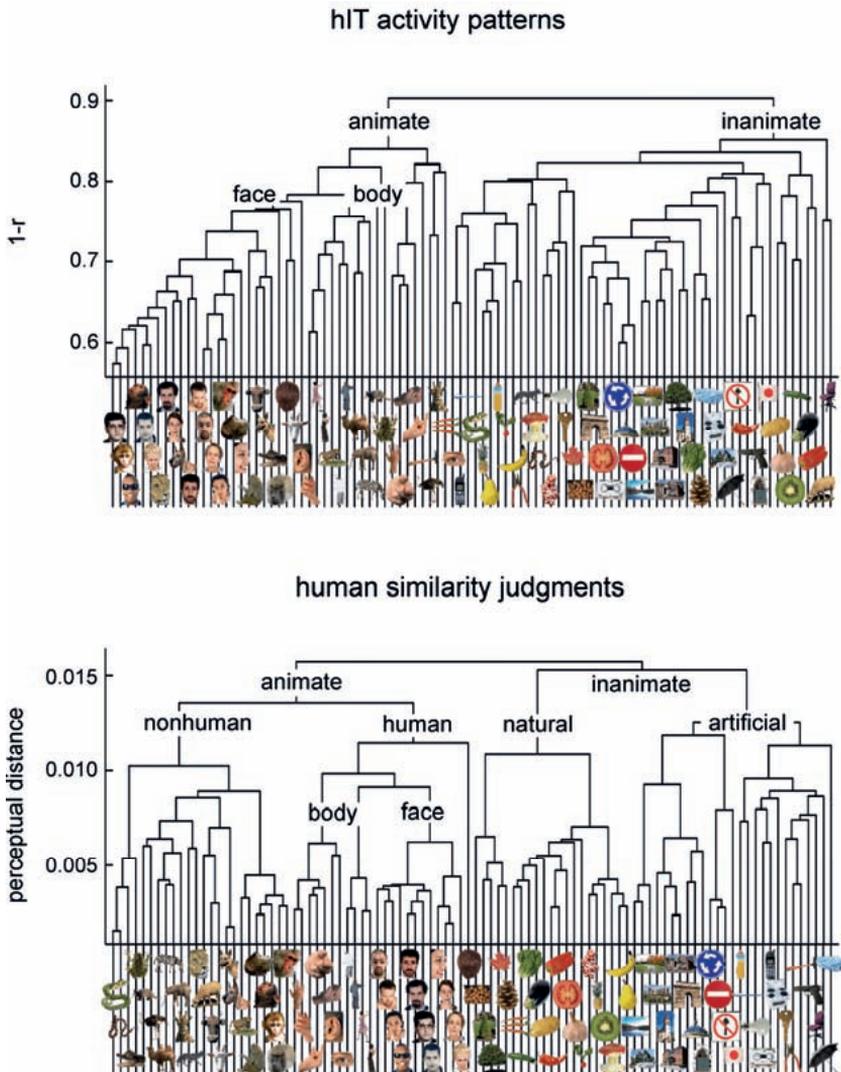


Figure 6.4 Hierarchical clustering of hIT activity patterns and human similarity judgments. hIT object-activity patterns have been shown to cluster according to natural categories (top panel) (Kriegeskorte et al., 2008a). In order to assess whether human object-similarity judgments show a similar categorical structure, we performed hierarchical cluster analysis on the similarity judgments (bottom panel). Hierarchical cluster analysis starts with single-image “clusters” and successively combines the two clusters closest to each other to form a hierarchy of clusters. The vertical height of each horizontal link reflects the average dissimilarity between the stimuli of two linked subclusters. hIT activity dissimilarity was measured as $1-r$ (where r is Pearson correlation coefficient), perceptual dissimilarity was measured as Euclidean distance (using the MA method). Text labels indicate the major clusters. Both hIT activity patterns and human similarity judgments cluster according to natural categories and show a top-level animate-inanimate distinction. However, the human similarity judgments show tighter subclusters and a different categorical hierarchy within the animate objects.

The strong categorical clustering of the perceptual representation is consistent with debriefing reports of the subjects. Fifteen out of sixteen subjects indicated that the most important guiding principle that they used during object arrangement was semantic category. The specific categories mentioned by the subjects correspond to the (sub)clusters shown in Figure 6.3b (e.g. human faces, monkeys/apes, fruits, tools). Most subjects indicated that they also used shape and color to arrange the objects, specifically within category clusters.

6.3.2 Do the hIT and perceptual similarity representations match in terms of continuous structure?

The preceding analysis indicated that the perceptual and hIT similarity representations share several key categorical distinctions. Here, we go one step further and compare the two representations in terms of their continuous similarity structure, i.e. the entire representational space and not only the categorical divisions that characterize the space. To quantify the match between the two continuous representations, we compared them at the level of the RDMS by correlating the dissimilarity estimates of corresponding object pairs. A one-sided condition-label randomization test showed that the two similarity representations were significantly correlated, both within all images and within most category subsets of images (Figures 6.5a, 6.6a). Perceptual and hIT dissimilarities were significantly correlated within the following category subsets: animate objects, inanimate objects, faces, (human) bodies, humans, nonhuman animates, natural objects, and artificial objects. The highest correlation coefficients between perceptual and hIT dissimilarities were found within humans ($r=0.60$), within faces ($r=0.40$), and within natural objects ($r=0.46$).

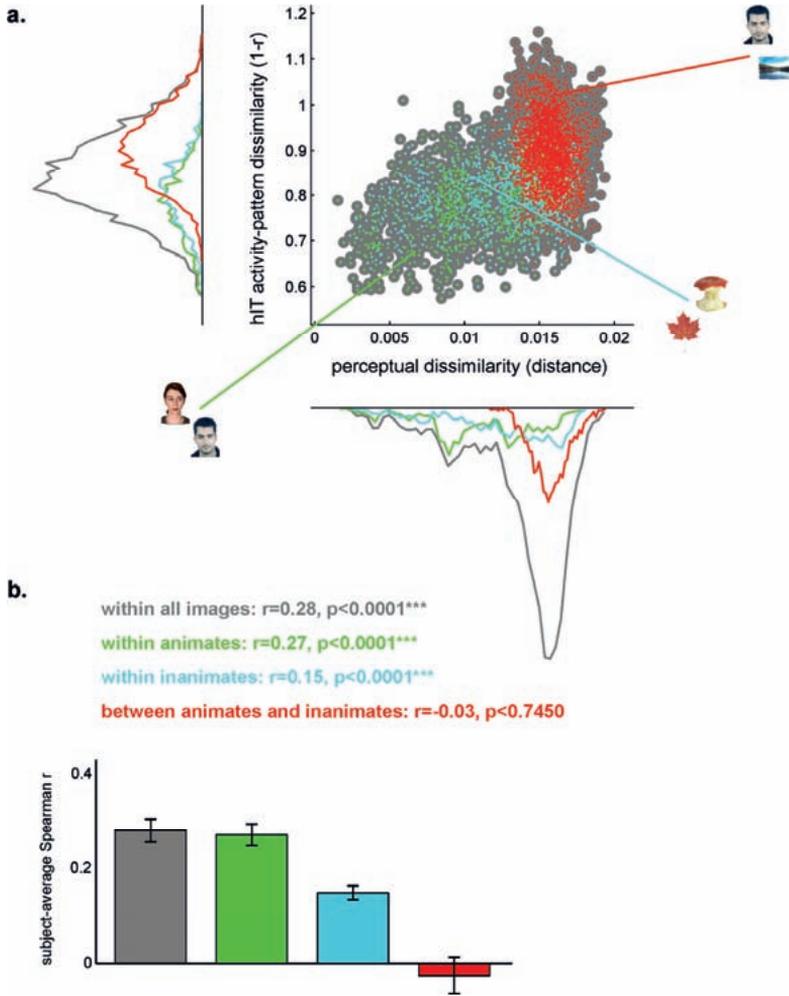


Figure 6.5 hIT activity-pattern dissimilarities and perceptual dissimilarities are significantly correlated within all images and within category subsets of images. (a) Scatter plot of hIT activity-pattern dissimilarities and perceptual dissimilarities taken from the subject-average RDMS shown in Figure 6.3a. A dot is placed for each stimulus pair based on its hIT activity-pattern dissimilarity and perceptual dissimilarity (three example stimulus pairs are shown). The large grey dots represent all possible stimulus pairs ($r=0.39, p<0.0001$; r is Spearman correlation coefficient). The smaller colored dots placed on top of the grey dots code for subsets of images: green dots represent animate object pairs ($r=0.34, p<0.0001$), cyan dots represent inanimate object pairs ($r=0.19, p<0.0001$), and red dots represent object pairs consisting of an animate and an inanimate object ($r=-0.16, p<0.9975$). Consistent with the results in Figure 6.3, the marginal histograms show that both perceptual and hIT dissimilarities are larger for object pairs that cross the animate-inanimate category boundary (red) than for object pairs that do not cross this boundary (green and cyan). **(b)** To test whether the match between hIT and perceptual dissimilarities would generalize to the population of similarity-judgment subjects, we computed the correlation of each single-subject perceptual RDM with the hIT RDM and tested whether the average of those correlations was significantly larger than zero, using a one-sample t-test. Bars show the average correlation between hIT and perceptual dissimilarities across subjects. Error bars show s.e.m.

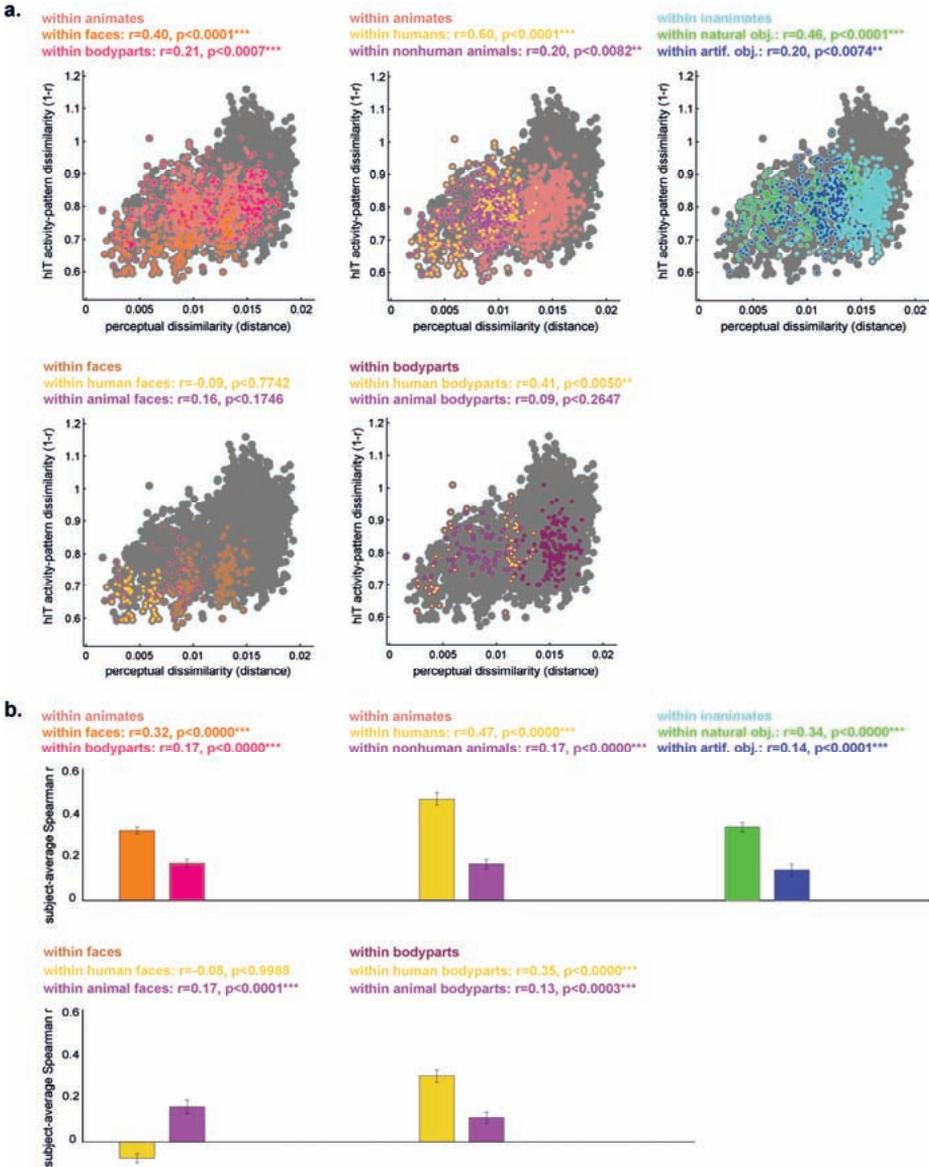


Figure 6.6 Correlation of hIT and perceptual dissimilarities for subsets of images. To investigate whether object-similarity judgments reflect the within-category similarity structure of hIT activity patterns, we correlated the hIT and perceptual dissimilarities for subsets of images. **(a)** Scatter plots of hIT and perceptual dissimilarities taken from the subject-average RDMs in Figure 6.3a. A dot is placed for each stimulus pair based on its hIT activity-pattern dissimilarity and perceptual dissimilarity. The large grey dots represent all possible stimulus pairs, the smaller colored dots placed on top of the grey dots code for subsets of images as indicated in the plot legends. Plot legends show Spearman correlation coefficients and associated p-values computed with a one-sided condition-label randomization test (10,000 randomizations). The hIT and perceptual similarity structures are significantly correlated within the following subsets of images: faces, (human) body parts, humans, non-human animals, natural objects, and artificial objects. **(b)** The within-category

match between hIT activity-pattern dissimilarities and perceptual dissimilarities generalizes to the population of similarity-judgment subjects. We computed the correlation of each single-subject similarity-judgment RDM with the hIT RDM and tested whether the average of those correlations was significantly larger than zero, using a one-sample t-test. Bars show the average correlation between hIT and perceptual dissimilarities across subjects. Error bars show s.e.m.

The match between the two representations was also found in a random-effects analysis across subjects (Figures 6.5b, 6.6b3). Again, perceptual and hIT dissimilarities were significantly correlated within all images and within most category subsets of images, including all subsets that were identified by the condition-label randomization test. This indicates that our results can be generalized to the population of similarity-judgment subjects. An impression of the inter-subject variability of similarity judgments can be found in the MDS plot shown in Figure 6.7a. This figure shows that each single-subject similarity representation is unique, but that, at the same time, the single-subject representations cluster together. One of the subjects (S1) falls outside of the cluster, showing a perceptual similarity representation more similar to simple models based on image features than to the perceptual similarity representations of the other subjects. This subject reported to have arranged objects by shape instead of semantic category. Consistent with the observation that single-subject representations cluster together, the average inter-subject correlation was reasonably high (Figure 6.8; average Spearman $r=0.33$; all but two of the 120 inter-subject correlations greater than zero as determined by condition-label randomization tests, and after False Discovery Rate correction for multiple comparisons). Figure 6.8 also displays inter-subject correlations for category subsets of images.

Other brain regions, including early visual cortex, the fusiform face area (FFA), and the parahippocampal place area (PPA), did not match the perceptual judgments as well as hIT (Figure 6.7). FFA showed a lower, but still significant correlation with the perceptual similarity representation ($r=0.22$, $p<0.0001$); for early visual cortex and PPA, the correlation was not significant. Computational models based on low-level and more complex natural image features also did not match the perceptual judgments as well as hIT (Figure 6.7). Among the models, simple models based on object color and shape, and a more complex model based on natural image features thought to be representative of primate V4 and posterior IT (Serre et al., 2005), showed the closest match to the perceptual similarity representation.

Error bars indicate the standard error of the deviation estimate. The standard error was estimated as the standard deviation of 100 deviation estimates obtained from bootstrap resamplings of the condition set. The p-value below each bar indicates whether the associated RDM is significantly related to the similarity judgment RDM (condition-label randomization test, 10,000 randomizations). hIT is the best match to the similarity judgments.

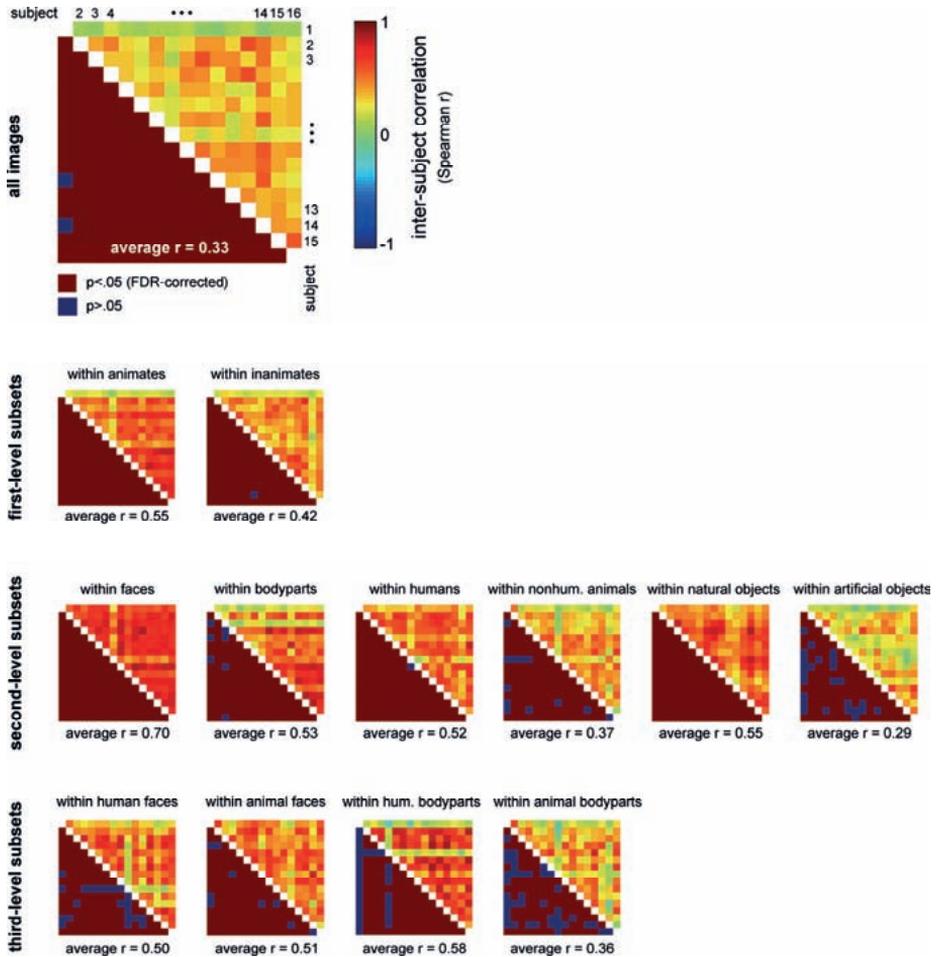


Figure 6.8 Correlation of similarity judgments across subjects for all images and for category subsets of images (same subsets as in Figures 6.5, 6.6). The upper triangle of each matrix shows all possible pairwise inter-subject RDM correlations (Spearman r). The mirror-symmetric entries in the lower triangle of each matrix show the corresponding p-values. P-values were computed using a condition-label randomization test with 10,000 randomizations and corrected for multiple comparisons using the False Discovery Rate. The average of all pairwise 120 inter-subject correlations is shown below each matrix.

6.4 Discussion

6.4.1 Human perception is categorical and reflects the primate IT object representation

We asked subjects to judge object similarity for a large set of real-world object images and investigated whether these similarity judgments reflected the IT object representation, including its hierarchy of category clusters. Our results show that human perception is categorical and reflects the two major categorical distinctions that characterize the primate IT object representation: the top-level animate/inanimate distinction and the face/body distinction among the animates (Kiani et al., 2007; Kriegeskorte et al., 2008a). The shared top-level animate/inanimate distinction relates to neuropsychological (Capitani et al., 2003; Warrington and Shallice, 1984), behavioral (Kirchner and Thorpe, 2006; New et al., 2007), and neuroimaging findings (Chao et al., 1999; Martin et al., 1996) that suggest a special status for the living/nonliving distinction. This special status might be explained in terms of evolutionary pressure towards fast and accurate recognition of animals (consistent with New et al., 2007). Recognizing animals, whether they were predator or prey, friend or foe, was of vital importance to our primate ancestors. Recognizing faces was key to survival and reproduction as well, since faces carry important information that can be used to infer the emotions, intentions, and identity of other animals. These important functions, recognizing animals and animal faces at the category level, appear to be achieved by the IT representation and are reflected in human perception.

Alternatively, one might argue that the categorical structure of both the similarity judgments and the IT object representation can be explained in terms of visual similarity. Animate and inanimate objects differ not only in terms of evolutionary relevance, but also in terms of visual properties. IT has long been known to be sensitive to visual object features (Tanaka, 1996). Combined with the fact that animate and inanimate objects differ in visual features, this could potentially lead to a categorical representation that is solely based on visual similarity. Moreover, previous studies have shown a relationship between perceived shape similarity and IT activity-pattern similarity for abstract object shapes (Haushofer et al., 2008; Kayaert et al., 2005; Op de Beeck et al., 2008). In order to test if our findings could be accounted for by visual similarity, we studied model representations of the stimuli. If visual similarity would explain our findings, we would expect model representations based on object shape to show a categorical structure similar to that of IT. However, we found that computational models based on object shape – a simple silhouette model and a more complex computational model based on natural image features at a level of complexity thought to approximately match V4 and posterior IT (Serre et al.,

2005) – do not show a clear categorical structure (Kriegeskorte et al., 2008a), and do not account for the similarity judgments as well as IT.

If it is not visual shape similarity, then what makes the IT representation, and human perception, categorical? It has been shown that visual features of intermediate complexity, which IT is sensitive to (Tanaka, 1996), are optimal for category discrimination (Ullman et al., 2002). However, sensitivity to visual features of intermediate complexity does not automatically lead to a categorical object representation. What may be needed is sensitivity to visual features that are most informative on category membership (Ullman et al., 2002). Indeed, several studies have shown that IT displays and can acquire sensitivity to category-discriminating visual features (Lerner et al., 2008; Sigala and Logothetis, 2002). Categories of high behavioral importance, including the evolutionary relevant categories of animals and faces (see also Mahon et al., 2009), are most likely to be represented by a feature set designed to emphasize their boundaries in shape space (Schyns et al., 1998).

Our results show that similarity judgments reflect not only the two major categorical distinctions of the IT representation, but also the IT within-category similarity structure. Given the functional properties of IT, this within-category match is likely to be based on visual similarity between objects that belong to the same category cluster. This explanation is consistent with the reports of our subjects, stating that they used object color and shape to arrange objects within category clusters. Furthermore, these findings are consistent with the previously reported relationship between perceived object-shape and IT activity-pattern similarities (Edelman et al., 1998; Haushofer et al., 2008; Kayaert et al., 2005; Op de Beeck et al., 2008).

6.4.2 Human perception transcends the primate IT object representation

Several features of the perceptual similarity representation cannot be explained by the IT representation. The perceptual representation shows stronger category clustering (consistent with Edelman et al., 1998) and introduces additional categorical distinctions, the most salient of which are the prominent human/nonhuman distinction among the animate objects and the natural/artificial distinction among the inanimate objects. These additional distinctions suggest that human perception is more flexible than the IT representation, reflecting species-specific (human/nonhuman) and evolutionarily more recent (natural/artificial) distinctions. Perceptual judgments appear to be shaped by semantic information (Tyler and Moss, 2001) and task instruction (Liu and Cooper, 2001) to a greater extent than the IT representation. This suggests a role for other brain regions, including prefrontal cortex, which has been shown to be

involved in category learning and rule-based categorization (Ashby and Ell, 2001; Freedman et al., 2001; Miller et al., 2003, but see Minamimoto et al., 2010).

The task instruction given to the subjects in our experiment was quite general (“Please arrange these objects according to their similarity”). Given that each object can be described by multiple properties, including color, shape, real-world size, function, and semantic category, and given that subjects were free to choose and weight these properties according to their subjective preferences, the instruction could be implemented in many different ways. Nevertheless, without explicit instruction, subjects exhibited a strong tendency to group the objects by semantic category. This finding underscores the importance of categorization in daily life and indicates that semantic information plays an important role in human perception of real-world object images (Barsalou et al., 2003). Subjects not only exhibited a strong tendency to group objects by semantic category, but also picked the same four main categories (human, nonhuman, natural, artificial) and tended to use shape and color to arrange the objects within category clusters. These tendencies shared across subjects might be driven by shared pre-existing knowledge about real-world object categories and can explain the fact that object-similarity judgments were significantly correlated across subjects. The inter-subject correlation that we found was higher than previously found for ambiguous two-dimensional shape contours (Haushofer et al., 2008) but lower than found for a small set of novel two-dimensional object prototypes (Op de Beeck et al., 2008). This suggests that perception effectively reduces the high-dimensional space that real-world object images reside in, but leaves space for flexibility (across subjects) as well.

6.4.3 Future directions

Our study is a first important step towards the identification of the neuronal substrates that give rise to high-level conscious similarity judgments of real-world object images, and can be used as a starting point for further studies. Our focus here was on the ventral-stream object representation. Future research should investigate the similarity representation in the entire brain (including frontal cortex), for example using a searchlight mapping approach (Kriegeskorte et al., 2006) to find the region that matches the similarity judgments most closely. A closer match to the similarity judgments might also be found by combining information from different brain regions.

Another avenue for future research would be to systematically investigate the effect of task instruction. Task instruction can be used to “bias” the subjects towards using certain object dimensions for judging object similarity, e.g. color, shape, real-world size, aesthetic appeal. It will be interesting to see to what de-

gree the perceptual representation reflects the task instruction and how task instruction modulates the explanatory contributions of different brain regions. Furthermore, the influence of task-instruction on inter-subject consistency could be investigated. A more specific task instruction might increase inter-subject consistency, but this might also depend on the object property mentioned in the task instruction (e.g. color vs. aesthetic appeal). One caveat of the current study is that the perceptual representation and the IT representation are based on different groups of subjects. It is encouraging to see that we still found a significant correlation between perception and IT object representations ($r=0.39$). Nevertheless, the fit between brain and behavior will likely be better when these two types of data are acquired in the same subjects (e.g. Op de Beeck et al., 2008; Haushofer et al., 2008).

6.4.4 Conclusion

We conclude that human perception of real-world object images reflects visual similarities and categorical distinctions of longstanding evolutionary relevance that are represented at the level of primate IT. At the same time, perception is influenced more strongly by semantic information and shows additional categorical distinctions. These features unexplained by the IT representation may reflect a prefrontal contribution allowing more flexible categorical distinctions and enabling humans to adequately adapt to changes in their environment.

Summary

We live in a rich visual environment, which is populated by many different kinds of objects. Recognition of these objects is essential for successful interaction with our environment. Object recognition is a computationally challenging task, and is accomplished by the primate brain in several processing steps along the ventral visual pathway. These processing steps result in high-level object representations at the level of inferior temporal (IT) cortex, which are fairly invariant to image transformations and form the basis for categorization and higher-order cognitive processes. These high-level object representations in IT are the focus of this thesis.

IT object representations have been studied in both humans and monkeys. While neurophysiological studies in monkeys investigated IT responses to individual object exemplars, neuroimaging studies in humans only assessed category-average IT responses. As a consequence, little is known about human brain responses to individual object exemplars. Furthermore, quantitative comparison of human and monkey data has been complicated by the need for defining the correspondency between measurement units. Similar correspondency problems complicate relating brain data to computational theory and behavior, hampering the development of an integrated theory of vision, and of a unified systems neuroscience in general. The work presented in this thesis addressed these issues by (1) assessing human IT responses to real-world object exemplars using fMRI, and (2) comparing the measured responses to data from monkey IT, computational models, and human behavior using the newly developed framework of representational similarity analysis (RSA). A summary of each chapter is given below, followed by a concise overall summary and discussion of the main findings.

Chapter summaries

The different fMRI analysis methods used in the presented work are described and compared in **Chapter 1**. This chapter focuses on pattern-information analysis, which in recent years has gained momentum in the field of neuroimaging. Pattern-information analysis aims to detect multivoxel activity-pattern differences between experimental conditions. These differences can be interpreted as reflecting differences in underlying neuronal population activity which is thought to represent mental content. Another technique that targets representational content is fMRI adaptation. This technique is based on the logic that fMRI stimulus-change effects in a specific brain region can be interpreted to indicate that the region contains neurons that represent the changed stimulus property.

fMRI adaptation has been widely used in the field of face perception to infer the presence of face-identity representations. Previous studies have suggested the

fusiform face area (FFA) as the locus of face-identity representations, but did not thoroughly investigate (1) the specificity of the effects to FFA and (2) the effect of low-level stimulus changes on the spatial extent of effects. **Chapter 2** reports that face-identity-change effects are not confined to face-selective regions: effects were also found in early visual cortex and the parahippocampal place area (PPA). To ensure that face-identity-change effects could not be attributed to low-level stimulus changes, we introduced viewpoint and illumination changes on both face-identity repetition and face-identity change trials. This led to a decrease in the spatial extent of effects, but did not eliminate effects outside of face-selective regions. Our findings could be interpreted as evidence for high-level face-identity representations in early visual cortex and PPA, but this seems unlikely given the known response properties of these regions. Alternatively, our effects could be explained by general attentional effects or carryover of activation from connected regions. Our results suggest that fMRI stimulus-change effects do not provide conclusive evidence for a neuronal representation of the changed stimulus property.

We therefore abandoned the fMRI-adaptation approach and moved to the more direct approach of simply measuring fMRI responses to individual object images using ungrouped-events designs (i.e. each object image is treated as a separate condition). Ungrouped-events designs feature in representational similarity analysis (RSA), which is introduced in **Chapter 3**. RSA is a new experimental and data-analytical framework that enables quantitative comparison of data from different branches of systems neuroscience by abstracting from activity patterns and relating the data at the level of similarities between activity patterns. Comparison of represented information at the level of activity-pattern similarity obviates the need for defining the correspondency between measurement units. Potential applications of RSA include validation of computational models by brain-activity data, relating brain representations between different species, analyzing representational connectivity between different brain regions, and exploring the representational content of brain regions using condition-rich ungrouped-events designs. Statistical inference is performed using randomization and bootstrap techniques. RSA was demonstrated by relating representations of real-world object images measured by fMRI in early visual cortex and FFA to a range of computational models. Consistent with existing literature, the object representation in early visual cortex was best explained by a simple silhouette-image model, and the representation in FFA was best explained by a conceptual face-animal-prototype model. Before moving to implementations of RSA in Chapters 5 and 6, we turn to Chapter 4.

Chapter 4 bridges a gap between classical fMRI analysis and RSA. Classical fMRI studies on object representation in IT investigated category-average activation of brain regions. RSA investigates object-exemplar pattern-information, incor-

porating two advances at once: from category-average to exemplar responses and from activation to pattern-information analysis. This creates two “gaps”. The first gap, category-average pattern-information analysis, has been addressed previously. What is missing is the analysis of activation of category-selective brain regions to individual object exemplars. In other words, would category-selectivity hold for individual objects? Bridging this second gap also establishes a link to neurophysiological studies in monkeys, which commonly measure neuronal activity in response to individual object exemplars. We measured fMRI activation of category-selective regions FFA and PPA to 96 real-world object images from a wide range of categories, including faces and places. We found no evidence of any images outside the preferred category eliciting a stronger response than any images inside the preferred category for either FFA or PPA. Regional-average activation might thus perfectly reflect category membership of individual objects. Within each category, individual images elicited different levels of activation, suggesting a graded rather than a pure step-function response profile. The 96-object fMRI data were subsequently analyzed for pattern-information within the RSA framework in the last two chapters.

Chapter 5 relates the IT representations of the same object images between human (fMRI) and monkey (cell recording). We found that IT activity patterns cluster according to natural categories: animate and inanimate objects formed top-level category clusters; faces and bodies formed subclusters within the animates. This hierarchical categorical structure inherent to IT matched between man and monkey. Within-category exemplar similarities also matched between the species. Results were robust against exclusion of category-selective regions FFA and PPA from analysis. Species-specific face analysis suggested an exception to the close match between man and monkey: IT better distinguished conspecific faces in each species. A range of low- and intermediate-level computational models could not account for the categorical representation observed in IT, indicating that our results cannot be explained by low-level feature similarity alone. In sum, these findings suggest that primate IT across species may host a common code, which combines a categorical and a continuous representation of objects.

The presence of IT activity-pattern clusters that correspond to well-known object categories suggests a link to human perception. This link was investigated in **Chapter 6** by relating the IT object representation to human object-similarity judgments of the same 96 real-world object images. Given our relatively large stimulus set, conventional methods for obtaining pairwise similarity judgments would be very time-consuming. We therefore developed a new multi-arrangement method for efficient and subject-tailored measurement of perceived similarity for large sets of stimuli. We found that human similarity judgments cluster the objects by category and reflect several features of the primate

IT object representation, including the basic distinction between animate and inanimate objects. Within-category exemplar similarities also matched between similarity judgments and IT object representations. These findings suggest that high-level conscious similarity judgments of real-world object images reflect visual similarities and categorical distinctions of longstanding evolutionary relevance that are represented at the level of primate IT. However, human similarity judgments transcended the IT representational stage in terms of a stronger categorical component and the introduction of species-dependent (human/nonhuman) and evolutionarily recent (natural/artificial) distinctions. These additional distinctions may reflect a prefrontal contribution allowing more flexible categorical distinctions.

Overall summary and discussion

Our findings suggest that category membership of individual objects is an important organizing principle of the primate IT object representation. This idea is not new: previous studies have shown that IT responses contain information about category membership (e.g. Tanaka, 1996; Tsao et al., 2006; Puce et al., 1995; Kanwisher et al., 1997; Haxby et al., 2001). However, we go beyond these studies using ungrouped-events designs and RSA, showing that the structure of the IT object representation is inherently categorical and hierarchically organized, with clusters reflecting object categories of longstanding evolutionary relevance that match between man and monkey. These “primate-default” categories represented at the level of IT were also reflected in high-level conscious human object-similarity judgments. A range of computational models containing low- and intermediate-level object representations could not explain the categorical structure observed in IT. Our findings suggest that IT models could be improved by the acquisition of category-discriminating features through supervised learning (Ullman et al., 2002; Sigala and Logothetis, 2002).

Our results indicate that IT does not only distinguish categories, but also individual object exemplars within each category. This finding is consistent with previous findings in the monkey (e.g. Young and Yamane, 1992; Hung et al., 2005) and with recent reports of pattern-information differences between exemplars of the same category in human IT (Kriegeskorte et al., 2007; Eger et al., 2008). We found that the within-category similarity representations match between species, and between brain data and similarity judgments. The within-category match is likely driven by visual similarities and dissimilarities between objects that belong to the same category cluster, and is consistent with the previously reported relationship between perceived object shape and primate IT activity-pattern similarities (Edelman et al., 1998; Haushofer et al., 2008; Op de Beeck et al., 2001; 2008).

The tightest activity-pattern cluster in both IT and FFA was formed by human faces, likely reflecting the high visual similarity between different face identities. Consistent with this observation, a previous attempt to detect face-identity pattern information in FFA failed; face-identity information was detected in anterior IT instead (Kriegeskorte et al., 2007). Cell recording from the middle macaque face patch, a suggested homologue of FFA, indicated some amount of face-identity information in its population response (Tsao et al., 2006), but the category effects explained most of the response variance. A recent fMRI study in humans detected face-identity information in anterior FFA and IT using spatio-temporal pattern analysis (Nestor et al., 2011). Our face-identity change findings would be consistent with this finding, but cannot provide conclusive evidence for the existence of face-identity representations in FFA. In sum, the current evidence suggests that face recognition is performed by a network of regions, including FFA and anterior IT. FFA might carry some amount of face-identity information, but its main function seems to be face detection (Puce et al., 1995; Kanwisher et al., 1997; Kriegeskorte et al., 2007).

In sum, our findings suggest that primate IT may host a common code across species, which combines a categorical and a continuous representation of objects. This code might be implemented by a continuous feature map containing several high-density clusters of related features tuned for the discrimination of categories of high behavioral and evolutionary relevance (e.g. faces) (Haxby et al., 2001). Such a map could explain the existence of category-selective regions, and would be consistent with our finding that activation of these regions appears to perfectly reflect category membership of individual objects, but nevertheless follows a graded instead of a pure step-function response profile. Furthermore, such a map would be consistent with our finding that activity patterns form clusters corresponding to conventional object categories despite the absence of step-function-like categorical responses in IT at the single-cell or single-voxel level. Step-function-like categorical single-cell responses have been reported for prefrontal cortex (Ashby and Ell, 2001; Freedman et al., 2001). Prefrontal cortex receives input from IT and might combine the information distributed in IT to explicate categories in a flexible task-dependent manner. This region might, in combination with IT, also contribute to high-level conscious human object-similarity judgments. Avenues for future research include investigation of the nature of the feature map(s) in IT using high-field fMRI (see Op de Beeck et al., 2008) and exploration of the representational connectivity between IT and prefrontal cortex within and across species using RSA.

Samenvatting

We leven in een rijke visuele omgeving waarin zich veel verschillende voorwerpen bevinden. Herkenning van deze voorwerpen is essentieel om succesvol te kunnen interacteren met onze omgeving. Het herkennen van voorwerpen is computationeel gezien een uitdagende taak. Deze taak wordt door het brein uitgevoerd in verschillende verwerkingsstappen in het ventrale visuele systeem. De verwerkingsstappen resulteren in hogere-orde representaties van voorwerpen op het niveau van de inferieure temporele (IT) cortex. Deze representaties zijn redelijk bestand tegen transformaties van de binnenkomende visuele informatie en vormen de basis voor het categoriseren van voorwerpen en andere hogere-orde cognitieve processen. Genoemde hogere-orde representaties van voorwerpen in IT vormen het onderwerp van dit proefschrift.

Representaties van voorwerpen in IT zijn onderzocht in zowel mensen als apen. In neurofysiologisch onderzoek in apen is het gebruikelijk om hersenresponsen op individuele voorwerpen te meten. In onderzoek met beeldvormende technieken in mensen is dit niet gebruikelijk: hersenresponsen worden over het algemeen gemeten voor een set van voorwerpen, waardoor de data de hersenactiviteit gemiddeld over verschillende voorwerpen weergeven. Ten gevolge hiervan is weinig bekend over de reactie van het menselijke brein op individuele voorwerpen. Daarnaast is het lastig om data die zijn verkregen in mensen op een kwantitatief niveau te vergelijken met data verkregen in apen omdat deze data gebaseerd zijn op verschillende meeteenheden waarvan de correspondentie onbekend is. Soortgelijke problemen spelen een rol bij het relateren van hersendata aan computationele modellen en gedragsmetingen. Deze correspondentieproblemen bemoeilijken de ontwikkeling van een integrale theorie van visuele perceptie, en, meer in het algemeen, de ontwikkeling van een overkoepelende neurowetenschappelijke systeemtheorie. Het werk dat in dit proefschrift beschreven wordt pakt deze zaken aan door (1) het meten van de fMRI respons in IT voor plaatjes van individuele voorwerpen uit het dagelijks leven (2) het vergelijken van de zo verkregen data met data gebaseerd op onderzoek in apen, computationele modellen, en menselijk gedrag, gebruik makend van “representational similarity analysis (RSA)” dat speciaal voor dit doel ontwikkeld werd. Hieronder volgt een samenvatting van elk hoofdstuk; deze samenvattingen worden afgesloten met een compacte samenvatting van het gehele proefschrift die tevens een discussie van de belangrijkste bevindingen bevat.

Samenvattingen per hoofdstuk

In **Hoofdstuk 1** worden de verschillende fMRI analyse methoden die in dit proefschrift gebruikt worden beschreven en vergeleken. Vooral patrooninformatie analyse komt aan de orde. Dit is een analysetechniek die gedurende de laatste jaren populair is geworden in de cognitieve neurowetenschap. Doel van

deze techniek is het detecteren van verschillen in multi-voxel patronen van hersenactiviteit, behorende bij verschillende experimentele condities. Deze verschillen zijn te interpreteren als een reflectie van verschillen in de onderliggende patronen van neurale activiteit, waarvan gedacht wordt dat ze de inhoud van onze waarneming en gedachten representeren. Een andere techniek die de inhoud van representaties onderzoekt is de fMRI-adaptatietechniek. Deze techniek is gebaseerd op de redenering dat fMRI-effecten die ten gevolge van een stimulusverandering worden waargenomen in een specifiek hersengebied erop wijzen dat het hersengebied neuronen bevat die de veranderde stimulouseigenschap representeren.

De fMRI-adaptatietechniek wordt veel gebruikt in onderzoek naar gezichts-waarneming, met name om de aanwezigheid van representaties van individuele gezichten aan te tonen. Verschillende fMRI adaptatiestudies hebben de “fusiform face area” (FFA) aangewezen als de locatie waar individuele gezichten gerepresenteerd worden, maar deze studies hebben niet grondig onderzocht (1) of de gevonden effecten alleen in FFA plaats vinden en (2) wat het effect is van lagere-orde stimulusveranderingen op de spatiële omvang van de waargenomen effecten. **Hoofdstuk 2** rapporteert dat het effect van een verandering van gezichtsidentiteit niet beperkt is tot FFA: effecten werden ook gevonden in vroege visuele gebieden en in de “parahippocampal place area” (PPA). Om er zeker van te zijn dat deze effecten niet konden worden toegeschreven aan lagere-orde veranderingen van de stimulus hebben we belichting en gezichtspunt gevarieerd voor zowel identiteitsveranderingen als repetities. Dit verminderde de spatiële omvang van de effecten, maar effecten waren nog steeds aanwezig buiten gebieden die selectief op gezichten reageren. Onze bevindingen zouden geïnterpreteerd kunnen worden als bewijs voor de aanwezigheid van hogere-orde representaties van gezichtsidentiteit in vroege visuele gebieden en PPA, maar dit is onwaarschijnlijk, gegeven de al bekende functionele eigenschappen van deze hersengebieden. Mogelijke alternatieve verklaringen behelzen algemene effecten van aandacht en overdracht van activiteit van andere hersengebieden, die verbonden zijn met de onderzochte gebieden. Onze resultaten suggereren dat fMRI-effecten die ten gevolge van een stimulusverandering worden waargenomen, geen sluitend bewijs vormen voor een onderliggende neurale representatie van de veranderde stimulouseigenschap.

Naar aanleiding van de bevindingen van Hoofdstuk 2 hebben we de fMRI-adaptatietechniek gelaten voor wat hij is en zijn we overgestapt op het simpelweg direct meten van fMRI activiteit gedurende het zien van individuele voorwerpen. Hiervoor hebben we gebruikt gemaakt van “ungrouped-events designs”, waarin elk voorwerp wordt behandeld als een aparte conditie. Ungrouped-events designs worden toegepast in “representational similarity analysis” (RSA), dat geïntroduceerd wordt in **Hoofdstuk 3**. RSA is een nieuw experimen-

teel en data-analytisch raamwerk, dat het mogelijk maakt om data van verschillende takken van de cognitieve neurowetenschap op een kwantitatief niveau met elkaar te vergelijken, door te abstraheren van de activatiepatronen en de data te vergelijken op het niveau van gelijkenissen tussen activatiepatronen. Door gerepresenteerde informatie te vergelijken op het niveau van gelijkenissen tussen activatiepatronen is het niet meer nodig om de correspondentie tussen meeteenheden te bepalen. Mogelijke toepassingen van RSA zijn validatie van computationele modellen aan de hand van hersenactiviteit, vergelijking van hersenrepresentaties tussen verschillende diersoorten, analyse van connectiviteit tussen hersengebieden op het niveau van gerepresenteerde informatie, en verkenning van de representatieve inhoud van hersengebieden met behulp van ungrouped-events designs die een groot aantal experimentele condities bevatten. Statistische inferentie binnen het RSA raamwerk berust op permutatie- en bootstraptechnieken. RSA wordt gedemonstreerd door representaties van alledaagse voorwerpen in vroege visuele gebieden en FFA, gemeten met behulp van fMRI, te vergelijken met representaties van dezelfde voorwerpen in een reeks van computationele modellen. De representatie van voorwerpen in vroege visuele gebieden kon het best verklaard worden aan de hand van een eenvoudig silhouet model. De representatie in FFA kon het best verklaard worden aan de hand van een conceptueel gezicht-dier-prototype model. Deze bevindingen zijn in overeenstemming met de functionele eigenschappen van deze gebieden als gevonden in eerdere studies. Voordat we ingaan op toepassingen van RSA in de Hoofdstukken 5 en 6, worden eerst de bevindingen van Hoofdstuk 4 besproken.

Hoofdstuk 4 slaat een brug tussen conventionele fMRI-analyse en RSA. Conventionele studies hebben representaties van voorwerpen in IT onderzocht door te kijken naar de activiteit van hersengebieden als geheel, waarbij de data de activiteit gemiddeld over voorwerpen behorende bij dezelfde categorie weergaven. RSA daarentegen onderzoekt patrooninformatie voor individuele voorwerpen, waardoor twee sprongen tegelijkertijd worden gemaakt: van activiteit gemiddeld over individuele voorwerpen naar activiteit voor elk voorwerp apart en van de analyse van activiteit van een hersengebied als geheel naar de analyse van informatie gerepresenteerd in patronen van activiteit binnen een hersengebied. Dit resulteert in twee gaten die gedicht moeten worden. Het eerste gat, patrooninformatie analyse van activiteit gemiddeld over voorwerpen, is al gedicht door eerder onderzoek. Wat ontbreekt is de analyse van activiteit van hersengebieden als geheel (met name hersengebieden die selectief reageren op een bepaalde categorie van voorwerpen) voor individuele voorwerpen. Met andere woorden, houdt eerder gemeten selectiviteit voor bepaalde categorieën stand voor individuele voorwerpen? Het dichten van dit tweede gat introduceert tevens een link met neurofysiologische studies in apen, waarin het gebruikelijk is om neurale activiteit voor individuele voorwerpen te meten. We hebben

fMRI activiteit in categorie-selectieve gebieden FFA en PPA gemeten voor 96 plaatjes van alledaagse voorwerpen behorende tot een groot aantal categorieën, inclusief gezichten en huizen. We vonden geen bewijs voor de hypothese dat er plaatjes van buiten de geprefereerde categorie zouden zijn die een sterkere respons zouden veroorzaken dan plaatjes van binnen de geprefereerde categorie. Dit gold zowel voor FFA als PPA. Activiteit van een gebied als geheel lijkt dus een perfecte reflectie te geven van de categorie waartoe een waargenomen individueel voorwerp behoort. Individuele plaatjes binnen elke categorie veroorzaakten verschillende niveaus van activiteit. Dit suggereert gradatie van activiteit in plaats van een puur stapsgewijs responsprofiel. De fMRI data verkregen voor de 96 plaatjes werden vervolgens in de laatste twee hoofdstukken geanalyseerd voor patrooninformatie binnen het RSA raamwerk.

Hoofdstuk 5 vergelijkt de IT-representaties van dezelfde plaatjes van voorwerpen tussen mensen (fMRI) en apen (electrofysiologie). Onze resultaten laten zien dat activatiepatronen in IT clusters vormen, die overeenkomen met natuurlijke categorieën: levende en niet-levende voorwerpen vormen topclusters, gezichten en lichaamsdelen vormen subclusters binnen de levende voorwerpen. Deze hiërarchische categorische structuur inherent aan IT kwam overeen tussen mens en aap. Ook binnen categorieën werd een significante gelijkheid gevonden tussen de twee soorten. Deze resultaten bleven onveranderd na verwijdering van categorie-selectieve gebieden FFA en PPA van de analyse. Een uitzondering op de gelijkheid tussen mens en aap werd gevormd door de representatie van individuele gezichten: de representaties van gezichten van de eigen soort waren beter van elkaar te onderscheiden dan die van de niet-eigen soort. De categorische structuur van de representatie in IT kon niet verklaard worden aan de hand van een reeks eenvoudige en meer complexe computationele modellen. Dit suggereert dat lagere-orde visuele gelijkheid tussen plaatjes niet afdoende is om onze bevindingen te verklaren. Alles bij elkaar suggereren onze resultaten dat de codering van voorwerpen in IT tot stand komt door een categorische en een continue representatie te combineren, en dat deze gecombineerde representatie gedeeld wordt door mens en aap.

De aanwezigheid van clusters van activatiepatronen die overeenkomen met natuurlijke categorieën suggereert een link tussen de IT-representatie en menselijke waarneming. Deze link werd onderzocht in **Hoofdstuk 6** door de IT-representatie uit Hoofdstuk 5 te relateren aan paarsgewijze gelijkheidsbeoordelingen voor dezelfde 96 plaatjes. Aangezien onze stimulusset relatief groot was, zou het erg tijdrovend zijn geweest om conventionele methoden voor het verkrijgen van paarsgewijze gelijkheidsbeoordelingen te gebruiken. Daarom hebben we een nieuwe “multi-arrangement” (MA) methode ontwikkeld voor het efficiënt meten van subject-specifieke waargenomen gelijknissen tussen grote aantallen stimuli. De zo verkregen gelijkheidsbeoordelingen van de 96 plaatjes

lieten een clustering van voorwerpen op basis van categorieën zien en deelden meerdere kenmerken met de IT-representatie, inclusief het primaire onderscheid tussen levende en niet-levende voorwerpen. Ook binnen categorieën werd een significante gelijkenis gevonden tussen IT en gelijkheidsbeoordelingen. Deze bevindingen suggereren dat hogere-orde bewuste gelijkheidsbeoordelingen van alledaagse voorwerpen gebaseerd zijn op visuele gelijknissen en evolutionair gezien belangrijke categorieën die gerepresenteerd zijn op het niveau van IT in mens en aap. Gelijkheidsbeoordelingen stijgen echter boven de IT-representatie uit in termen van een sterkere categorische component en de introductie van soort-specifieke (mens/dier) en evolutionair recente (natuurlijk/artificieel) categorieën.

Algemene samenvatting en discussie

Onze bevindingen suggereren dat de categorie waartoe een individueel voorwerp behoort een belangrijk organisationeel principe vormt van de IT-representatie in primaten. Dit is geen nieuw idee: eerdere studies hebben aangetoond dat hersenresponsen in IT informatie bevatten over de categorie waartoe waargenomen voorwerpen behoren (o.a. Tanaka, 1996; Tsao et al., 2006; Puce et al., 1995; Kanwisher et al., 1997; Haxby et al., 2001). Wij gaan echter verder dan deze studies door het gebruik van ungrouped-events designs en RSA, en laten zien dat de structuur van de representatie van voorwerpen in IT inherent categorisch en hiërarchisch georganiseerd is, met clusters van voorwerpen die evolutionaire relevantie reflecteren, en die overeenkomen tussen mens en aap. Een reflectie van deze “primate-default” categorieën, die gerepresenteerd zijn op het niveau van IT, werd gevonden in gelijkheidsbeoordelingen van alledaagse voorwerpen. De categorische structuur van de representatie in IT kon niet verklaard worden aan de hand van een reeks eenvoudige en meer complexe computationele modellen. Onze resultaten suggereren dat modellen van IT verbeterd kunnen worden door het implementeren van sensitiviteit voor objecteigenschappen die informatief zijn voor het onderscheiden van categorieën, door middel van training onder supervisie (Ullman et al., 2002; Sigala and Logothetis, 2002).

Onze resultaten tonen aan dat IT niet alleen categorieën onderscheidt, maar ook individuele voorwerpen. Deze bevinding komt overeen met eerdere bevindingen in apen (o.a. Young and Yamane, 1992; Hung et al., 2005) en met recent gerapporteerde verschillen in patrooninformatie in IT voor individuele voorwerpen behorende tot dezelfde categorie (Kriegeskorte et al., 2007; Eger et al., 2008). Onze data laten zien dat de representatieve structuur binnen categorieën een significante gelijkenis vertoont tussen mens en aap, en ook tussen hersendata en gedragsmetingen van menselijke waarneming. De gelijkenis bin-

nen categorieën wordt waarschijnlijk veroorzaakt door visuele overeenkomsten en verschillen tussen voorwerpen die tot dezelfde categorie behoren, en is consistent met de eerder gerapporteerde relatie tussen waargenomen vorm van voorwerpen en activatiepatroon gelijkenissen in IT (Edelman et al., 1998; Haushofer et al., 2008; Op de Beeck et al., 2001; 2008).

De sterkste clustering van activatiepatronen in zowel IT als FFA werd gevonden voor menselijke gezichten, waarschijnlijk ten gevolge van de hoge mate van visuele gelijkenis tussen verschillende individuele gezichten. Overeenkomend met deze observatie rapporteerde een eerdere studie dat geen detecteerbare patrooninformatie over gezichtsidentiteit aanwezig was in FFA. Deze informatie werd wel gevonden in de anterieure IT cortex (Kriegeskorte et al., 2007). Electrofysiologische metingen van de “middle macaque face patch”, een hersengebied in apen dat functioneel homologo lijkt te zijn aan FFA, toonden aan dat de activiteit van populaties neuronen in dit gebied informatie over gezichtsidentiteit bevat (Tsao et al., 2006), maar het grootste deel van de variatie in respons werd verklaard door effecten van categorie. Een recente fMRI studie in mensen toonde de aanwezigheid van informatie over gezichtsidentiteit aan in de anterieure FFA en IT met behulp van spatiële-temporele patrooninformatie analyse (Nestor et al., 2011). Onze bevindingen van Hoofdstuk 2 lijken overeen te komen met deze recente bevindingen, maar kunnen geen sluitend bewijs leveren voor de aanwezigheid van representaties van individuele gezichten in FFA. Samenvattend: het huidige bewijs suggereert dat gezichtsherkenning wordt uitgevoerd door een netwerk van hersengebieden, waaronder FFA en de anterieure IT. Responsen in FFA bevatten misschien een bepaalde hoeveelheid informatie over de identiteit van waargenomen gezichten, maar de belangrijkste functie van dit gebied lijkt detectie van gezichten te zijn (Puce et al., 1995; Kanwisher et al., 1997; Kriegeskorte et al., 2007) en niet identificatie.

Onze bevindingen suggereren dat IT in primaten een code bevat die gedeeld wordt door mens en aap, en dat deze code tot stand komt door de combinatie van een categorische en een continue representatie van voorwerpen. Deze code zou geïmplementeerd kunnen zijn als een continue “kaart” van voorwerpeigenschappen met meerdere clusters van gerelateerde eigenschappen, die afgestemd zijn op het discrimineren van categorieën met een hoge mate van gedragsmatige en evolutionaire relevantie (bv. gezichten) (Haxby et al., 2001). Deze kaart zou het bestaan van categorie-selectieve gebieden kunnen verklaren, en zou consistent zijn met onze bevinding dat activiteit van deze gebieden een perfecte reflectie lijkt te geven van de categorie waartoe een waargenomen individueel voorwerp behoort, maar tegelijkertijd ook gradatie laat zien in plaats van een puur stapsgewijs responsprofiel. Daarnaast zou zo’n kaart consistent zijn met onze bevinding dat activatiepatronen clusters vormen die overeenkomen met natuurlijke categorieën, ondanks het ontbreken van stapsgewijze categorische respon-

sen in IT op het niveau van individuele neuronen of voxels. Stapsgewijze categorische responsen zijn wel gerapporteerd voor neuronen in de prefrontale cortex (Ashby and Ell, 2001; Freedman et al., 2001). De prefrontale cortex ontvangt informatie van IT en combineert mogelijk de patrooninformatie uit IT om scheidingen tussen categorieën te verduidelijken op een taakafhankelijke, flexibele wijze. De prefrontale cortex draagt misschien, in combinatie met IT, ook bij aan hogere-orde bewuste gelijkheidsbeoordelingen van voorwerpen. Richtingen voor toekomstig onderzoek behelzen nader onderzoek naar de aard van de kaart(en) in IT met behulp van hoogveld fMRI (zie Op de Beeck et al., 2008) en het exploreren van de representatieve connectiviteit tussen IT en de prefrontale cortex met behulp van RSA.

References

- Afraz SR, Kiani R, Esteky H (2006) Microstimulation of inferotemporal cortex influences face categorization. *Nature* 442, 692-695.
- Aguirre GK, Zarahn E, D'Esposito M (1998) An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron* 21, 373-383.
- Aguirre GK, D'Esposito M (1999) Topographical disorientation: a synthesis and taxonomy. *Brain* 122, 1613-1628.
- Aguirre GK (2007) Continuous carry-over designs for fMRI. *Neuroimage* 35, 1480-1494.
- Aguirre GK, Thomas A, Hu D, Kerr W (in preparation). Dissociable representation of face features at coarse and fine neural scales.
- Andrews TJ, Ewbank MP (2004) Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage* 23, 905-913.
- Anzai A, Peng X, Van Essen DC (2007) Neurons in monkey visual area V2 encode combinations of orientations. *Nat Neurosci* 10, 1313-1321.
- Ashby FG, Ell SW (2001) The neurobiology of human category learning. *Trends Cogn Sci* 5, 204-210.
- Avidan G, Hasson U, Hendler T, Zohary E, Malach R (2002) Analysis of the neuronal selectivity underlying low fMRI signals. *Curr Biol* 12, 964-972.
- Baker CI, Behrmann M, Olson CR (2002) Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat Neurosci* 5, 1210-1216.
- Bandettini PA, Wong EC, Hinks RS, Tikofsky RS, Hyde JS (1992) Time course EPI of human brain function during task activation. *Magn Res Med* 25, 390-397.
- Bandettini PA, Cox RW (2000) Event-related fMRI contrast when using constant interstimulus interval: Theory and experiment. *Magn Reson Med* 43, 540-548.
- Bandettini, PA, and Ungerleider, LG (2001) From neuron to BOLD: new connections. *Nat Neurosci* 4, 864-866.
- Barsalou LW, Simmons WK, Barbey AK, Wilson CD (2003) Grounding conceptual knowledge in modality-specific systems. *Trends Cogn Sci* 7, 84-91.
- Bedny M, Aguirre GK, Thompson-Schill SL (2007) Item analysis in functional magnetic resonance imaging. *Neuroimage* 35, 1093-1102.
- Bell AH, Hadj-Bouziane F, Frihauf JB, Tootell RBH, Ungerleider LG (2009) Object representations in the temporal cortex of monkeys and humans as revealed by functional magnetic resonance imaging. *J Neurophysiol* 101, 688-700.
- Bellgowan PSF, Bandettini P, van Gelderen P, Martin A, Bodurka J (2006) Improved BOLD detection in the medial temporal region using parallel imaging and voxel volume reduction. *Neuroimage* 29, 1244-1251.
- Bentin S, Moscovitch M (1988) The time course of repetition effects for words and unfamiliar faces. *J Exp Psychol Gen* 117, 148-160.
- Bentin S, Peled BS (1990) The contribution of task-related factors to ERP repetition effects at short and long lags. *Mem Cognit* 18, 359-366.
- Bernard FA, Bullmore ET, Graham KS, Thompson SA, Hodges JR, Fletcher PC (2004) The hippocampal region is involved in successful recognition of remote and recent famous faces. *Neuroimage* 22, 1704-1714.
- Birn RM, Cox RW, Bandettini PA (2002) Detection versus estimation in event-related fMRI: Choosing the optimal stimulus timing. *Neuroimage* 15, 252-264.
- Blanz V, Vetter T (1999) A morphable model for the synthesis of 3D faces.
- Bodurka J, Ledden P, van Gelderen P, Chu R, de Zwart JA, Duyn J (2004) Scalable multichannel MRI data acquisition system. *Magn Reson Med* 51, 165-171.
- Bodurka J, Ye F, Petridou N, Murphy KM, Bandettini P (2007) Mapping the MRI voxel volume in which thermal noise matches physiological noise-implications for fMRI. *Neuroimage* 34, 542-549.
- Borg I, Groenen PJF (2005) *Modern multidimensional scaling - Theory and applications* (Second edition). New York: Springer.
- Boynton GM, Engel SA, Glover GH, Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci* 16, 4207-4221.

- Boynton GM, Finney EM (2003) Orientation-specific adaptation in visual cortex. *J Neurosci* 23, 8781-8787.
- Breiter HC, Etcoff NL, Whalen PJ, Kennedy WA, Rauch SL, Buckner RL, Strauss MM, Hyman SE, Rosen BR (1996) Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* 17, 875-887.
- Bruce V, Young A (1986) Understanding face recognition. *Br J Psychol* 77, 305-327.
- Buckner RL (1998) Event-related fMRI and the hemodynamic response. *Hum Brain Mapp* 6, 373-377.
- Busigny T, Robaye L, Dricot L, Rossion B (2009) Right anterior temporal lobe atrophy and person-based semantic defect: A detailed case study. *Neurocase* 15, 485-508.
- Capitani E, Laiacona M, Mahon B, Caramazza A (2003) What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cogn Neuropsychol* 20, 213-261.
- Carlson TA, Schrater P, He S (2003) Patterns of activity in the categorical representations of objects. *J Cogn Neurosci* 15, 704-717.
- Chan D, Fox NC, Scahill RI, Crum WR, Whitwell JL, Leschziner G, Rossor AM, Stevens JM, Cipelotti L, Rossor MN (2001) Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Ann Neurol* 49, 433-442.
- Chao LL, Haxby JV, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat Neurosci* 2, 913-919.
- Chao LL, Weisberg J, Martin A (2002) Experience-dependent modulation of category-related cortical activity. *Cereb Cortex* 12, 545-551.
- Cheng K, Waggoner RA, Tanaka K (2001) Human ocular dominance columns as revealed by high-field functional magnetic resonance imaging. *Neuron* 32, 359-374.
- Cooke T, Jakel F, Wallraven C, Bulthoff HH (2007) Multimodal similarity and categorization of novel, three-dimensional objects. *Neuropsychologia* 45, 484-495.
- Cortese JM, Dyre BP (1996) Perceptual similarity of shapes generated from Fourier descriptors. *J Exp Psychol Hum Percept Perform* 22, 133-143.
- Cox DD, Savoy RL (2003) Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261-270.
- Cutzu F, Edelman S (1996) Faithful representation of similarities among three-dimensional shapes in human vision. *Proc Natl Acad Sci USA* 93, 12046-12050.
- Cutzu F, Edelman S (1998) Representation of object similarity in human vision: psychophysics and a computational model. *Vision Res* 38, 2229-2257.
- Czarnecki K, Duffy JR, Nehl CR, Cross SA, Molano JR, Jack Jr CR, Shiung MM, Josephs KA, Boeve BF (2008) Very early semantic dementia with progressive temporal lobe atrophy. *Arch Neurol* 65, 1659-1663.
- Damasio H, Grabowski TJ, Tranel D, Hichwa RD, Damasio AR (1996) A neural basis for lexical retrieval. *Nature* 380, 499-505.
- Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughhead JW, Gur RC, Langleben DD (2005) Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* 28, 663-668.
- David SV, Gallant JL (2005) Predicting neuronal responses during natural vision. *Network* 16, 239-260.
- Dennett D (1987) *The Intentional Stance*. Cambridge, MA: MIT Press / A Bradford Book.
- Denys K, Vanduffel W, Fize D, Nelissen K, Peuskens H, Van Essen D, Orban GA (2004) The processing of visual shape in the cerebral cortex of human and nonhuman primates: a functional magnetic resonance imaging study. *J Neurosci* 24, 2551-2565.
- Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci* 4, 2051-2062.
- DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11, 333-341.
- Dobbins IG, Schnyer DM, Verfaellie M, Schacter DL (2004) Cortical activity reductions during repetition priming can result from rapid response learning. *Nature* 428, 316-319.

- Dougherty RF, Koch VM, Brewer AA, Fischer B, Modersitzki J, Wandell BA (2003) Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *J Vision* 3, 586-598.
- Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. *Science* 293, 2470-2473.
- Downing, PE, Chan, AW-Y, Peelen, MV, Dodds, CM, Kanwisher, N (2006). Domain specificity in visual cortex. *Cereb Cortex* 16, 1453-1461.
- Dricot L, Sorger B, Schiltz C, Goebel R, Rossion B (2008a) The roles of "face" and "non-face" areas during individual face perception: Evidence by fMRI adaptation in a brain-damaged prosopagnosic patient. *Neuroimage* 40, 318-332.
- Dricot L, Sorger B, Schiltz C, Goebel R, Rossion B (2008b) Evidence for individual face discrimination in non-face selective areas of the visual cortex in acquired prosopagnosia. *Behav Neurol* 19, 75-79.
- Drucker DM, Aguirre GK (2009) Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cereb Cortex* 19, 2269-2280.
- Duda RO, Hart PE, Stork DG (2001) *Pattern Classification*. New York, NY: John Wiley and Sons.
- Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behavior. *Trends Cogn Sci* 14, 172-179.
- Duong TQ, Kim DS, Ugurbil K, Kim S-G (2001) Localized cerebral blood flow response at submillimeter columnar resolution. *Proc Natl Acad Sci USA* 98, 10904-10909.
- Edelman S (1995) Representation of similarity in three-dimensional object discrimination. *Neural Comput* 7, 408-423.
- Edelman S (1997) Computational theories of object recognition *Trends Cogn Sci* 1, 296-304.
- Edelman S, Duvdevani-Bar S (1997a) A model of visual recognition and categorization. *Philos Trans R Soc Lond B Biol Sci* 352, 1191-1202.
- Edelman S, Duvdevani-Bar S (1997b) Similarity, connectionism, and the problem of representation in vision. *Neural Comput* 9, 701-721.
- Edelman S (1998) Representation is representation of similarities. *Behav Brain Sci* 21, 449-498.
- Edelman S, Grill-Spector K, Kushnir T, Malach R (1998) Toward direct visualization of the internal shape space by fMRI. *Psychobiology* 26, 309-321.
- Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap*. Chapman and Hall.
- Eger E, Henson RNA, Driver J, Dolan RJ (2004) BOLD repetition decreases in object-responsive ventral visual areas depend on spatial attention. *J Neurophysiol* 92, 1241-1247.
- Eger E, Schweinberger SR, Dolan RJ, Henson RN (2005) Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *Neuroimage* 26, 128-1139.
- Eger E, Ashburner J, Haynes J-D, Dolan RJ, Rees G (2008) fMRI activity patterns in human LOC carry information about object exemplars within category. *J Cogn Neurosci* 20, 356-370.
- Epstein R, Kanwisher N (1998). A cortical representation of the local visual environment. *Nature* 392, 598-601.
- Epstein R, Harris A, Stanley D, Kanwisher N (1999) The parahippocampal place area: Recognition, navigation or encoding? *Neuron* 23, 115-125.
- Epstein R, Graham KS, Downing PE (2003) Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron* 37, 865-876.
- Epstein RA, Parker WE, Feiler AM (2008) Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms. *J Neurophysiol* 99, 2877-2886.
- Evans JJ, Heggis AJ, Antoun N, Hodges JR (1995) Progressive prosopagnosia associated with selective right temporal lobe atrophy. *Brain* 118, 1-13.
- Fang F, Murray SO, Kersten D, He S (2005) Orientation-tuned fMRI adaptation in human visual cortex. *J Neurophysiol* 94, 4188-4195.
- Fischl B, Sereno MI, Tootell RBH, Dale AM (1999) High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp* 8, 272-284.
- Földiák P (1991) Learning invariance from transformation sequences. *Neural Comput* 3, 194-200.
- Földiák P, Xiao D, Keyser C, Edwards R, Perrett DI (2004) Rapid serial visual representation for the determination of neural selectivity in area STSa. *Prog Brain Res* 144, 107-116.

- Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312-316.
- Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J Neurosci* 23, 5235-5246.
- Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. *Nature* 443, 85-88.
- Friston KJ, Jezzard P, Turner R (1994) Analysis of functional MRI time-series. *Hum Brain Mapp* 1, 153-171.
- Friston KJ, Holmes AP, Poline J-B, Grasby PJ, Williams SCR, Frackowiak RSJ, Turner R (1995a). Analysis of fMRI time-series revisited. *Neuroimage* 2, 45-53.
- Friston KJ, Holmes AP, Worsley KJ, Poline J-P, Frith CD, Frackowiak RSJ (1995b) Statistical parametric maps in functional imaging: A general linear approach. *Hum Brain Mapp* 2, 189-210.
- Friston K, Chu C, Mourao-Miranda J, Hulme O, Rees G, Penny W, Ashburner J (2008) Bayesian decoding of brain images. *Neuroimage* 39, 181-205.
- Gärdenfors P (2000) *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.
- Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000a) Expertise for cars and birds recruits brain areas involved in face recognition. *Nat Neurosci* 3, 191-197.
- Gauthier I, Tarr MJ, Moylan J, Skudlarski P, Gore JD, Anderson AW (2000b) The fusiform 'face area' is part of a network that processes faces at the individual level. *J Cogn Neurosci* 12, 495-504.
- Gegenfurtner KR, Kiper DC, Levitt JB (1997) Functional properties of neurons in macaque area V3. *J Neurophysiol* 77, 1906-1923.
- George N, Dolan RJ, Fink GR, Baylis GC, Russell C, Driver J (1999) Contrast polarity and face recognition in the human fusiform gyrus. *Nat Neurosci* 2, 574-580.
- Gobbini MI, Leibenluft E, Santiago N, Haxby JV (2004) Social and emotional attachment in the neural representation of faces. *Neuroimage* 22, 1628-1635.
- Goebel R, Singer W (1999) Cortical surface-based statistical analysis of functional magnetic resonance imaging data. *Neuroimage* 9, S64.
- Goebel R, Esposito F, Formisano E (2006) Analysis of functional image analysis contest (FIAC) data with BrainVoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum Brain Mapp* 27, 392-401.
- Goebel R (2007) Localization of brain activity using functional magnetic resonance imaging. In C Stippich (Ed.) *Clinical Functional MRI - Presurgical Functional Neuroimaging*. Heidelberg: Springer.
- Goldstone R (1994) An efficient method for obtaining similarity data. *Behav Res Methods Instrum Comput* 26, 381-386.
- Gorno-Tempini ML, Price CJ, Josephs O, Vandenberghe R, Cappa SF, Kapur N, Frackowiak RSJ (1998) The neural systems sustaining face and proper-name processing. *Brain* 121, 2103-2118.
- Gorno-Tempini ML, Price CJ (2001) Identification of famous faces and buildings. A functional neuroimaging study of semantically unique items. *Brain* 124, 2087-2097.
- Grabowski TJ, Damasio H, Tranel D, Boles Ponto LL, Hichwa RD, Damasio AR (2001) A role for left temporal pole in the retrieval of words for unique entities. *Hum Brain Mapp* 13, 199-212.
- Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. *Vision Res* 41, 1409-1422.
- Grill-Spector K, Malach R (2001) fMRI-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol* 107, 293-321.
- Grill-Spector K, Knouf N, Kanwisher N (2004) The fusiform face area subserves face perception, not generic within-category identification. *Nat Neurosci* 7, 555-562.
- Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci* 10, 14-23.
- Gross CG, Rocha-Miranda CE, Bender DB (1972) Visual properties of neurons in inferotemporal cortex of the macaque. *J Neurophysiol* 35, 96-111.

- Hanson SJ, Matsuka T, Haxby JV (2004) Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage* 23, 156-166.
- Harel N, Ugurbil K, Uludag K, Yacoub, E (2006) Frontiers of brain mapping using fMRI. *J Magn Reson Imaging* 23, 945-957.
- Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634-1640.
- Haushofer J, Livingstone M, Kanwisher N (2008) Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biol* 6:e187.
- Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. *Trends Cogn Sci* 4, 223-233.
- Haxby JV, Gobbini MI, Furoy M, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2425-2430.
- Haynes J-D, Rees G (2005a) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci* 8, 686-691.
- Haynes JD, Rees G (2005b) Predicting the stream of consciousness from activity in human visual cortex. *Curr Biol* 15, 1301-1307.
- Haynes J-D, Rees G (2006) Decoding mental states from brain activity in humans. *Nat Rev Neurosci* 7, 523-534.
- Haynes J-D, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE (2007) Reading hidden intentions in the human brain. *Curr Biol* 17, 323-328.
- Hegd  J, Van Essen DC (2000) Selectivity for complex shapes in primate visual area V2. *J Neurosci* 20, RC61.
- Henson R, Shallice T, Dolan R (2000) Neuroimaging evidence for dissociable forms of repetition priming. *Science* 287, 1269-1272.
- Henson RNA, Shallice T, Gorno-Tempini ML, Dolan RJ (2002) Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cereb Cortex* 12, 178-186.
- Henson RN, Mouchlianitis E (2007) Effect of spatial attention on stimulus-specific haemodynamic repetition effects. *Neuroimage* 35, 1317-1329.
- Hesterberg TC (2007) *Bootstrap*, <http://home.comcast.net/~timhesterberg/articles/tech-encyclopedia.pdf> under review.
- Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *J Physiol* 195, 215-243.
- Humphreys GW, Forde EME (2001) Hierarchies, similarity, and interactivity in object recognition: "Category-specific" neuropsychological deficits. *Behav Brain Sci* 24, 453-509.
- Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863-866.
- Hyde JS, Biswal BB, Jesmanowicz A (2001) High-resolution fMRI using multislice partial k-space GR-EPI with cubic voxels. *Magn Reson Med* 46, 114-125.
- Johnson SC (1967) Hierarchical Clustering Schemes. *Psychometrika* 2, 241-254.
- Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8, 679-685.
- Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the visual cortex. *Curr Biol* 16, 1096-1102.
- Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci* 17, 4302-4311.
- Kanwisher N, Stanley D, Harris A (1999) The fusiform face area is selective for faces not animals. *Neuroreport* 10, 183-187.
- Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. *Nature* 452, 352-355.
- Kayaert G, Biederman I, Vogels R (2005) Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cereb Cortex* 15, 1308-1321.
- Kiani R, Esteky H, Tanaka K (2005) Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces. *J Neurophysiol* 94, 1587-1596.

- Kiani R, Esteky H, Mirpour K, Tanaka K (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J Neurophysiol* 97, 4296-4309.
- Kiehl K, Laurens KR, Duty TL, Forster BB, Liddle PF (2001) An event-related fMRI study of visual and auditory oddball tasks. *J Psychophysiol* 15, 221-240.
- Kietzmann TC, Lange S, Riedmiller M (2008) Computational object recognition – a biologically motivated approach. *Biol Cybern* 100, 59-79.
- Kirchner H, Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye-movements: visual processing speed revisited. *Vision Res* 46, 1762-1776.
- Koch C (1999) *Biophysics of computation: Information processing in single neurons*. New York: Oxford University Press.
- Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Phys Rev E* 69, 066138.
- Kravitz DJ, Kriegeskorte N, Baker CI (2010) High-level visual object representations are constrained by position. *Cereb Cortex* 20, 2916-2925.
- Kreiman G, Koch C, Fried I (2000) Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* 3, 946-953.
- Krekelberg B, Boynton GM, Van Wezel RJA (2006) Adaptation: from single cells to BOLD signals. *Trends Neurosci* 29, 250-256.
- Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103, 3863-3868.
- Kriegeskorte N, Bandettini P (2007a) Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage* 38, 649-662.
- Kriegeskorte N, Bandettini P (2007b) Combining the tools: Activation- and information-based fMRI analysis. *Neuroimage* 38, 666-668.
- Kriegeskorte N, Formisano E, Sorger B, Goebel R (2007) Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci USA* 104, 20600-20605.
- Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008a) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126-1141.
- Kriegeskorte N, Mur M, Bandettini PA (2008b) Representational similarity analysis –connecting the branches of systems neuroscience. *Front Syst Neurosci* doi:10.3389/neuro.06.004.2008.
- Kriegeskorte N, Cuzack R, Bandettini P (2010) How does an fMRI voxel sample the neuronal activity pattern: Compact kernel or complex spatiotemporal filter? *Neuroimage* 49, 1965-1976.
- Kruskal JB, Wish M (1978) *Multidimensional scaling*. Beverly Hills, CA: Sage Publications.
- Ku Sp, Gretton A, Macke J, Logothetis NK (2008) Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. *Magn Reson Imaging* 26, 1007-1014.
- Kwong KK, Belliveau JW, Chesler DA, Goldberg IE, Weisskoff RM, Poncelet BP, Kennedy DN, Hoppel BE, Cohen MS, Turner R, et al. (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc Natl Acad Sci USA* 89, 5675-5679.
- Laakso A, Cottrell GW (2000) Content and cluster analysis: Assessing representational similarity in neural systems. *Philos Psychol* 13, 47-76.
- LaConte S, Strother S, Cherkassky V, Anderson J, Hu X (2005) Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 26, 317-329.
- Lamp I, Ferster D, Poggio T, Riesenhuber M (2004) Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J Neurophysiol* 92, 2704-2713.
- Lane RD, Chua PML, Dolan RJ (1999) Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia* 37, 989-997.
- Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J Empirical Finance* 10, 603-621.
- Lehky SR, Sereno AB (2007) Comparison of shape encoding in primate dorsal and ventral visual pathways. *J Neurophysiol* 97, 307-319.

- Lerner Y, Epshtein B, Ullman S, Malach R (2008) Class information predicts activation by object fragments in human object areas. *J Cogn Neurosci* 20, 1189-1206.
- Leveroni CL, Seidenberg M, Mayer AR, Mead LA, Binder JR, Rao SM (2000) Neural systems underlying the recognition of familiar and newly learned faces. *J Neurosci* 20, 878-886.
- Levitt JB, Kiper DC, Movshon JA (1994) Receptive fields and functional architecture of macaque V2. *J Neurophysiol* 71, 2517-2542.
- Li N, DiCarlo JJ (2008) Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321, 1502-1507.
- Liu T, Cooper LA (2001) The influence of task requirements on priming in object decision and matching. *Mem Cognit* 29, 874-882.
- Liu T, Pestilli F, Carrasco M (2005) Transient attention enhances perceptual performance and fMRI response in human visual cortex. *Neuron* 45, 469-477.
- Loffler G, Gordon GE, Wilkinson F, Goren D, Wilson HR (2005) Configural masking of faces: evidence for high-level interactions in face-perception. *Vision Res* 45, 2287-2297.
- Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. *Curr Biol* 5, 552-563.
- Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150-157.
- Logothetis NK (2008) What we can do and what we cannot do with fMRI. *Nature Rev* 453, 869-878.
- Mahon BZ, Milleville SC, Negri GAL, Rumiati RI, Caramazza A, Martin A (2007) Action-related properties shape object representations in the ventral stream. *Neuron* 55, 507-520.
- Mahon BZ, Anzellotti S, Schwarzbach J, Zampini M, Caramazza A (2009) Category-specific organization in the human brain does not require visual experience. *Neuron* 63, 397-405.
- Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden PJ, Brady TJ, Rosen BR, Tootell RBH (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc Natl Acad Sci USA* 92, 8135-8139.
- Marotta JJ, Genovese CR, Behrmann M (2001) A functional MRI study of face recognition in patients with prosopagnosia. *Neuroreport* 12, 1581-1587.
- Martin A, Wiggs CL, Ungerleider LG, Haxby JV (1996) Neural correlates of category-specific knowledge. *Nature* 379, 649-52.
- Martin A (2007) The representations of object concepts in the brain. *Annu Rev Psychol* 58, 25-45.
- Mazard A, Schiltz C, Rossion B (2006) Recovery from adaptation to facial identity is larger for upright than inverted faces in the human occipito-temporal cortex. *Neuropsychologia* 44, 912-922.
- McClelland JL, Rogers TT (2003) The parallel distributed processing approach to semantic cognition. *Nat Rev Neurosci* 4, 310-322.
- Miller EK, Nieder A, Freedman DJ, Wallis JD (2003) Neural correlates of categories and concepts. *Curr Opin Neurobiol* 13, 198-203.
- Minamimoto T, Saunders RC, Richmond BJ (2010) Monkeys quickly learn and generalize visual categories without lateral prefrontal cortex. *Neuron* 66, 501-507.
- Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X (2004) Learning to decode cognitive states from brain images. *Mach Learn* 57, 145-175.
- Mourao-Miranda J, Bokde ALW, Born C, Hampel H, Stetter M (2005) Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on fMRI data. *Neuroimage* 28, 980-995.
- Movshon JA, Lennie P (1979) Pattern-selective adaptation in visual cortical neurones. *Nature* 278, 850-852.
- Muckli L, Kohler A, Kriegeskorte N, Singer W (2005) Primary visual cortex activity along the apparent-motion trace reflects illusory perception. *PLoS Biol* 3, e265.
- Mueller JR, Metha AB, Krauskopf J, Lennie P (1999) Rapid adaptation in visual cortex to the structure of images. *Science* 285, 1405-1408.
- Mummery CJ, Patterson K, Price CJ, Ashburner J, Frackowiak RSJ, Hodges JR (2000) Semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Ann Neurol* 47, 36-45.

- Murray SO, Wojciulik E (2004) Attention increases neural selectivity in the human lateral occipital complex. *Nat Neurosci* 7, 70-74.
- New J, Cosmides L, Tooby J (2007) Category-specific attention for animals reflects ancestral priorities, not expertise. *Proc Natl Acad Sci USA* 104, 16598-16603.
- Nestor A, Plaut DC, Behrmann M (2011) Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proc Natl Acad Sci USA* 108, 9998-10003.
- Ng M, Ciaramitaro VM, Anstis S, Boynton GM, Fine I (2006) Selectivity for the configural cues that identify gender, ethnicity, and identity of faces in human cortex. *Proc Natl Acad Sci USA* 103, 19552-19557.
- Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* 15, 1-25.
- Nicolelis MAL, Dimitrov D, Carmena JM, Crist R, Lehew G, Kralik JD, Wise SP (2003) Chronic, multisite, multielectrode recordings in macaque monkeys. *Proc Natl Acad Sci USA* 100, 11041-11046.
- Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10, 424-430.
- O'Craven KM, Downing PE, Kanwisher N (1999) fMRI evidence for objects as the units of attentional selection. *Nature* 401, 584-587.
- Ogawa S, Lee TM, Kay AR, Tank DW (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci USA* 87, 9868-9872.
- Ogawa S, Tank DW, Menon R, Ellerman JM, Kim S-G, Merkle H, Ugurbil K (1992) Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proc Natl Acad Sci USA* 89, 5951-5955.
- Op de Beeck H, Wagemans J, Vogels R (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat Neurosci* 4, 1244-1252.
- Op de Beeck HP, Torfs K, Wagemans J (2008) Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J Neurosci* 28, 10111-10123.
- Orban GA, Van Essen D, Vanduffel W (2004) Comparative mapping of higher visual areas in monkeys and humans. *Trends Cogn Sci* 8, 315-324.
- O'Toole AJ, Jiang F, Abdi H, Haxby JV (2005) Partially distributed representations of objects and faces in ventral temporal cortex. *J Cogn Neurosci* 17, 580-590.
- Palermo R, Rhodes G (2007) Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia* 45, 75-92.
- Patterson K, Nestor PJ, Rogers TT (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat Rev Neurosci* 8, 976-987.
- Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199-S209.
- Pessoa L, Padmala S (2006) Decoding near-threshold perception of fear from distributed single-trial brain activation. *Cereb Cortex* 17, 691-701.
- Poirson AB, Wandell BA (1993) Appearance of colored patterns: pattern-color separability. *J Opt Soc Am A* 10, 2458-2470.
- Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes retrieval during memory search. *Science* 310, 1963-1966.
- Pourtois G, Schwartz S, Seghier ML, Lazeyras F, Vuilleumier P (2005) Portraits or people? Distinct representations of face identity in the human visual cortex. *J Cogn Neurosci* 17, 1043-1057.
- Probic G, Jefferies E, Lambon Ralph MA (2007) Anterior temporal lobes mediate semantic representation: Mimicking semantic dementia by using rTMS in normal participants. *Proc Natl Acad Sci USA* 104, 20137-20141.
- Prüssmann KP (2004) Parallel imaging at high field strength: Synergies and joint potential. *Top Magn Reson Imaging* 15, 237-244.
- Puce A, Allison T, Gore JC, McCarthy G (1995) Face-sensitive regions in human extrastriate cortex studied by functional MRI. *J Neurophysiol* 74, 1192-1199.

- Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102-1107
- Raizada RDS, Tsao FM, Liu HM, Kuhl PK (2010) Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cereb Cortex* 20, 1-12.
- Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RBH (2011) The "Parahippocampal Place Area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biol* 9(4), e1000608.
- Reichardt W (1969). Movement perception in insects. In Reichardt W (Ed.) *Processing of optical data by organisms and by machines*. New York: Academic Press.
- Rieke F, Warland D, De Ruyter van Steveninck R, Bialek W (1999) *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.
- Riesenhuber M, Poggio T (2002) Neural mechanisms of object recognition. *Curr Opin Neurobiol* 12, 162-168.
- Risvik E, McEwan JA, Colwill JS, Rogers R, Lyon DH (1994) Projective mapping: A tool for sensory analysis and consumer research. *Food Qual Prefer* 5, 263-269.
- Rossion B (2008) Constraining the cortical face network by neuroimaging studies of acquired prosopagnosia. *Neuroimage* 40, 423-426.
- Rotshtein P, Henson RNA, Treves A, Driver J, Dolan RJ (2005) Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat Neurosci* 8, 107-113.
- Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323-2326.
- Sato N, Nakamura K (2003) Visual response properties of neurons in the parahippocampal cortex of monkeys. *J Neurophysiol* 90, 876-886.
- Sawamura H, Orban GA, Vogels R (2006) Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the fMRI adaptation paradigm. *Neuron* 49, 307-318.
- Schiltz C, Sorger B, Caldara R, Ahmed F, Mayer E, Goebel R, Rossion B (2006) Impaired face discrimination in acquired prosopagnosia is associated with abnormal response to individual faces in the right middle fusiform gyrus. *Cereb Cortex* 16, 574-586.
- Schyns PG, Goldstone RL, Thibaut JP (1998) The development of features in object concepts. *Behav Brain Sci* 21, 1-54.
- Serences JT, Boynton GM (2007) The representation of behavioral choice for motion in human visual cortex. *J Neurosci* 27, 12893-12899.
- Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RBH (1995) Borders of multiple visual areas in humans revealed by functional MRI. *Science* 268, 889-893.
- Sergent J, Ohta S, MacDonald B (1992) Functional neuroanatomy of face and object processing. *Brain* 115, 15-36.
- Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual cortex. In: *Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, USA, June 2005.
- Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci USA* 104, 6424-6429.
- Shepard RN, Chipman S (1970) Second-order isomorphism of internal representations: Shapes of states. *Cogn Psychol* 1, 1-17.
- Shepard RN, Kilpatrick DW, Cunningham JP (1975) The internal representation of numbers. *Cogn Psychol* 7, 82-138.
- Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. *Science* 210, 390-398.
- Sigala N, Logothetis NK (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415, 318-320.
- Spiridon M, Kanwisher N (2002) How Distributed Is Visual Category Information in Human Occipito-Temporal Cortex? An fMRI Study. *Neuron* 35, 1157-1165.
- Stringer SM, Rolls ET (2000) Position invariant recognition in the visual system with cluttered environments. *Neural Netw* 13, 305-315.

- Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, Frutiger S, Muley S, LaConte S, Rotenberg D (2002) The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *Neuroimage* 15, 747-771.
- Sugiura M, Kawashima R, Nakamura K, Sato N, Nakamura A, Kato T, Hatano K, Schormann T, Zilles K, Sato K, Ito K, Fukuda H (2001) Activation reduction in anterior temporal cortices during repeated recognition of faces of personal acquaintances. *Neuroimage* 13, 877-890.
- Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. *Nat Neurosci* 11, 1004-1006.
- Talairach J, Tournoux P (1988) *Co-planar stereotaxic atlas of the human brain*. New York: Thieme Medical Publishers.
- Tanaka K (1996). Inferotemporal cortex and object vision. *Annu Rev Neurosci* 19, 109-139.
- Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319-2323.
- Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. *Nature* 381, 520-522.
- Tolias AS, Keliris GA, Smirknakis SM, Logothetis NK (2005) Neurons in macaque area V4 acquire directional tuning after adaptation to motion stimuli. *Nat Neurosci* 8, 591-593.
- Tootell RBH, Hadjikhani NK, Vanduffel W, Liu AK, Mendola JD, Sereno MI, Dale AM (1998) Functional analysis of primary visual cortex (V1) in humans. *Proc Natl Acad Sci USA* 95, 811-817.
- Tootell RB, Tsao D, Vanduffel W (2003) Neuroimaging Weighs In: Humans Meet Macaques in "Primate" Visual Cortex *J Neurosci* 23, 3981-3989.
- Torgerson WS (1958) *Theory and Methods of Scaling*. New York: Wiley.
- Tovee, MJ, Rolls, ET, Treves, A, Bellis, RP (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. *J Neurophysiol* 70, 640-654.
- Tranel D, Damasio H, Damasio, AR (1997) A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia* 35, 1319-1327.
- Tsao DY, Freiwald WA, Knutsen TA, Mandeville JB, Tootell RBH (2003). Faces and objects in macaque cerebral cortex. *Nat Neurosci* 6, 989-995
- Tsao DY, Freiwald WA, Tootell RBH, Livingstone MS (2006). A cortical region consisting entirely of face-selective cells. *Science* 311, 670-674.
- Tyler LK, Moss HE (2001) Towards a distributed account of conceptual knowledge. *Trends Cogn Sci* 5, 244-252.
- Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. *Nat Neurosci* 5, 682-687.
- Ullman S (2007) Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn Sci* 11, 58-64.
- Ungerleider LG, Mishkin M (1982) Two cortical visual systems. In DJ Ingle, MA Goodale, RJW Mansfield (Eds.) *Analysis of Visual Behavior*. Cambridge, MA: MIT Press.
- Ungerleider LG, Haxby JV (1994) 'What' and 'where' in the human brain. *Curr Opin Neurobiol* 4, 157-165.
- Van Essen DC, Lewis JW, Drury HA, Hadjikhani N, Tootell RBH, Bakircioglu M, Miller MI (2001) Mapping visual cortex in monkeys and humans using surface-based atlases. *Vision Res* 41, 1359-1378.
- Van Essen DC, Dierker DL (2007) Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron* 56, 209-25.
- Van Horn JD, Wolfe J, Agnoli A, Woodward J, Schmitt M, Dobson J, Schumacher S, Vance B (2005) Neuroimaging databases as a resource for scientific discovery. *Int Rev Neurobiol* 66, 55-87.
- Vogels R (1999) Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur J Neurosci* 11, 1239-1255.
- Von Luxburg U (2007) A Tutorial on Spectral Clustering. *Stat Comput* 17, 395-416. (See also Technical Report 149, Max Planck Institute for Biological Cybernetics, 2006.)
- Vuilleumier P, Henson RN, Driver J, Dolan RJ (2002) Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nat Neurosci* 5, 491-499.

REFERENCES

- Wager TD, Nichols TE (2003) Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage* 18, 293-309.
- Wandell BA, Dumoulin SO, Brewer AA (2007) Visual field maps in human cortex. *Neuron* 56, 366-383.
- Warrington EK, Shallice T (1984) Category specific semantic impairments. *Brain* 107, 829-853.
- Williams MA, Berberovic N, Mattingley JB (2007a) Abnormal fMRI adaptation to unfamiliar faces in a case of developmental prosopamnesia. *Curr Biol* 17, 1259-1264.
- Williams MA, Dang S, Kanwisher N (2007b) Only some spatial patterns of fMRI response are read out in task performance. *Nat Neurosci* 10, 685-686.
- Williams MA, Baker CI, Op de Beeck HP, Shim WM, Dang S, Triantafyllou C, Kanwisher N (2008) Feedback of visual object information to foveal retinotopic cortex. *Nat Neurosci* 11, 1439-1445.
- Winston JS, Henson RNA, Fine-Goulden MR, Dolan RJ (2004) fMRI-Adaptation reveals dissociable neural representations of identity and expression in face perception. *J Neurophysiol* 92, 1830-1839.
- Wojciulik E, Kanwisher N, Driver J (1998) Covert visual attention modulates face-specific activity in human fusiform gyrus: fMRI study. *J Neurophysiol* 79, 1574-1578.
- Worsley KJ, Evans AC, Marrett S, Neelin P (1992) A three-dimensional statistical analysis for CBF activation studies in human brain. *J Cereb Blood Flow Metab* 12, 900-918.
- Worsley KJ, Friston KJ (1995) Analysis of fMRI time-series revisited – again. *Neuroimage* 2, 173-181.
- Worsley KJ, Poline J-B, Friston KJ, Evans AC (1997) Characterizing the response of PET and fMRI data using multivariate linear models. *Neuroimage* 6, 305-319.
- Yacoub E, Duong TQ, Van De Moortele PF, Lindquist M, Adriany G, Kim SG, Ugurbil K, Hu X (2003) Spin-echo fMRI in humans using high spatial resolutions and high magnetic fields. *Magn Reson Med* 49, 655-664.
- Young MP, Yamane S (1992) Sparse population coding of faces in inferotemporal cortex. *Science* 256, 1327-1331.
- Yovel G, Kanwisher N (2005) The neural basis of the behavioral face-inversion effect. *Curr Biol* 15, 2256-2262.
- Zhang X, Wandell BA (1997) A spatial extension of CIELAB for digital color image reproduction. *SID Journal*.

Acknowledgments

I would like to thank Niko for his inspiring supervision and friendship. Thank you for teaching me how to perform research, for showing me how much fun science can be, for answering my questions, for pushing me just a little further than I thought I could go. I feel lucky our paths crossed and converged into a fruitful collaboration. In Maastricht, I would like to thank Peter de Weerd for his constant support and wisdom. I am only realizing now how invaluable your advice has been at crucial moments in my professional life. I would like to thank Rainer for his never-ending enthusiasm, his vision, and his guidance from the early start of my journey into the world of science. At the NIH, I would like to thank Peter Bandettini for offering me the opportunity to perform research in his lab, for sharing his excitement about research, and for giving honest advice at the right moments. I would also like to thank Doug, who has shaped my view on what makes a good scientist and who was always there to help; Jerzy, who contributed substantially by sharing his positive attitude and his vast knowledge on MRI physics and design; Mirjam, who taught me by teaching her, and who became a close friend and housemate; Ning, whom it was a pleasure to collaborate with; and Kay and Dorian for their considerate and efficient help in administrative matters. I would further like to thank the GPP staff, especially Caroline and Sharon, for their continuous interest in and support for my graduate training and education. In Maastricht, I would like to thank Annemie, Christl, and Riny for their valuable administrative support, especially during the last phases of my PhD trajectory. Finally, I would like to thank Wil Botden for her adequate career advice.

I would like to thank my colleagues, both at the NIH and in Maastricht, for creating a stimulating and positive atmosphere. Bob, Ziad, Pat, Sean, Adam, Rasmus, Kevin, and Dan, I am still missing our lunch discussions and jokes. Wish I could just teleport myself over to SFIM for lunch once in a while. About time I send you some more black licorice and spice cakes. Job and Michelle, it has been a pleasure to share an office with you, to organize CN outings together, to talk about research, teaching, and university politics, and chase the winter away with cosy Christmas lights. 4.777(a) rules! Valerie, I have much enjoyed our conversations and collaboration over the past two years. Federico, I will always remember the wonderful dinners at your place, and our earnest conversations. Joao, you have been such a great housemate and friend. I really miss our diner conversations about science and everything else. Michael, where would I have been without your 'taartjes' and climbing advice? Kamil, thanks for adding your positive vibe to the group. All the other CN colleagues, thank you for the good times at international dinner parties, CN weekends, and conferences. Members of ProVUM and the University Council, it has been a unique experience, thank you for teaching me about university politics, and for joining forces to make things work.

Next, I would like to thank de kipjes! Life is definitely more fun because you are around. Aline, thank you for staying in touch when I was in the US, for being there to listen and give advice, and for convincing me to go out and party. Marin, thank you for making me see things in perspective, for sharing your thoughts and feelings, and for helping out when needed. Anne, thank you for sharing frustrations on research and relationships during late-night conversations. Floor, thank you for your directness, energy, and enthusiasm. Lauran, Jasper, Marnix, Tim, and Chun, what would de kipjes have done without you guys? Thanks for all the nice parties, late breakfasts, tv-nights, visits (especially Chun!), and good times we have had. Also thanks to the Delft crew for organizing fun parties and trips. These good times were the best recipe against work-related stress.

Helen, Nina, Christianne, Saskia, Rilana, and Sanae, thank you for the fun dinners and movie nights, and for coming to visit me in the US. Nina and Helen, I have really enjoyed our relaxing horse back rides through the Maastricht countryside. Lauri, thank you for your friendship over the past years, talking to you always makes me feel better. Ewout, thank you for showing me that shopping can actually be fun. Manuel, your text messages on work and life have often made me smile.

I would like to thank Jo, Brie, Kimbo, and Meghan, my family away from home, for teaching me about US culture and making me feel comfortable in a new place. Christina, thank you for all the times we talked and hung out, and for being there when times were a bit rough. Peter, thank you for showing me DC, and for teaching me about science and life. Jonathan, I really enjoyed our day trips and cooking adventures. Heng, my dear and oldest friend, thank you for being the first to make me feel at home, and for staying in touch ever since.

Finally, I would like to thank my family for their unconditional support and understanding. It is difficult to put into words how important this has been to me. I would like to thank my mam for her endless patience and wise advice. You probably know me better than I do myself. I would like to thank my dad for his indispensable help in technical matters, and for understanding my late working hours. Carine, lievelingszusje, thank you for holding up a mirror when I most needed it and for making me feel that everything will be all right. Carianne, Wim, Lennart, Niek, tante Ien, and oom Niek: your warmth and interest in my endeavors has been a source of strength. I also would like to thank Peter for his unconditional help in so many matters, Tiny for wonderful Friday night dinners, and Toon for his humour and positive attitude towards life. Last but not least, Zohar, thank you for your care and respect, for your optimism, for not being distracted by the fact that I moved to a different country, and for everything else.

Curriculum Vitae

Marieke Mur studied psychology and cognitive neuroscience at Maastricht University, and graduated cum laude in 2006. Her master thesis on the neural correlates of face recognition was based on a research internship at the National Institutes of Health (NIH) in the US, where she worked with Dr. Kriegeskorte and Dr. Bandettini. She entered a joint PhD program involving both the NIH and Maastricht University in 2006, continuing the investigation of high-level visual object representations in the brain using functional magnetic resonance imaging (fMRI). She was supervised by Dr. Kriegeskorte and Dr. Bandettini at the NIH, and by Prof. De Weerd and Prof. Goebel at Maastricht University. She is currently working as a postdoc at the MRC Cognition and Brain Sciences Unit in Cambridge, UK, on a two-year Rubicon grant from the Netherlands Organization for Scientific Research. She is investigating the influence of top-down attentional modulation on high-level visual object representations, under the supervision of Prof. Duncan. Please see below for a list of publications.

Publications

Mur M, Ruff DA, Bodurka J, Bandettini PA, Kriegeskorte N (2010) Face-identity change activation outside the face system: “Release from adaptation” may not always indicate neuronal selectivity. *Cereb Cortex* 20, 2027-2042, doi: 10.1093/cercor/bhp272.

Mur M, Bandettini PA, Kriegeskorte N (2009) Revealing representational content with pattern-information fMRI – an introductory guide. *Soc Cogn Affect Neurosci* 4, 101-109. doi: 10.1093/scan/nsn044.

Kriegeskorte N, **Mur M**, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126-1141.

Kriegeskorte N, **Mur M**, Bandettini P (2008) Representational similarity analysis – connecting the branches of systems neuroscience. *Front Syst Neurosci* 2, doi: 10.3389/neuro.06.004.2008.

Broers NJ, **Mur MC**, Bude L (2005) Directed self explanation in the study of statistics. In Burrill G, Camden M (Eds.) *Curricular development in statistics education*. Voorburg, The Netherlands: International Statistical Institute.

Manuscripts submitted or in preparation

Mur M, Ruff D, Bodurka J, De Weerd P, Bandettini P, Kriegeskorte N. Single-image activation of category-selective regions in human inferior temporal cortex. In revision, *J Neurosci*.

Mur M, Meys M, Bodurka J, Goebel R, Bandettini P, Kriegeskorte N. Human object-similarity judgments reflect and transcend primate IT categorical object representations. In revision, *Front Psychology*.

Goffaux V, Schiltz C, **Mur M, Goebel R.** The saliency of local cues determines the strength of holistic face processing: behavioural and neuroimaging evidence. Submitted.

Kriegeskorte N, **Mur M.** Inverse MDS: inferring dissimilarity structure from multiple item arrangements. In preparation.

Abstracts and oral presentations

Liu N, Kriegeskorte N, **Mur M, Hadj-Bouziane F, Tootell RBH, Ungerleider LG** (2010) Patterns of fMRI response elicited by individual faces in macaque cerebral cortex. Society for Neuroscience Annual Meeting, San Diego, California, USA.

Mur M, Meys M, Bodurka J, Bandettini P, Kriegeskorte N (2009) Relating neural object representations to perceptual judgments with representational similarity analysis. Vision Science Society Annual Meeting, Naples, Florida, USA. *Oral presentation.*

Mur M, Ruff D, Bodurka J, Bandettini P, Kriegeskorte N (2008) Ranking 96 object images by their activation of FFA. Vision Science Society Annual Meeting, Naples, Florida, USA.

Kriegeskorte N, **Mur M, Ruff D, Kiani R, Bodurka J, Bandettini P** (2007) Matching categorical object representations in inferotemporal cortex of man and monkey. 13th Human Brain Mapping Conference, Chicago, Illinois, USA.

Mur M, Ruff D, Bodurka J, Bandettini P, Kriegeskorte N (2006) Recognizing a person by face: dissociating brain regions involved in perceptual and conceptual components of person identification. 12th Human Brain Mapping Conference, Florence, Italy.