Neural pattern similarity and visual perception

Thesis by

Jonathan Harel

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2012

(Defended May 21, 2012)

To my parents, Nily and William

Acknowledgments

Although my path at Caltech to the present moment has been circuitous, confusing, and difficult, along the way, I have had the privilege of being rewired and transformed by countless interactions with so many thoughtful people who challenged and impressed me. I would like to humbly offer here my deepest gratitude to my advisers, to whom I am forever indebted: Robert McEliece, for his unusual clarity of thought and crystalline explanations of information theory, for his seeing some potential in and offering me my first opportunity to research a problem (the practical workings of poset belief propagation), for inviting me to pursue a Ph.D. and generously continuing to support me even after I had decided on a change of direction; Pietro Perona, whose work in vision first interested me in the topic, and for our lengthy discussions deconstructing surprise and saliency; Christof Koch, who first interested me in neuroscience, for his enormous doctor-fatherly patience with me as I wended my way from visual attention to an interest in decoding brain activity, for his inspiring and invigorating full-throttle passion in the search for truth from all branches of human endeavor; Ralph Adolphs, whose talent in understanding, organizing, and managing of people is exemplary, for his generosity in guiding me through my last two years, and for his focused creativity and fruitful mentorship which contributed most significantly to originating and shaping the content of this thesis. I would also like to thank the chairman of my thesis committee, Jehoshua (Shuki) Bruck, for his time, effort, and support of my unconventional thesis topic for the department of Electrical Engineering.

I wish to thank some friends who have had a particularly important role in my intellectual life: Jeremy Thorpe, for teaching me about linear algebra, dynamic programming, and many matroids in between, and for encouraging my interest in coding theory – I can still feel his influence on my thinking today, twelve years after those first lessons; Demetri Spanos, for his principled contrarianism, for his putting up with my incessant questions, and whose firm understanding of mathematical ideas was often enlightening – his ideas informed early stages of Chapter 4 of this thesis, as well as my opinions on everything from stock investing to how to optimally shoot aliens; Ueli Rutishauser, who taught me among many other things the importance of ridge regression, nonparametric hypothesis tests, and a patience with and reverence for the delicate, incremental findings born in the progression of experimental neuroscience.

Ronnie Bryan, a great and imaginative friend, directly contributed his code (keypoint annotation and face morphing), time, and ideas to the work in Part I, and Catherine Holcomb competently and diligently facilitated many, many hours of brain scans for me. Alex Huth deserves credit for introducing me to a beautiful experiment which would change my course towards neuroimaging[1], for guiding me through my early encounters with its application, for helping me collect and retinotopize data for Chapter 4, and above all for expanding my social horizons and being an awesome friend. Cendri Hutcherson introduced me to SPM; Julien Dubois, Dan Kennedy, Jed Elison, and Mike Tyszka helped organize my thoughts about the results in Chapter 3 (Mike also has come to the rescue on many occasions with technical wizardry magnetic resonant); Melissa Saenz mentored me in my work with synesthetes, which became the basis for Chapter 5. I am thankful for these valuable contributions, without which this thesis would never have been possible.

I greatly enjoyed fraternizing with, and am most grateful for the encouragement I received from, Mortada Mehyar, Xiaodi Hou, Moran Cerf, and Kalen Jordan. These friendships have kept me sane and I will fondly remember all the foolish times together.

I would also like to thank the members of the Klab and the Adolphs Lab, and all the other quirky, clever, and adventurous people I've had the pleasure of meeting while at Caltech, including Sotiris Masmanidis, Himanshu Mishra, Uri Maoz, Rosemary Rohde, Costas Anastassiou, Amy Chou, Dirk Neumann, Alice Lin, Shuo Wang, Florian Mormann, Naotsugu Tsuchiya, Damian Stanley, Fatma Imamoglu, Kerry Fender, and Craig Countryman.

I wish to thank the subjects of my fMRI experiments, who bravely volunteered hours of their time to lie still in a dark, claustrophobic, and cacophonous tunnel, while I both annoyed and bored them with flashing images, in hopes of possibly understanding just a little bit more about what goes on in the brain during visual perception.

Finally, I wish to express my heartfelt gratitude to my family: to Anat and Kevin for all their love and warmth throughout the years, and for their intoxicating *joie de vivre*, and most importantly to my parents, whose scientific worldview set me on this trajectory long before I was able to appreciate it, and whose boundless love and endless support I am embarrassingly fortunate to have received.

[1] this one: [53]

Abstract

This thesis addresses the question of whether people actually see the same visual stimuli somehow differently, and under what conditions, if so. It is an experimental contribution to the basic understanding of visual and especially face perception, and its neural correlates, with an emphasis on comparing patterns of neural activity driven by visual stimuli across trials and across individuals. We make extensive use of functional magnetic resonance imaging (fMRI); all inferences about neural activity are made via this intermediary. The thesis is organized into two parts:

In Part I, we investigate the nature of face familiarity and distinctiveness at perceptual and neural levels. We first address the question of how the faces of those people personally familiar to a viewer appear different than they would to an unfamiliar viewer. The main result is that they appear more distinctive, i.e., dissimilar to and distinguishable from other faces, and more so the higher the level of familiarity. Having established this connection between face familiarity and distinctiveness, we ask next what is different about the perception of such faces, as compared with indistinct and unfamiliar faces, at the level of brain activation. We find that familiar and distinctive faces are represented more consistently: compared with indistinct faces, which evoke slightly different patterns of activity with each new presentation, these faces evoke slightly similar patterns. Combined with the observation that consistency can enhance memory encoding (a result reported by Xue et al. [102]), this suggests a cyclic process for the learning of unfamiliar faces in which consistent representation and the presence of newly formed memories mutually feedback on each other.

Whereas in Part I we focus on individual *differences* in neural activity, principally by experimentally manipulating stimulus familiarity, in Part II, we shift our focus to *similarities* across individuals and extend our investigation beyond faces to the perception of visual objects in general and moving images. We begin with an experiment involving the perception of static images selected from 44 object categories, where we find that the distances between these categories, induced from activity in cortical visual object areas, correlate highly between subjects, and also to distances inferred from a behavioral clustering task, and that this correlation remains significant even among subsets of closely related categories. We also show that one subject's brain activity can be accurately modeled using another's, and that this allows us to predict which image a subject is viewing based on his/her brain activity. Then, in a different experiment investigating the perception of dynamic/video stimuli, we find evidence that when watching videos with sound, visual attention is likely blurred at times and transferred to audition; subjects relatively temporally decorrelate in visual areas compared to the muted case, in which the patterns of neural activity correlate across subjects at an average of 78% the level found with oneself later in time.

The findings reported in this thesis thus offer quantitative lower bounds on how similarly different individuals neurally experience visual stimuli, and an explanation for how they perceptually and neurally diverge when familiarity with a (face) stimulus varies, suggesting a possible mechanism for the encoding of new visual objects into memory. We conclude with a discussion of some of the questions raised by this work and directions for future research.

Contents

Chapter 1

Introduction

Continually fluctuating streams of multicolored photons flow into the eye of an observer, lighting a grid of firing patterns among photoreceptive retinal neurons, sparking a chain reaction which spreads out into midbrain and through primary visual cortex, to higher reaches of the brain, where eventually a belief somehow emerges about the contents and configuration of the world. In between the cryptic and specific signals at the retina and the sweeping abstractions represented by individual neurons at high-level regions in the brain is a vast array of organized and back-feeding neural networks of increasingly general purpose. This thesis quantifies and compares the measurable activity of this process of *vision,* across time and across individuals.

One motivation for implementing this particular research program was to come closer to understanding how visual qualia, those internal experiences of conscious visual states, e.g., how the color red seems, compare across people. We can begin to provide some clues to the answer by quantitatively comparing visual cognition at its various levels, starting from its behavioral output and moving down to its cortical manifestations. To the extent that these subprocesses are similar across individuals, it is plausible that the accompanying qualia are too, as it seems unlikely that such similar physical processes could lead to very different mental states; however, this relates to what David Chalmers calls the hard problem of consciousness [15], and philosophers may disagree on whether the question is ultimately tractable, especially in this way. But if we take the practicable view that qualia are defined by their physical or informatic relationships to other mental states, as in Tononi's formalism [91], this style of investigation may prove fruitful.

To this end, and also to understand how subjects perceptually and neurally diverge, in particular when viewing faces, this thesis makes extensive use of functional magnetic resonance imaging (fMRI). fMRI has revolutionized the conduct of modern neuroscience by providing experimentalists with a tool to measure patterns of neural activity distributed across the entire brain in fully conscious

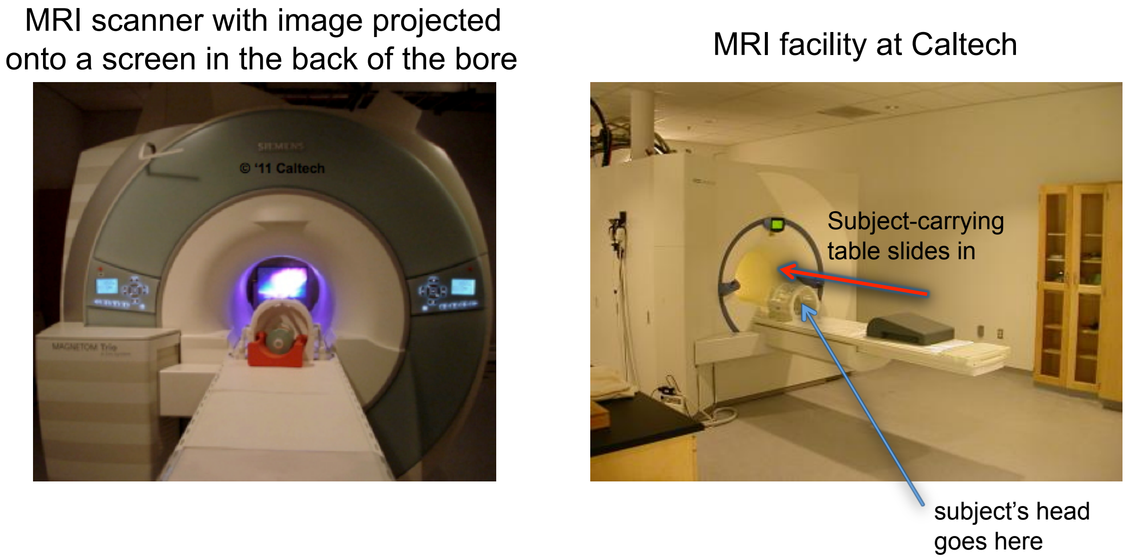

MRI scanner with image projected onto a screen in the back of the bore
SIEMENS
© '11 Caltech
MAGNETOM Trio
MRI facility at Caltech
Subject-carrying table slides in
subject's head goes here

Figure 1.1: Photographs of the human MR scanners at Caltech

humans and other animals. It is currently the best tool for achieving such whole-brain simultaneous coverage.

1.1 Neuroimaging with fMRI; visual cortex; faces

fMRI fMRI data are collected in MR scanners like the ones shown in Figure 1.1. An imaging subject enters the bore of the powerful and noisy superconducting magnet in the supine position for periods of 5 minutes to 1 hour at a time. During this time, bursts of radio-frequency energy are transmitted from a coiled antenna around the subject's head at a frequency resonant with the precession of hydrogen nuclei inside the magnetic field. The energy in these bursts temporarily excites the hydrogen nuclei out of alignment with the main magnetic field, the aggregate effect of which is magnetic flux through the transceiver head coil, inducing currents with characteristic relaxation time constants (such as T2*) which depend on the chemical environment (including relative concentration of deoxygenated hemoglobin) in which the hydrogen nuclei are embedded. Together with a systematic varying of the strength of the main magnetic field depending on the position in the scanner, and a frequency coding scheme, this endows the induced currents with information sufficient to reconstruct an image of the brain, one 2D slice at a time.

Although the functional MRI signal, also known as the BOLD (blood-oxygen-level-dependent) signal, is at base an estimation of a magnetic relaxation constant (T2*), which can be used to infer concentrations of deoxygenated hemoglobin, which change with blood flow to meet the metabolic demands of the brain, it has nonetheless been repeatedly shown to significantly correlate with the actual activity of neurons, specifically their local field potentials, especially in the gamma range, and to a lesser extent with spiking activity (see, e.g., Logothetis [62], [61] for careful explanation). We will take this relationship between the fMRI signal and neural activity as an assumption when interpreting our results.

For our purposes, the useful data from an fMRI scan of one subject can be summarized as a matrix β where the rows index individual "**voxels**" (3D pixels), i.e. discrete points in the brain, each of which may be assigned a regional label (such as "FFA" or "STS"), and the columns index either stimuli (Chapters 3 and 4) or equivalently time points for a single changing/dynamic stimulus (Chapter 5). Then β_{ij} = response to stimulus j at voxel i: a β or "beta" value is an estimate of the hemodynamic response amplitude at a particular location (viz., a voxel) in the brain to a particular stimulus. For this reason, beta values will also be called response amplitudes, or response magnitudes. A column of this β matrix may be considered a *spatial neural pattern*, and a row may be considered a *temporal neural pattern* (used only in Chapter 5).

Visual cortex For the work presented in this thesis, fMRI neuroimaging is mainly used to record activity in visual cortex. Some of the distinct regions inside visual cortex can be organized into a hierarchy, like the one shown in Figure 1.2. This includes primary visual cortex (V1), where an individual neuron responds only to visual input from a very restricted part of the visual field, and specifically to simple features within this field such as a bar or grating oriented at a specific angle (see [45] for original work and [69] for a theoretical model of it). V2, V3, and V4 are outwardly spread from V1 on the cortical surface, and are associated with incrementally larger and more flexible receptive fields. In order to gain some intuition about the kinds of features represented by intermediate-level neurons, see Figure 1.3; Gallant et al. [28] found neurons in monkey V4 highly selective for such stimuli.

Faces are special We devote half of this thesis to just one category of visual stimulus, focusing on differences in perceptions between people viewing exemplars from this category. This category is faces, arguably the most important category of visual stimulus. From the moment we are born [101, 85, 66], we begin to recognize them, and their importance in social functioning grows. Failure to recognize an individual [19], or the failure to recognize an emotion in a face [1], can have very high social cost. They are crucial for interpersonal interaction and communication, both verbal and not. They are perceived both holistically [20] and categorically [4]. In the brain, faces are processed faster

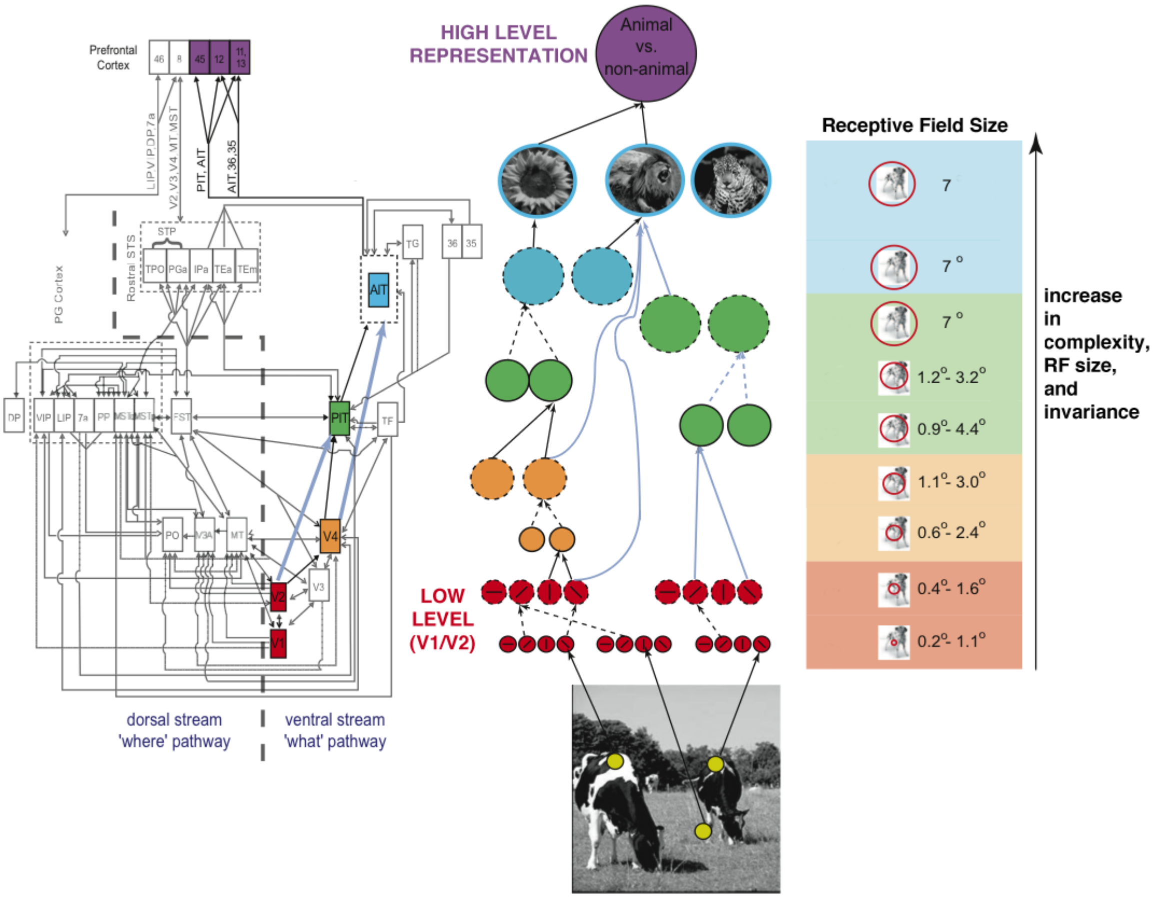


Figure 1.2: A schematic representation of the hierarchical processing stages through various regions in the primate visual cortex (adapted from [82]). The neurons in primary/low-level visual cortex (V1/V2) fire when simple visual features (such as oriented lines) occur in highly specific, small regions of the visual field (their receptive field sizes are small, $0.2 - 1.1°$ of visual angle). These combine into progressively more generalized features until somewhere high in cortex (such as the prefrontal cortex), a specific label such as "animal" can be applied to the input.

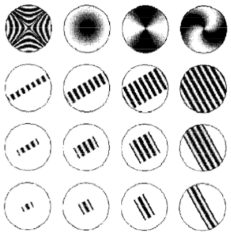


Figure 1.3: Example stimuli (nonstandard gratings) used by Gallant et al. [27, 28] to characterize the tuning of neurons in macaque monkey visual area V4.

than other types of stimuli [90], and they have been found to have specialized modules devoted just to their detection and individuation. These areas include the fusiform face area (FFA) [50] (though some view FFA as an expertise module, see [29]), and a network of face "patches" through temporal,

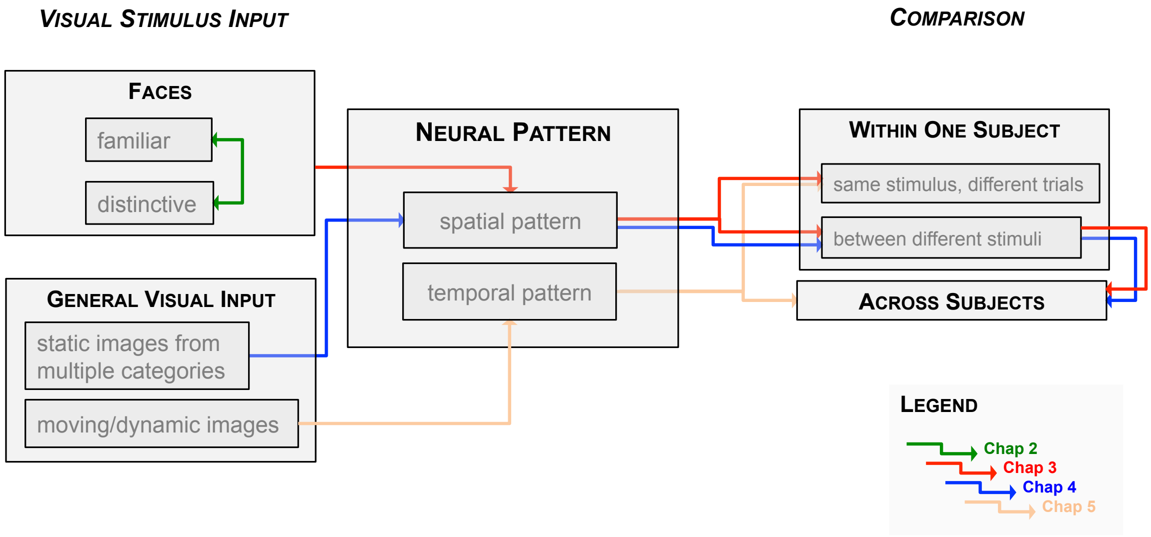


Figure 1.4: The visual stimuli and analyses used in each chapter are connected by logical arrows, meant to aid the reader to infer relationships such as "*In Chapter 4, the spatially distributed neural patterns elicited by different static images are compared within a subject, then these differences are compared across subjects.*" or "*In Chapter 2, a connection is established between familiar and distinctive faces.*"

occipital, and frontal lobes (see [93, 94, 95, 65]). A very wide range of other brain areas have been found to be involved in the perception of faces as well [39, 31, 49, 55, 59, 60], including hippocampus and amygdala [26, 58], and even primary auditory cortex [43]. A stronger case could not be made for any other category of visual stimulus.

1.2 Organization of thesis

Figure 1.4 schematizes the organization of this thesis, including which kinds of neural patterns are compared and how, in each chapter. Face perception is the focus of Part I of this, and in particular the perception of distinctive and familiar faces. General static and moving images are investigated in Part II. **Notes about types of neural pattern similarity:** In order to compare neural patterns elicited by static images across subjects, we first compute neural distances between different stimuli (summarized in *representational dissimilarity matrices*, or RDMs), and then compare those across subjects, an approach endorsed by Kriegeskorte [56, 74]. This static image comparison is used in Chapters 3 and 4. In Chapter 3, we perform another kind of neural pattern similarity analysis, in which the spatially distributed response pattern to a single stimulus is compared across its different presentation trials: this kind of neural similarity is what we term *consistency.* In Chapter 5, we use moving image stimuli, which allow direct comparison of neural response patterns across subjects, by using one element per time point then comparing the resultant temporal patterns in analogous brain locations.

1.2.1 Important brain regions mentioned throughout thesis

For reference, we provide a list of some important brain regions, their abbreviations, and associated functions:

- ◇ **V1**, primary visual cortex, also known as striate cortex [45, 18];
- ◇ **V2**, prestriate cortex, adjacent to and receiving feedforward connections from V1 [84];
- ◇ **V3**, third visual complex, a part of early visual cortex immediately adjacent to V2 [84];
- ◇ **V4**, part of extrastriate visual cortex, associated with intermediate visual features including gratings [27, 28];
- ◇ **MT**, originally "medial temporal" (but not in humans), an area of visual cortex associated with the representation of visual motion [92];
- ◇ **LO** (or **LOC**), lateral occipital cortex/complex, associated with visual objects [35];
- ◇ **Fusi**, the fusiform gyrus, a ventral stream part of visual cortex associated with high level object representation including faces and words [63];
- ◇ **FFA**, the fusiform face area, an area in fusiform gyrus which is face-selective [50];
- ◇ **STS**, superior temporal sulcus, associated with many functions, including audio-visual integration [5];
- ◇ **IT**, inferior temporal cortex, associated with face processing and other high level vision [56];
- ◇ **Cuneus**, associated with basic visual processing and inhibition [36];
- ◇ **Precuneus**, associated with self-perception and memory [14];
- ◇ **PostCing**, posterior cingulate cortex, associated with many functions, including awareness, memory retrieval, and familiarity [88];
- ◇ **AntCing**, anterior cingulate cortex, associated with motivation and attention [12];
- ◇ **IPL**, inferior parietal lobe (including lobule), associated with the many functions, including visual perception of emotion in faces [39, 72, 73].

Part I

Face distinctiveness and neural pattern similarity across trials

Chapter 2

Personally familiar faces appear to be more distinctive

Based on experiments involving ten human subjects, we conclude that faces which are familiar to a viewer appear to look more different from other faces than they do to unfamiliar viewers: that is, they appear more **distinctive**. Furthermore, we find evidence that faces which are entirely unfamiliar, merely similar in appearance to a familiar face, can be more easily distinguished than ones distant from any familiar face. These two effects taken together constitute a warping in the entire perceptual space around a learned face, pushing all faces in this region away from each other. We additionally show that the fact of familiarity with a face is predictable from performance level on a visual distinguishability task, and provide some preliminary results characterizing how facial features weigh differently when comparing familiar faces.

2.1 Previous work

The behavioral psychology and neuroscience literature on face perception is rich, and important within it is the notion of two distinct types of face encoding in the brain: on one side, there is norm-based coding, wherein faces are represented by their deviations from the average, and on the other, exemplar-based coding, wherein faces are represented by their distances to previously learned, i.e., familiar, faces. The most cited evidence supporting norm-based coding is the observation that caricatures, or exaggerated renderings of a face, are recognized more rapidly and more easily than the undistorted faces themselves [77, 57, 6]. Figure 2.1 shows an illustration of a “face space”, of the kind imagined by the norm-based coding hypothesis, in which deviating faces are scattered in various directions away from the origin, at which we find the average and least distinctive face. The figure is adapted from a paper by Leopold [59], in which he presents data characterizing the response

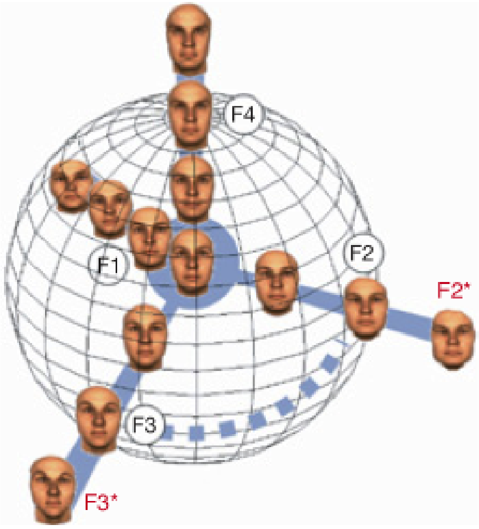


Figure 2.1: An illustration of a 3-dimensional "face space" (adapted from Leopold [59]), including the origin, at which we find the most average looking face, and several faces along orthogonal directions. Caricatures are defined as faces extrapolated beyond their true position in this space to a position more distant from the origin but along the same direction: e.g., F2* is a caricature of F2, and F3* is a caricature of F3.

patterns of single neurons in the anterior infero-temporal cortex of macaque monkeys, supporting the norm-based view.

Several authors have written on the relationship between familiarity and distinctiveness. Valentine and Bruce [97] wrote, in the context of an earlier hypothesis: "If judgements of distinctiveness depend upon some mean of a large population of faces, familiarity with a particular face should not alter its perceived distinctiveness." However, in their study, familiarity and distinctiveness were rated by two mutually exclusive sets of subjects respectively, so any reported relationship would bear on only memorability as an intrinsic property of a face, not on actual familiarity: they go on to provide data showing a positive but insignificant correlation between this independently rated familiarity of a face and its distinctiveness. Vokey and Read [99] found that familiarity and distinctiveness were anti-correlated; however, in their study familiarity was actually *entirely imagined* (subjects were misled to believe that some face stimuli were of people at the same school).

Two relevant studies have specifically examined the individual differences which arise in perception as a result of familiarity with a face. In the first, by Beale and Keil [4], in an experiment exploring the "categorical" perception of faces, that is, their tendency to not look as though they are from some undifferentiated continuum, the degree of familiarity with a pair of famous faces was highly correlated with the extent to which a kind of ordering performance was peaked at the center of the morphing continuum between them, and in the second, Ryu [79] showed that adaptation to a familiar face created a greater orientation after-effect than an unfamiliar face. As we shall see, these last two studies are compatible with the findings reported in this chapter, insofar as familiar faces are associated with a kind of heightened perceptual acuity.

Under a strictly norm-based model of face coding, one would expect little or no systematic relationship between the distinctiveness of a face and its degree of familiarity. Perhaps as a face is in the process of being learned, its position in face space would slowly converge on a final position from an initially noisy estimate, but, on average, one would not expect it to end up systematically more nor less distant from the origin, that is, different in distinctiveness, than it started. Also, its contribution to shifting the global experiential average, and thus its effect on the perception of other faces, would be very minor or negligible. However, under the exemplar-based coding model, after a face is learned, it can be used to help resolve subtle differences between faces which were previously distant from any reference, by adding a new one inside a neighborhood of comparability.

The results in this chapter show what happens to the perceptual face space around a familiar face.

2.2 Basic experimental setup

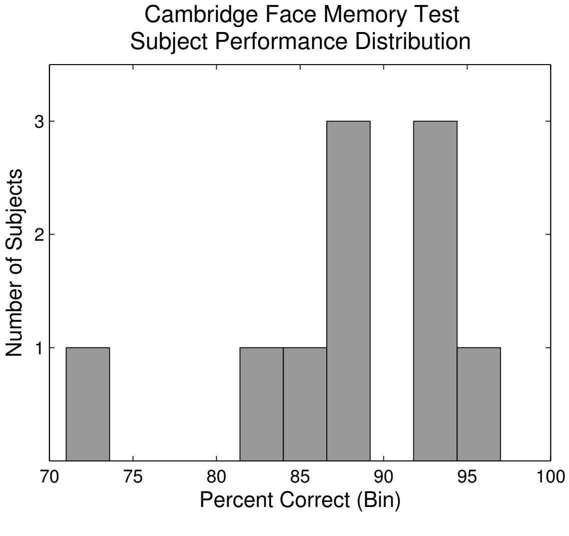


Figure 2.2: All ten participants scored in the normal range (score > 60) of face-recognition ability, as assessed by the Cambridge Face Memory Test ([23], http://www.faceblind.org/facetests/). Note, however, that one subject's (S0's) performance is a bit unusually low relative to the others.

Ten subjects viewed pairs of face stimuli, selected from a set of forty, presented in rapid temporal succession, on a computer screen in a laboratory at Caltech, providing similarity judgments between them using keyboard inputs, explicitly and implicitly (see below). The subjects were all Caucasian females, a selection intended to eliminate gender and race effects, which are known to influence face perception and which have already been extensively studied [13, 32, 68, 96, 3].

2.2.1 Participants

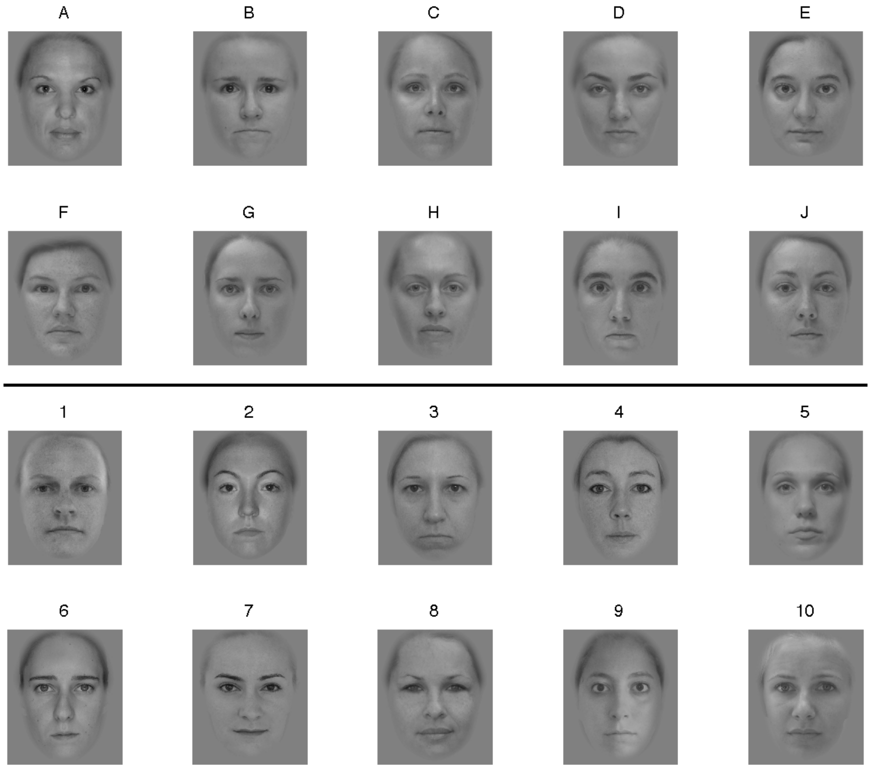


Figure 2.3: Each subject (A, B, ... J) is shown in the top half, with her sister occurring somewhere in the bottom half (1, 2, ... 10). Can you guess the 1:1 mapping between subjects and sisters? The answer key is given in Figure 2.48.

Ten healthy Caucasian adult females (range 23 - 29, mean 25.7 ± 0.55 years) participated in the series of experiments. All subjects had normal or corrected-to-normal vision, nine were right-handed, and all tested in the normal face recognition ability range as assessed by the Cambridge Face Memory test on synthetic faces (by Duchaine and Nakayama [23], http://www.faceblind.org/facetests/); see Figure 2.2 for the distribution of scores on this test. We note that one subject's score was unusually low given the distribution of the others' scores, though still in the normal range.

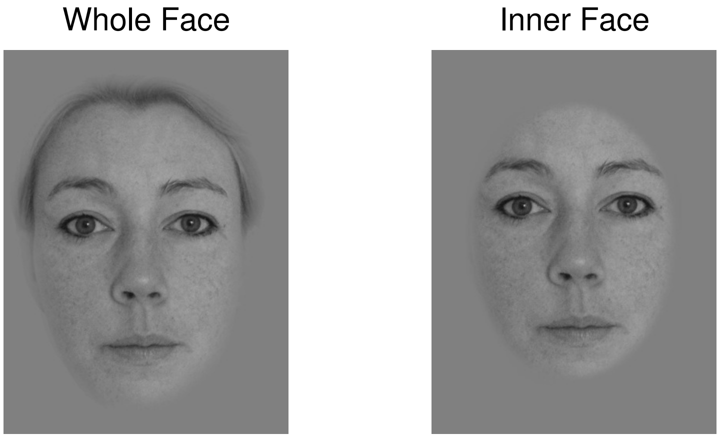


Figure 2.4: An example face from the dataset masked to include hair and jawline (left), and to only reveal inner features (right). The SimRate task was performed with both kinds of masks, but MorphDiscrim was performed only with inner faces.

2.2.2 Face stimuli

The face stimulus set consisted of forty "base" faces (the word "base" is used to distinguish these faces from "morph" faces, blended from these base faces and used in a later experiment described in Chapter 3). The faces were all emotionally neutral, front-facing, directly-gazing, Caucasian females organized as follows:

- ◇ 10 of the images were of the subjects themselves,
- ◇ 10 were of their sisters (one full biological sister per subject, age range 23 - 29, mean 26.3 ± 0.67 years), and
- ◇ 20 "*extra stimuli*": eighteen were photographed in controlled conditions at Caltech[1], and two were selected from the PUT Face Database [51].

All women in the stimulus set appeared to be in their twenties or early thirties at the time of their photograph. The pictures of the subjects' sisters were obtained after sending them written instructions aimed at matching the conditions of the rest of the photographs (head-leveled camera, 5-6 feet away). Figure 2.5 shows all forty base faces used in the experiment, using a preparation hiding jaw and hairline, that is, including only inner face features (referred to later as "inner only"), Figure 2.3 shows the 10 subjects and 10 sisters stimuli[2] in a preparation including jaw and hairline ("whole face"), and Figure 2.4 shows one stimulus in both preparations side-by-side. Importantly, 83 keypoints (e.g. left eye lateral extremity, or nose tip) were manually annotated on each face.

[1] thanks to Jan Glaescher for contributing some of these
[2] thanks to Prof. McEliece for suggesting the guessing game in this figure

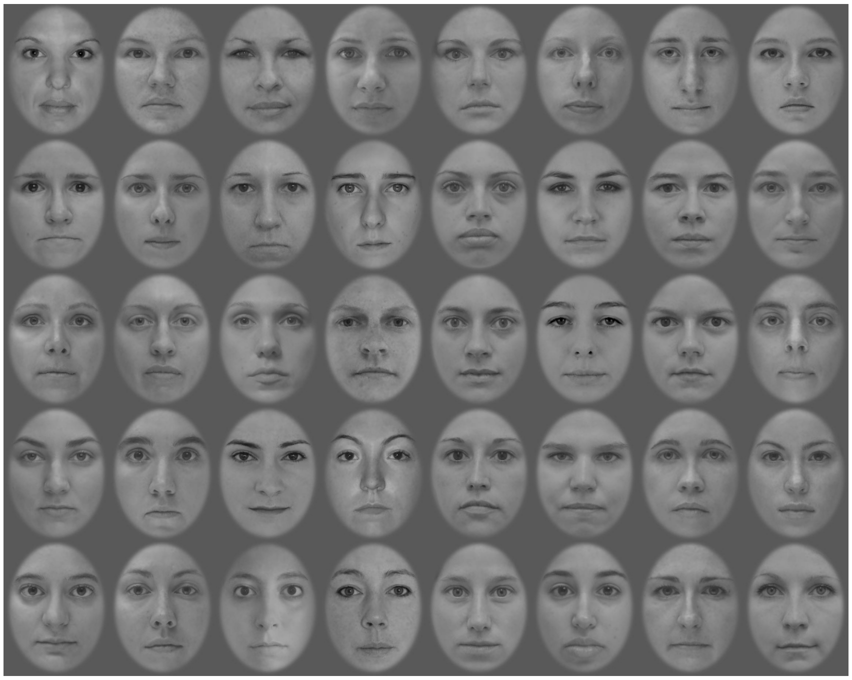


Figure 2.5: The forty base face stimuli, resized for equal proportions, and normalized for mean and standard deviation in luminance. Normalization beyond this (e.g., local contrast) was not carried out in order to allow some variability in low-level features. Subjects only viewed one face at a time at full scale (~355 pixels tall); faces are only combined here into a single image for spatial compactness.

Together with the pixel content of an image, this allowed us to compute objective distances between faces based on image content alone. See section 2.10.2 for additional details on stimulus preparation.

2.2.3 In this thesis, familiarity = personal familiarity

In this chapter and the next, we will refer to faces as familiar to a viewer, or reciprocally, a viewer as familiar with a face. Without exception, familiarity will herein mean, *personal familiarity*: the viewer will have had personal acquaintance with and knowledge of the person whose face they are said to be familiar with. All stimuli in the experiments are ultimately familiar in the less strict sense of having ever seen before in any context. The nature of the experiments was such that viewers saw all the faces many times by the end. The number of experimental trials during which a subject saw a face was not at all factored into level of familiarity at any point in the analysis.

Measuring degree of familiarity Because eight of the ten subjects and eighteen of the twenty of women in the "extra stimuli" set were members of Caltech community, which is relatively small and insular, subjects were on average familiar with a few faces in the stimulus set (2, 3.5, and 5 for 25th, 50th, and 75th percentiles of number of familiar faces among subjects including self and sister). The degree or level of personal familiarity each subject had with each person represented in the stimulus set was acquired through an interview with the author. Personal familiarity was assigned a

◇ 10 for self,

◇ 9 for sister,

◇ 8 for friend seen very frequently (every day),

◇ 7 for friend seen slightly less frequently, and so on, down to

◇ 1 for possibly seen around campus, and

◇ 0 for do not recall having ever seen before.

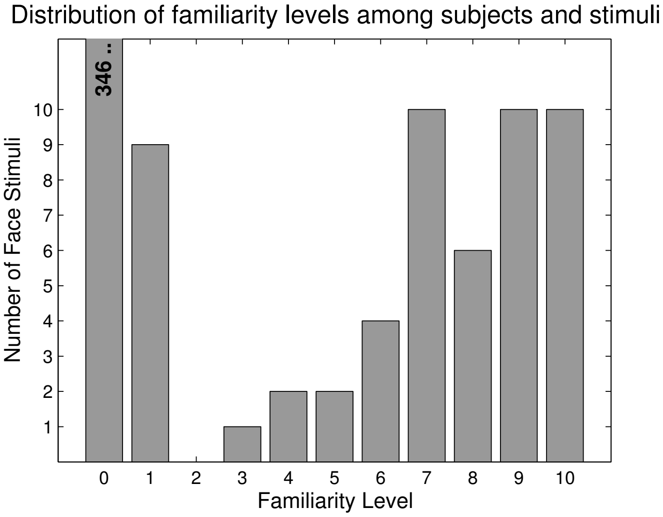


Figure 2.6: Number of face stimuli, summed across subjects, at each level of familiarity. There are 54 total having level ≥ 1, 45 having level > 1, and 36 total having level ≥ 7.

In this chapter, unless otherwise specified, "familiar" as a category will mean familiarity > 1, and unfamiliar will mean familiarity ≤ 1 (i.e., the "possibly seen around campus" condition was considered too weak to count as familiar).

2.2.4 Experimental Tasks

Two experimental tasks were employed, each measuring the visual similarity between face pairs as perceived by the subjects; we herein call these tasks "**SimRate**", for *Similarity Rating*, and "**MorphDiscrim**", for *Morph Discrimination*.

In **SimRate**, the similarity between faces was **explicitly** provided by subjects using a numerical score from 1 (least) to 8 (most similar). In **MorphDiscrim**, the similarity between a pair of faces was **implicitly** provided in the form of a confusability rate, measuring how frequently subjects confused two distinct morphs (between a pair of faces) for two presentations of an identical face: the task was to tell whether a pair of morph stimuli were the same or different, when in half the trials they were actually the same. The rationale for this was that the more different the two base faces (from which the morphs were generated) appeared to look to the viewer, the easier a same/different discrimination between intermediate morphs would be to the subject, leading to a lower confusability rate.

In both experiments SimRate and MorphDiscrim, face pairs were never viewed simultaneously; instead, a base face or face morph centered on the screen was flashed for a brief interval (200 ms or less), followed by another face stimulus after a brief intervening visual "mask" to clear the space in between, followed by an interval during which the subject keyed in a response. See Section 2.10.4 for exact details about the trial structure used in these experiments.

Similarity scores across face pairs were Z-scored for each subject independently (that is, normalized to have sample mean 0 and sample standard deviation 1; see Section 2.10.1.1), and confusability scores were inferred from error rates in morph discrimination (see Eq. 2.2 for details). Among all subjects and face pairs, the similarity scores ranged from a minimum of -2.10 to a maximum 2.51. The minimum confusability score among all subjects and all face pairs was 0, and the maximum was 1.33. Distinguishability was defined as negative confusability (i.e., ranging from a minimum of -1.33 to a maximum 0).

All 780 unique pairs of the 40 faces received explicit similarity scores by each subject in the SimRate task, and a subset consisting of 118 face pairs received implicit similarity (confusability) scores in the MorphDiscrim task. Importantly, two variants of the SimRate task were employed: in one, subjects saw whole faces, including jaw and hairline, and in the other, these outer face features were masked out. See Figure 2.4 for an example of both kinds of faces, and refer to Section 2.10.4 for more details on this. In MorphDiscrim, only the inner face stimuli versions were employed (to increase difficulty and range of performance results). Unless otherwise stated, the SimRate results presented will be based on whole faces.

2.2.5 RSMs and RDMs

The result of the SimRate and MorphDiscrim tasks can be completely summarized in what we will refer to as representational similarity matrix, RSM, or representational dissimilarity matrix, RDM (one for each subject, for each task). In the case of MorphDiscrim, the values are confusability for an RSM and distinguishability for an RDM. We will use these two somewhat interchangeably in some contexts as the only difference between them is that one is the negative of the other. The RSM is a symmetric matrix M whose entry M_{ij} is the similarity (or confusability) measure between stimulus i and stimulus j.

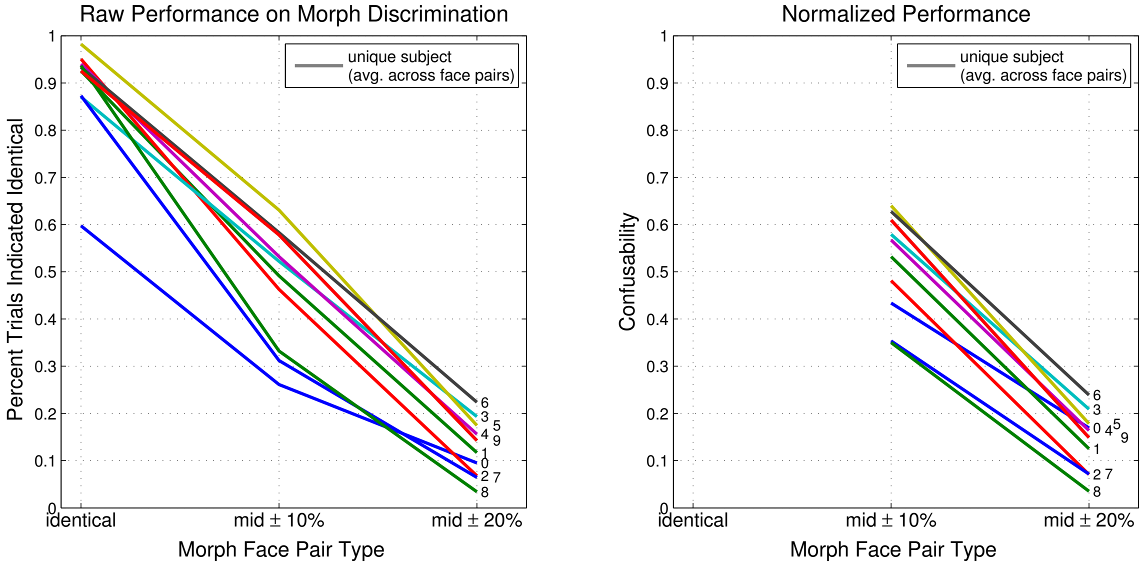


Figure 2.7: Each subject's raw performance on MorphDiscrim is shown as an single line (subject number indicated at terminus), computed as the average over face pairs. *Left*: Each individual subject's performance in the morph discrimination task is shown for the three levels used: (1) identical (faces actually the same), (2) mid ± 10%, wherein each face was 10% removed from the midpoint between the pair (in opposite directions), and (3) mid ± 20%, wherein each face in the pair was 20% removed from the midpoint. Note that subject S0 appears to be an outlier in "identical condition"; she is also the outlier in Cambridge Memory – see Figure 2.32 for another comparison of Cambridge Memory score and experimental task performance. *Right:* A confusability metric is derived by dividing by fraction of trials indicated identical under the actually identical condition.

2.3 Raw performance results

In this section, we briefly discuss some of the raw performance results from SimRate and MorphDiscrim, and show how they compare. More basic results are provided in the supplementary results to this chapter: Section 2.9.

Figure 2.7 confirms that subjects can more easily distinguish between faces on a morphing continuum which are further apart. When the face morph stimuli are very similar (only 20% separated, ± 10% from the midpoint between a base pair), subjects perform at roughly chance level on average, indicating in about half the trials in which the faces are presented that they believe them to be identical. In the right panel of the same figure, we introduce a ***confusability*** metric, which we define as the *percentage of trials indicated identical when the faces are different, divided by the percentage indicated identical when they actually are* (see Eq. 2.2). This is computed separately for each face. We note here that when we refer to distinguishability or confusability, we will be referring to performance in the MorphDiscrim task, and when we refer to dissimilarity or similarity, we will be referring to performance in the SimRate task.

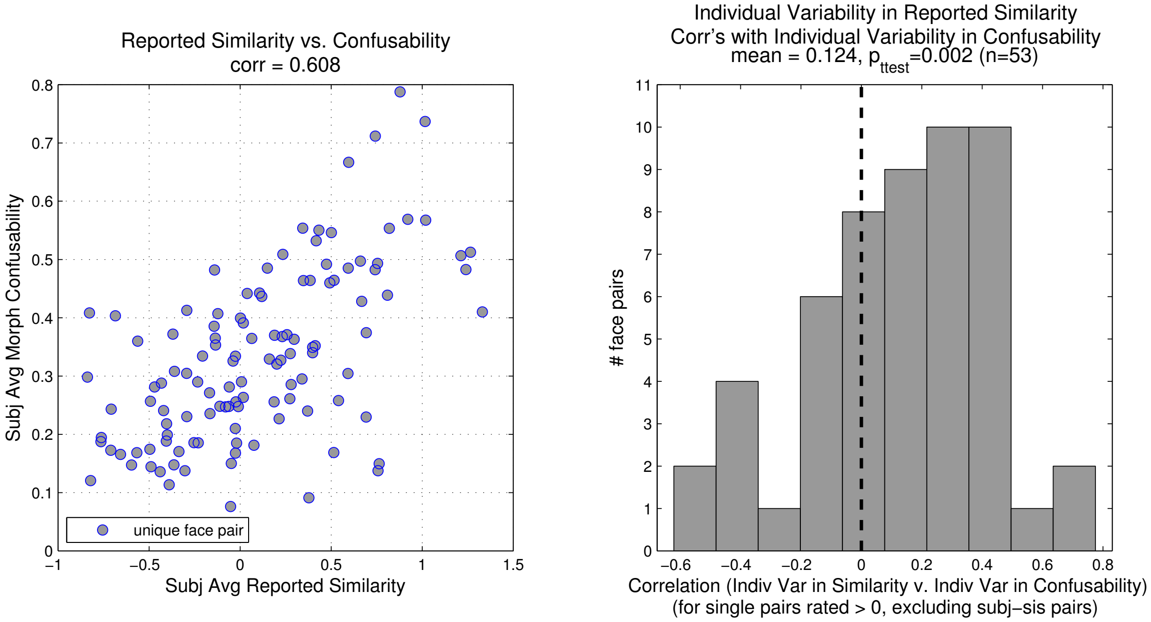


Figure 2.8: Subjects are consistent across the two types of face pair similarity tasks. *Left:* relationship between subject-averaged similarity and confusability. *Right:* each point in the left panel corresponds to an average of 10 x- and y-coordinates, one per subject; each may be assigned a single-face-pair correlation between (i) subject variability in SimRate and (ii) subject variability in MorphDiscrim; the population of such correlations is shown for face pairs rated more similar than average (similarity in SimRate > 0) − it is significantly to the right of 0 by 1-sided t-test.

The left panel of Figure 2.8 shows that the subject-averaged similarity for a face pair in the SimRate task was correlated with the subject-averaged morph confusability, across the 118 face pairs for which both were measured, at a level of .608, with significance $p < 1.9 \times 10^{-11}$, computed by an empirical significance test in which each trial used to form the null distribution corresponded to a random shuffling of the rows of the SimRate-based RSM. **Empirical significance tests** are used extensively in this chapter and in this thesis; their p-values will often be denoted $\mathbf{p_{shuff}}$. The general method is described in Section 2.10.1.3. Figure 2.9 shows the correlation between subject-averaged dissimilarity and subject-averaged confusability across face pairs, and compares this with distances based on low-level stimulus features.

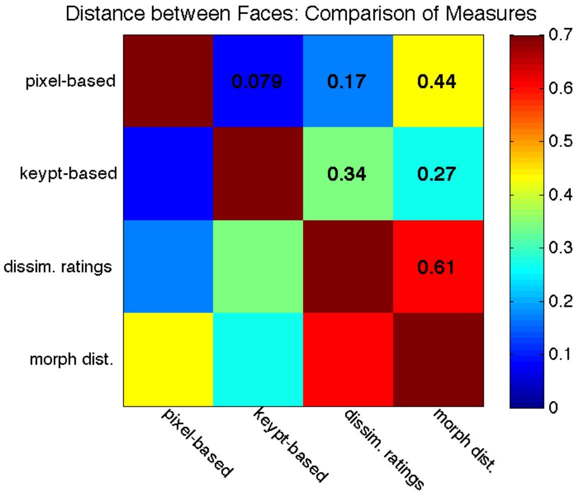


Figure 2.9: The relationship between different measures of inter-stimulus distance (first subject-averaged). We show the correlation between each such measure across face stimuli pairs. Distinguishability and dissimilarity are most correlated (0.608), whereas pixel-based distance and keypoint-based distance are least (0.079). See Section 2.10.6 for methodological details and Section 2.9.4 for a comparison of distinctiveness measures.

We were also interested in whether, irrespective of the subject-averaged similarity rating of a pair, one subject's rating a particular face pair as more similar than another subject in SimRate would correlate with that subject's confusing the faces less than the other subject in MorphDiscrim. This is what the right panel of Figure 2.8 addresses. Conceptually, one can think of each point in the left panel as hiding a constellation of 10 constituent points, one per subject, each with its own x-coordinate (similarity value) and y-coordinate (confusability value). The right panel assesses the correlation between this variability in x- and y-coordinates, ignoring their average for each face pair, which is what the left panel correlates. We find that among faces which are rated as less similar to each other than average, that is with SimRate score < 0, there is no relationship between individual variability similarity rating and distinguishability: performance in the distinguishability task hits a ceiling in that regime. However, for faces which were more similar (SimRate score $>$ 0), distinguishability performance significantly correlated with this the explicit similarity score. We left out of this analysis (right panel of Figure 2.8) the 10 pairs of subject-and-sister faces, due to an "insider information" effect we shall discuss later (in Section 2.4.2.1). If subject-sister pairs are included in the right panel of Figure 2.8, the mean inter-individual correlation drops from 0.124 to .096, and the 1-sided t-test drops from $p < 0.002$ to $p < 0.01$.

2.4 Familiar faces are more distinct

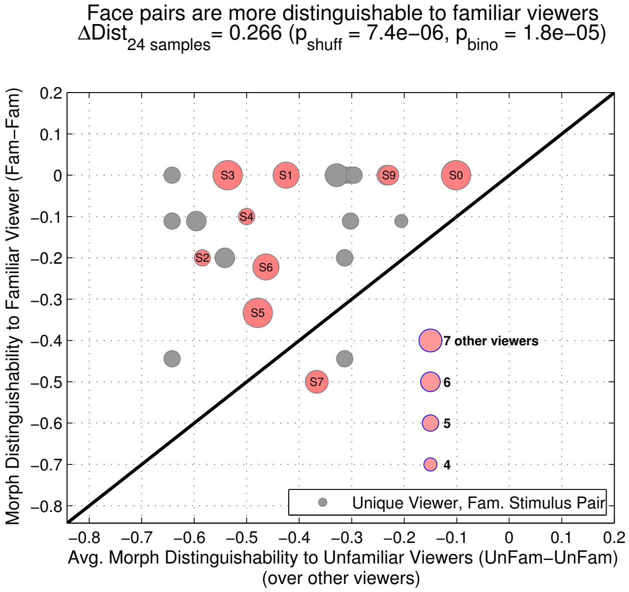


Figure 2.10: A pair of faces, both familiar to a viewer, is more easily distinguished in MorphDiscrim than it is by viewers unfamiliar with either face in the pair (i.e., data falls above y=x line on average). The sizes of the circles correspond to the number of unfamiliar viewers averaged together for the x-coordinate (this number varies due to the familiarity structure between subjects and stimuli). The red colored circles correspond to the (self,sister) pairs of familiar faces, and are labeled with the corresponding subject number. The mean displacement above the diagonal across data point is 0.266 and is reported in the title.

In this section, we show that pairs of faces are more easily confused (inside a morphing continuum) by unfamiliar viewers than by familiar viewers. We also show that viewers explicitly rate faces as appearing more different from each other when they are familiar with them, compared to ratings by viewers unfamiliar with the faces. Both of these trends are true even when only one of the faces is familiar to the viewer. Lastly, based on our limited data set, we show that pairs of *entirely unfamiliar faces* are perceived to be more dissimilar and distinguishable when they are merely close to a familiar face. These effects all point to the enhanced apparent distinctiveness of familiar faces.

2.4.1 Effect on Morph Distinguishability

Figure 2.10 is the first in a series illustrating the enhanced distinctiveness of familiar faces. We see that pairs of faces, when they are both familiar to a viewer, are distinguished much more easily than when neither one of the faces if familiar to a viewer: the data lie above the y=x diagonal. In the same figure, we see that for all but subject S7, the pair of (self face, sister face) can be more easily distinguished by the subject viewing herself than by subjects unfamiliar with either.

Understanding the statistical significance values: The magnitude of the overall effect is highly significant, with an empirical significance of $p_{shuff} < 7.4 \times 10^{-6}$ (again, see Section 2.10.1.3); the null distribution is generated by randomly shuffling the occupied entries of the RDM while holding familiarity relationships constant. We tally the number of shuffled trials where at least as great an effect magnitude (in this case $\Delta Dist_{24\, samples}= 0.266$) is observed, establishing an empirical significance measure. For such plots, we also report a "binomial" significance value, p_{bino}, which is the probability of observing at least as many data points with y-coordinate>x-coordinate (in this case, distinguishability-to-familiar-viewer > distinguishability-to-unfamiliar-viewer) as we actually do, assuming the probability of each such event were 50%.

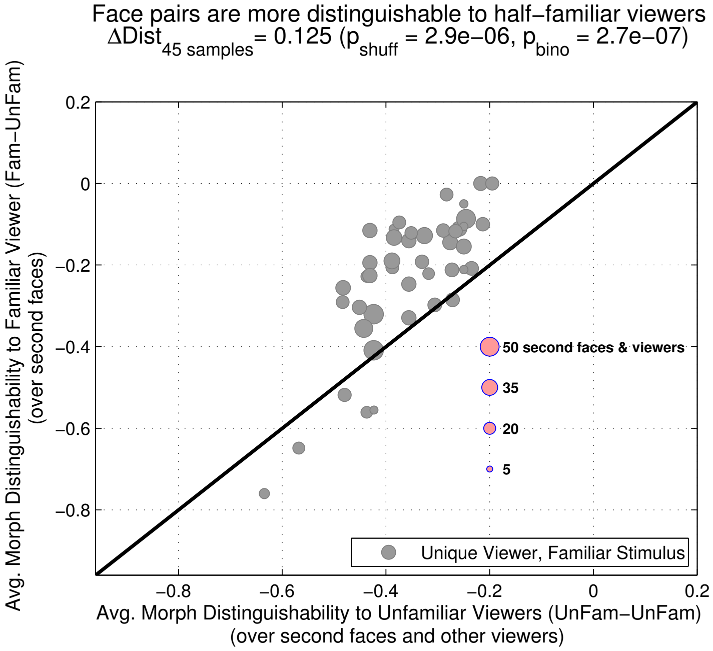


Figure 2.11: A pair of faces, with only one face familiar to a viewer, is more easily distinguished in MorphDiscrim than it is by viewers unfamiliar with either face in the pair. The sizes of the circles correspond to the number of unique combinations of unfamiliar viewers and unfamiliar faces averaged together for the x-coordinate (each circle/point corresponds to a unique (viewer, familiar face) pair) .

We next investigate whether pairs of faces can be more easily distinguished even when only one face in the pair is familiar to the viewer; this is what we will refer to as a half-familiar viewer (or viewer only half-familiar with the pair). Figure 2.11 shows that such face pairs are indeed more easily distinguished, though the boost in distinguishability falls from 0.266 in the case of fully familiar pairs to 0.125 in the case of half familiar pairs. However, the effect is highly significant under both empirical ($p_{shuff} < 2.9 \times 10^{-6}$) and binomial tests ($p_{bino} < 2.7 \times 10^{-7}$).

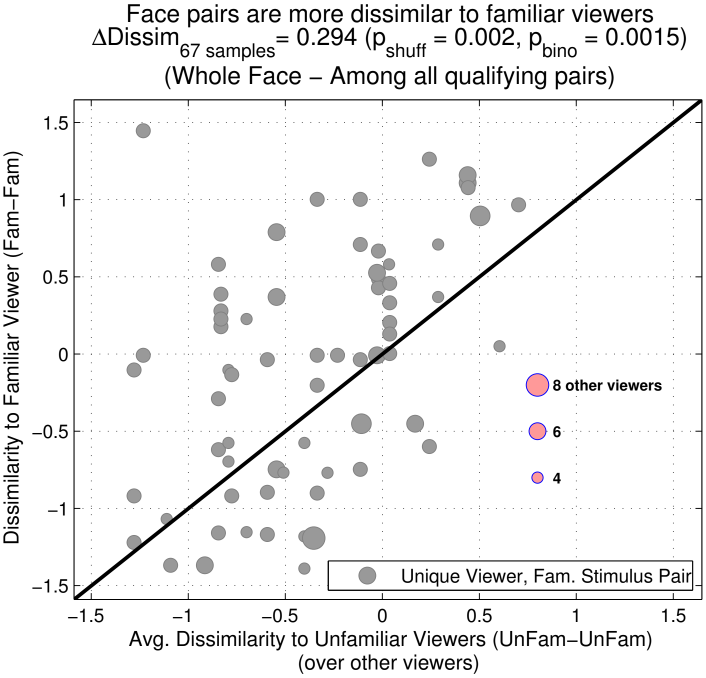


Figure 2.12: A pair of faces, both familiar to a viewer, is rated more dissimilar than by viewers unfamiliar with either face in the pair. The sizes of the circles correspond to the number of unfamiliar viewers averaged together for the x-coordinate. (self,sister) pairs are excluded from this analysis due to “insider information” bias.

2.4.2 Effect on Explicit Similarity Judgment

Having established that a face can be more easily distinguished from another by a familiar viewer, we next test whether a face is also explicitly rated as more dissimilar by a familiar viewer. In Figure 2.12, it is shown that pairs of faces are indeed rated as more dissimilar when they are both familiar to a viewer than when neither of them is. The average boost in dissimilarity, relative to the dissimilarity rated by unfamiliar viewers, is 0.294 (in Z-score range). That is, pairs of familiar faces just look more different from each other than they would otherwise. We note that in this section, we discuss SimRate results for ratings of whole faces, i.e. including outer features.

2.4.2.1 Insider information: influenced by the expectation of familial similarity

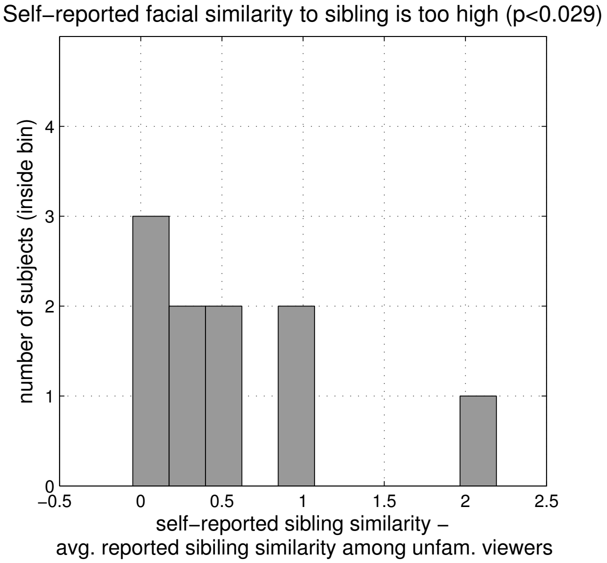


Figure 2.13: "Insider information" The similarity between a subject's own face and her sister's is rated higher (to the right of 0 in the plot) than it is by entirely unfamiliar viewers likely due to knowledge of the familial relationship not known to the unfamiliar viewers. This effect is significant by two-sided t-test ($p < 0.029$).

In Figure 2.12, pairs of subject-and-sister faces are specifically excluded. Included, the average dissimilarity boost drops from 0.294, to 0.189, although the effect remains on the whole statistically significant. In fact, whereas pairs of faces both familiar are usually rated as more dissimilar than they would be by unfamiliar viewers, this trend goes in exactly the *opposite direction* when the face pair is the subject's own face and her sister's. This is what we will refer to as "insider information": the subjects themselves know that they are related to their sisters, and this information biases them to report the faces as more similar looking, having knowledge that sisters *typically do* look alike – they essentially know what the reporting *should* be. Despite this supposed perception of sibling likeness, subjects exhibit a clear distinguishability advantage in separating their own faces from those of their sisters (Figure 2.10); in that sense, the faces explicitly reported to look similar are implicitly very different. Entirely unfamiliar viewers do not have knowledge of familiarity, and make the dissimilarity judgment based on visual information alone. No other pairs of faces in our stimulus set are affected by this bias (even if a subject is familiar with both faces in a pair, unless they are sisters, they have no reason to believe a priori that they should look alike or not, on average).

2.4.2.2 Half-familiar face pairs are rated as more dissimilar

Next, we see in Figure 2.14 that, just as with the distinguishability of faces, their explicitly rated dissimilarity to others is greater even when only one of the two faces being compared is familiar to the viewer; that is, the half-familiar viewer perceives a pair of faces to be more different looking than an entirely unfamiliar viewer. The average dissimilarity boost is smaller, at only 0.146, than was obtained for pairs of faces both familiar (0.294), but is highly significant: the probability of observing so many (viewer, familiar stimulus) combinations with trend going in the same bias direction is $p_{bino} < 0.00019$. We note that for this analysis, face pairs with dissimilarity $>$ -0.25 (as rated by others) are excluded. As we shall discuss later in Section 2.5.2, this is because fair pairs which are already very dissimilar looking do not become *even more* dissimilar looking once when of them is familiar. However, this qualification is not necessary for the boost to be significant when considering ratings between inner faces only. See the Supplementary Figure 2.37 for this.

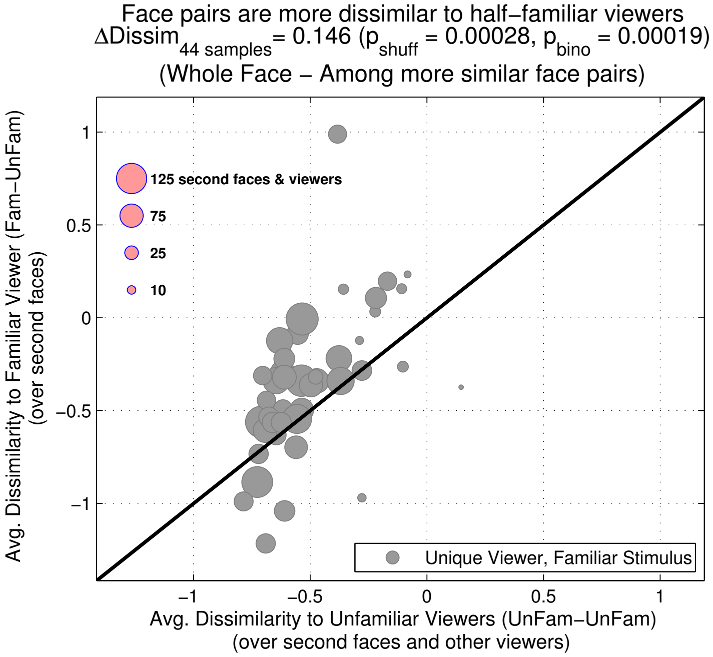


Figure 2.14: A pair of faces, with only one face familiar to a viewer, is rated as more dissimilar (than by viewers unfamiliar with either face in the pair). The sizes of the circles correspond to the number of unique combinations of unfamiliar viewers and unfamiliar faces averaged together for the x-coordinate (each circle/point corresponds to a unique (viewer, familiar face) pair). Face pairs rated greater than -0.25 in dissimilarity are excluded from analysis because they are less affected by familiarity.

2.4.3 Effect on neighborhood of familiar faces

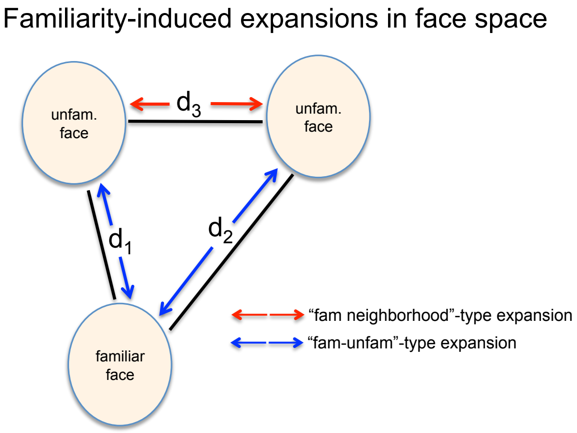

Familiarity-induced expansions in face space
unfam. face
unfam. face
familiar face
d_3
d_1
d_2
"fam neighborhood"-type expansion
"fam-unfam"-type expansion

Figure 2.15: Illustration of changes in face space due to familiarity with a face stimulus. The unfamiliar faces are assumed to be in the neighborhood of faces which look similar to the familiar one, which we refer to as the "familiar anchor" of this region of face space, below.

We were interested in whether the expansion in face space around a familiar face (the perceptual distancing, in dissimilarity judgment and distinguishability) extends measurably beyond comparisons directly involving a familiar face. Figure 2.15 illustrates a second kind of change in face space we might expect: that between two unfamiliar faces merely close to a familiar face (termed a "**fam neighborhood**" expansion and denoted with red arrows). The problem with assessing this kind of effect is establishing a perfect control subject, or contrasting case, because, based on the experimental limitations, we could not know whether two faces were actually in the neighborhood of a face familiar to a viewer, but one which was not in our stimulus set. Nonetheless, we could at least guarantee that the control subject not have a familiar face, from our stimulus set, close to the candidate pair of faces. We refer to this as the **imperfect control condition**. Below, we will refer to a viewer of a face pair, who is unfamiliar with both faces in the pair, but familiar with a face close to the pair (by some metric), as a **quasi-familiar viewer.**

We begin by examining the fam neighborhood effect on similarity judgments. Despite the imperfect control, Figure 2.16 confirms that, indeed, unfamiliar faces in the neighborhood of a familiar one are perceived to look more dissimilar to one another relative to the judgments of non-quasi-familiar viewers. In this case, non-quasi-familiar-viewers were viewers (i) unfamiliar with both faces in the pair being compared (as was the quasi-familiar viewer), who (ii) did not have a familiar (level ≥ 7) face stimulus in the neighborhood of either one, where the neighborhood was any face rated within

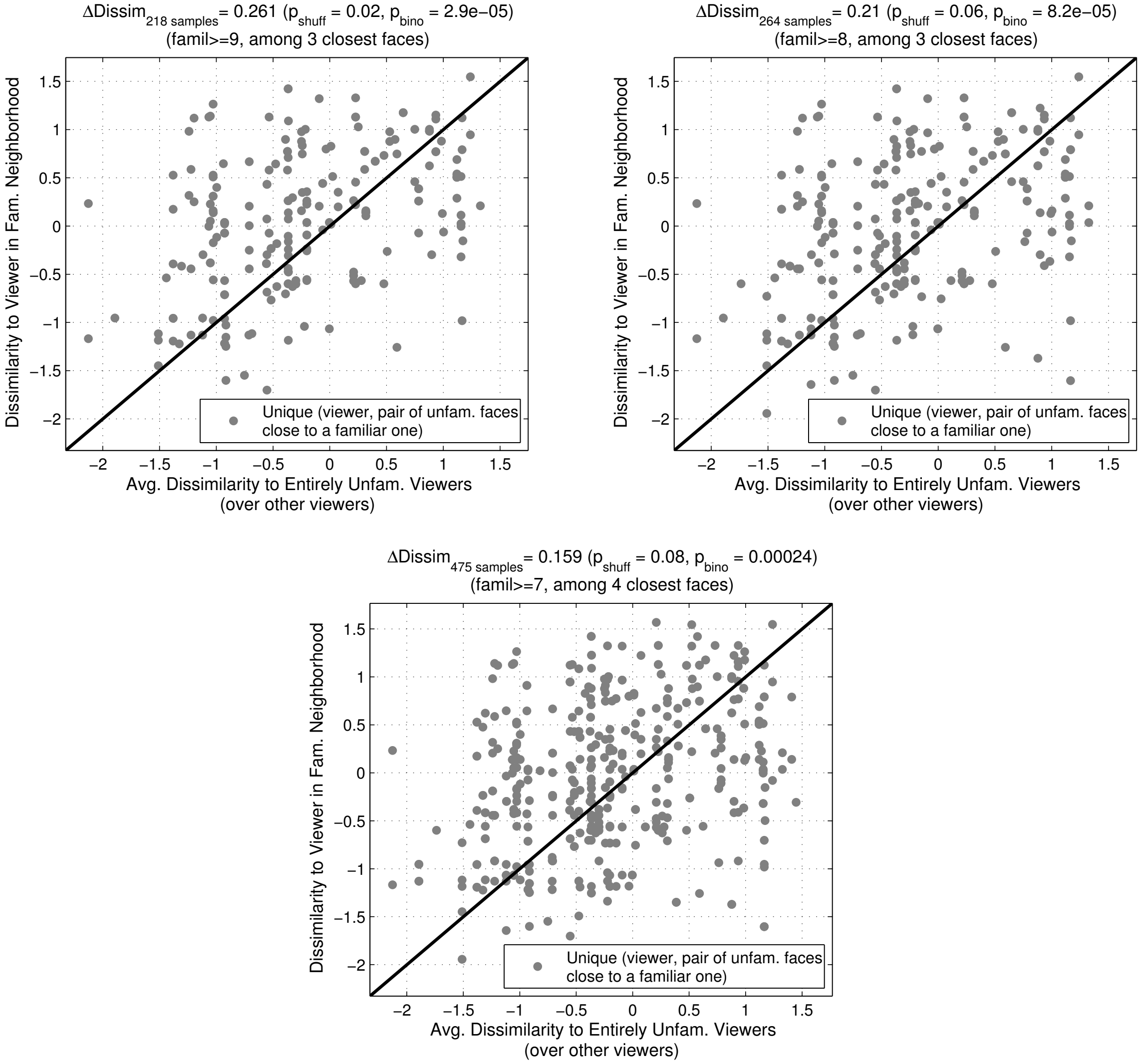


Figure 2.16: The dissimilarity between a pair of faces is greater when rated by a quasi-familiar viewer than by a properly selected unfamiliar control subject. The boost in dissimilarity shrinks for increasingly loose definitions of a familiarity neighborhood, within which a viewer is said to be quasi-familiar. The top left panel is most strict (effect size=0.261), top right less strict (effect size=0.210), bottom least strict (effect size=0.159).

the top 60% of similarity, averaged across the 8 remaining subjects (not non- or quasi-familiar being compared). The effect is most robust in the neighborhood of self face or sister's face (effect size: 0.261), corresponding to familiarity ≥ 9, presumably because people have an especially great level of expertise in this familial territory of face space, perhaps having other similar looking relatives whose faces they learned. The effect size smoothly falls off to 0.210 when we include close friends and to 0.159 when we loosen the quasi-familiarity level further.

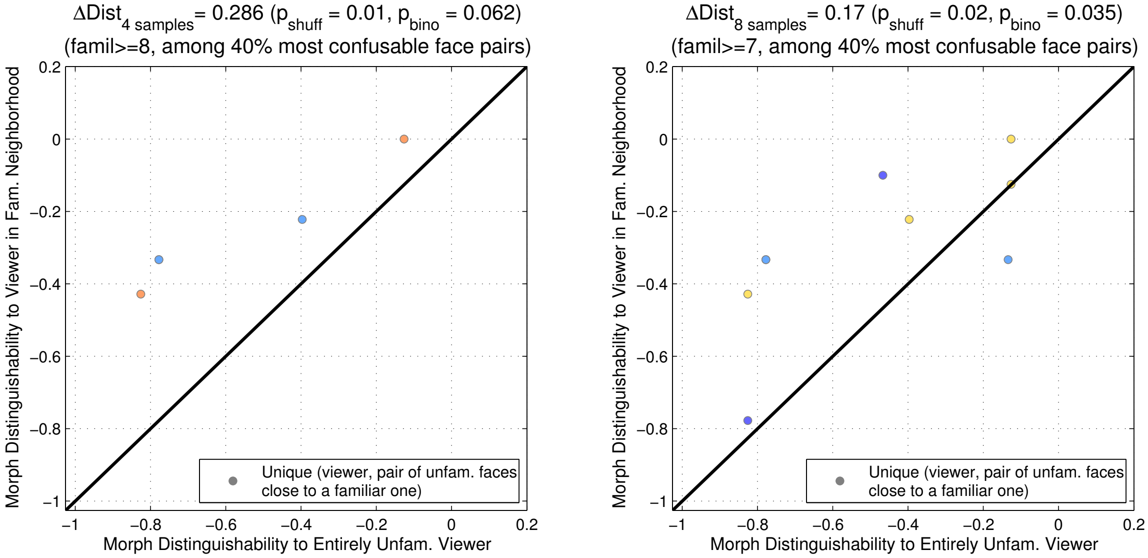


Figure 2.17: Distinguishability is relatively enhanced in the neighborhood of a familiar face. *Left*: Faces pairs are selected for analysis if the are each close to a familiarity-level 8 or greater face, where close means among 40% most confusable faces. *Right:* Faces pairs are selected for analysis if the are each close to a familiarity-level 7 or greater face. Each data point corresponds to a unique selection of (contrasting viewer, pair of faces being compared), but multiple data points can be produced by the same quasi-familiar viewer's distinguishability; points are colored by unique quasi-familiar viewer.

Next we examine effects on distinguishability. Because subjects only performed the MorphDiscrim task on a subset consisting of 118 face pairs, the number of qualifying comparisons available was rather small. Nonetheless, in Figure 2.17, it is shown that when two faces are both among the most confusable with a familiar face, the quasi-familiar viewer has an advantage in distinguishing between them. We show that as we move from a more strict definition of familiarity neighborhood (familiar anchor at familiarity level of at least 8), to a less strict one (anchor at familiarity level of at least 7), this boost in distinguishability falls from 0.286 to 0.170. These effects are both directionally as we expect (having positive sign), although only weakly significant by the previously employed RDM-shuffling empirical significance test. The limited number of face pairs for which we have distinguishability levels and the imperfect control condition contribute to weakening the effect size and significance. However, we found that by holding the RDM constant and instead shuffling the mapping between subject and stimulus familiarity, the empirical significance values improved: those are the ones reported in Figure 2.17 (and Figure 2.16). This is likely because, compared to shuffling the RDM, shuffling the familiarity mapping more completely disrupts the structure of distinctiveness biases. For each pair of faces in the fam neighborhood, a non-quasi-familiar viewer (represented by the x-coordinate in Figure 2.17) was qualified exactly as with SimRate above, except the face pair was required to not be in the top 25%, instead of 60%, of similarity in comparison with any face familiar to her (as rated by the other 8 subjects). The requirement was loosened in order to admit enough data points for analysis.

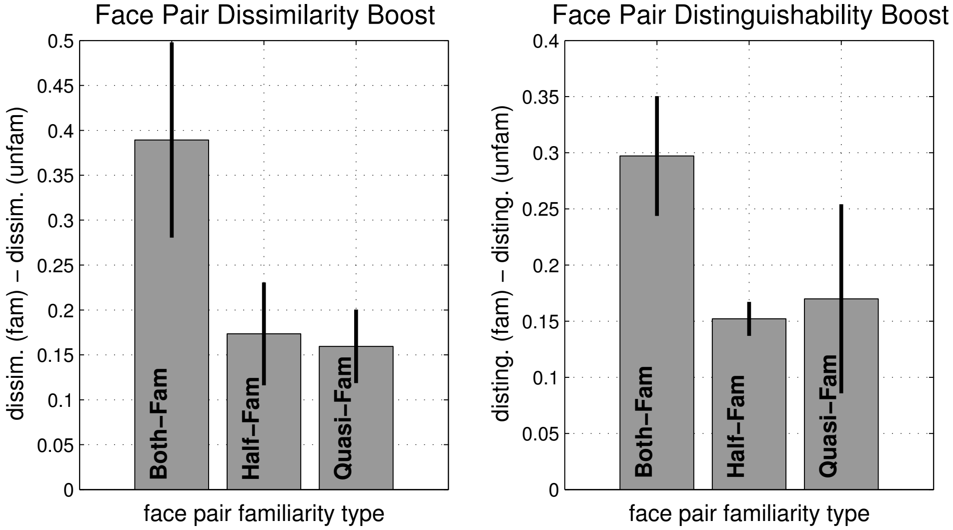


Figure 2.18: A summary of dissimilarity and distinguishability boosts for different categories of familiar (or quasi-familiar) face pairs. Standard error bars are computed over unique subject-stimulus combinations (e.g., “unique viewer, familiar stimulus”), as described in the preceding section. **Important note:** To make the comparisons across all three categories of face pairs meaningful, familiarity was defined in each case (including quasi-familiar) to mean level ≥ 7. Note that is different from the scatter plots shown in the preceding section, which also include faces with lower familiarity levels, for completeness.

2.5 Factors affecting the enhanced distinctiveness of familiar faces

2.5.1 Degree of familiarity and distinctiveness boost

To test whether the degree or level of familiarity with a face has an influence on its relatively enhanced perceived distinctiveness, we grouped faces by type of familiarity relationship and plotted

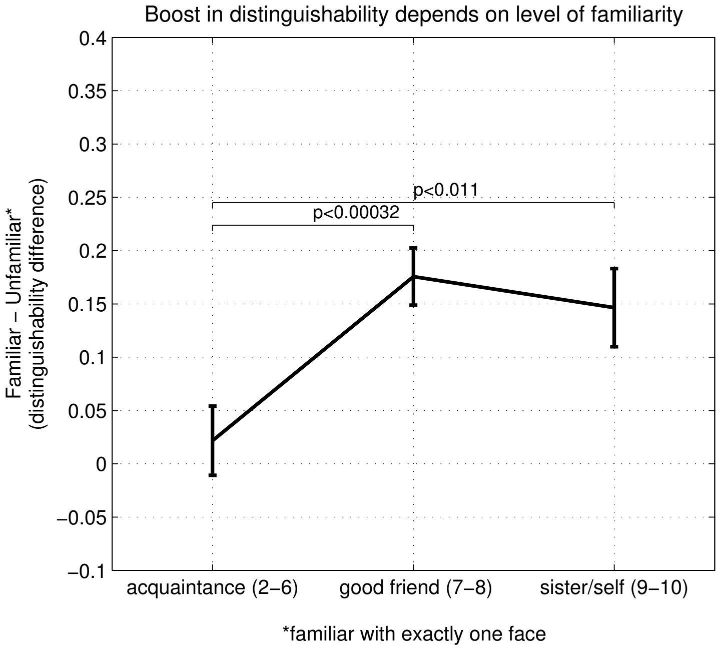


Figure 2.19: The degree of familiarity enhances distinguishability of faces (analysis is based on pairs of half-familiar pairs). p-values are based on 2-sided t-tests.

the boost in distinguishability when comparing with an unfamiliar face. We find that, on average, faces of oneself or one's sister, or that of a good friend (familiarity 7 - 8), receive a larger boost in distinctiveness relative to less familiar faces. The result is shown in Figure 2.19, which is essentially the average displacement above the diagonal in Figure 2.11 by familiarity type, except leaving unique second viewers as unique data points for more statistical power. We note that we left out face pairs which had confusability score of less than 0.2 (33rd percentile), rated by the other 8 viewers, because, as we shall discuss shortly, face pairs which are already easily distinguished do not gain much in distinguishability from familiarity with one of them.

We find the same effect of greater familiarity level on face pair dissimilarity judgments. This is shown in Figure 2.20. As with distinguishability, we find that the more a face is familiar, the more its dissimilarity to other unfamiliar faces is enhanced. This boost is of greater magnitude for whole faces, due likely to more complete recognition of the face, but the secondary trend of increasing with familiarity level is weaker. One reason for this might be that, with outer features revealed, even weakly familiar faces are easily recognized and thus perceived differently. The analysis is done exactly analogously to the distinguishability analysis, including leaving out face pairs which are already too dissimilar, as they are not affected by the familiarity of one of the faces (we leave out faces below the 60th percentile of similarity judgment by the other 8 viewers).

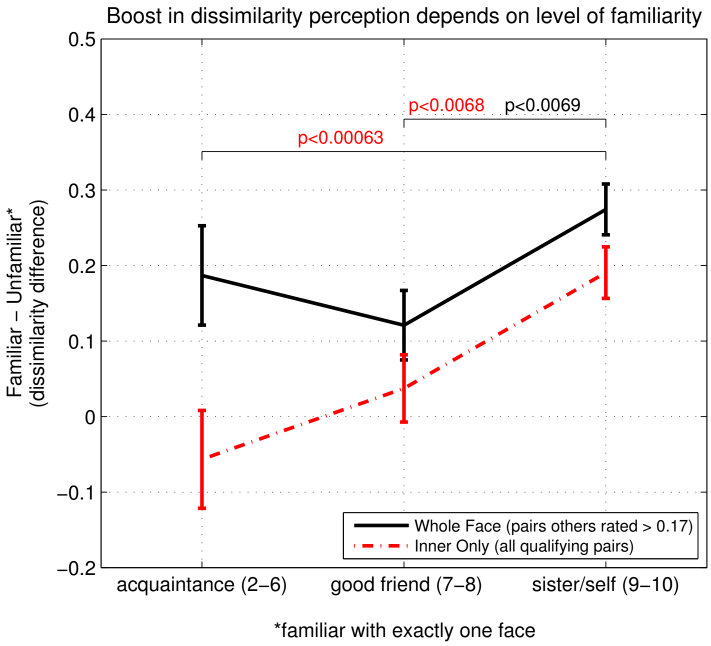


Figure 2.20: The degree of familiarity enhances dissimilarity judgments (analysis is based on pairs of half-familiar faces). The dissimilarity boost is generally larger for judgments between whole faces (including outer features) than for pairs showing only inner features, possibly because faces are recognized more readily and completely with jaw and hairline cues. p-values are based on 2-sided t-tests.

2.5.2 Baseline similarity/confusability and distinctiveness boost

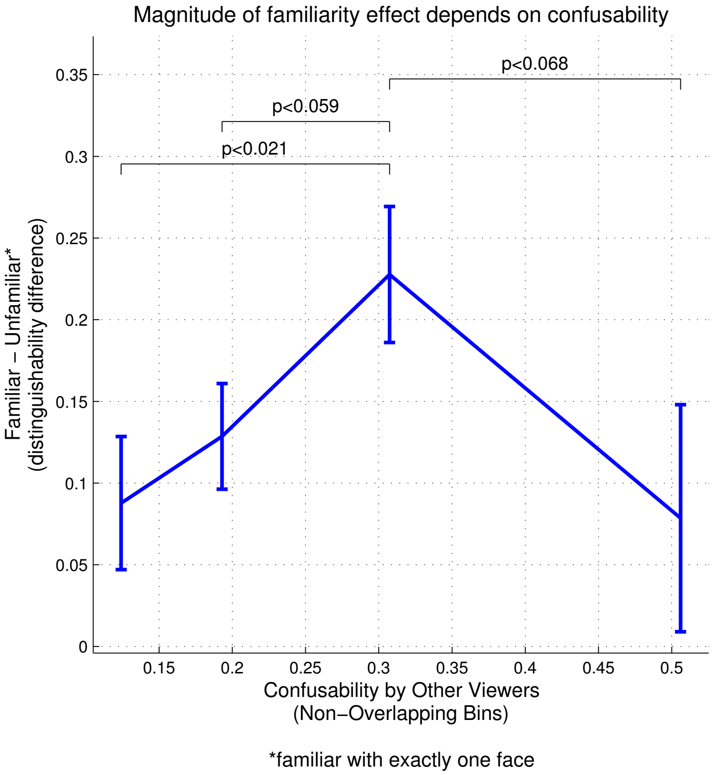


Figure 2.21: The extent to which a face pair becomes more distinguishable to a familiar viewer depends on the underlying confusability of the pair.

As mentioned above, face pairs which are already very dissimilar or distinguishable, even without the benefit of familiarity, do not gain much perceived dissimilarity or distinguishability when one of the faces in the pair is familiar. This is reflected in the left part of Figure 2.21 and Figure 2.22, showing a relatively smaller distinctiveness boost for the *least* similar face pairs. For these analyses, the baseline similarity and confusability were established by taking the average among the 8 remaining subjects not used to compute the difference (between the half-familiar's and fully unfamiliar viewer's similarity ratings or confusability level). However, whereas the most *confusable* face pairs show a relatively *smaller* distinctiveness boost (compared to ones in the middle of the baseline confusability range), the most *similar* face pairs are the ones which show the *greatest* average distinctiveness boost. We thought perhaps the difference in some of the experimental parameters (outer vs. inner faces, different numbers of face pairs across experiments) might be responsible for this discrepancy, but the same general trends hold, even when these differences are eliminated. One explanation is that the MorphDiscrim task was simply too difficult for subjects near the high end of baseline face pair confusability, and so performance there floored, whereas face pairs near the middle of the confusability levels just admitted a broader performance range within which the familiarity effect could be measured: this would suggest that perhaps the most similar looking face pairs *actually are* most affected by familiarity, both in terms of dissimilarity perception *and* distinguishability, but that the nature of our experimental setup prevented the accurate measuring of the latter. In either case, it is clear that the least confusable or similar face pairs are perceptually affected least from familiarity.

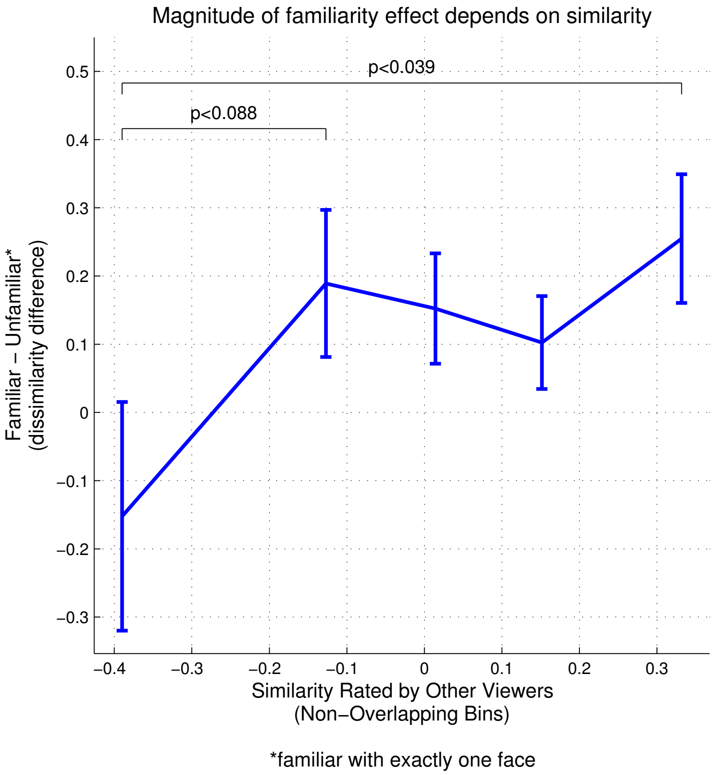


Figure 2.22: The extent to which a face pair becomes more dissimilar to a familiar viewer depends on the underlying similarity of the pair.

2.6 Predicting familiarity from distinguishability

The results presented above suggest an interesting possibility: that from the performance of a subject in discriminating between morphs on a continuum between a pair of faces, one could predict whether the subject was familiar with one of the two base faces. We used a simple linear model which modeled the binary familiarity with one of the faces in the pair (-1 for not, 1 for familiar above some threshold level) as a combination of the confusability of the pair, as determined from unfamiliar subjects, and

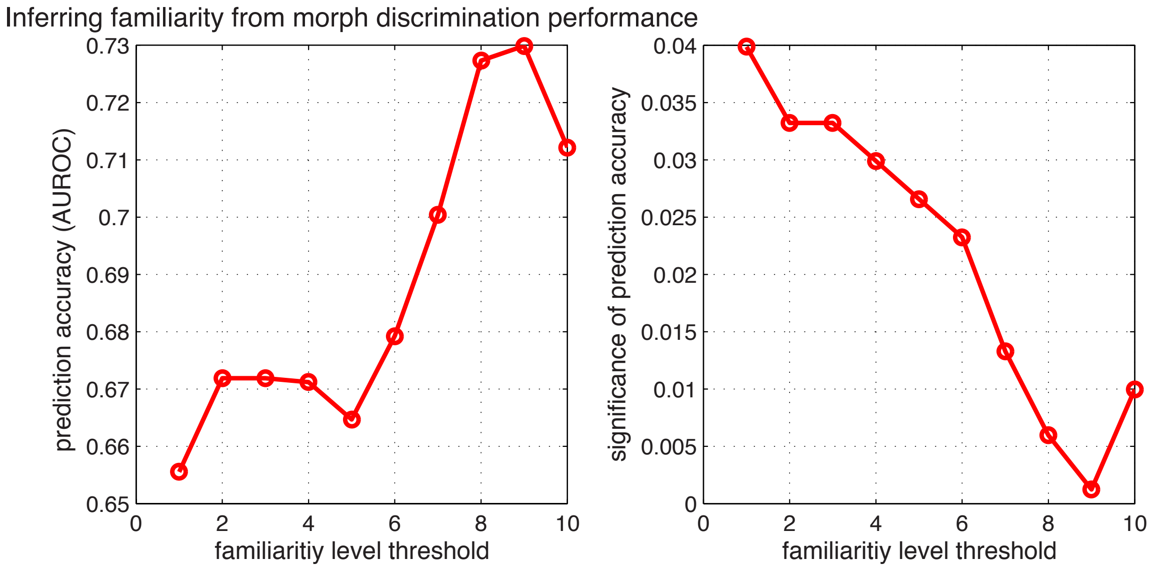


Figure 2.23: Based on the distinguishability of a pair of faces, it is possible to accurately predict, using a linear model, whether they are familiar at or above some threshold level. The familiarity of a pair is taken to be the maximum of the familiarity with either one of the faces in the pair. *Left:* Average (across subjects) area-under ROC-curve showing the prediction performance level for each familiarity threshold, *Right:* statistical significance of each performance level determined using random shufflings of familiarity relationships.

the confusability by the test subject with possible familiarity:

$$M = [C_{others,mean}\, C_{others,min}\, C_{self}\, \mathbf{1}], \text{ model: } Mw = F$$

where $C_{others,mean}$ is a column vector, one entry per face pair, containing the sample mean confusability among subjects unfamiliar with the pair, $C_{others,min}$ is the minimum confusability among these subjects, and C_{self} is the confusability as determined by the test subject. F is the column vector of binarized familiarity. For each subject, we form a training data set by using data from all the remaining subjects, left out one at a time (leaving 8 possible others), as test subjects to form C_{self}, and then find the optimal w which minimizes $||F - Mw||$ via standard linear regression. We then apply this learned w vector to the left-out subject's ratings and assess the predictiveness of Mw as an estimate of F. The result is shown in Figure 2.23. The prediction performance associated with the real-valued estimate Mw of the binarized familiarity F was determined by an area-under the ROC-curve (AUROC) analysis in which, for each thresholding of Mw, a true and false positive rate was established, yielding an ROC curve swept out by varying the threshold (Figure 2.24 shows the individual ROC curves for the 10 subjects at threshold familiarity 7). The AUROC prediction performance ranges from around 0.65 to 0.71, and is statistically significant as determined by trying to predict familiarity under a shuffled mapping of subject to stimulus familiarity. Higher levels of familiarity are easier to predict from distinguishability performance, but prediction accuracy peaks at separating self/sister ($\geq$9) from all other types of familiarity ($\leq$ 8). Separating the self face as familiar from sister as unfamiliar, corresponding to threshold level 10, weakens prediction accuracy slightly. As one would intuitively expect, the average weight vector w effectively subtracts the confusability of the test viewer from that of unfamiliar viewers: the greater this difference, the more likely the subject is actually familiar: the entry in w corresponding to $C_{others,mean}$ is positive on average (0.4), and the one corresponding to C_{self} is negative (-0.8).

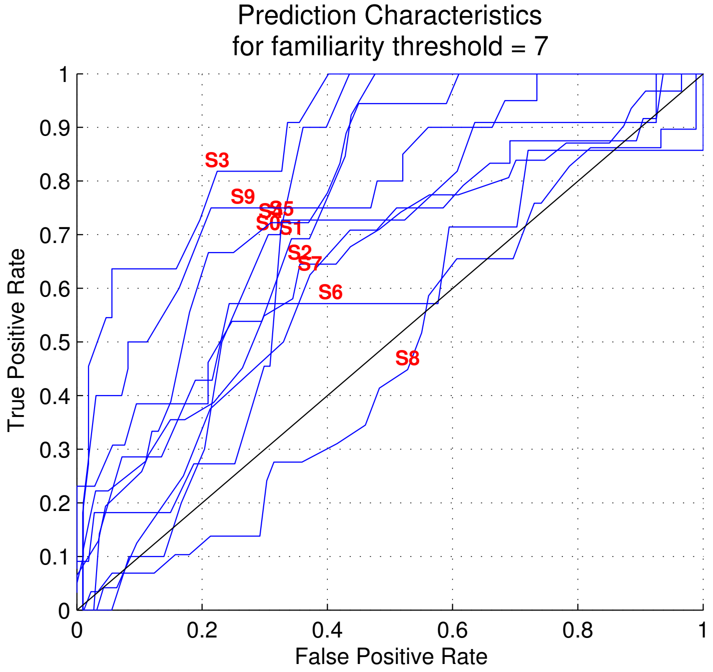


Figure 2.24: ROC performance curves for predicting the familiarity ($\geq$ 7) of each individual subject from distinguishability performance. Subject S8's familiarity level is most difficult to predict.

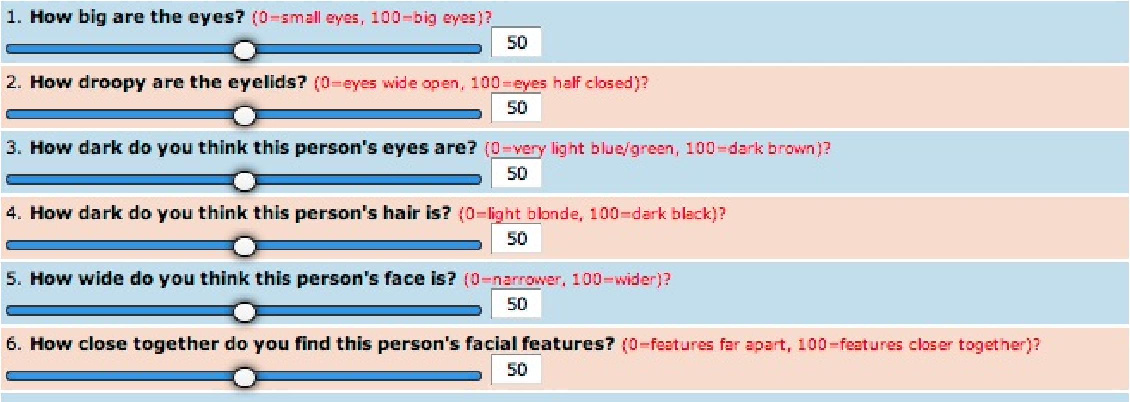
Predictability of familiarity from distinguishability and average familiarity effect size: There were some individual differences in the extent to which a subject exhibited a boost in dissimilarity judgments from familiarity with a face. If we rank subjects from most affected to least affected, and also rank their predictability in the above analysis (AUROC) at threshold level 7, we find a positive correlation between the two rankings (0.30), which is consistent with the positive correlation, reported earlier and shown in Figure 2.8, between individual differences in dissimilarity and distinguishability on single face pairs.

2.7 Factors contributing to face similarity perception

Although stimuli were relatively small (9° × 12° of visual angle) and briefly displayed (200 ms), thus discouraging both the need for and possibility of eye movements despite the experimental instructions to stay fixated at the center, naturally different subjects would occasionally move their eyes or covertly attend to different facial properties. In this section, we investigate how facial features, both holistic and localized, weigh differently in the perception of familiar faces relative to unfamiliar ones. We do this by modeling dissimilarity judgments between faces as driven by the inter-stimulus distances between single features, such as eye position or attractiveness. Refer to Section 2.10.5.2 for a detailed description of how a single feature is assigned a corrected significance level in modeling the dissimilarity judgments.

2.7.1 Independently-rated face properties and dissimilarity

An entirely separate set of 10 healthy adult subjects (7 male) participated in an online experiment, **FaceRate**, in which each of the 40 face stimuli in our data set were rated for 18 holistic and localized features such as femininity, largeness of the eyes relative to average, and attractiveness. The face features were rated by subjects on a computer, in their own home, taking as much time as they wanted for each face, while being presented with the face stimulus (inner features only) on the screen, and a sliding-bar interface like the one shown in Figure 2.25.



1. **How big are the eyes?** (0=small eyes, 100=big eyes)?
50

2. **How droopy are the eyelids?** (0=eyes wide open, 100=eyes half closed)?
50

3. **How dark do you think this person's eyes are?** (0=very light blue/green, 100=dark brown)?
50

4. **How dark do you think this person's hair is?** (0=light blonde, 100=dark black)?
50

5. **How wide do you think this person's face is?** (0=narrower, 100=wider)?
50

6. **How close together do you find this person's facial features?** (0=features far apart, 100=features closer together)?
50

Figure 2.25: A snapshot of part of sliding bar interface used by subjects in FaceRate. This was rendered in a browser and used by 10 participants to provide holistic and featural judgments about the 40 face stimuli.

Each feature was to be rated between 0 (much less than average) and 100 (much greater than average), and was initialized at 50. These ratings were then individually Z-scored (across stimuli) for each rater. Finally, the 10 raters' data were averaged together to form a single averaged Z-score

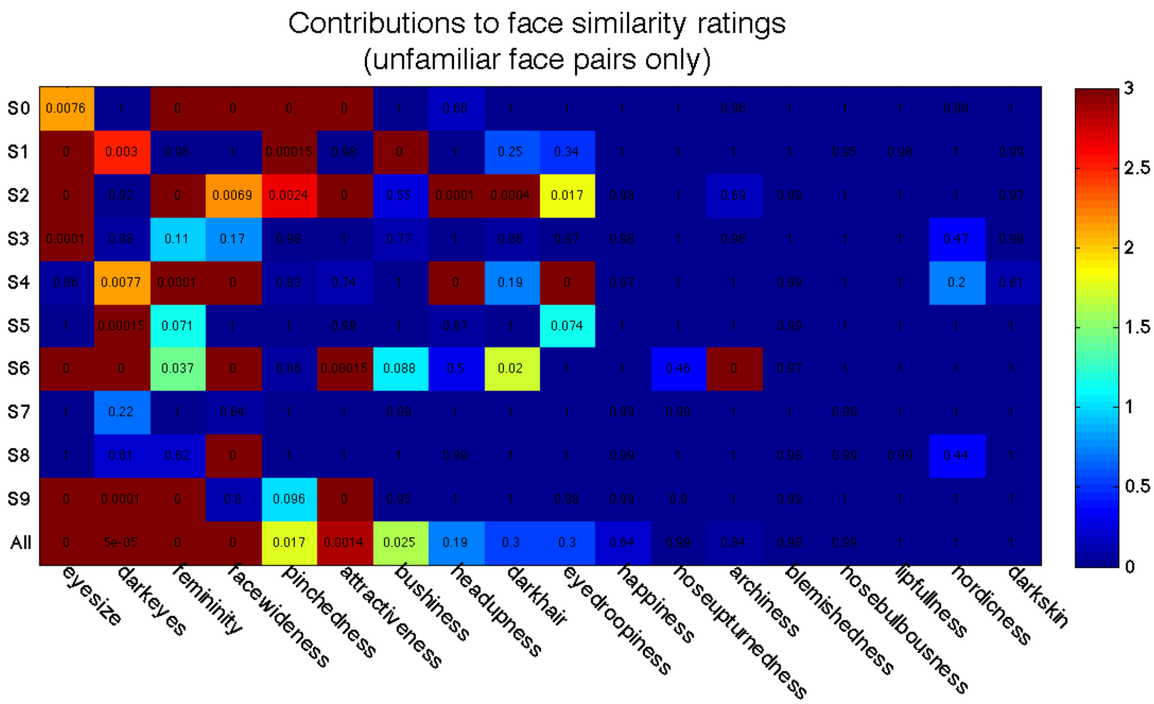
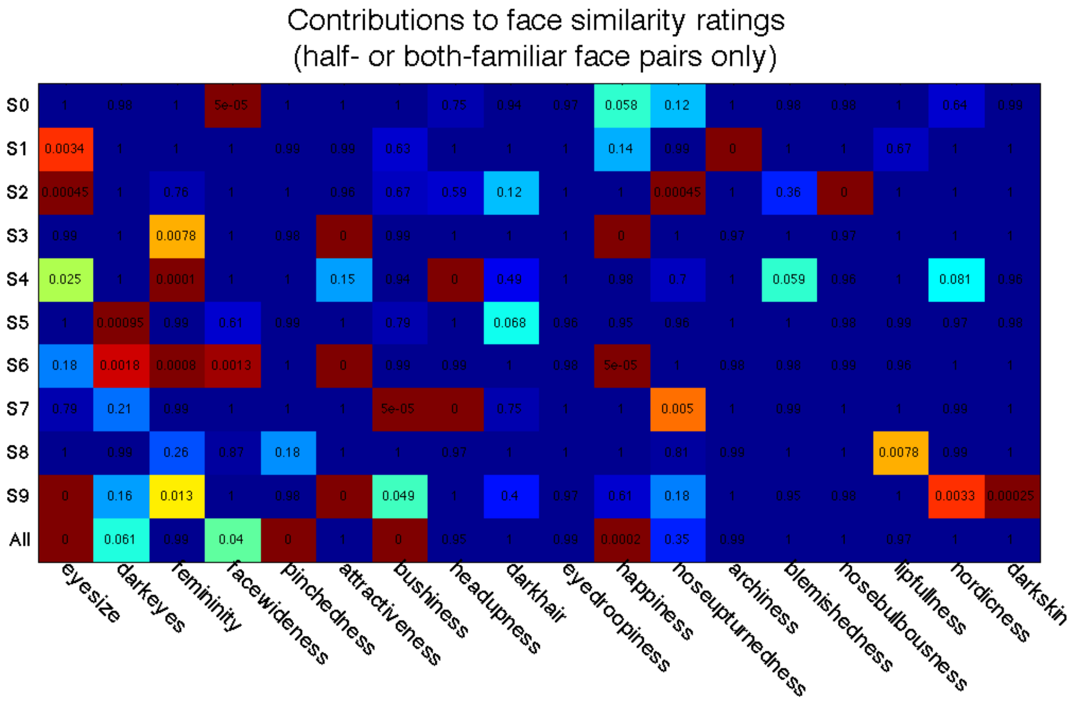


Figure 2.26: Face dissimilarity judgments, **between pairs of entirely unfamiliar faces**, can be statistically significantly modeled using distances between single features. The significance values are written in the table and the color corresponds to $-log_{10}(p)$. The values in the "All" row are determined by modeling the subject-averaged dissimilarities.

metric for each of the face features. The rated face features are described in detail in Section 2.10.5.1, but they have mostly self-explanatory names.

Figures 2.26 and 2.27 show the result of modeling dissimilarity judgments in SimRate using the 18 face features provided by participants in FaceRate. Whereas in Figure 2.26 we model only those dissimilarity judgments between pairs of unfamiliar faces, in Figure 2.27 we model only the dissimilarity judgments between pairs in which at least one face is familiar. We see that for both kinds of comparisons, the size of the eyes relative to the face is an important: if one face has relatively smaller eyes, and the other relatively larger, this will on average be associated with a higher dissimilarity. This is likely in part explained by the fact that subjects were asked to fixate between the eyes.

The weights ($-log_{10}p$) of features for the unfamiliar face pairs are on the whole positively correlated with the weights of features for familiar face pairs: there is an average individual subject correlation across conditions of 0.23, and the subject-averaged weights correlate across conditions at 0.28. However, there are also some important differences. Comparisons involving familiar faces tend to use a broader range of face features than those involving only unfamiliar faces: this is seen



	eyesize	darkeyes	femininity	facewideness	pinchedness	attractiveness	bushiness	headupness	darkhair	eyedroopiness	happiness	noseupturnedness	archiness	blemishedness	nosebulbousness	lipfullness	nordicness	darkskin
S0	1	0.98	1	5e-05	1	1	1	0.75	0.94	0.97	0.058	0.12	1	0.98	0.98	1	0.64	0.99
S1	0.0034	1	1	1	0.99	0.99	0.63	1	1	1	0.14	0.99	0	1	1	0.67	1	1
S2	0.00045	1	0.76	1	1	0.96	0.67	0.59	0.12	1	1	0.00045	1	0.36	0	1	1	1
S3	0.99	1	0.0078	1	0.98	0	0.99	1	1	1	0	1	0.97	1	0.97	1	1	1
S4	0.025	1	0.0001	1	1	0.15	0.94	0	0.49	1	0.98	0.7	1	0.059	0.96	1	0.081	0.96
S5	1	0.00095	0.99	0.61	0.99	1	0.79	1	0.068	0.96	0.95	0.96	1	1	0.98	0.99	0.97	0.98
S6	0.18	0.0018	0.0008	0.0013	1	0	0.99	0.99	1	0.98	5e-05	1	0.98	0.98	0.99	0.96	1	1
S7	0.79	0.21	0.99	1	1	1	5e-05	0	0.75	1	1	0.005	1	0.99	1	1	0.99	1
S8	1	0.99	0.26	0.87	0.18	1	1	0.97	1	1	1	0.81	0.99	1	1	0.0078	0.99	1
S9	0	0.16	0.013	1	0.98	0	0.049	1	0.4	0.97	0.61	0.18	1	0.95	0.98	1	0.0033	0.00025
All	0	0.061	0.99	0.04	0	1	0	0.95	1	0.99	0.0002	0.35	0.99	1	1	0.97	1	1

Figure 2.27: Face dissimilarity judgments, **between half- or both-familiar face pairs**, can be statistically significantly modeled using distances between single features. The significance values are written in the table and the color corresponds to $-log_{10}(p)$. The values in the "All" row are determined by modeling the subject-averaged dissimilarities (among those who are familiar).

qualitatively by examining the relatively higher scattering of brightness values on the right side of table showing weights for familiar comparisons. Also, whereas happiness does not significantly model unfamiliar comparisons ($p < 0.64$), it does for familiar comparisons ($p < 0.0002$). Although the faces were all supposed to be emotionally neutral, inevitably subtle micro-expressions leaked through conveying some kind of nonzero emotional valence. The increased weighing of happiness in familiar comparisons is consistent with the hypothesis that we are more sensitive to detecting the emotional state of familiar faces.

2.7.2 Individual facial keypoints and dissimilarity

Using the locations of the annotated keypoints in the scale-normalized faces (i.e., the stimuli the subjects viewed), we performed an analysis exactly analogous to the one on facial features above, except with scalar distances replaced with Euclidean distances in 2D combining the x- and y-coordinates of the keypoints. We used inter-stimulus distances between keypoints to model the dissimilarity judgments of subjects. The results are shown on the average face (equally blended morph of all 40 stimuli) in Figures 2.28 and 2.29.

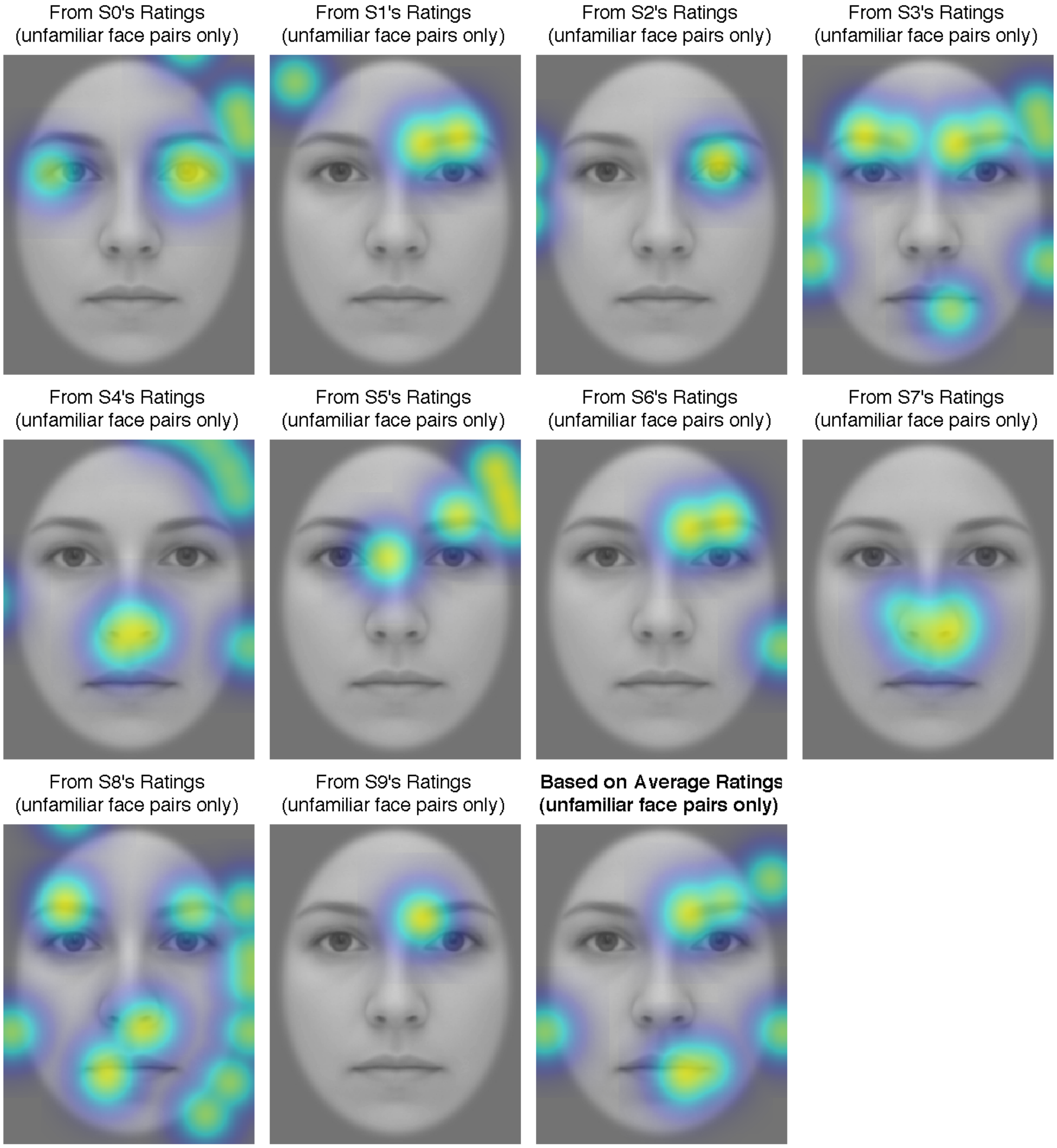


Figure 2.28: Face dissimilarity judgments, **between pairs of entirely unfamiliar faces**, can be statistically significantly modeled using distances between single features. The locations of keypoints which significantly model (viz., having $p \leq 0.01$) dissimilarity judgments between **unfamiliar faces** are highlighted with a circular Gaussian kernel for illustration purposes. The standard deviation of this kernel (0.5° of visual angle) was chosen to allow some fuzziness in the location of the underlying driving features, and to allow overlap.

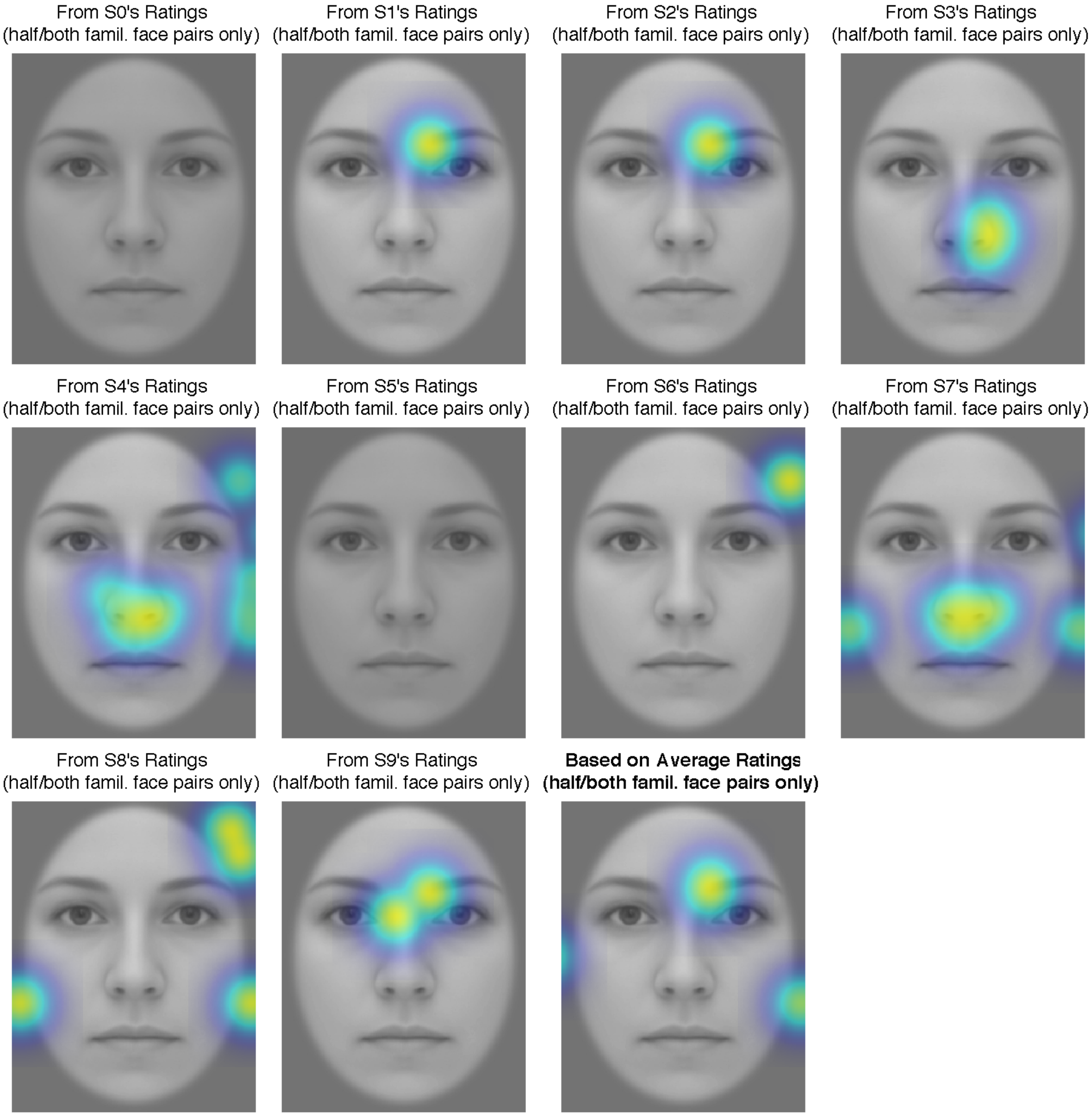


Figure 2.29: Face dissimilarity judgments, **between half- or both-familiar face pairs**, can be statistically significantly modeled using distances between single features, shown highlighted where $p \leq 0.01$. However, compared to judgments between unfamiliar faces, there are fewer significant keypoint locations, possibly due to more holistic processing of familiar faces.

For both comparisons involving familiar faces, and those which do not, we find individual facial keypoints which can statistically significantly model dissimilarity judgments. However, we find on average many more of these when modeling unfamiliar face comparisons. One possible simple reason might be that we have fewer data samples of dissimilarity judgments involving familiar pairs, so the average ratings are noisier and the individual ratings are sparse, making a significant fit more difficult to achieve. However, this is unlikely because there were many examples of face features, rated in FaceRate and shown above, which significantly modeled only the familiar comparisons, not the unfamiliar ones, such as "nordicness" for subject S9, or "blemishedness" for subject S4, or happiness for several others. A more plausible explanation is that familiar faces are perceived and compared more holistically than unfamiliar faces, thereby eliminating a significant effect of any one single keypoint location as our analysis here attempts to find for each individually. A related finding was reported by Heisz et al. [40], in which it was found that familiar faces (though not *personally* familiar) under free viewing (3 seconds, compared to our ≤ 0.2 seconds) were scanned by eye movement less in some contexts (identity recall but not recognition of familiarity). If we suppose that eye movements are more difficult to suppress or are otherwise more frequent for unfamiliar faces, this might mean that locally salient visual information plays a larger role in unfamiliar faces than for familiar faces.

The facial keypoint weightings in this subsection may be related to the independently-rated face features weightings, shown in the previous subsection; we list here several examples of consistency between the two: (1) subject S7 heavily weighs both nose-upturned-ness above, and nose keypoints here, in the case of familiar face pairs, (2) subject S6 weighs "archiness" (extent to which eyebrows are arched) heavily above, and eyebrow keypoints here, in the case of unfamiliar face pairs, (3) for subject S0, eye size is significant in unfamiliar face pairs above, but it is not for familiar ones, and the corresponding results are found with respect to eye keypoints here, (4) subject S8 heavily weighs face wideness above, and here keypoints along the side of the face which would cue face wideness, for unfamiliar face pairs, (5) the most significant rated feature above for subject S1 in the case of familiar face pairs is eyebrow archiness, and the only keypoint location we find significance for here is the tip of one of the eyebrows.

2.8 Discussion

In this chapter, we showed that familiarity with a face enhances its distinguishability from and dissimilarity to others, and we found evidence that such perceptual distancing extends into the entire neighborhood of faces around a familiar one. We summarize these effects as a "distinctiveness boost" given to familiar faces. We also showed that it is possible to accurately predict whether a face is familiar based on a person's ability to distinguish it from others. Lastly, we provided data suggesting that familiar faces are processed more holistically and with more attention to emotion.

We began the chapter with a discussion of norm-based coding vs exemplar-based coding in the brain. We discussed the implication that norm-based coding does not allow for familiar faces to look progressively more distinctive as they are learned. Now, it is easy to imagine how our results are consistent with the predictions of *exemplar-based coding* of faces, wherein each new face is represented by its distances to familiar faces. If we assume that as a face is being learned its weighting in the distances-to-references representation increases, thus adding a new and stretching dimension to this face space, all faces would gradually push away from it, and from each other, leading to a perceptual expansion of the face space around it, consistent with the one we experimentally observe. That the effect should be limited to only the neighborhood of a familiar face can be accounted for if we assume that reference faces beyond some threshold distance do not factor (e.g., one might not use an Asian male face reference to distinguish between two Caucasian females); analogously, in the brain, a neuron encoding a likeness to a reference face too dissimilar to the visual input may not fire at all, thus not contributing to its representation. However, although our results are *not* compatible with *strictly* norm-based coding of faces in visual cortex, it is our view that visual cortex likely accommodates neurons implementing both norm-based and exemplar-based coding of faces.

It is easy to argue that the perception of familiar faces as more different looking confers an evolutionary advantage: for example, this would make it easier to pick out a familiar face in a crowd, thus facilitating more efficient acquisition of critical social information among competing or cooperating animals. The relative ease of visual search for such a stimulus was demonstrated in a classic paper by Duncan et al. [24], wherein it was shown that "[search] difficulty increases with increased similarity of targets to nontargets". So a target face which is perceived to look dissimilar to nontargets would be found more easily. A related finding specific to faces was reported by Pilz et al. [71], who found that faces which were studied in motion, rather than as still images, were subsequently located more quickly in a visual search array.

Future directions: In defining familiarity, we did not account for number of stimulus presentations. A follow-up longitudinal study could investigate the relationship between distinctiveness and familiarity as they coevolve over the course of the experiment. Also, it may be interesting to investigate whether non-visual familiarity with a person may similarly enhance visual distinctiveness.

For instance, in a laboratory setting it would be possible to present a subject with two face images: both are *seen* for equal times; however, the subject may then have a phone conversation with one of the two people whose faces she just saw (chosen randomly). One would then test whether the face corresponding to the person the subject felt she talked to appeared relatively more distinctive. Also, all the results presented in this chapter are based on adults with face recognition in the normal range. This raises the question of whether individuals with prosopagnosia, or those at the highest end of face recognition ability, would show a similar boost in distinctiveness perception of familiar faces.

2.9 Supplementary results

In this section, we provide supplementary results which test various low-importance hypotheses. The figures and captions will be mostly self-explanatory based on the text earlier in the chapter. Note that, after the supplementary results, there is also an Appendix (Section 2.10), which includes experimental methods.

2.9.1 There is consensus across subjects within tasks

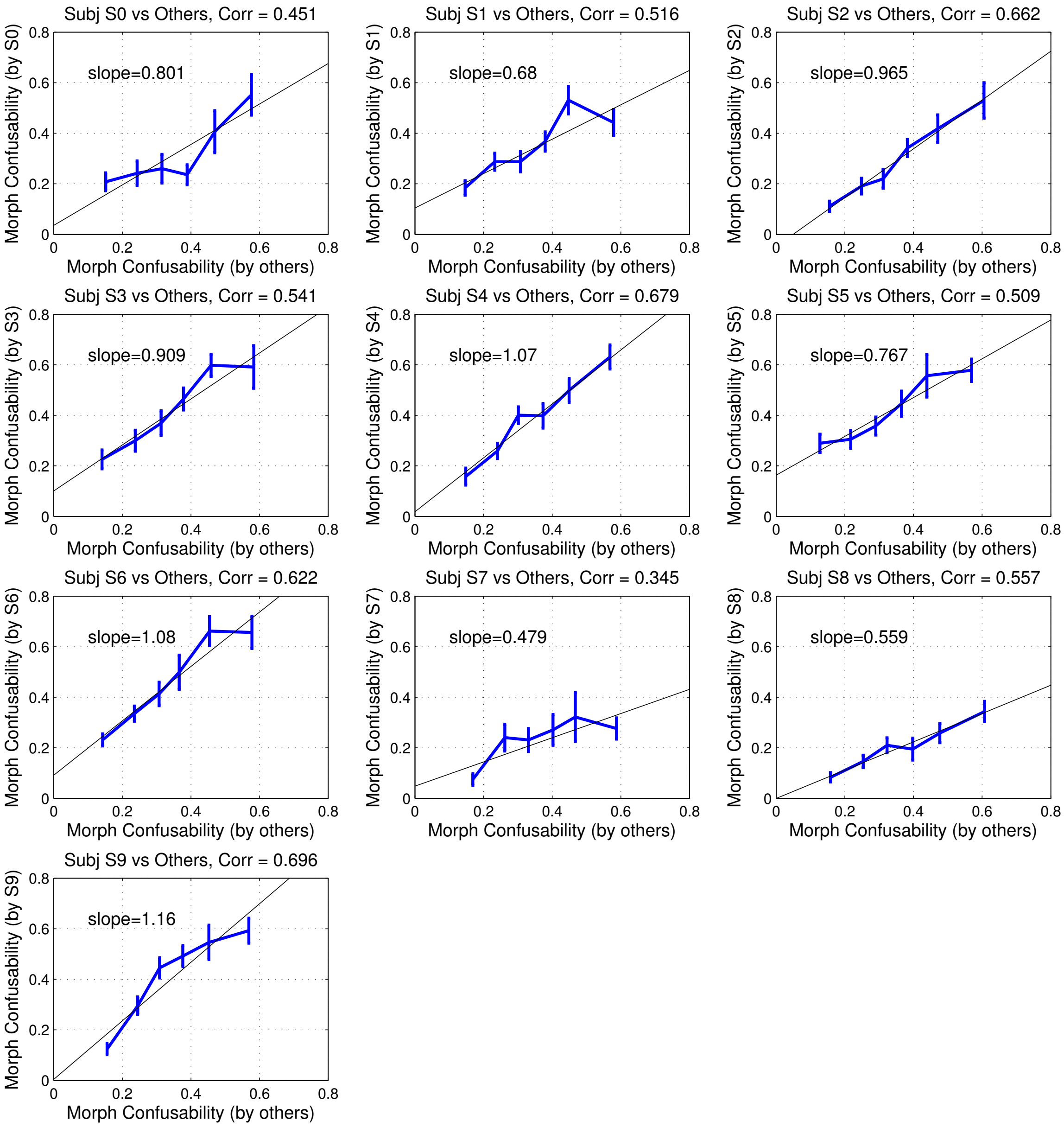


Figure 2.30: Each subject's confusability between face pairs increases on average with the confusability of face pairs as determined by other subjects (averaged across them), i.e., subjects are consistent in confusing face pairs in the task MorphDiscrim.

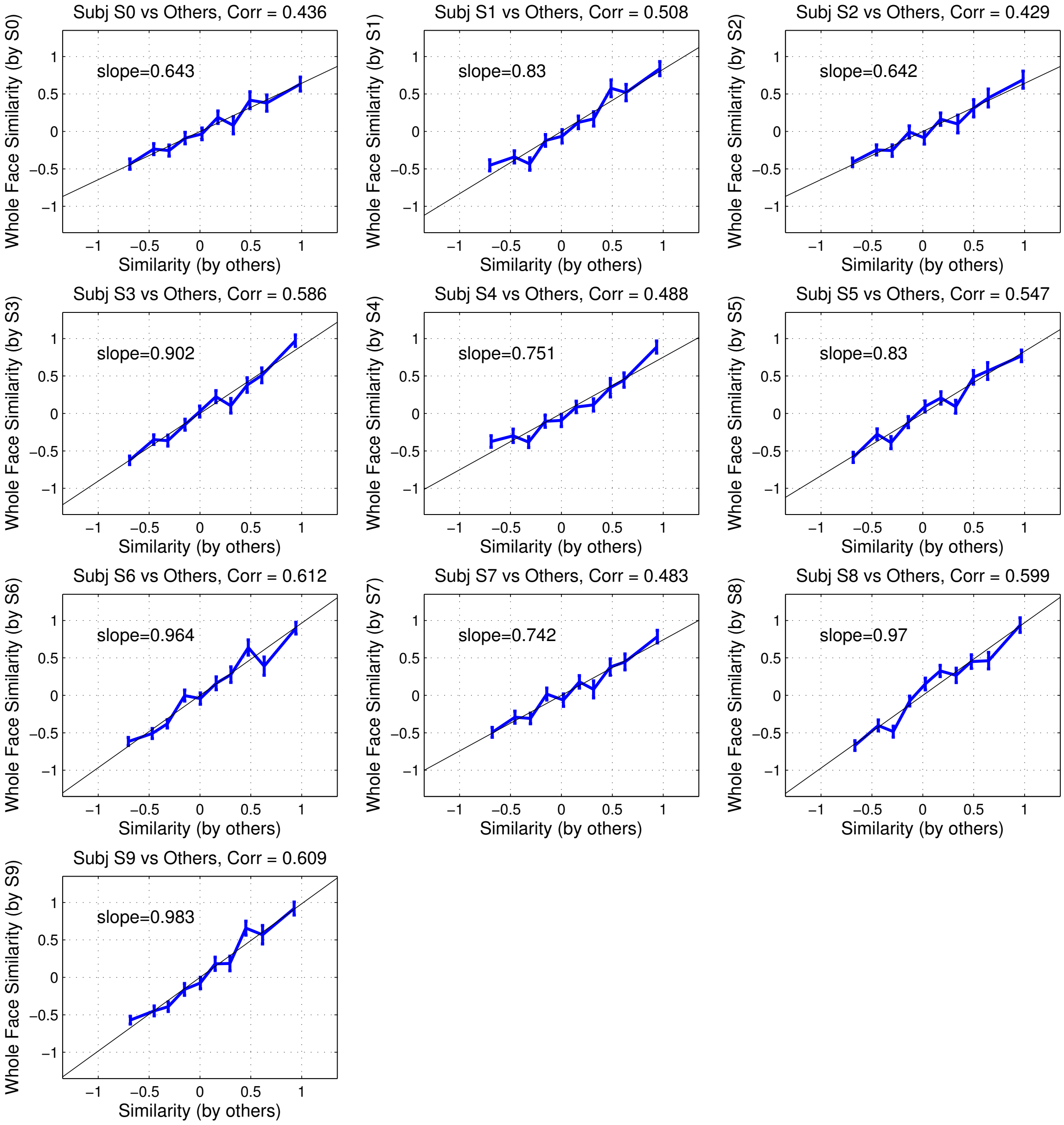


Figure 2.31: Each subject's similarity judgments between face pairs increases on average with the similarity of face pairs as judged by other subjects (averaged across them), i.e., subjects are consistent in judging similarity between face pairs in the task SimRate.

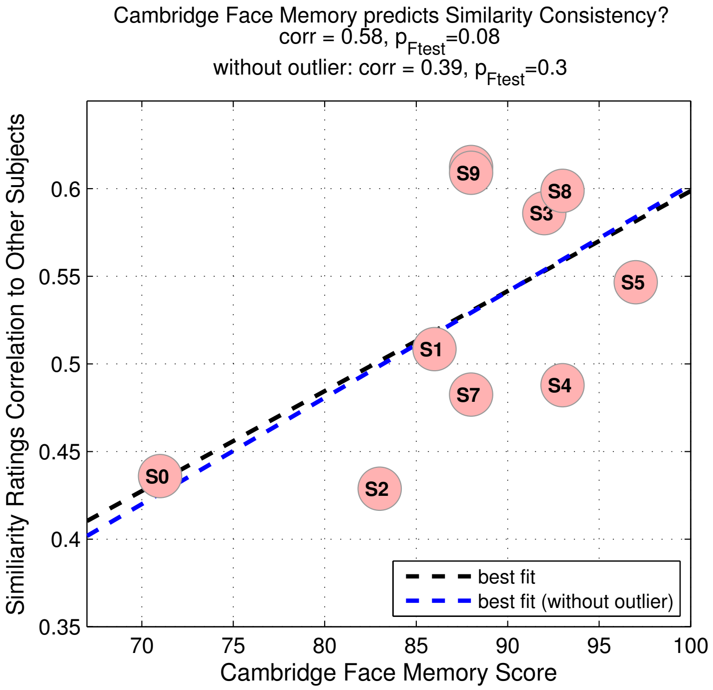


Figure 2.32: We tested whether a subject's correlation to others' ratings in SimRate (shown in the title of each panel in Figure 2.31) was related to their performance on the Cambridge Face Memory task, which tests face recognition ability. The two are positively correlated, even if we leave out subject S0, who was an outlier in the Cambridge task. However, the few data points do not reach statistical significance. For confusability, cross-subject consistency was not related to the Cambridge performance; however, anecdotally, subject S0 who scored lowest on the Cambridge task also had the fewest trials correct in labeling identical face pairs as such.

2.9.2 Unique subject pairs are similar across face similarity tasks

In this section, we test whether, if subjects A and B are provide more similar ratings in one experimental session of SimRate than some other subject pair C and D, they will on average also have more similar ratings in another session with entirely different face pairs. This is confirmed in Figure 2.33, and is consistent with the idea that subjects had a specific strategy when providing judgments, possibly involving covertly attending to specific features. We also find that in the only batch shown in the figure where the subjects' faces themselves were *not* included (Extra (i.e., non-subject/non-sister) v Sisters), consistency across subjects was the lowest. This is probably because those batches were more boring to subjects, so their attention and performance suffered, leading to more noisy judgments.

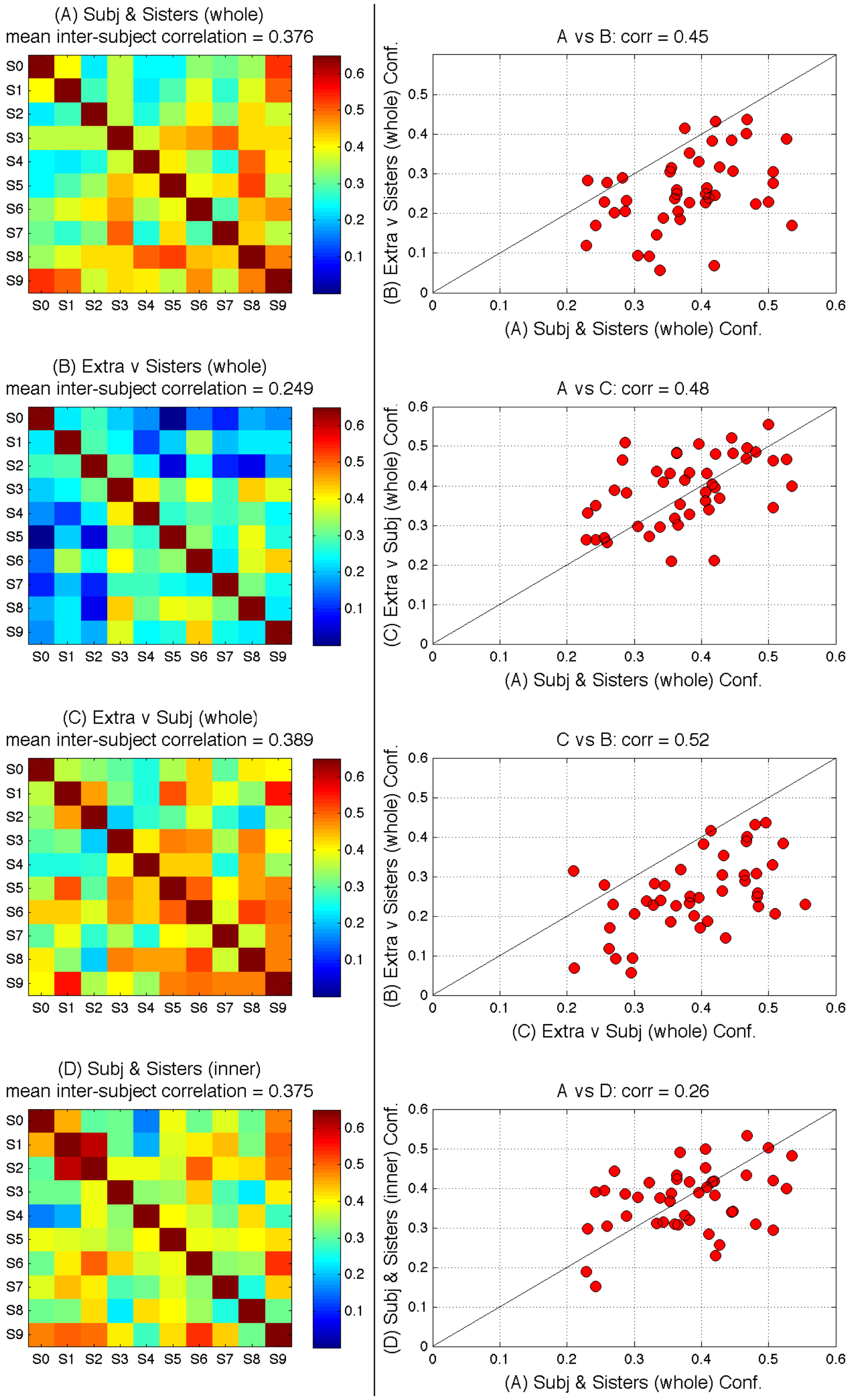


Figure 2.33: *Left:* Correlation (across faces) between individual subjects' ratings in individual batches (labeled A-D) of SimRate. *Right:* The cross-face correlation in similarity judgments between unique pairs of subjects is correlated across batches.

2.9.3 Siblings look alike

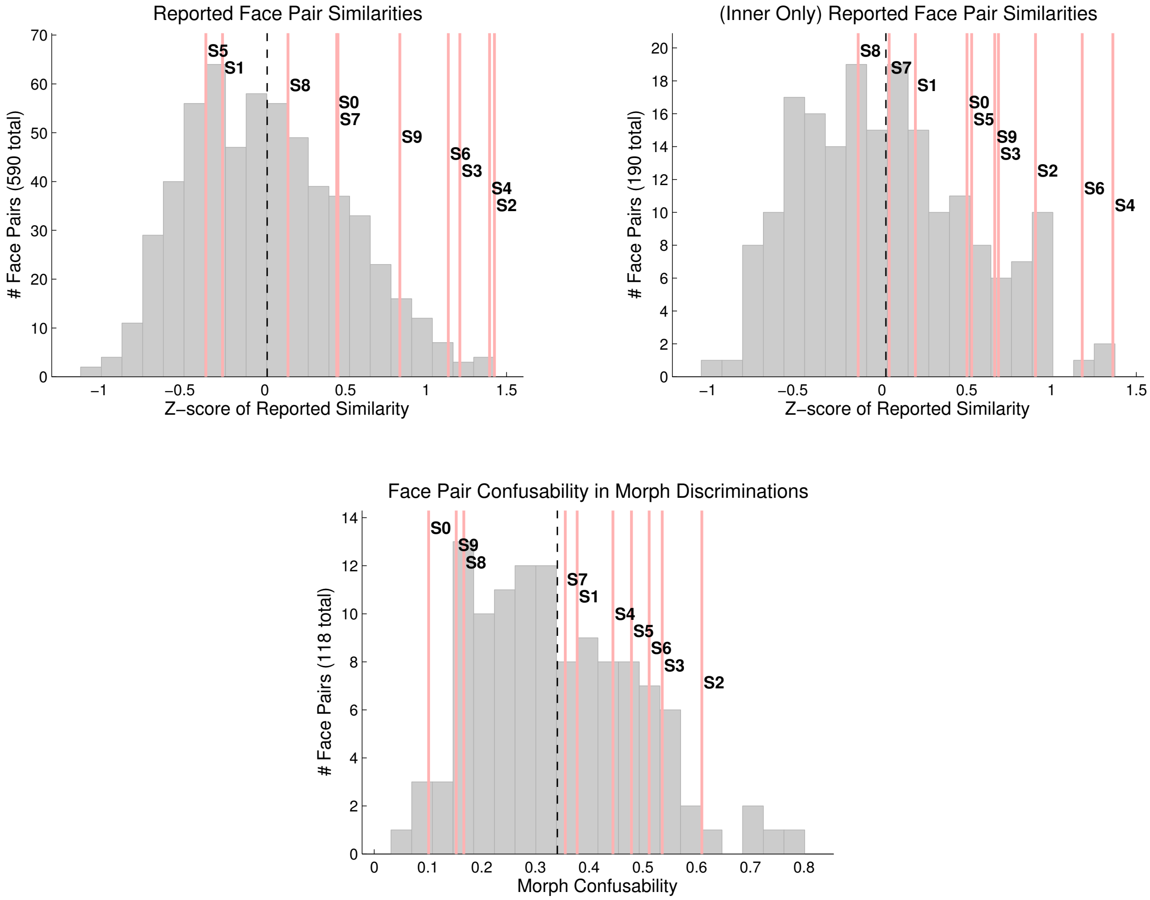


Figure 2.34: As rated by unfamiliar viewers, sister pairs are more similar looking than random pairs of faces. This histograms in gray show the distribution of similarity/confusability ratings, and the pink lines indicate the where the similarity between a particular subject (labeled) and her sister falls. Most of the pink lines lie to the right of the mean (shown in a dashed black line).

2.9.4 Comparison of stimulus distinctiveness measures

For this section, it is important to understand how pixel-based, keypoint-based (these last two being governed entirely by the visual content of the stimuli and not filtered through the brain), similarity-rating-based, and morph-distinguishability-based stimulus distances, and distinctiveness values, are computed. For details on this, refer to Section 2.10.6. Also, see Figure 2.9 for a comparison of stimulus *distance* metrics (rather than distinctiveness).

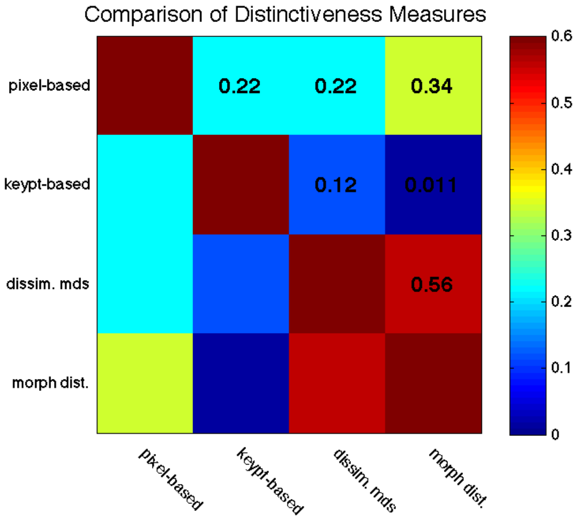


Figure 2.35: The relationship between different measures of stimulus distinctiveness. We show the correlation between each such measure across face stimuli. Distinguishability-based and dissimilarity-based distinctiveness are most correlated (0.56), whereas distinguishability-based and keypoint-based distinctiveness are least (0.011).

2.9.5 Familiarity effect on explicit similarity ratings: inner face only

The familiarity distinctiveness boost is observed on similar ratings based on inner-only face presentations too.

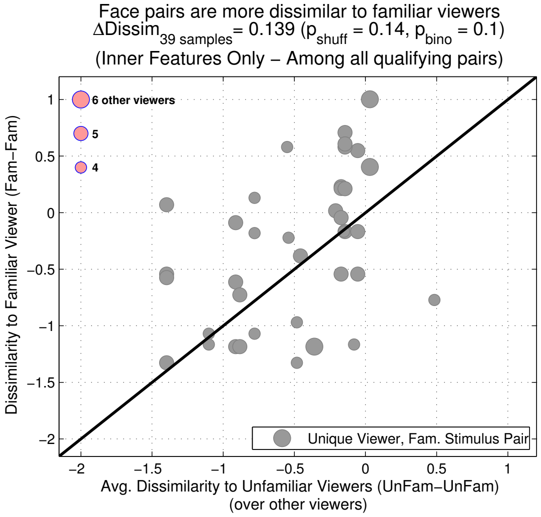


Figure 2.36: A pair of faces, with only inner features exposed, both familiar to a viewer, is rated more dissimilar than by viewers unfamiliar with either face in the pair. The sizes of the circles correspond to the number of unfamiliar viewers averaged together for the x-coordinate. (self,sister) pairs are excluded from this analysis due to “insider information” bias.

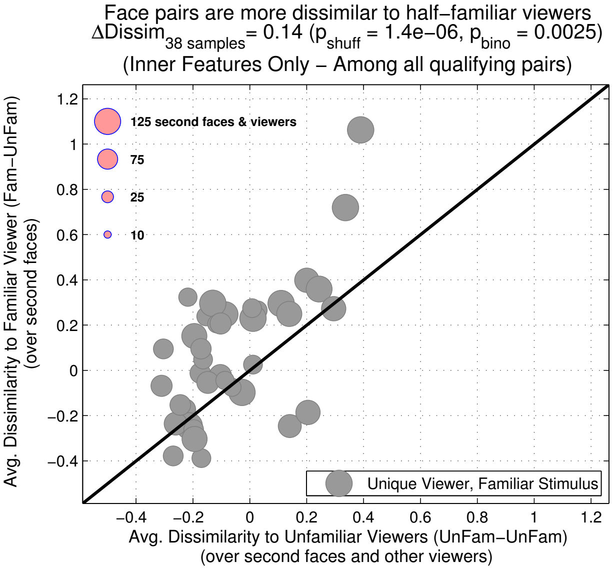


Figure 2.37: A pair of faces, with only inner features exposed, and with only one face familiar to a viewer, is more rated dissimilar than by viewers unfamiliar with either face in the pair. The sizes of the circles correspond to the number of unique combinations of unfamiliar viewers and unfamiliar faces averaged together for the x-coordinate (each circle/point corresponds to a unique (viewer, familiar face) pair). Note that unlike the complementary analysis in Figure 2.14, here we do not exclude face pairs which are more dissimilar.

2.9.6 Independently-rated face properties, orthogonalized

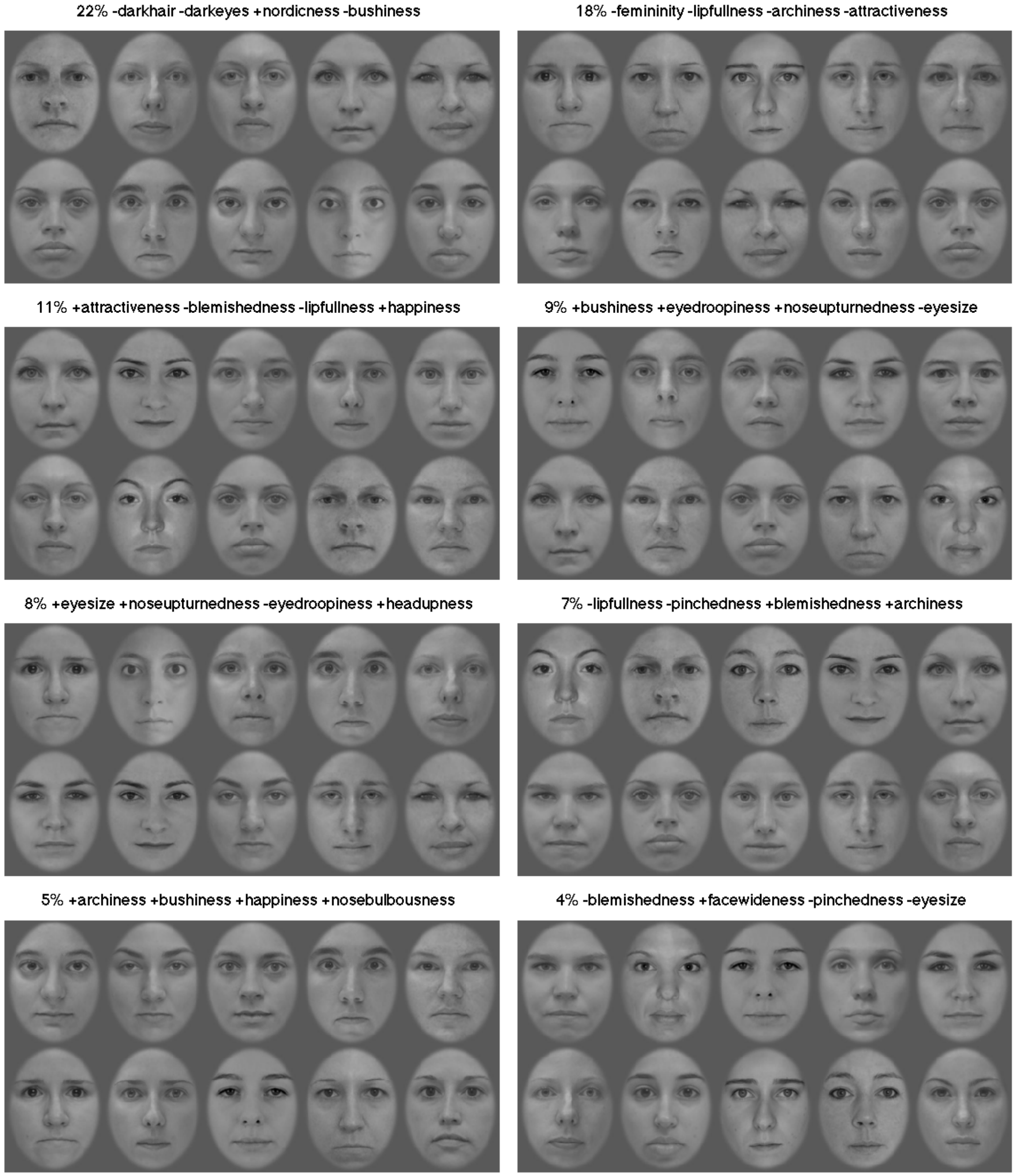

22% -darkhair -darkeyes +nordicness -bushiness
18% -femininity -lipfullness -archiness -attractiveness
11% +attractiveness -blemishedness -lipfullness +happiness
9% +bushiness +eyedroopiness +noseupturnedness -eyesize
8% +eyesize +noseupturnedness -eyedroopiness +headupness
7% -lipfullness -pinchedness +blemishedness +archiness
5% +archiness +bushiness +happiness +nosebulbousness
4% -blemishedness +facewideness -pinchedness -eyesize

Figure 2.38: The top 8 principal components are illustrated above, together with the percent of the variance they each account for (84% combined). Each component is named after its 4 highest-weighted contributing properties (with sign in front indicating positive or negative). For instance, the first component heavily weighs nordicness positively, and dark hair/eyes, and eyebrow bushiness negatively. For each component, we show the 5 faces which maximize it (top row), and the 5 which minimize it (bottom).

We were interested in whether the contributions to dissimilarity ratings are more easily or accurately explained by a PCA-orthogonalized version of the independently-rated face properties, rather than the properties themselves, as many of the properties are correlated, such as femininity and attractiveness. To this end, we derived all 18 principal components (not dimensionally reduced, though "orthogonalized" for 0 covariance), and computed, as before with the original properties, their contributions to dissimilarity ratings of familiar and unfamiliar face pairs. The top 8 components are shown in Figure 2.38, and the contributions are found in Figures 2.39 and 2.40. As with the original analysis, information concentrated around the eyes dominates both types of face pair judgments: the combined feature which best explains the dissimilarity ratings is "+bushiness +eyedroopiness +noseupturnedness -eyesize". However, unlike the previous analysis which suggested that familiar face pairs are viewed more holistically, this one does not seem to suggest (or refute) that. The basic result here is that the orthogonalized face properties are less scientifically informative about the differences in perception of familiar vs unfamiliar faces than the original face property set.

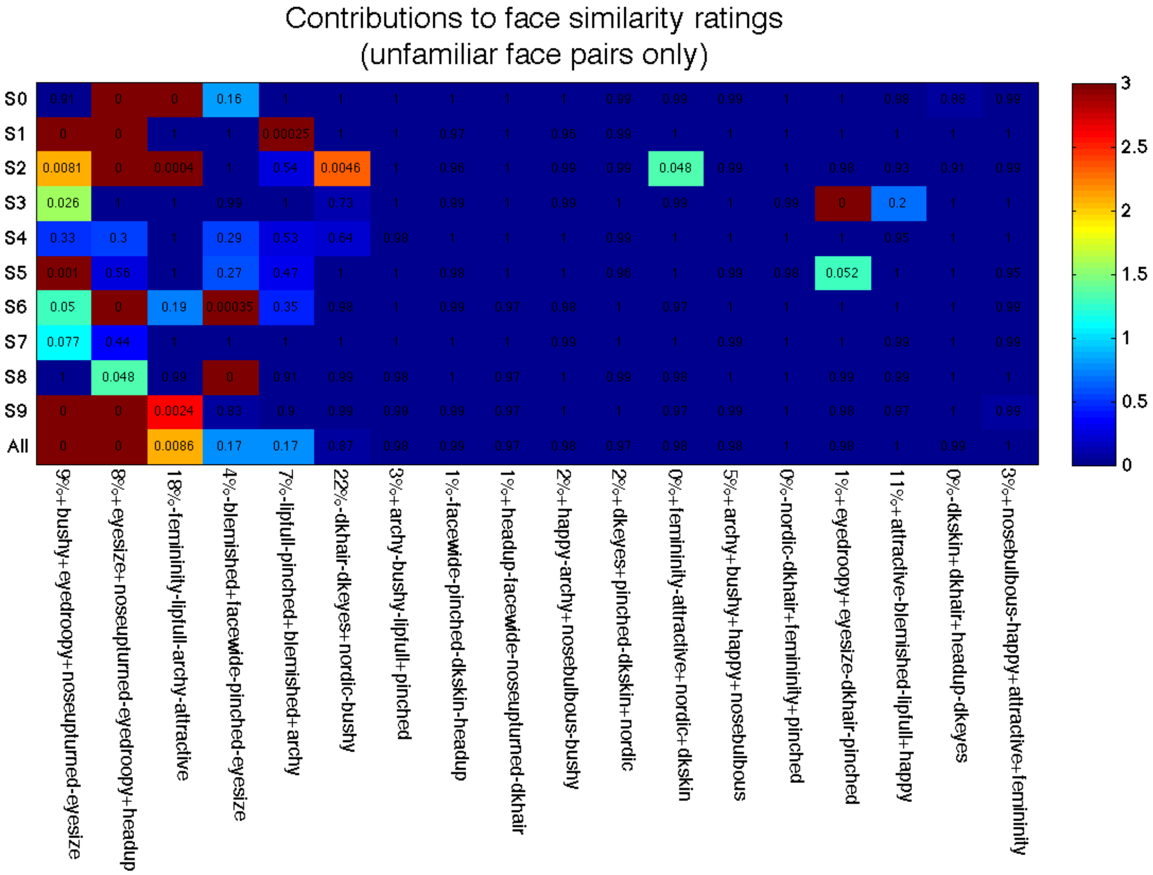


Figure 2.39: The significance of the ability of the orthogonalized face properties to model dissimilarity ratings, among unfamiliar face pairs, is shown. Color corresponds to $-log_{10}p$.

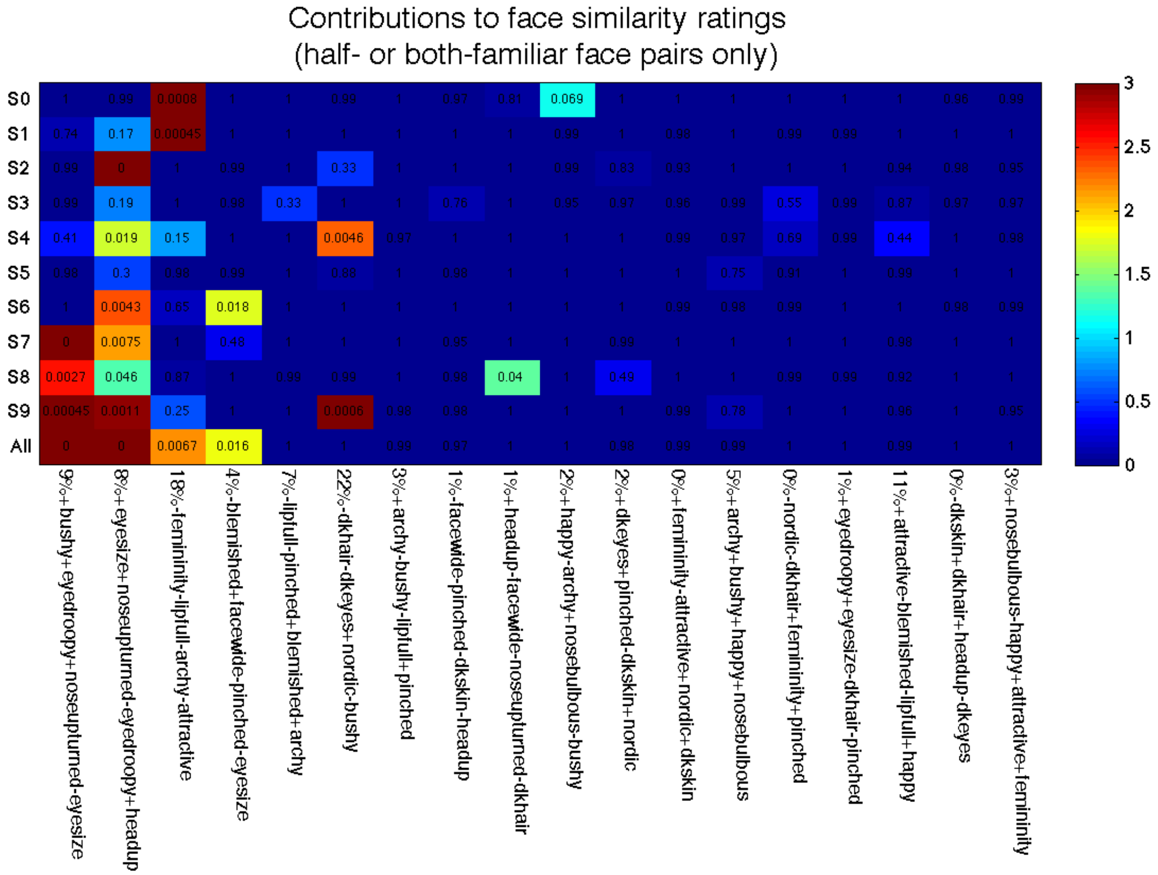


Figure 2.40: The significance of the ability of the orthogonalized face properties to model dissimilarity ratings, among half- or both-familiar face pairs, is shown. Color corresponds to $-log_{10}p$.

2.10 Appendix

Subjects provided informed written consent prior to the experiments. The Caltech Institutional Review Board approved all experimental procedures.

2.10.1 Basic computational methods

2.10.1.1 Z-score

Suppose we have a set of real values $\{x_1, x_2, \ldots x_n\}$. The Z-scored values x_i' are

$$x_i' = \frac{x_i - \bar{x}}{s}, \text{ where}$$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ is the sample standard deviation.

2.10.1.2 Principal component analysis (PCA)

Suppose we have an $m \times n$ matrix A with rows summing to zero, where the columns index sample vectors (e.g., each corresponding to one of n images), and the rows index their components (e.g., each corresponding to one of m pixels). The idea is that when we project A onto some subspaces of $\mathbb{R}^m$, we observe more variance among the sample vectors than when we project onto other subspaces. Projecting A onto a dimensionally small subspace in which much of the variance is still present is considered a reasonable means of dimensionality reduction. PCA finds an ordered list of orthogonal directions in which the data are decreasingly variant [47]. This is done as follows:

We compute the singular value decomposition $A = U\Sigma V^T$ (where U and V are orthonormal and Σ is diagonal). This provides us with the principal components: the i^{th} column of U is the i^{th} most variant direction in the data set, and for some desired number of reduced dimensions $k < m$, the reduced data are in the $k \times n$ matrix $A_k = (U_k)^T A$ where U_k consists of only the first k columns of U. In this thesis, when we refer to a q-reduced PCA representation of A, we mean A_k with k being the minimal one such that $\sum_{i=1}^{k} \sigma_i \geq q \sum_{i=1}^{m} \sigma_i$ where σ_i are the diagonal components of Σ.

2.10.1.3 Empirical significance level

In many experiments, we obtain a particular experimental value whose statistical significance level, or p-value, we wish to estimate. In order to do this without too many assumptions, we empirically compute a null distribution by randomly perturbing some part of the computation in order to assess the probability of obtaining the experimental value, or something even more favorable, under different, “dummy” circumstances, such as a shuffling of stimulus order, but with all the rest of the underlying structure left completely intact, together with systemic biases should they be present. Suppose our experimental value is x and that the set of dummy values we obtain under the perturbed settings is $\{d_1, d_2, ... d_n\}$, and for simplicity that we wish for x to be greater than would be observed at chance. We define an empirical p-value as follows:

$$p_{empirical} \triangleq \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(d_i \geq x) & \text{if at least one } d_i \text{is } \geq x \\ 1 - normcdf((x-\bar{d})/s) & \text{otherwise} \end{cases} \tag{2.1}$$

where comparison is an operation yielding 1 if true and 0 if false, $\bar{d}$ is the sample mean of the dummy variables, s is the sample standard deviation, and $normcdf()$ is the normal (Gaussian) cumulative distribution function.

2.10.2 Stimulus image preparation

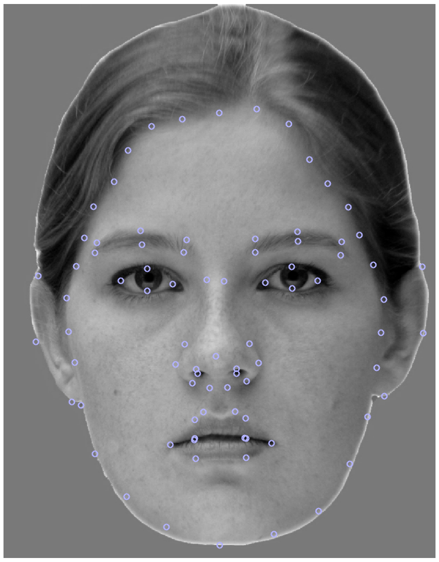


Figure 2.41: One of the 40 faces from the stimulus set (in a pre-normalized/pre-cropped version), with the 83 keypoints which were manually annotated

Images were first manually annotated for facial keypoints (e.g., “left eye outer”, “left nose bridge top”) by the author and Ronnie Bryan. We used 53 keypoints inside the face and another 30 along the boundary of the face. Figure 2.41 shows the locations of these keypoints.

The images were then resized such that the average of the following metrics was equalized to 250 pixels across stimuli: distance between lateral extremes of eyes, distance between the tip of the nose and the left extremity of the mouth, distance between the ears, and distance between the top of the forehead and bottom of the chin. The images were then centered (viz., eyes at the vertical center, midpoint between eyes at horizontal center), converted to grayscale, normalized to have the same mean and standard deviation of luminosity across pixels, and then pixel values less than .03 away from black or .01 from white were set to .03 and .99, respectively. After this normalized rescaling, the average distance between the lateral extremes of the eyes was 179 pixels.

Different elliptical masks were used to hide features outside of the facial interior: one which revealed the face including the hair and jawline, and which which did not. Figure 2.4 shows the result of both masking types.

2.10.3 Morphing faces

To form a blend between two faces, a linear combination between their keypoints was calculated, a Delaunay triangulation was formed on the set of keypoints, and within each triangle, a linear transformation was calculated mapping each base face's two-dimensional coordinates to the intermediate face's coordinates, which was used to compute intensities for each point in the morphed face based on the corresponding locations in each constituent base face. We used code from Ronnie Bryan for this purpose. See Figure 2.42 to see the result of blending each subject with her sister.

2.10.4 Details of Experimental Tasks

In both experiments SimRate and MorphDiscrim, pairs of face stimuli were shown in rapid temporal succession. Subjects were always asked to fixate on a white point in the middle of the screen, and face stimuli quickly flashed in around this fixation point, centered between the eyes.

2.10.4.1 Experiment SimRate

The purpose of this experiment was to get explicit ratings of similarity between pairs of faces. The subjects performed this experiment using both whole faces and just inner faces (see Figure 2.43 for comparison of results). In each case, subjects viewed pairs of different base faces in rapid temporal succession, and were asked to key in a number between 1 and 8 (on a standard computer keyboard) indicating how similar the presented faces in the pair were to each other (1=least similar, 8=most similar). Subjects performed several training runs first (on a separate set of faces), and were instructed to try to use the full range of similarity values.

Trial Structure: Each face in the pair was presented for 200 ms, with a 100 ms intervening

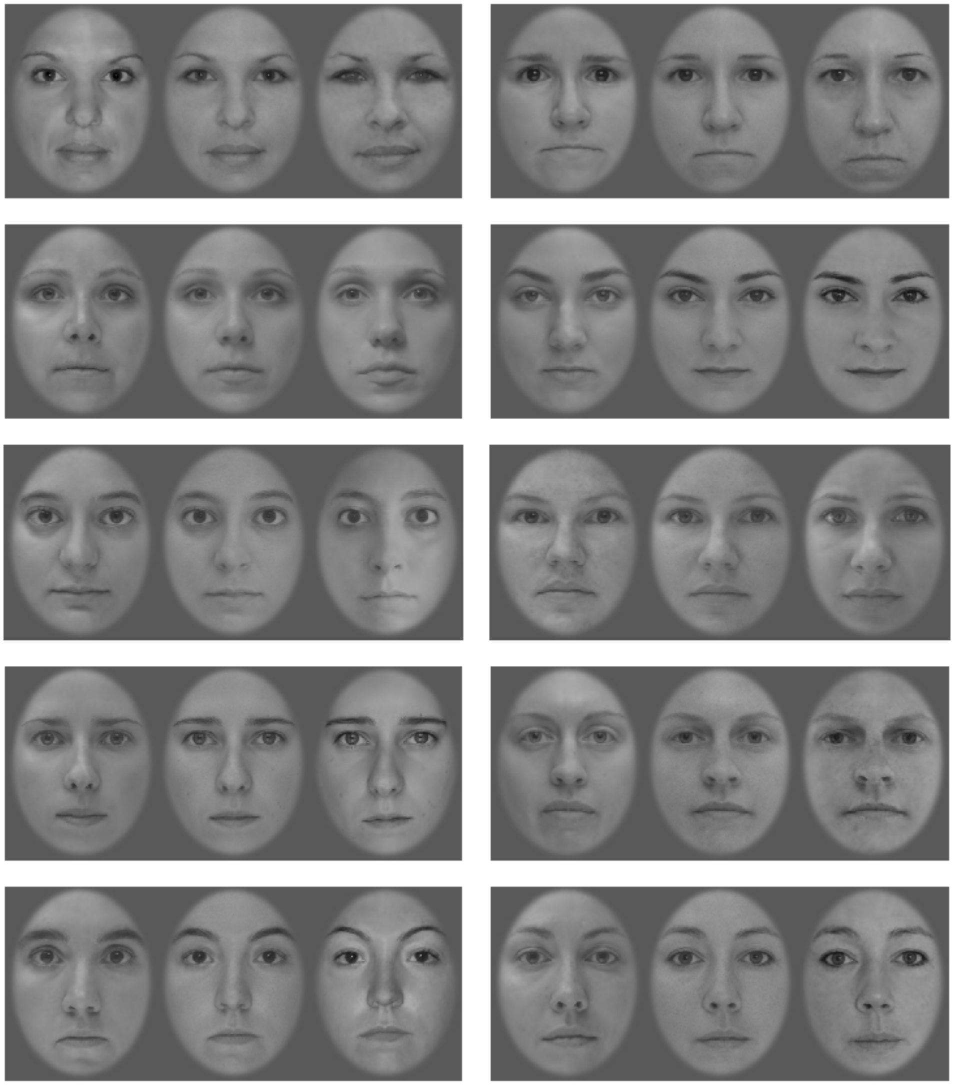


Figure 2.42: The ten subjects, their sisters, and 50/50 morphs between them. Each panel consists of subject, morph, and sister in that order, respectively.

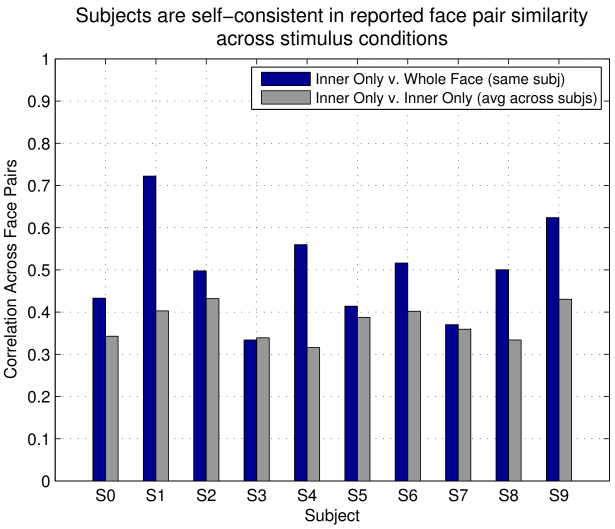


Figure 2.43: Consistency across different types of tasks SimRate: Except for subject S3, the correlation between the similarity ratings (across face pairs) was higher between (i) the subject's own ratings (whole-face ratings vs. inner-only), shown in blue, than between (ii) the subject's inner-only ratings and one other subject's inner-only ratings; the gray bar represents the average inter-subject correlation among 9 others for each subject.

mask matched for low-level spectral content (a "Fourier mask"). Subjects then had up to 5 seconds to enter a response, and a random delay uniformly distributed between 0.9 and 1.5 seconds followed that before the subsequent trial. Subjects all entered responses on all trials. See Figure 2.44 for a graphical representation of this time-course.

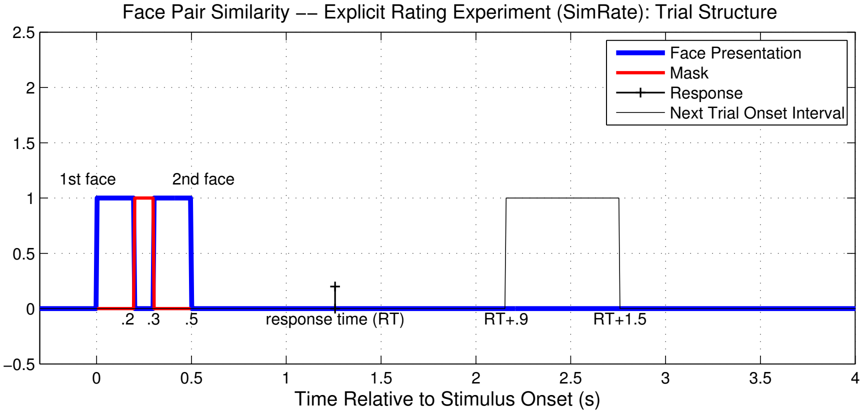


Figure 2.44: The time-course of a trial in the SimRate experiment, wherein subjects explicitly rated the similarity between a pair of faces. A response is shown occurring randomly inside its allowable range.

To discourage subjects from using very low-level image queues, and instead to pay more attention to the structure of the face, independent **Gaussian noise** (see Figure 2.45) was added to each image pixel with standard deviation .05. Subjects were asked to fixate on a point in the center of the

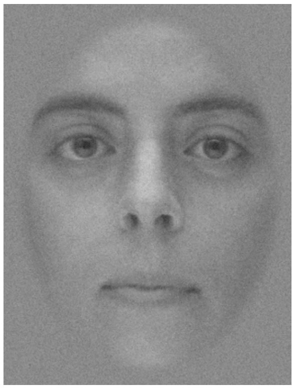


Figure 2.45: A face stimulus as it was seen for 200ms at a time during the SimRate task, with Gaussian noise added.

screen; face stimuli were shown such that this point was centered between the eyes. Faces subtended approximately 9°x12° of visual angle on the computer monitor.

Trial Blocks: All 780 unique pairs among the 40 faces were evaluated by each subject in a series of experimental sessions occurring over several days according to subject availability. Each experimental session was designed to last about 30 minutes (actual time depended on rate of subject response and jittered delays), and repeated each unique pair exactly three times throughout the session in order to assess subjects' self-consistency with respect to this measure. The order of the pairs, and order of faces within a pair, was randomly selected and different for each subject. Furthermore, the order of experimental blocks was randomized and counter-balanced among subjects. The 3 similarity scores in each session were averaged together, then Z-scored among all other face pairs in that session (i.e., the sample mean was subtracted then the sample deviation was divided).

Inner and Outer SimRate batches: All 780 unique pairs of faces were ultimately rated in SimRate, and some of them were rated using both inner-only faces and whole-faces. However, the face pairs of (non-subject/non-sister, non-subject/non-sister) stimuli were rated only using inner-face presentations, and the face pairs consisting of the (subject or sister, non-subject/non-sister) stimuli were rated only using whole-face presentations. The two kinds of ratings are generally pretty comparable, and where data from all 780 pairs is necessary (such as in modeling dissimilarity using featural distances), these two kind of face presentations are combined into unified hybrid RDM, which uses inner-only data where available and whole-face data elsewhere.

2.10.4.2 Experiment MorphDiscrim

The purpose of this experiment was to infer perceptual similarity between pairs of faces, or in other words, to get implicit similarity judgments. To this end, subjects would see pairs of faces morphed to various levels between a pair of base faces, in rapid temporal succession, and enter a key press (on a standard computer keyboard), indicating whether they thought the two faces were identical or at all different. The more poorly subjects performed at this task, the more similar the base faces were inferred to look to the subject: equivalently, a subject who could easily distinguish between the morphs was inferred to perceive greater differences between the base faces.

Experiments were set up such that in exactly half the trials the faces were in fact identical (selected uniform-randomly from the four positions in the morphing continuum 10 and 20% away from the midpoint), and in the remaining half, they were different. When face pairs were different, they were evenly selected from one of two types: mid ± 10%, wherein each of the two morphs was a mere 10% removed from the midpoint between the two faces (difficult trials), and mid ± 20%, wherein each of the two morphs was 20% removed from the midpoint (less difficult trials, as these were easier to tell apart).

Confusability and Distinguishability: For each unique pair of base faces, subjects had to complete 10 trials in the identical condition, 5 trials in the ± 10%, and 5 trials in the ± 20% condition. From these 20 trials, a "confusability" score between the base pair of faces was computed:

$$Confusbility \triangleq \frac{1}{2}\frac{(f_{10}+f_{20})}{f_{ident}},\ Distinguishability \triangleq -Confusability \tag{2.2}$$

where f_{ident} was the fraction (out of 10) trials indicated identical (i.e., correctly) when the faces were identical, f_{10} was the fraction (of 5) indicated identical (i.e, incorrectly) when the faces where in fact ± 10% away from the midpoint, and f_{20} was the fraction (of 5) indicated identical (also incorrectly) when the faces where in fact ± 20% away from the midpoint. The closer this score is to 0, the more distinguishable the faces would appear to the subject. A score of 1 would mean that the faces being different did not lower the rate at which the subject indicated they were different, suggesting the differences in the faces fell below some perceptual threshold. Other more sophisticated measures (including an estimation of psychophysical thresholds) were also tried but this simple definition of confusability and distinguishability yielded the cleanest most self-consistent results.

Trial Structure: In each trial, two faces (either identical or slightly different) were presented. Each face was presented for 110 ms, with a 110 ms intervening mask matched for low-level spectral content (a "Fourier mask"). Faces were presented for a relatively shorter time than in the SimRate task so that performance was more variable and not perfect. Subjects then had up to 5 seconds to enter a response, and a random delay uniformly distributed between 0.9 and 1.5 seconds followed that before the subsequent trial. Subjects all entered responses on all trials. See figure 2.46 for a graphical representation of this time-course.

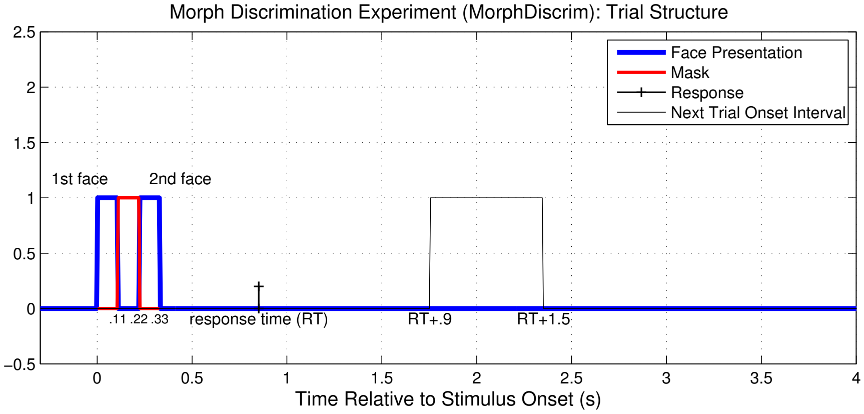


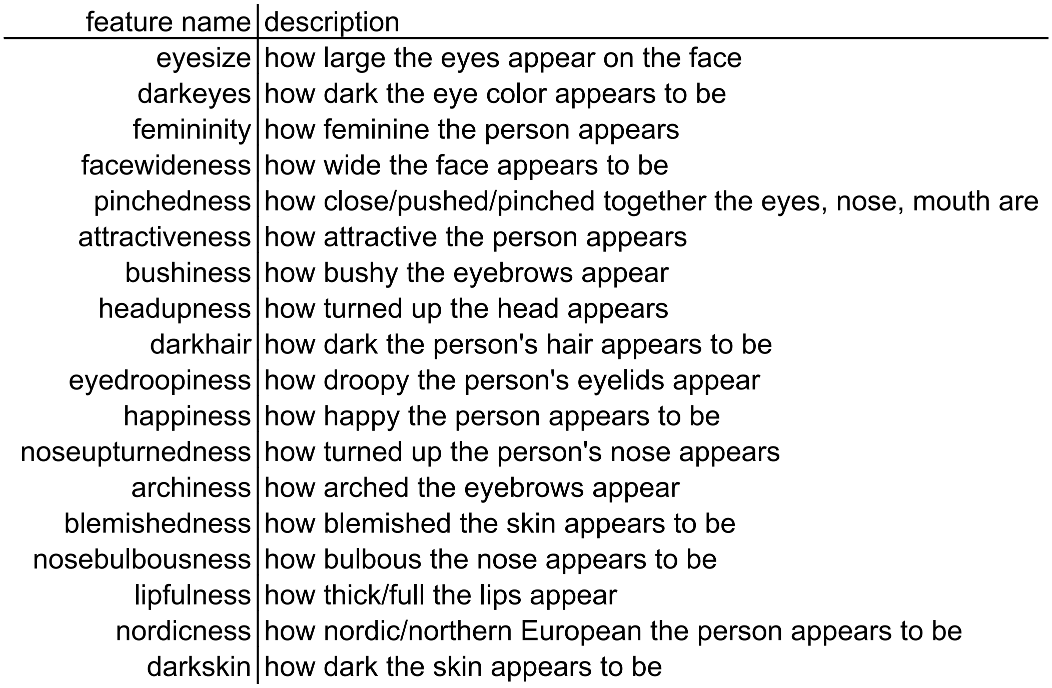
Figure 2.46: The time-course of a trial in the MorphDiscrim experiment, wherein subjects implicitly provided information about the similarity between a pair of faces. A response is shown occurring randomly inside its allowable range.

To discourage subjects from using very low-level image queues, and instead to pay more attention to the structure of the face, pixel-wise independent **Gaussian noise** (see Figure 2.45) was added to each image (independently) with standard deviation .05. Subjects were asked to fixate on a point in the center of the screen; face stimuli were shown such that this point was centered between the eyes. Faces subtended approximately 4.9°x6.6° of visual angle on the computer monitor. As in the shorter trial times, faces were relatively smaller than in the SimRate task so that performance was more variable and not perfect.

Trial Blocks: Of 780 available unique pairs among the 40 faces, only 118 were evaluated in experimental sessions by the subjects in MorphDiscrim, with the same exact pairs for each subject, due to practical constraints (at about 20-30 seconds per pair, subjects would get too exhausted doing all 780). These pairs included subject-sister pairs, and 108 others selected to span a range of similarity levels. Blocks of trials were organized into roughly 30-minute chunks, occurring over several days according to subject availability, randomized and counter-balanced among subjects, and within which trial order (including face order within a trial) was randomized independently for each subject.

2.10.5 Rated face features (FaceRate): description and modeling of dissimilarity

2.10.5.1 Description rated face features



feature name	description
eyesize	how large the eyes appear on the face
darkeyes	how dark the eye color appears to be
femininity	how feminine the person appears
facewideness	how wide the face appears to be
pinchedness	how close/pushed/pinched together the eyes, nose, mouth are
attractiveness	how attractive the person appears
bushiness	how bushy the eyebrows appear
headupness	how turned up the head appears
darkhair	how dark the person's hair appears to be
eyedroopiness	how droopy the person's eyelids appear
happiness	how happy the person appears to be
noseupturnedness	how turned up the person's nose appears
archiness	how arched the eyebrows appear
blemishedness	how blemished the skin appears to be
nosebulbousness	how bulbous the nose appears to be
lipfulness	how thick/full the lips appear
nordicness	how nordic/northern European the person appears to be
darkskin	how dark the skin appears to be

All ratings were asked to be made relative to the average, directly-gazing, neutral-expression Caucasian female. The features listed above are in order of average weight, decreasing from the top, determined by fitting to the subject-averaged dissimilarity ratings (including familiar and unfamiliar face pairs). Many of the features were a bit difficult to infer from the grayscale, inner-only, face stimuli used. For example, face wideness is conveyed only by subtle variations in the arrangement of inner face features and shading.

2.10.5.2 Modeling dissimilarity using distance between facial features

Suppose we have a real-valued candidate feature f_s for face stimulus s. This might be, for example, the x-coordinate of the nose tip in the normalized face coordinates, or the Z-scored extent to which the face features holistically appear to be close together relative to average (“pinchedness”). We assign the candidate feature a corrected statistical significance value using a three-step procedure:

(1) We determine the p-value of the F-statistic measuring the goodness of the linear fit:

$$M_k = \left[||f_i - f_j||^2 \;\; 1 \right], \text{ model: } Mw = D,$$

where M is a matrix having two columns (each row is M_k), the second only containing the constant 1, the first containing, for each unique pair of stimuli $k = (i, j)$, the inter-stimulus (squared) distance between the candidate feature values, and D is a column vector containing in each entry the corresponding dissimilarity (rated in SimRate) for stimulus pair (i, j). The goal is to determine whether the dissimilarity between faces can be significantly explained by only the differences measured in the candidate feature, for example, differences in nose tip location. The F-statistic is the standard textbook one comparing the explained variance to the unexplained variance.

(2) In order to further control for spurious fitting of the feature distances to dissimilarity values, and to create a more fair comparison between features, we Z-score the negative log p-value, by generating a large sample set of empirical ones, derived by fitting the same features, only randomly shuffled with respect to stimulus identity, to D. We call this measure **Z-score-p**. This provides us with a measure roughly telling us how unusually good the fit is relative to what we would expect at chance given the dynamic range of the candidate feature in the stimulus set.

(3) Because we test a larger number of candidate features at once (K=18 rated face features in one case, or K =83 annotated keypoint coordinates in the other), we finally assign the candidate feature a significance value by determining the probability of observing a value of **Z-score-p** or greater extremity in a population of K normally distributed values.

2.10.6 Computing inter-stimulus distances and distinctiveness values

See Figure 2.9 for a comparison of stimulus distance metrics, and Figure 2.35 for a comparison of stimulus distinctiveness measures.

2.10.6.1 Computing inter-stimulus distances

Pixel-based distances: Face stimuli were first resampled (with bicubic interpolation) to 25% of their original dimensions, yielding faces with a distance between the lateral extremes of the eyes of 45 pixels. This was intended to effectively reduce the importance of high spatial frequency information. Then, each pixel intensity in each image was Z-scored relative to the other stimuli in the stimulus set. We then treated each face stimulus as a vector with m pixels, one per component, and computed an 0.8-reduced PCA representation (see above, Section 2.10.1.2) of these face vectors, yielding only 37 components per face. The Euclidean distance between these reduced vectors is what was used as the pixel-based inter-stimulus distance. The correlation between this pixel-based metric and the simplest possible one (mean pixel-wise square-difference between full-sized images) is 0.68. Whereas the simplest pixel-based metric correlates with the dissimilarity ratings at a level of 0.10, this method correlates at a level of 0.17.

Keypoint-based distances: Using the coordinates of the annotated keypoints of each face (in the normalized, centered, space) we Z-scored each x- and y-coordinate independently relative to the other stimuli in the set. This yielded a matrix of dimensions 166×40, with one column per face, and one row for each x- and y- coordinate of the annotated keypoints (83 of them). We then computed an 0.8-reduced PCA representation of this, yielding 21 components per face. The euclidean distance between these reduced vectors is what was used as the keypoint-based inter-stimulus distance.

2.10.6.2 Computing stimulus distinctiveness values

Pixel-based distinctiveness: The mean pixel-based distance to the other faces is used (see above).

Keypoint-based distinctiveness: The Euclidean distance between each of the candidate face's keypoints and the corresponding ones, averaged across all the faces, is computed, then Z-scored relative to the population of corresponding distances among the other stimuli. For instance, the nose tip may be one standard deviation away from where the average nose tip is. This yields 83 Z-scored facial keypoint distinctiveness measures per face. The keypoint-distinctiveness of a face is taken to be the average among these.

Morph-distinguishability-based ("morph dist.") distinctiveness: The subject-averaged distinguishability (negative confusability) to other faces (among those which were compared) is averaged across other faces.

Dissimiliarity-ratings-based distinctiveness ("dissim. MDS"): In order to compute distinctiveness values based on dissimilarity ratings, we first used the dissimilarity RDM as the input to MATLAB's default multidimensional scaling (MDS) algorithm (*mdscale*, all parameters default), with 2 dimensions. This finds a locally optimal assignment of 2 coordinates to each face, such that the euclidean distance between them is tightly correlated with the dissimilarity. In general, of course, RDMs will not lend themselves to perfect reconstruction from an underlying 2D space; nonetheless, this is a good way of simplifying the dissimilarities between the faces. The result of running MDS on whole-face dissimilarity ratings and inner-face dissimilarity ratings is shown in Figure 2.47. The distinctiveness of a face is taken to be its distance from the origin in the 2D MDS space (the MDS algorithm's convention is that the origin is always the average of all the x- and y-coordinates). This turns out to correlate better with explicitly-rated stimulus distinctiveness (provided in an experiment in Chapter 3) than simply taking the mean dissimilarity to other faces as we do with morph-distinguishability-based distinctiveness for lack of having a full set of face pairs. The MDS-based distinctiveness values correlated with these explicit ratings of distinctiveness at a level of 0.56 among the base 40 faces, compared with a correlation of 0.45 using the mean dissimilarity to other faces.

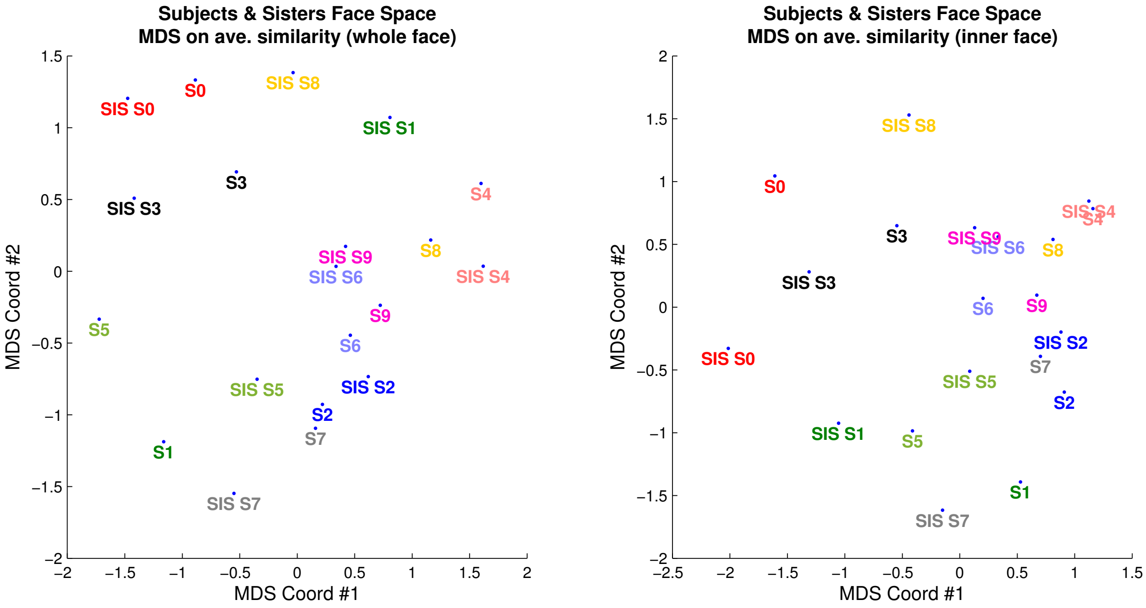


Figure 2.47: The result of reducing the subject-averaged dissimilarity judgments in SimRate to 2D, using multidimensional scaling, and only 20 of the stimuli. *Left:* Based on judgments of whole faces. *Right:* based on judgments of inner faces only. Sister pairs are color coded; notice that they are close together on average. The distinctiveness of a face is taken to be its distance from (0,0) in this space.

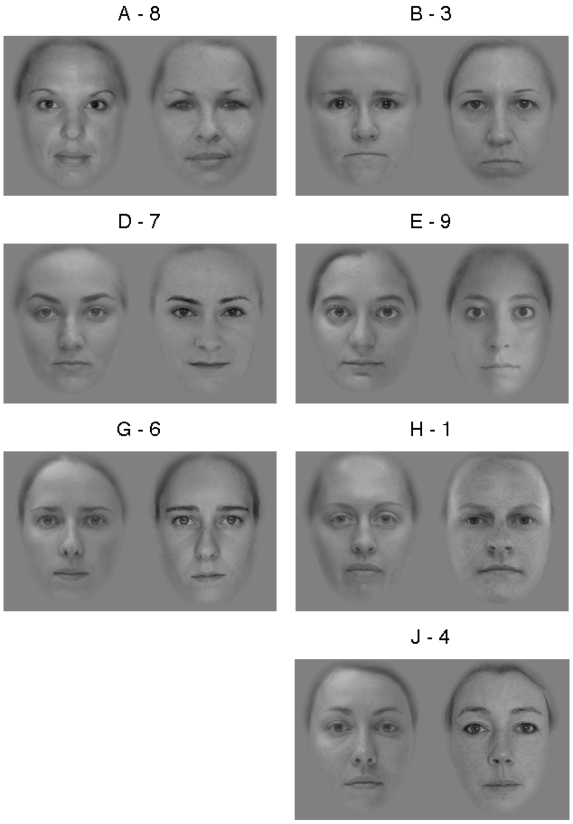


Figure 2.48: Answer key to Figure 2.3: subjects are shown with their corresponding sisters.

Chapter 3

Distinctive and personally familiar faces elicit more consistent patterns of neural activity across trials

In the last chapter, we showed that faces of those personally familiar to a viewer appear to look more distinctive than they actually are, as perceived by unfamiliar viewers. In this chapter, we further develop this relationship between distinctiveness and familiarity. We show that, in the brain, both are associated with increased cross-trial[1] neural pattern similarity (below, "pattern consistency", "neural consistency" or simply "consistency"). Among familiar faces, those which are especially distinctive to the viewer are more consistent than those which are not. Furthermore, individuals who rated an unfamiliar face as especially distinctive in the experiments in Chapter 2 are, as found in the experiments of this chapter (which began several weeks later), more likely to have consistent patterns of activation in response to it, in amygdala and hippocampus. Finally, we find that the distances between faces, as induced by spatially distributed patterns of activity, are positively correlated across subjects in visual cortex. Throughout, neural pattern consistency is shown to be a robust measure, independent of aggregate neural activity localized to the region in which the consistency is computed; consistency is enhanced even when local activation magnitude is decorrelated from it, and even when the activation in a region is slightly depressed.

Many authors have previously investigated familiar and distinctive faces and their representation in the brain. It has been shown that personally familiar faces elicit relatively enhanced neural activity in areas including amygdala [31], inferior parietal lobule, middle frontal gyrus, middle temporal gyrus, and supramarginal gyrus [72], that they are recognized more quickly [75], and that they are represented more invariantly in some face-related areas of the brain, in terms of being less

[1] "trial", here, is meant to encompass both individual experimental trials and averages of them within a single experimental session (more on this in Section 3.9.7)

sensitive to changes in face orientation [25]. Loffler et al. [60] showed that the distinctiveness of a synthetic/cartoon grayscale face, directly varied by computer manipulation, was positively correlated with magnitude of activation in the fusiform face area (FFA). Relatedly, Golby et al. [32] showed in a group of European and African American men that same-race faces are remembered more easily and elicit a greater magnitude of response in the fusiform face area.

However, neural pattern similarity, of the kind computed in this chapter, is a relatively new method of analysis, popularized by Kriegeskorte [56], and not many studies have employed it. Of those which have, the most relevant for this chapter is an experiment by Xue et al. [102], in which, using a subsequent memory paradigm developed by Brewer [11] and Wagner [100], neural pattern consistency was linked to memory encoding strength, tested 1 to 6 hours after scanning, for both faces and words, in areas including lateral occipital complex and inferior parietal lobule. In this chapter, we attempt to understand the origins of neural pattern consistency, and provide several lines of evidence suggesting that distinctiveness, neural pattern similarity/consistency, and memory are closely related in visual cognition.

3.1 Basic experimental setup and preliminaries

3.1.1 Participants and face stimuli

Participants and stimuli were are all identical to those described in Section 2.2, except with the addition of forty morphed faces, blended from the original forty “base” stimuli. This brings the total number of face stimuli to **80**. All results in this chapter are based on this set of 80 faces unless otherwise noted. The motivation for introducing an additional forty faces was twofold: (1) to increase the number of stimuli and therefore number of data points on which to perform statistical analyses, (2) to introduce a new set of faces, interleaved randomly with the base faces, which have lower distinctiveness on average (by construction), in order to increase the range of distinctiveness among the stimuli. The forty morph faces were constructed as follows: (1) ten of the morphs were 50% blends between a subject and her sister, (2) twenty nine of the morphs were 50% blends between two faces not personally familiar to any of the subjects, and chosen to range in distinctiveness, and the remaining one face stimulus was equally blended from all forty base faces, we here call it the “average face”. Figure 3.1 shows these additional forty faces used in this experiment (Figure 2.5 shows the other, “base” faces). We note that subjects only viewed inner faces for this experiment, and that the familiarity level of a morph was defined for a viewer as the average familiarity among

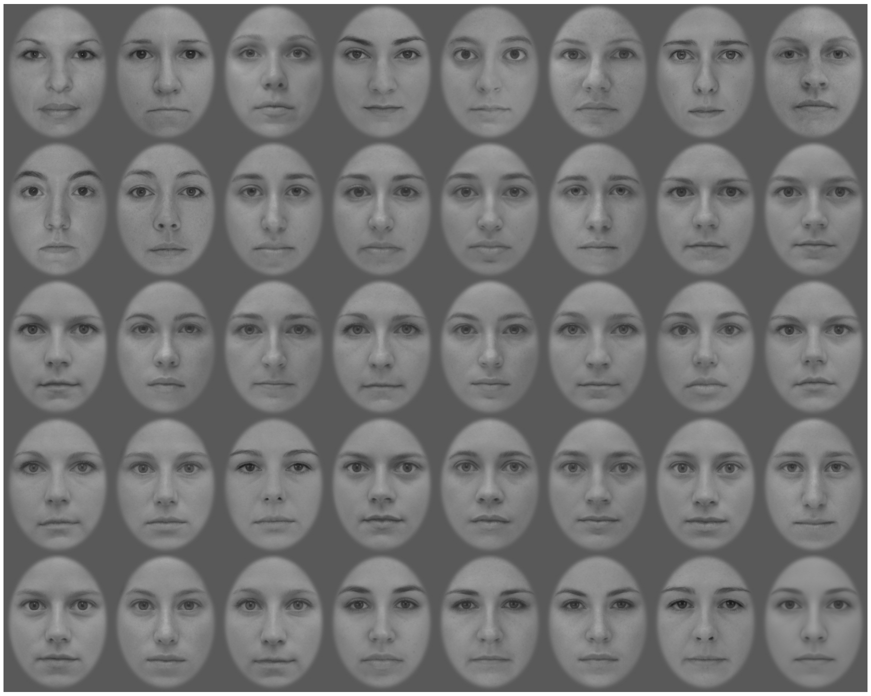


Figure 3.1: The forty morphed face stimuli used in the fMRI experiment (FaceView). The first ten (left-to-right, top-to-bottom) are the subject-sister morphs, the next twenty-nine are the unfamiliar faces morphs, and the one in the lower right is the morph of all forty base faces ("the average face").

the constituent faces.

3.1.2 Experimental Tasks

In the main experiment, which we call **FaceView** here, subjects viewed faces presented one at a time in an fMRI scanner (see Section 3.9.1 for details). After the presentation of each face, they were instructed to enter a 1 (least) to 6 (most) distinctiveness rating using keypad controllers in both hands (3 buttons per hand). Guidelines were given for distinctiveness prior to scanning as follows:

> *"What we mean by distinctiveness is: How unusual looking is this face? How much would this face stand out to you in a crowd? How different is this face from a normal face? It can be unusually attractive, or unattractive, or unusual in any other way. How striking is this face? Try to base your judgment on a holistic impression of the face."*

In order to minimize the effect of eye movements, subjects were instructed to fixate, faces were only up for 500 ms at a time, and stimuli were sufficiently small to mostly fit inside foveal and parafoveal vision (5°×6.7°). Each face in our stimulus set was viewed in two independent scanning sessions, occurring a median of 2.24 days apart, and multiple times (average 4) within a session, at an average delay of 2 minutes 47 seconds. Additionally, subjects participated in two “functional localizer” tasks, which were used to determine the location of face-selective clusters in their brains (see Section 3.9.4).

3.1.3 Categorical familiarity and distinctiveness

First, we note that, in the scanner, subjects rated faces which were personally familiar to them as more distinctive, on average, than viewers unfamiliar with those faces. If we take the definition of unfamiliar as familiarity level 0, and familiar as having familiarity > 7, the boost in distinctiveness is $+0.23$ standard deviations of ratings, and the population of ratings contrasts (familiar - unfamiliar), of which this is the mean, lies to the right of 0 by 1-sided t-test with $p < 2.5 \times 10^{-4}$. If we loosen the definition of unfamiliar as having familiarity level ≤ 3, and loosen the definition of familiar to ≥ 7, we find an average boost of $+0.13$ standard deviations with a significance of $p < 0.011$. These results from explicitly-reported distinctiveness values are in line with what we found earlier through an analysis of face pair ratings in Chapter 2.

In this chapter, much of our analysis will depend on categorizing faces into three disjoint sets (familiarity levels the same as defined in Chapter 2, Section 2.2.3):

1. **familiar** – these are defined as those faces for which the subject is familiar at a level of 7 or greater

2. **distinct (and unfamiliar)** – these are defined as faces having familiarity 3 or less, and being in the top 40% of distinctiveness ratings of all faces according to some distinctiveness metric

3. **indistinct (and unfamiliar)** – these are defined as faces having familiarity 3 or less, and being in the bottom 40% of distinctiveness ratings of all faces according to some distinctiveness metric

We modified the categorical familiarity thresholds relative to Chapter 2 (unfamiliar is $\leq 1 \rightarrow \leq 3$, familiar is $\geq 2 \rightarrow \geq 7$) because we wanted to restrict “familiar” to mean, at a minimum, good friend seen frequently, in order to emphasize the neural differences. Refer to Section 3.9.2 for an explanation of different distinctiveness measures explored. We note that, unless otherwise stated, the default distinctiveness measure used was **pixel+scanner**, a combination of a subject’s in-scanner distinctiveness rating of a face, together with a pixel-based measure of distinctiveness. This measure unifies the most high-level measure (explicitly provided rating) possible together with the most low-level measure possible. This measure is very highly correlated (≥ 0.84) with its two constituent components, and so results based on using one of the two base metrics are essentially the same.

Also, although not always stated, distinct and indistinct faces are always unfamiliar (familiarity ≤ 3) as described above.

3.1.4 Brief explanation of analyses used in this chapter

3.1.4.1 Two neural measures: magnitude and consistency

Throughout this chapter, we will extensively discuss two neural measures, each defined per subject, per brain region, per face stimulus:

1. **magnitude**: the aggregate (based on median) response magnitude across voxels in the region (see Section 3.9.6 for details)

2. **consistency**: a measure of similarity between the spatially distributed neural response pattern at one time, and that to the same stimulus at another time (see Section 3.9.7 for details)

These values are residualized for experimental artifacts such as the delay between trials and the variability in button presses. See Section 3.9.9 for details on residualization. Most importantly, the consistency values are always *magnitude-residualized*. This is because, if the overall level of activation within a region is enhanced, the SNR is higher at a constant noise level, and so measures of consistency would naturally be expected to increase as well. To eliminate this correlation we have in every case[2] regressed out response magnitudes from consistency measures within a region. Finally, these measures are always shown and referred to **Z-scored** relative to the set of 800 (subject, stimulus) values. The consistency measure used combines both cross-session consistency, and cross-trial consistency, which is based on estimates of face responses at individual face presentation trials within an fMRI scan. These two types of consistencies were correlated (see Supplementary Figure 3.39 and Equation 3.3), results were similar between them, and combining them provided us with more statistical power.

3.1.4.2 Two types of brain region exploration: ROI-wise and map-type

We perform two basic types of analyses in this chapter, which we term **ROI-wise** and **map-type**. Both rely on identifying a specific region within the brain, meaningfully defined across all subjects, and then pooling together magnitude and consistency values across subjects and faces in this region, and, for example, computing their average.

[2] except where noted otherwise, in a few select figures for which this was specifically undesirable

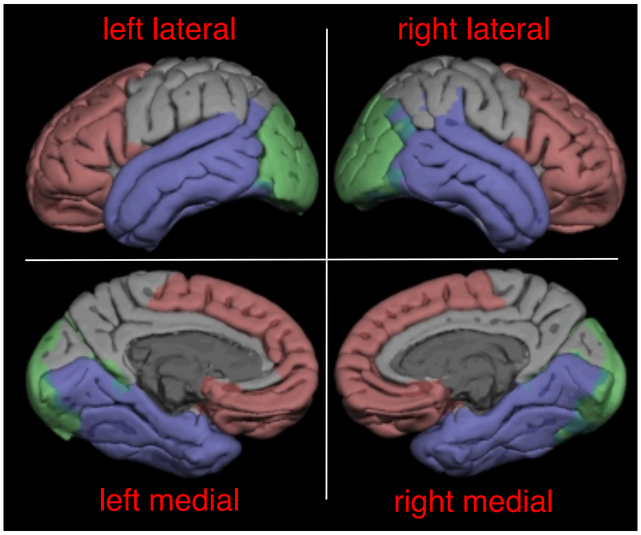


Figure 3.2: The MNI template brain (Montreal Neurological Institute, based on the work of Collins [16]) on which results are shown throughout the chapter. For reference, *occipital lobe is colored in green*, *temporal lobe in blue/lavender*, and *frontal lobe red*. The uncolored portions correspond to parietal lobe and limbic lobes. The lateral and medial views of each hemisphere of the brain are shown on corresponding sides of the figure. Note that in order to bring visual cortex (which extends along the inferior surface of the temporal lobe) into better view, all brains are actually rotated along the horizontal axis into the page by 30°, such that the topmost part of the lateral views is actually closer to inferior parietal lobe, and actual superior parietal lobe is tucked slightly out of view.

1. **ROI-wise** analyses are based on anatomically and functionally defined regions and are shown as bar plots. Only two types of functional regions are used: **FFA** and **Face**. FFA consists of face-selective clusters in the fusiform gyrus. "Face" consists of face-selective clusters all over the brain and including the fusiform gyrus. See Section 3.9.4 and Figure 3.35 for more information on these clusters and how they were localized.

2. **Map-type** analyses are based identifying a large collection (1078 total) of comparable regions (**"spheres"** of voxels) spanning the whole brain, then combining them all together onto into a single map, shown on a cortical surface like the one labeled in Figure 3.2.

ROI-wise analyses are advantageous because they allow us to focus on one specific region with clear function and anatomical boundaries. Map-type analyses are advantageous because they allow us to visualize how a particular effect varies smoothly as we move over the entire brain. Refer to Section 3.9.8 for details on how these are computed. The surface maps are always shown with the left hemisphere on the left, and the right hemisphere on the right, and rotated around to show both the lateral and medial surfaces.

3.1.4.3 Two kinds of statistical significance tests: empirical and t-test

For a specific region, and a particular measure (magnitude or consistency), we will compute two types of statistical significance values and will occasionally refer to one or the other or both depending on the context:

1. **empirical significance level** (denoted below p or p_{shuff}): the estimated probability of observing an effect size of equal or greater extremity, pooled across subjects, under a randomly shuffled mapping of values (magnitude or consistency) to stimuli (with an identical shuffling order for each subject). This is the preferred and default test used in this chapter (i.e., unless noted p values will derive from this test) as it leverages the large number of data points driving the trend and is suited for this style of experiment in which a very thorough analysis is carried out on a relatively small number of subjects (10).

2. **t-test significance** (denoted below p_t): the probability of observing an effect size of equal or greater extremity, based on the distribution of exactly and only the 10 individual subject values, using the standard 1-sided t-test based on sample mean and variance. Although this type of test is more well suited for an experiment (unlike our own) with a very large number of subjects and a relatively small number of data points per subject (as it collapses across them), it does provide us with another view of the statistical reliability of the results.

3.2 Personally familiar and distinct faces elicit more consistent neural patterns in visual cortex

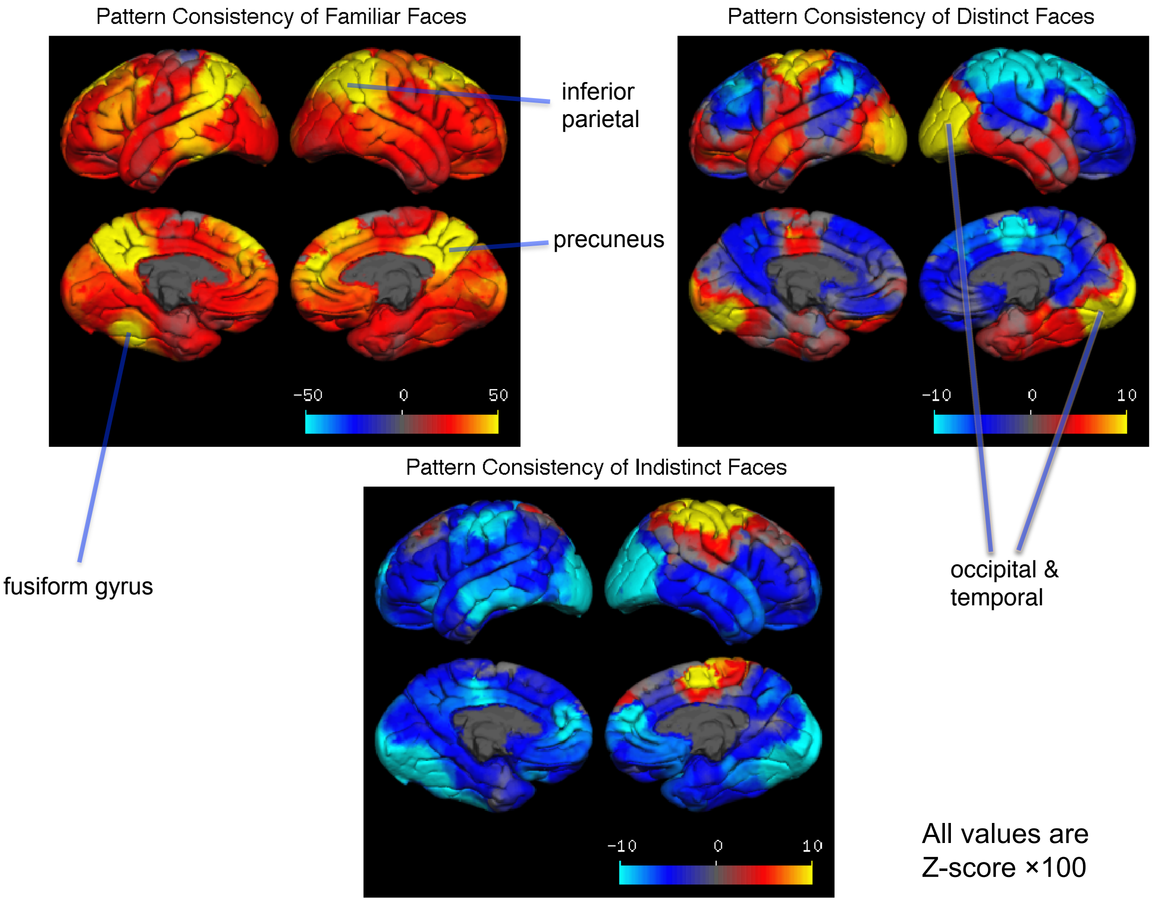

Pattern Consistency of Familiar Faces
Pattern Consistency of Distinct Faces
inferior parietal
precuneus
-50
0
50
-10
0
10
Pattern Consistency of Indistinct Faces
fusiform gyrus
occipital & temporal
-10
0
10
All values are Z-score ×100

Figure 3.3: All values are **Z-score**×**100**; middle of the range is zero (light blue is negative). *Top left:* Familiar faces are represented very consistently all throughout the brain (especially inferior parietal, precuneus, and fusiform gyrus); note scale bar goes to 50 (Z-score 0.5) *Top right:* Distinct faces are represented more consistently than average in occipital and temporal lobe; note scale bar goes to 8 (Z-score 0.10). *Bottom:* Indistinct faces are less consistent than average in occipital and temporal lobe. *Note that **left** sensorimotor cortex (seen clearly at the boundary of parietal and frontal lobes laterally, at the central sulcus) has high consistency for **distinct** faces, and **right** sensorimotor cortex has high consistency for **indistinct** faces, because the contralateral hands were used to enter high and low distinctiveness ratings respectively (see Section 3.9.11 for more careful explanation). This is purely an artifact of the experimental design and scientifically unimportant.*

In this section, we focus our analysis mainly on visual cortex, namely occipital lobe as a whole, LO (lateral occipital cortex), and V1/V2.

We begin with an analysis of the consistency of familiar faces, unfamiliar but distinct faces (below simply "distinct" or "distinctive"), and unfamiliar indistinct faces ("indistinct"). Inspecting Figure 3.3 reveals a simple trend: familiar > distinct > indistinct. Familiar faces are represented most consistently, distinct faces are more consistent than average, but less so than familiar faces, and indistinct faces are represented least consistently. It is important to understand that figures like Figure 3.3 represent results pooled across all subjects. For instance, the color of a point in the "Familiar Faces" brain shown in that figure represents the average consistency among the subset of the 800 (subject, face stimulus) pairs which satisfy the familiarity condition.

The contrast between distinct and indistinct faces is clearly seen in occipital and temporal areas, including fusiform gyrus. For a statistical map of pairwise comparisons between stimulus categories, see Figure 3.4. Averaging consistency over all occipital voxel spheres[3], we find that familiar faces ($c_{fam} = 0.31$) are more consistent than distinct faces ($c_{dct} = 0.070$: $c_{fam} > c_{dct}$ with $p < 0.0034$, $p_t < 0.011$) and indistinct faces ($c_{ind} = -0.10$, $c_{fam} > c_{ind}$ with $p < 3.1 \times 10^{-6}$, $p_t < 0.0022$), and distinct faces are more consistent than indistinct faces ($c_{dct} > c_{ind}$ with $p < 8.5 \times 10^{-5}$, $p_t < 0.034$). To compare consistency types, the analogous results based on *cross-session* consistency alone are: $c_{fam} > c_{dct}$ with $p < 0.011$, $c_{dct} > c_{ind}$ with $p < 5.0 \times 10^{-5}$, and the results based on *cross-trial* consistency alone are $c_{fam} > c_{dct}$ with $p < 0.0020$ and $c_{dct} > c_{ind}$ with $p < 0.11$. These are representative results, in the sense that the combined consistency metric provides effects which are directionally equivalent to the cross-trial and cross-session consistencies, but often with improved significance.

Consistency in early visual cortex in the absence of increased activation: Notably, whereas familiar faces are considerably more consistently represented than distinct faces in the occipital lobe, they do not activate this part of cortex more than distinct faces. The average magnitude of response across the occipital spheres for familiar faces is $m_{fam} = 0.029$, compared with an average magnitude of response to distinct faces of $m_{dct} = 0.082$: in fact familiar faces activate occipital cortex *less* than distinct faces (though not significantly), despite being much more consistently represented there. See Supplementary Figure 3.14 for the average magnitude of response to familiar, distinct, and indistinct faces, mapped across the entire brain. Enhancements in neural consistency in regions where there is no enhanced aggregate activation can also been seen in Figure 3.5. In

[3] voxel spheres based on (i.e., centered in or stepped from the center of) a region which is labeled by freesurfer as being occipital. See Section 3.9.8.

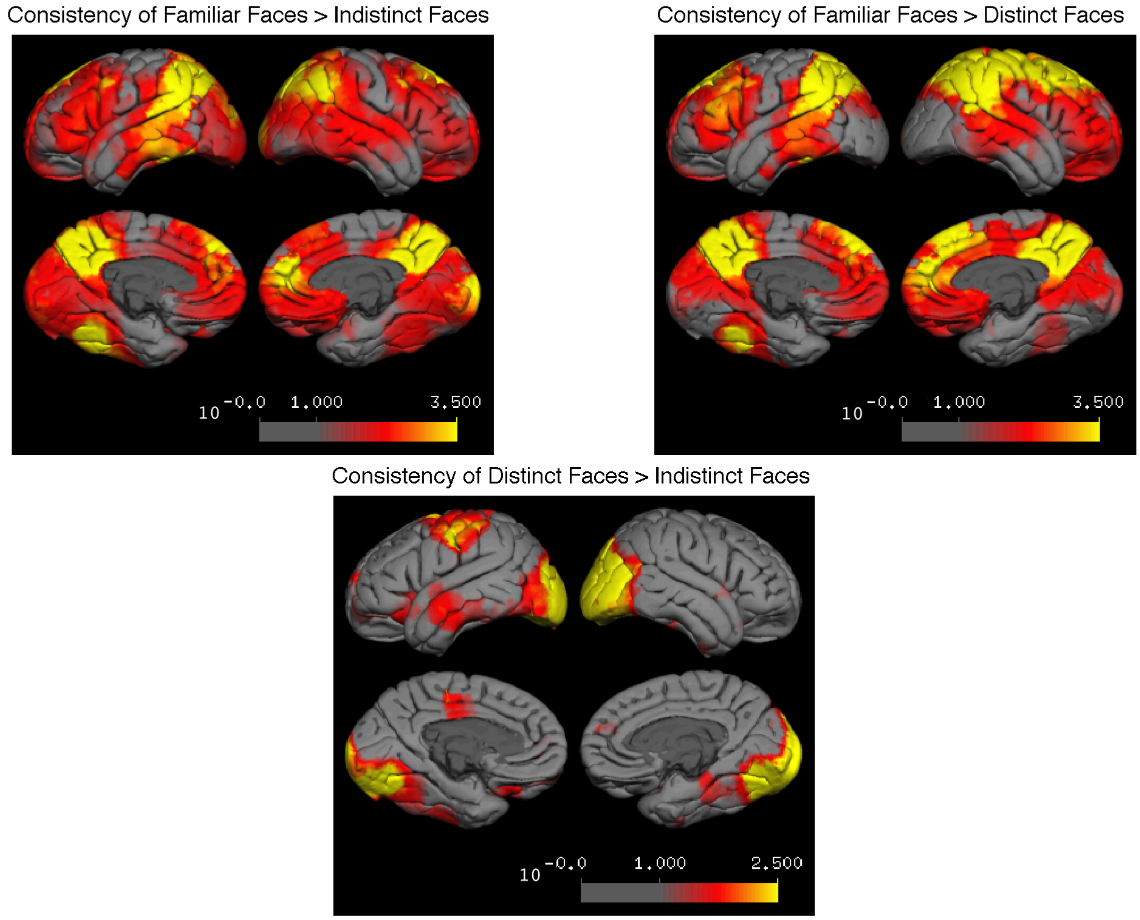

Consistency of Familiar Faces > Indistinct Faces
$10^{-0.0}$ 1.000 3.500
Consistency of Familiar Faces > Distinct Faces
$10^{-0.0}$ 1.000 3.500
Consistency of Distinct Faces > Indistinct Faces
$10^{-0.0}$ 1.000 2.500

Figure 3.4: Voxels which, on average over the spheres which contain it, have significantly different consistencies for a category of face stimulus (familiar, distinct, or indistinct) are highlighted. Note the region of statistical significance for the comparison Distinct > Indistinct in occipital and posterior temporal regions. We map the average empirical statistical significance values using $-log_{10}p$ scale.

V1/V2, distinct faces are more consistent that indistinct faces with $p < 0.001$, despite not having an enhanced response magnitude there. Also in V1/V2, familiar faces are more consistent than both distinct and indistinct faces with $p < .001$, and also despite showing no significant increase in activation magnitude. The same relatively-consistent-but-not-relatively-activated trend is present in LO, for familiar faces.

In line with Loffler's finding [60], we do find that activation magnitude is greater for distinctive faces in FFA, together with consistency. Interestingly, in Supplementary Figure 3.18, it is shown that compared to indistinct faces, distinct faces are relatively activated ($p < 0.01$) and consistent ($p < 0.1$) in hippocampus, suggesting that they may be better encoded into memory across trials or experimental sessions.

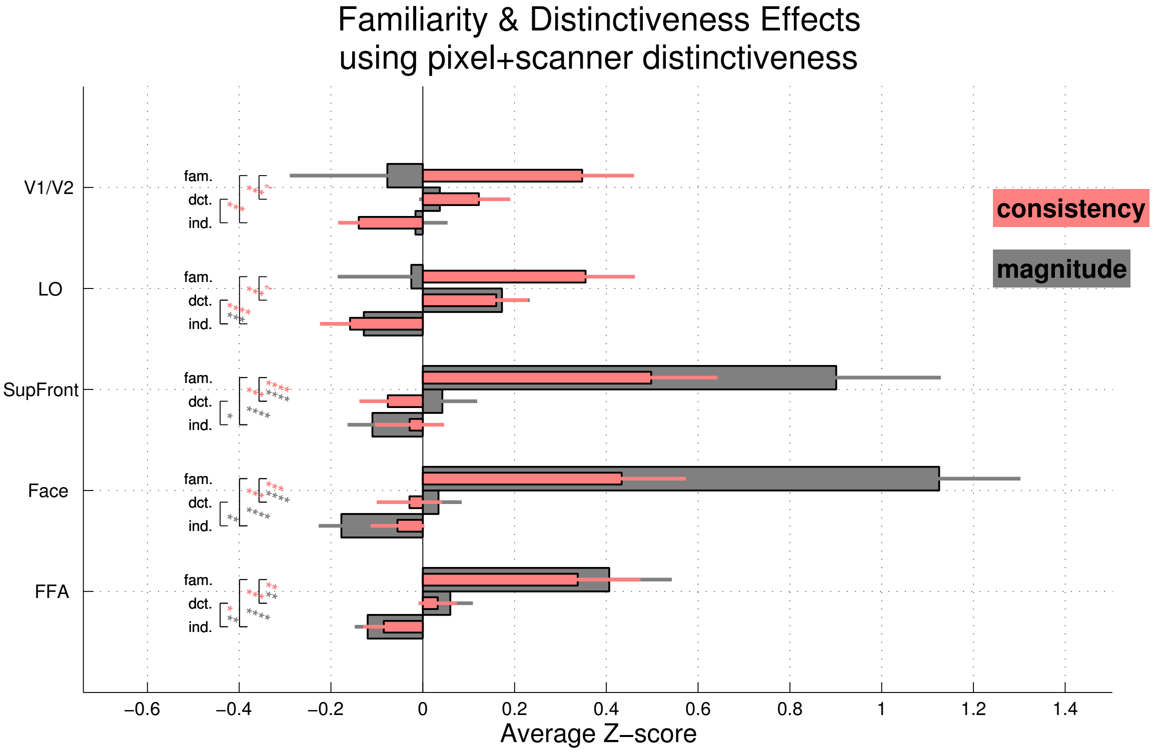


Figure 3.5: Average (across subjects and stimuli) response magnitude and consistency are shown for several regions of interest (SupFront is superior frontal lobe), including FFA and the union of all face selective clusters ("Face"). In each region, we show the average neural measure for each category of face stimulus: familiar, distinct, and indistinct. To provide a graphical representation of the variability across subjects, standard error bars are computed over the distribution of 10 individual subjects. The empirical significance of pairwise comparisons is indicated according to the convention described in Table 3.1 and used in all subsequent figures. Notice that familiar faces do not have enhanced activation magnitude (indicated with gray bars) in V1/V2, or LO (lateral occipital cortex), though they do have enhanced consistency in those regions. An extended version of this figure, with more ROIs, is available as Supplementary Figure 3.18.

symbol	p-value range
~	$(0.05, 0.1]$
*	$(0.01, 0.05]$
**	$(.001, .01]$
***	$(.0001, .001]$
****	$(0, .0001]$

Table 3.1: Significance values in bar plots will be denoted with the symbols above.

Familiar faces are generally very highly consistently represented in visual cortex and throughout the entire brain, showing a global enhancement in consistency of response but not a global enhancement in magnitude of response. Three regions are conspicuously more consistent for familiar faces: these are precuneus, inferior parietal cortex, and left fusiform gyrus. We first consider the effect in fusiform gyrus, which contains FFA. Of course, it is not surprising that this is a region of especially high effect size as it is a region specifically responsible for encoding faces. We find that, in left FFA, familiar faces have an average consistency value of $c_{fam} = 0.44$, whereas distinct faces have $c_{dct} = 0.014$, and indistinct faces have $c_{ind} = -0.10$. The average consistency values in the right hemisphere FFA are 0.24, 0.052, and -0.067 for familiar, distinct, and indistinct faces, respectively. The fact that familiar faces are ones which evoke memories of specific names, involving language, and that language is a left-hemisphere-dominant function, may explain why this effect of consistency in FFA is left-dominant: unresidualized[4] consistencies for familiar faces in FFA are greater in left than right hemisphere with $p_t < 0.087$ over the set of 10 subject differences. Language areas may preferentially feedback to visual areas within the same hemisphere, to reinforce a more consistent representation. We will revisit this in Section 3.3, below. There is no clear laterality effect in activation magnitude of familiar faces (Supplementary Figure 3.14).

Next, we consider precuneus, which is both relatively activated and relatively consistent bilaterally for familiar faces. The precuneus has been implicated in a wide variety of high-level processes, including self-consciousness and episodic memory [14]. Taylor et al. [89] reported an increased activation in precuneus for self-face and partner's face relative to baseline neural activity. Although we do not find that it is especially strongly activated for self-face stimuli as compared with other types of familiar stimuli (see Supplementary Figure 3.25), familiar faces do certainly evoke more memories and require more self-reflection (e.g., "this is *my* sister" or "this is *my* friend") than totally unfamiliar faces, so the strong response there is compatible with earlier studies.

Finally, we consider inferior parietal cortex (which we term **IPL** below for inferior parietal lobe, and including inferior parietal lobule). As with precuneus, it is both relatively activated and relatively consistent bilaterally for familiar faces. This brain region has long been associated with the encoding of faces [39], including specifically familiar ones [72], and a recent study by Radua et al. [73] found that, in normal adults, activity in IPL was significantly correlated with the emotional state conveyed in the eyes of a face stimulus. We found in Chapter 2 that subjects are likely more sensitive to the emotional state of familiar faces. These results are mutually self-consistent and in

[4] the artifacts normally residualized out affect both hemispheres identically; residualization in this case diminishes some signal and leaves $p_t < 0.19$ (effect has same direction/sign).

line with the idea that familiar face processing involves IPL in particular.

3.2.1 Non-categorical distinctiveness and consistency

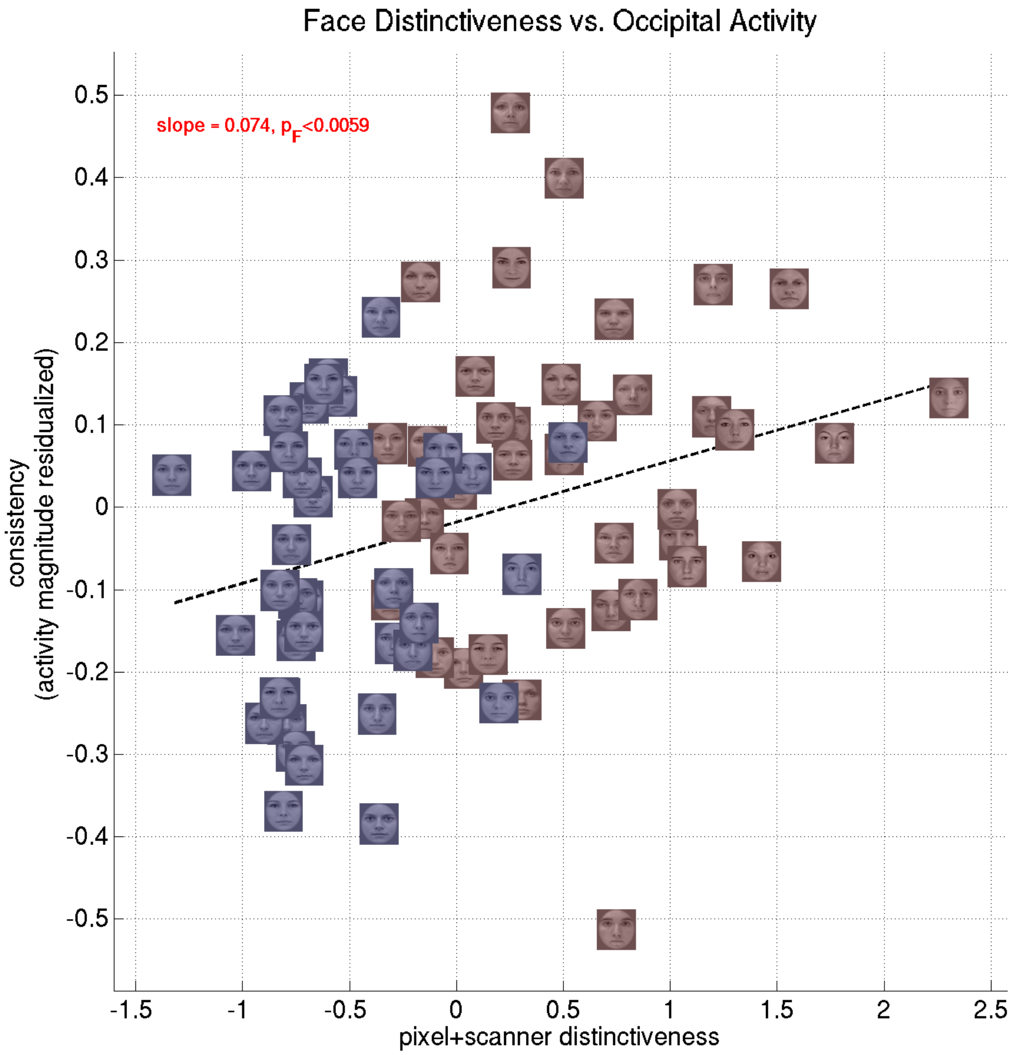


Figure 3.6: The subject-averaged (excluding familiar viewers) neural pattern consistency in occipital lobe (i.e., averaged across occipital spheres) is plotted against the corresponding subject-averaged distinctiveness for each of the 80 face stimuli. Morphs are shaded blue and base faces are shaded red (base faces are more distinctive). Consistency can be linearly modeled with distinctiveness, yielding a goodness-of-fit F-test significance level of $p_F < 0.0059$.

To test whether the greater consistency of distinct faces was a purely categorical effect, only significant when comparing the top 40% to the bottom 40% of unfamiliar faces, as in the analyses above, or whether it was an effect sufficiently reliable to fit in a linear relationship with individual distinctiveness values, we linearly modeled the subject-averaged consistency of a face, averaged across **occipital** spheres, with the subject-average distinctiveness (excluding pairs of subjects and stimuli which met the familiarity condition). The result is shown in Figure 3.6. We find that indeed, with a regression F-test based significance level of $p_F < 0.0059$ the average neural consistency of face can be related to its distinctiveness. The equivalent relationship between distinctiveness and the activation magnitude, also in occipital regions, is weaker ($p_F < 0.041$, see Supplementary Figure 3.22). This *intermediate* effect size is compatible with the results reported earlier, namely the combination of (i) an absence of magnitude effect of distinctiveness in V1/V2, and (ii) the presence of such an effect in LO: that is, two regions both in the occipital lobe, but with a range of sensitivities. We also find this relationship between consistency in fusiform gyrus (i.e., averaged across spheres in the fusiform gyrus) and distinctiveness, with $p_F < 0.0038$ (see Supplementary Figure 3.23). However, this trend evaporates once we exit visual cortex: e.g., in the frontal lobe, there is no positive correlation at all and the significance drops to $p_F < 0.78$ (see Supplementary Figure 3.24).

3.3 Activity magnitude and occipital consistency

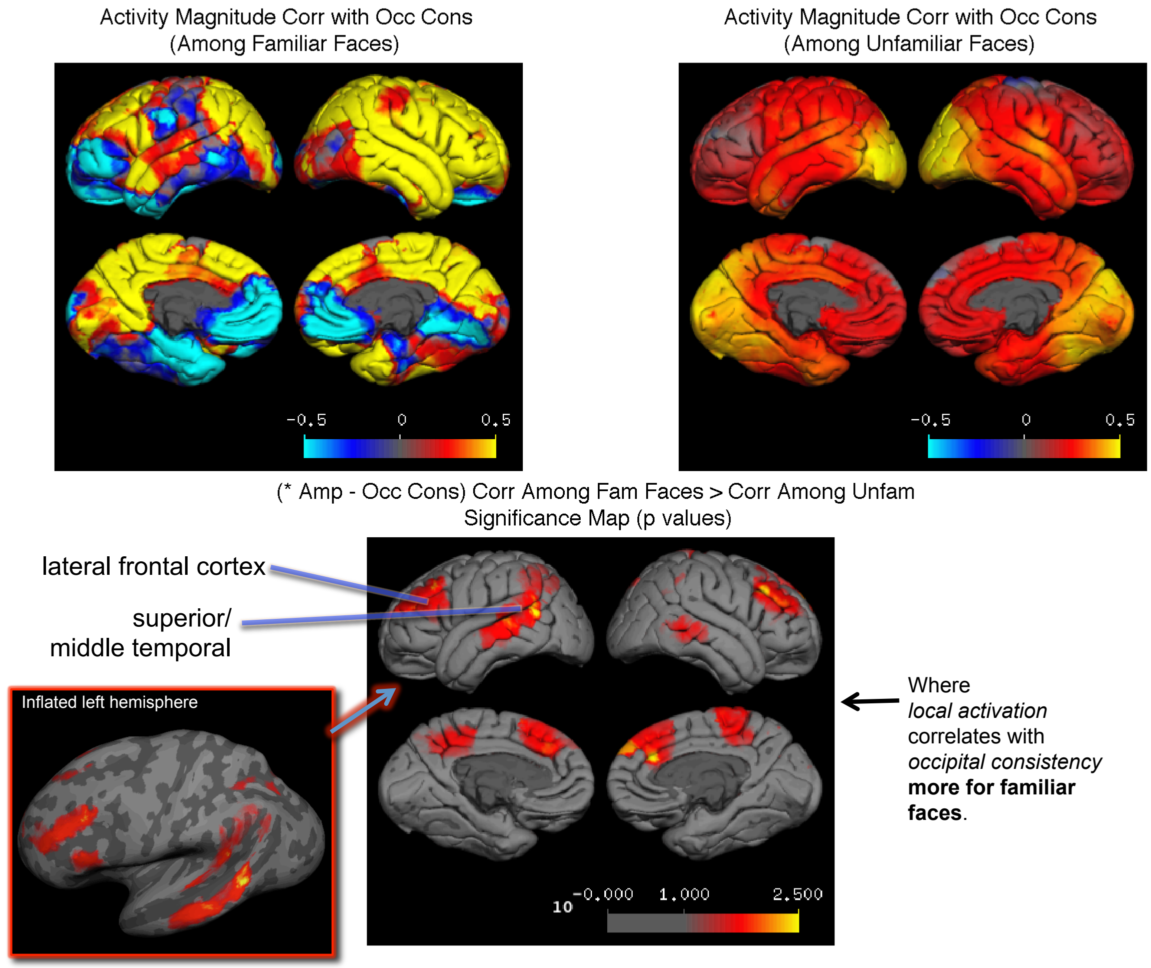


Figure 3.7: *Top left:* Subject-averaged correlation between occipital consistency (averaged across occipital spheres) and local activation magnitude among faces familiar to a viewer, $corr_{fam}$. *Top right:* Subject-averaged correlation between occipital consistency and local activation magnitude among faces unfamiliar to a viewer, $corr_{unfam}$. *Bottom:* The significance (in $-log_{10}p$ scale) of a 1-sided t-test, on the set of 10 individual subject value pairs, that $corr_{fam} > corr_{unfam}$; voxel significance values are averaged over containing spheres as before.

Guided by the intuition that the enhanced consistency in visual cortex of familiar and distinctive faces may be caused by enhanced activation magnitude in some secondary regions feeding back into visual cortex and reinforcing previously formed representations, we created a whole brain map of the correlation between local activation magnitude, across different regions, and occipital neural pattern consistency (i.e., averaged across occipital spheres).

We computed the correlation separately for each hemisphere. The result is shown in Figure 3.7. The analysis was done twice, once over each of two disjoint sets of stimuli (including morphs) for each subject: (1) familiar, i.e. having familiarity ≥ 7, and (2) unfamiliar, i.e. having familiarity ≤ 3. The correlation was computed using each subject's data separately, then averaged over subjects. Finally, we compared the resulting correlation maps to see where they differed significantly. Not shown in Figure 3.7 is the result that no areas, on average, were significantly *more* correlated with occipital activity for *unfamiliar* faces than for *familiar* faces. Although it may appear that, for instance, left temporal pole is anti-correlated among familiar faces whereas it is positively correlated for unfamiliar faces, the average value maps do not show the variance in the data, and in fact there is no significant difference there.

There are several regions in which the magnitude correlation with occipital consistency is greater for familiar faces than for unfamiliar faces. These are seen in the bottom panel of Figure 3.7. In the left hemisphere, these regions, including middle temporal gyrus, occur very close to and overlap with the two areas of the brain associated with language: Wernicke's area (in superior temporal gyrus), associated with the understanding of language, and Broca's area (in inferior lateral frontal cortex), associated with its production and speech. Because, unlike unfamiliar faces, familiar faces have names associated with them, the visual perception of familiar faces may evoke some language processing, e.g., the viewer may recall the name of the person whose face they're viewing, specific words exchanged with that person, or otherwise have internal thoughts requiring linguistic representation. The results suggest that when viewing a set of familiar faces, the relatively *increased linguistic thought (such as name recollection)* may strengthen the consistency of the *visual* representation in the brain. On the right hemisphere, we find significantly greater correlations for familiar faces in the homologous areas found in the left hemisphere, but with diminished extent. Medially, we also find activity correlates in precuneus and middle frontal lobe.

Laterality of familiar faces: It should noted that familiar faces activate left languages areas more than right language areas, as expected: the magnitude of activation, among familiar faces, in Broca's area (i.e. BA44/45, left) was greater than the activation in the right hemisphere homologue with $p_t < 0.037$, using the 10 subject inter-hemispheric differences [5]. This left-laterality for familiar faces activation magnitude was also found in IPL (close to Wernicke's area), hippocampus, cingulate cortex, precuneus, and STS, each with $p_t < 0.028$.

For a detailed exploration of how neural pattern consistency in each parcellated anatomical region

[5] based on unresidualized magnitudes, as before for the laterality of FFA, because the normally-residualized artifacts affect both hemispheres identically.

correlates with the consistency or magnitude of another, across faces, refer to Appendix, Section 3.8.5.

3.4 Among familiar faces, more distinct ones are represented more consistently

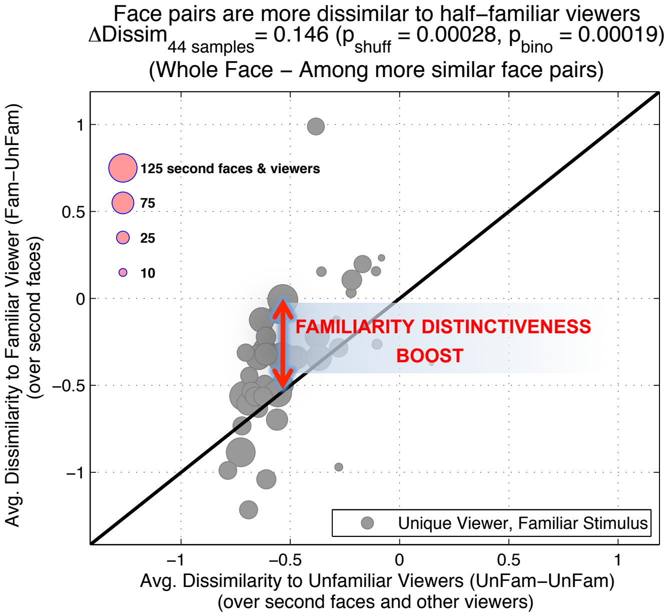


Figure 3.8: The familiarity distinctiveness boost is defined as the extent to which a face appears more dissimilar than it does to a set of unfamiliar viewers. This can be visualized in the kind of figures we showed in Chapter 2 (adapted from Figure 2.14).

Having established that, as category, familiar faces are represented more consistently than distinct but unfamiliar faces, we next investigated whether, among only familiar faces (familiarity level ≥ 7, only 36 out of 800 (subject, stimulus) pairs qualify – we ignore morphs for this section), those which showed the largest distinctiveness boost in Chapter 2 were more consistently represented than those which showed a smaller boost. The result is shown in Figure 3.9. The methodology for computing familiarity distinctiveness boost is described briefly in Figure 3.8 and described at the end of this subsection in more detail. We find that, indeed, across the 36 familiar stimuli, there is a strong positive and empirically statistically significant correlation across a wide variety of brain regions. In the figure, we show the results from 11 out of the 15 regions tried; we show the ones which are significant (all positive). Not shown are IT, hippocampus, amygdala, and FFA – none of which have

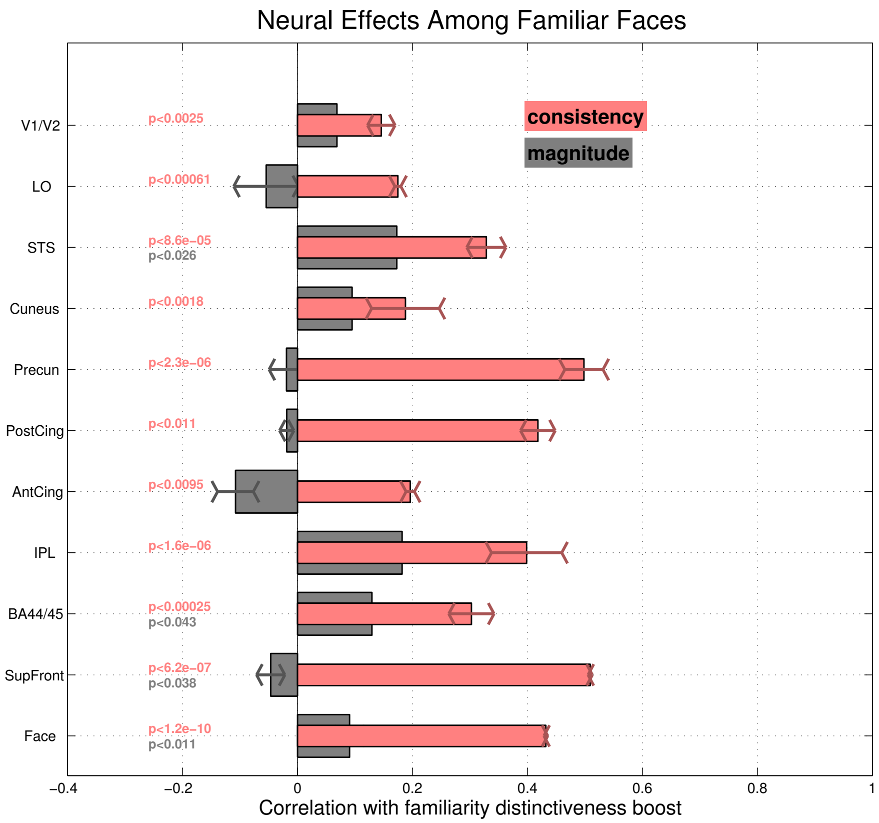


Figure 3.9: The correlation (across the 36 familiar base faces) between familiarity distinctiveness boost and the neural measures is shown. Whereas neural consistency is highly correlated across a variety of regions, the activation magnitude is not. Empirical significance values are provided, based on randomly shuffling the mapping of familiar stimulus to neural measure, for each subject independently (different faces are familiar to different viewers). *Notes: 1. STS is the superior temporal sulcus; 2. Cingulate cortex is divided into posterior and anterior portions (PostCing, AntCing); 3. BA44/45 is also known as Broca's area, linked to speech/language production in the left hemisphere; 4. To provide an additional measure of the reliability of the results, the effect size from each hemisphere is indicated in addition to their average, with* $<$ *and* $>$ *for left and right respectively.*

significantly positive or negative trends.

Even under the extremely conservative null hypothesis that an attempted region would have an empirically significant effect with probability 0.1 (and ignoring the actual *extent* of significance in each region beyond this level), the probability of having 11 or more significant regions out of 15 corresponds to $p_{pooled} < 9.3 \times 10^{-9}$. In contrast, among the 15 regions tried, only 3 showed a significant ($p < .05$) correlation between *neural activation magnitude* and familiarity distinctiveness boost, and of those, only 1 was positively correlated. This is one line of evidence which suggests that the distinctiveness of a face, at the neural level, is somehow more strongly associated with the consistency of its representation than its relative activation level.

Next, we tested whether this relationship between distinctiveness boost and consistency was strong enough to survive even at the single-subject level. That is, we have already seen that across subjects and stimuli, i.e. across all 36 familiar faces, there is a significant effect. But would the effect hold even among the handful of familiar faces per subject? We expect the results to be weaker, since subjects are each only familiar with a few face stimuli (25th, 50th, and 75th percentiles of number of familiar faces per subject is 2, 3.5, and 5). We show the results in Figure 3.10. We find that, averaged across subjects, but computed for each independently, the correlation between familiarity distinctiveness boost and neural pattern consistency tends to be positive, but not significantly so by 1-sided t-test among the 10 subject values in any region except IPL (inferior parietal lobe), in which we find $p_t < 0.034$, and in which we previously found an especially strong effect of magnitude and consistency for familiar faces. Considering all 15 interrogated regions together (including the 11 shown in Figure 3.10), we find that the distribution of their correlations with consistency is statistically significantly positive ($avecorr_{consistency} = 0.15$ with $p_t < 0.0012$) and so is their distribution of correlations with activation magnitude ($avecorr_{magnitude} = 0.12$ with $p_t < 0.0047$), though slightly less so.

To put these results in context, it helps to describe some similar analyses which failed; we will describe three. First, one can derive a similar familiarity distinctiveness boost based on the in-scanner ratings: simply take the difference between the familiar viewer's rating and the average among the unfamiliar viewers. Using this definition fails to produce a similar result, viz., a correlation between this in-scanner distinctiveness boost and neural pattern consistency (same for absolute in-scanner ratings among familiar faces). Second, if we simply use level of familiarity (7, 8, 9, or 10), instead of the familiarity distinctiveness boost, we again fail to find an effect. Lastly, if instead of taking the familiarity distinctiveness *boost*, which is a comparison value between subjects, we instead use the *absolute* distinctiveness value (inferred from MDS on the ratings in SimRate, see Section 3.9.2) and correlate that with consistency among the familiar faces, we fail yet again to find an effect.

The failure of the scanner-ratings method may be related to the following observation: it may be unusual or confusing for a subject to capture with a single number how distinctive they actually find a face, especially a highly familiar one; in contrast, using the experiments of Chapter 2, we can infer this from careful analysis involving comparisons with many other faces. Similarly, the familiarity level is a single and somewhat arbitrary number (e.g., the difference between degree 7 and 8; subjects not guaranteed to see sisters (9) more frequently than friends (8), etc.), unrelated to

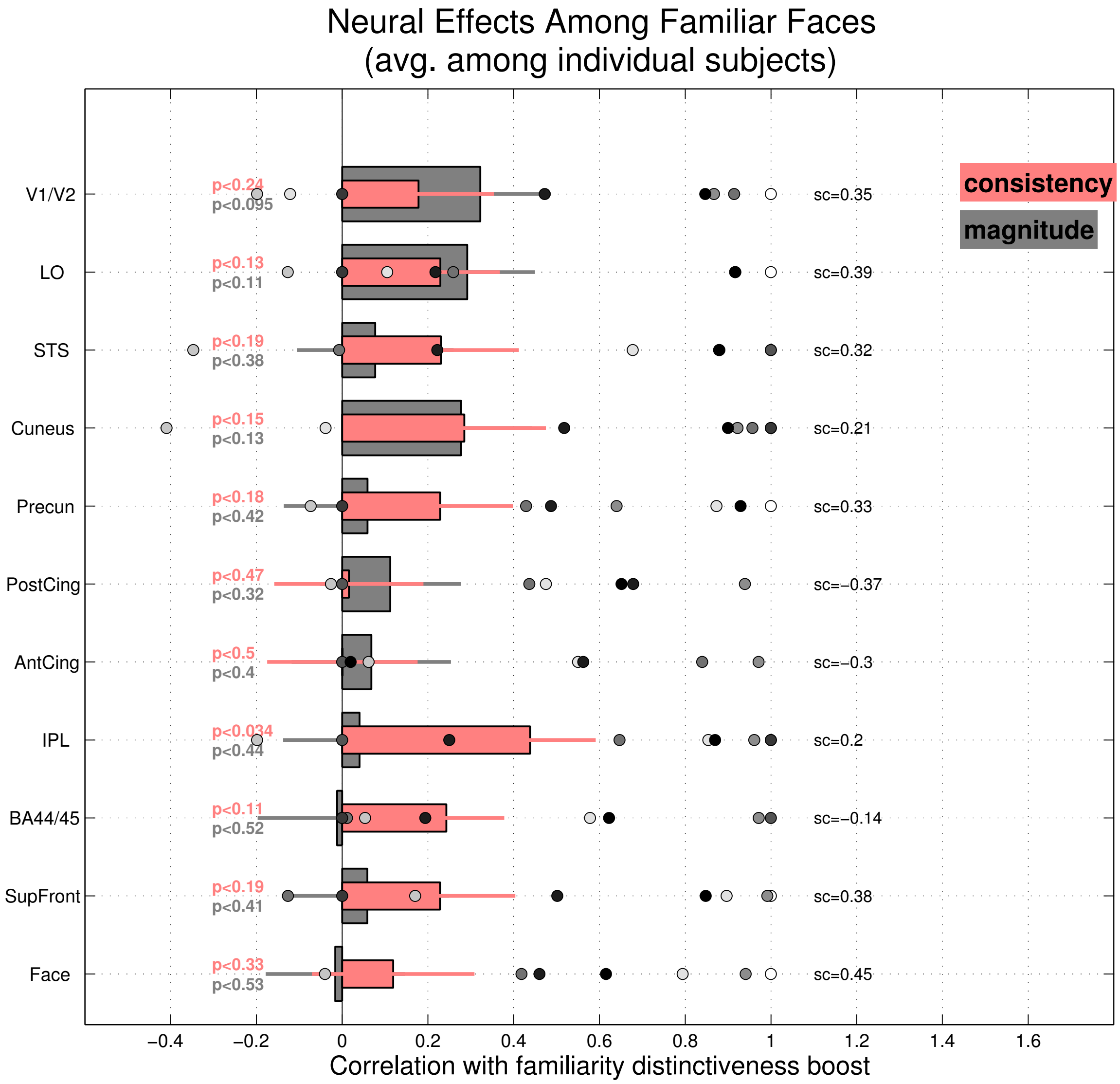


Figure 3.10: The subject-averaged correlation (across the handful of familiar faces per subject) between familiarity distinctiveness boost and the neural measures is shown, together with the standard error among the 10 subjects, and the result of a 1-sided t-test to the right of zero. We also plot the 10 individual subject values (some are cropped off to the left for convenience), colored by their percentile-rank for average magnitude, among familiar faces, of familiarity distinctiveness boost (white highest; black lowest). The correlation between (a) the average distinctiveness boost for the subject, and (b) the correlation with neural consistency, is written as a value "sc" for (subject correlate). We find a positive correlation between this tendency to have a large distinctiveness boost and tendency to have the boost relate to consistency in most regions (11/15 regions total, 7/10 shown here).

any direct perceptual evidence, whereas the distinctiveness boost is based on a thorough perceptual study, so we expect that the perceptual data correlate better with the brain data. Although we did find in Chapter 2 that greater familiarity was associated with a greater distinctiveness boost, this was only on average: the perceptual measurement is still a much stronger correlate of the brain data. Lastly, the failure of the absolute distinctiveness measure may arise from the fact all familiar faces were on average more distinctive to their viewer, and so subtle differences at the high, compressed, end of the distinctiveness range are effectively more noisy. By using the relative, distinctiveness boost, measurement, we add information to this measure, allowing us to distinguish between faces which are distinctive looking to everyone and those which are especially distinctive looking specifically to the subject – so it is a more sensitive measure.

Methods notes for this section We defined the familiarity distinctiveness boost using the SimRate ratings on whole faces, restricted to those pairs having dissimilarity less than -0.25 (as in Chapter 2, because highly dissimilar pairs are unaffected by familiarity), and qualified unfamiliar viewers as having familiarity ≤ 3, in line with the other analyses in this chapter. The boost was defined as the average dissimilarity boost for a stimulus by a subject, averaged over unfamiliar face comparisons and contrasting unfamiliar viewers (i.e., the distance from the diagonal indicated in Figure 3.8).

3.5 Neural pattern consistency and individual differences in distinctiveness perception among unfamiliar faces

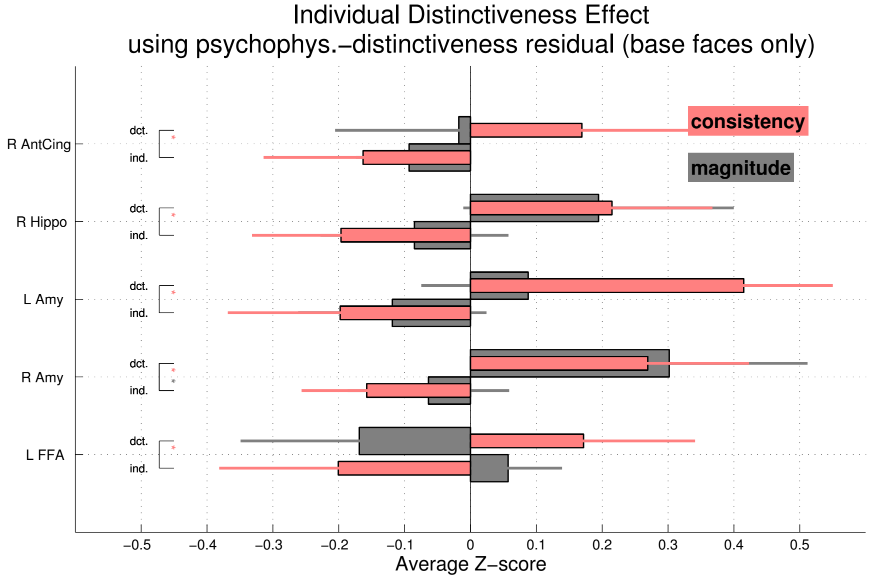


Figure 3.11: We compare subjects who found an unfamiliar face more distinctive to subjects who found the same face indistinct. The subjects who found the face more distinctive represent it more consistently in amygdala, bilaterally, and in right hippocampus. *Notes: For a particular face, only the subject finding it most distinctive among the 10 subjects contributes neural data to the "distinct" category here. The two subjects finding the same face least distinctive among the 10 subjects contribute neural data to the "indistinct" category. Bars represent the average and standard error, across subjects (each computed relative to the face stimuli they were qualified for). Significance tests are 1-sided t-tests to the right among the population of 10 subject values. Refer to Section 3.5.1 for methods details.*

Whereas in the last section we focused on distinctiveness among familiar faces, we now return to distinctiveness among unfamiliar faces. In the first main results section of this chapter, Section 3.2, the analysis was such that the selected distinct stimuli, and also the selected indistinct stimuli, were often the same for each subject (differences due to variable scanner ratings and unfamiliarity). We find that, averaged over unique subject pairs, the stimulus sets overlapped by 54% on average for distinct faces (mostly base faces), and by 69% on average for indistinct faces (mostly morphs), where we define overlap as $|A \cap B|/|A \cup B|$. Viewed from the other direction, but illustrating the same point, we find that 66% of the stimuli were selected into the distinctive category by either 1 or fewer, or 9 or more, subjects' data, i.e. most stimuli were nearly unanimously distinctive or not, and the same calculation reveals that 75% of stimuli were chosen into the indistinct category in such near unanimity. One disadvantage of subjects' having roughly the same (largely overlapping) sets of stimuli in each distinctiveness category is that it hides/obscures the individual differences in neural measures which may arise when a stimulus flips distinctiveness categorization from one subject to another, because the effects may be dominated by the stimuli which do not (and they are).

In this section, we examine the neural correlates of such individual differences in the perception of unfamiliar faces by performing an analysis which eliminates this overlap. Each stimulus contributes in a balanced way to the distinct and indistinct categories: the subject who found it most distinctive has her neural data contribute to the distinctive measure bar, and the two subjects who found it least distinctive have their data contribute to the indistinct bar. We found this method gave us most statistical power, by excluding subjects who were nearer the consensus/average judgment of distinctiveness. We note that distinctiveness here is based on the dissimilarity ratings of Chapter 2, and normalized, in the way distinctiveness boost was, by the distinctiveness ratings of others, and only using the 40 "base" faces studied in Chapter 2. Refer to Section 3.5.1 below for more methods details. The results are shown in Figure 3.11. The result is that, of 30 regions tried (the same 15 as before interrogated independently by hemisphere), 5 showed a significant increase in neural measure (magnitude or consistency), with all 5 showing a significant increase in consistency, and one of those (right amygdala) also showing an increase in magnitude. The pooled significance of having 5 out of 30 comparisons yield significance $p_t < .05$ is $p_{pooled} < 0.016$. Only 1 area showed a significantly lower signal, but this does not survive the multiple comparisons correction. We find a significant increase in consistency in amygdala bilaterally, right hippocampus, right anterior cingulate cortex, and left FFA. We will summarize this as *limbic* areas and left FFA.

To understand this result, we must recall that the distinctiveness measure used here is based on data collected in experiments described in Chapter 2, and which were conducted several weeks prior to subjects' brain scans. So, the proper interpretation is that a subject who finds a stimulus distinctive over a series of experiments at time t_0 will have more consistent responses in time t_1, several weeks after t in these limbic areas and left FFA. It will also help our understanding to quickly review some of the functions of these limbic areas: It is well known that the hippocampus is heavily involved in memory formation [86]. It has also been shown to be specifically involved in the recognition of faces [26] and similar high-level visual categories [54]. The amygdala has been implicated in a wide variety of functions including face recognition [44, 58], the recognition of emotion in facial expressions [1], the visual recognition of fear [2], and memory consolidation following emotionally charged events [81]. The anterior cingulate cortex has also been found to be involved in the processing of faces [39]. Taken together, a reasonable interpretation is that consistent representation in these limbic areas at t_1 for those who found a face distinctive at t_0 may result from an enhanced memory of the face from the earlier psychophysical experiments of Chapter 2. At t_1, those faces may be in the early stages of being transferred to the subject's long-term memory.

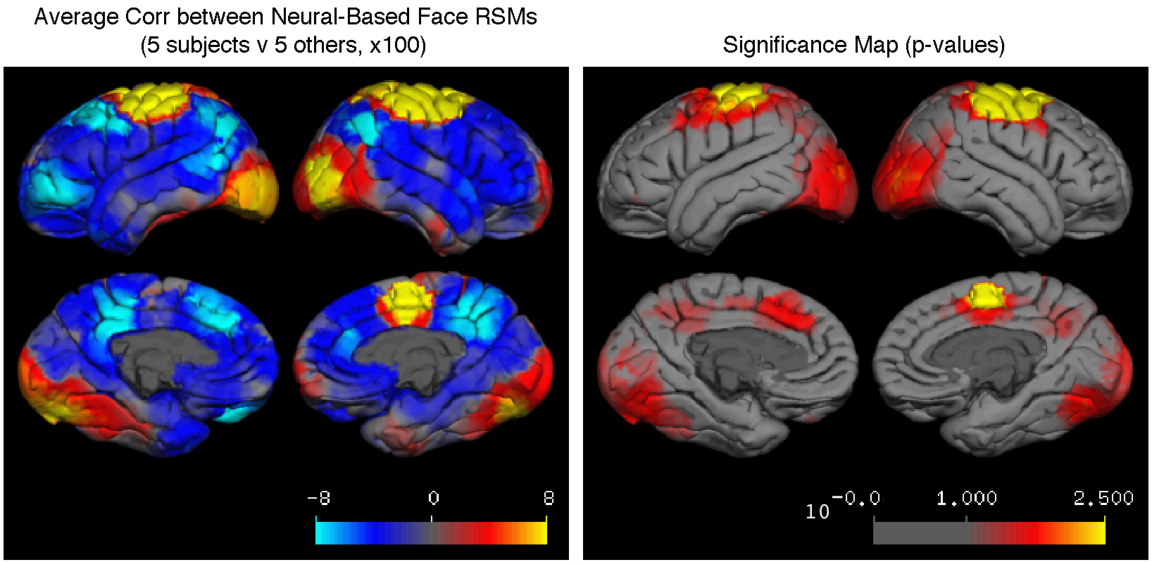
Repeating the exact same analysis methodology, except with the simultaneously recorded in-scanner distinctiveness ratings, reveals that, compared to subjects who found a face indistinct, those who found it especially distinctive in-scanner have significantly greater activation magnitude in 17/30 regions (with all individual $p_t < 0.1$, the pooled significance is $p_{pooled} < 3.3 \times 10^{-10}$), and decreased activation or consistency nowhere (0/30). We see a minor consistency enhancement in left hippocampus ($p_t < 0.075$), but it is weak and so possibly spurious. We can conclude that, in the scanner, the instantaneous perception of greater distinctiveness than other unfamiliar viewers resulted from greater global brain activation, probably with increased attention.

3.5.1 Methods notes for this section

Separating "distinct" from "indistinct" while balancing subjects in each category for each stimulus: A (subject, stimulus) pair qualified for the distinct condition if the subject was unfamiliar (≤ 3) and found the stimulus more distinctive than the other 9 subjects. A (subject, stimulus) pair qualified for the indistinct condition if the subject was unfamiliar and found the stimulus less distinctive than 8 or 9 of the other subjects. Of the 40 possible pairs which could have satisfied the distinct condition, ignoring familiarity, 30 qualified. Of the 80 possible pairs which could have satisfied the indistinct condition, ignoring familiarity, 75 qualified. The number of stimuli used per subject in the distinct case was [2 5 1 2 2 3 3 3 3 6] and the number per subject in the indistinct case was [14 7 9 6 8 8 4 4 8 7].

Distinctiveness in this case was based on subtracting, from the individual's psychophys.-based distinctiveness (see Section 3.9.2) , the mean of the other 9 subjects' (hence residual in "psychophys.-distinctiveness residual" in the title of Figure 3.11), to find out how much more less distinctive they found the stimulus than the others. This is analogous to the subtraction performed in computing the familiarity distinctiveness boost in the last section. Finally, we Z-score each subject's residual relative to the other stimuli.

3.6 Inter-subject neural pattern similarity in face representation


Average Corr between Neural-Based Face RSMs (5 subjects v 5 others, x100)

Significance Map (p-values)

Figure 3.12: *Left:* Values are correlation×100 (so value 8 is correlation .08) The distances (equivalently here, neural pattern similarities) between faces induced from neural activity correlate weakly but significantly across subjects in parts of visual cortex. Notice red and yellow streaks of positive correlation across parts of temporal and occipital cortex. *Right:* Average significance values of correlation (null hypothesis: correlation is among independent normal variates). Neural RSMs are first averaged over 5 random subjects, and then over the 5 remaining. The resulting two RSMs have a peak average correlation of 0.11 in lateral occipital cortex, and 0.095 in fusiform gyrus (both right hemisphere). Note that sensorimotor cortex correlates highly across subjects because subjects tended to rate faces on the same half of the distinctiveness scale, so with the same hand (see 3.6.1 for details) .

Although we have so far focused on ways in which subjects are *different* in their perception of and neural responses to faces, when they vary in familiarity or distinctiveness, subjects are nonetheless more alike than they are different in these behavioral and neural measures. In the first section of the supplement to Chapter 2 (Section 2.9.1), we found that subjects are remarkably consistent in rating the differences between faces: any one subject's behavior in a task was shown to be very well modeled by the average of the others'. In the beginning of this chapter, we reported that in-scanner distinctiveness ratings diverge for familiar faces; however, subjects are very consistent in their distinctiveness judgments otherwise. The average correlation among subject-pair-wise in-scanner distinctiveness ratings is 0.57 ± 0.03 (standard error); if we exclude morphs, subjects correlate at 0.48 ± 0.02; the average one-subject-against-others correlation of in-scanner distinctiveness ratings is 0.73 ± 0.04 (individual one-subject-against-others values can be seen in the top row of Figure 3.32).

At the neural level, the perception of faces is also quite similar across subjects. For instance, all 10 of our subjects have face-selective clusters within fusiform gyrus (see Figure 3.35). Here, we take that a step further and test how similar the distances between individual faces are, as inferred from neural patterns of activity alone. We note that, although relative to the space of all visual stimuli, our stimulus set was extremely restricted, having only grayscale, static, images, of equal spatial extent, shaped like an oval and having low contrast (when projected onto the back of a screen in the MR room), containing only one kind of visual object, faces, of the same gender and race, and always wearing approximately the same neutral expression and gazing straight ahead, there were nonetheless some discernible and reliable differences between the representations of individual faces in the brain, differences which served as the basis for the consistency calculation used all throughout this chapter.

We find a weak but positive correlation between the neural-based RSMs across subjects. To gain statistical power, we compare the average among 5 subjects to the average among the remaining 5 (100 trials). The RSMs have a peak correlation of 0.11 in lateral occipital cortex and 0.095 in fusiform gyrus.

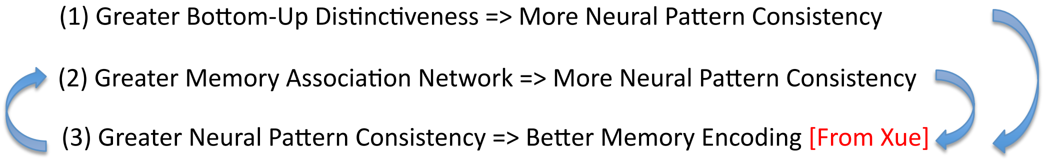
3.6.1 Methods and notes

RSMs are only computed over sets of faces viewed within the same scanning session to eliminate the artifact of same-scan faces being more correlated than different-scan faces (as faces were partitioned into scans equally for all subjects). Cross-subject correlations are then averaged across scans. For each set of faces (one per scan), subjects are partitioned into two sets of 5 randomly in 100 trials. In each trial, the correlation between the two RSMs, and the significance of the correlation (under the null hypothesis that the unique/symmetric RSM entries are normally distributed and independent of each other), is computed. Results are averaged together over trials (with p values averaging in $-log_{10}p$ scale).

Sensorimotor artifact: We find a very high correlation in sensorimotor cortex because, there, the distances between faces rated by button controllers in opposite hands are very large relative to the distances between faces rated by controllers in the same hand; therefore, any similarity in the behavioral ratings task gives rise to a sensorimotor RSM correlation by experimental artifact (namely, that subjects all used the same button layout).

3.7 Discussion and interpretation of results

The relationship between distinctiveness and memory, which gives rise to familiarity, is subjectively clear: our feeling that an unfamiliar face is more memorable is coincident with the feeling that it is more distinctive. The results in this chapter may provide a novel framework within which to understand how memories of faces, or perhaps even memories in visual objects in general, are formed, as a result of distinctiveness, and how these two relate to neural pattern consistency. The framework is summarized in the three interrelated experimentally-supported claims below:



(1) Greater Bottom-Up Distinctiveness => More Neural Pattern Consistency

(2) Greater Memory Association Network => More Neural Pattern Consistency

(3) Greater Neural Pattern Consistency => Better Memory Encoding [From Xue]

Claim (1) is supported by our finding that distinctive faces are associated with greater neural pattern consistency in early visual cortex. Suppose we begin with a face which, for some reason, is perceived to be more distinctive to its viewer. We found that such a face is likely to enhance activation in secondary visual areas, including LO, IT, and FFA, but not primary visual cortex (see Figure 3.5). However, the response to the face is on average more consistent in primary visual cortex. It may be that distinct stimuli are ones which significantly activate intermediate or high-level visual features, such as those found in secondary visual areas, more strongly, and that this activity feeds back into primary visual cortex, enforcing a consistent pattern of activity there, without enhancing aggregate activity magnitude. Figure 3.13, adapted from Douglas and Martin [22], illustrates a simple recurrent neural network model compatible with this idea: a soft winner-takes-all (soft-WTA) network. If we suppose that the activity in V1 depends in part on feedback input from secondary visual areas, and if some of its networks implement a soft-WTA functionality, then the overall output activity level (e.g., the area under the solid curve in Figure 3.13) may be relatively constant irrespective of the feedback profile. This may explain why we find no significant difference in overall activation magnitude between distinct and indistinct faces in V1/V2. However, if the feedback profile is more consistent under one stimulus condition, the soft-WTA output would be

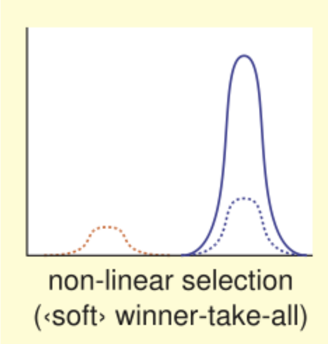


Figure 3.13: A soft winner-takes-all (soft-WTA) neural network, adapted from Douglas and Martin [22]. There is a network of neurons (indexed along the x-axis), whose activity is shown in the solid line, and which receive input indicated by the dotted line.

more consistent under that condition as well, picking the same winner more often.

Claim (2) is supported by our finding that familiar faces are associated with greater neural pattern consistency, in visual cortex and well beyond. Familiar faces are by definition ones which involve a greater network of associated memories than unfamiliar faces, which are associated with zero or few memories. The same feedback mechanism described above may explain why familiar faces are more consistently represented in V1/V2, despite not increasing activation there, except with the feedback signals originating, not only in secondary visual areas, but also from frontal cortex and language areas.

Claim (3) is based on a result from Xue et al. [102]. They showed that greater neural pattern consistency in response to faces in an fMRI scan was associated with better recognition and recall 1 to 6 hours after scanning.

Main interpretation: Claims 1-3 suggest that there may be a cyclic process for the learning of unfamiliar faces in which consistent representation and the presence of newly formed memories mutually feedback on each other, with distinctiveness progressively increasing, as described in Chapter 2. To further interpret the results, a distinctive face may feel more memorable because, every time it is seen, activity in V1/V2 may be in a state more similar to the one visited the last time the face was seen (than it would be for an indistinct face), giving an impression at the lowest neural level of a familiar state. This may be related to (i) top-down/guided covert attention to the same features on the face, (ii) the same cells within redundant subpopulations of neurons firing, which would be a form of sensory memory, or, alternatively or complementarily, (iii) more temporally similar firing patterns. This result may also be connected to the finding of Goard and Dan [30]: basal forebrain activation, related to arousal and increased attention, increased the reliability of firing patterns in response to movies of natural scenes in rat visual cortex.

It should be noted that another possible source of neural pattern consistency in V1/V2 to familiar and distinctive faces is more consistent eye-movement patterns for these faces. Although we instructed subjects to try not to move their eyes, and ensured that they were fixated at stimulus onset by employing a detection task at the fixation point, and the stimuli were only briefly displayed (500ms), and small enough ($5^\circ \times 7^\circ$) to discourage large saccades, it is nonetheless possible that these faces are explored, and more consistently across trials than unfamiliar faces and indistinct faces. In Chapter 2, Section 2.7.2, we found that familiar faces are compared more holistically, and briefly discussed the possibility that familiar faces are in fact explored less (thus more consistently), based on a free viewing (3 seconds) memory experiment by Heisz et al. [40]. However, van Belle et al. [98] found no significant differences in either the number of fixations or duration of fixations to familiar faces compared to unfamiliar faces. One further reason to discount this as the explanation is that we find enhanced consistency in regions far beyond V1/V2, regions which should be insensitive to exact eye position on such a small stimulus: for instance, familiar faces are relatively more consistent in LO (see Figure 3.5), and the consistency boost among especially distinctive familiar faces seems to span the whole brain (see Figure 3.9).

Possible application: The primary scientific value of this chapter is in its contribution to the basic understanding of face perception and memory. However, Section 3.5 suggests a possible application for the methods used here. It was shown that, although an *insignificant* number of brain regions exhibited an activation *magnitude* boost at time t_1 in individuals who found a face especially distinctive at time t_0 several weeks before t_1, a significant number of regions *did* show a *consistency* boost. We hypothesized that the enhanced consistency, especially in amygdala and hippocampus, was likely due to early memory formation – the distinctive faces being the ones more likely remembered by the subjects. If so, this suggests a kind of neural lie detector test, one which could help build a case for the memory of a face only briefly seen. Conventional activation-magnitude-only analysis methods would potentially fail to find a signal in such cases. It is clear, however, that such applications are ethically dubious and would need to be applied only sparingly and prudently if at all. Furthermore, the relationship between short exposures to people and the hippocampal and amygdala consistencies in response to photographs of their faces would need to be more rigorously characterized than was done for the purposes of this study.

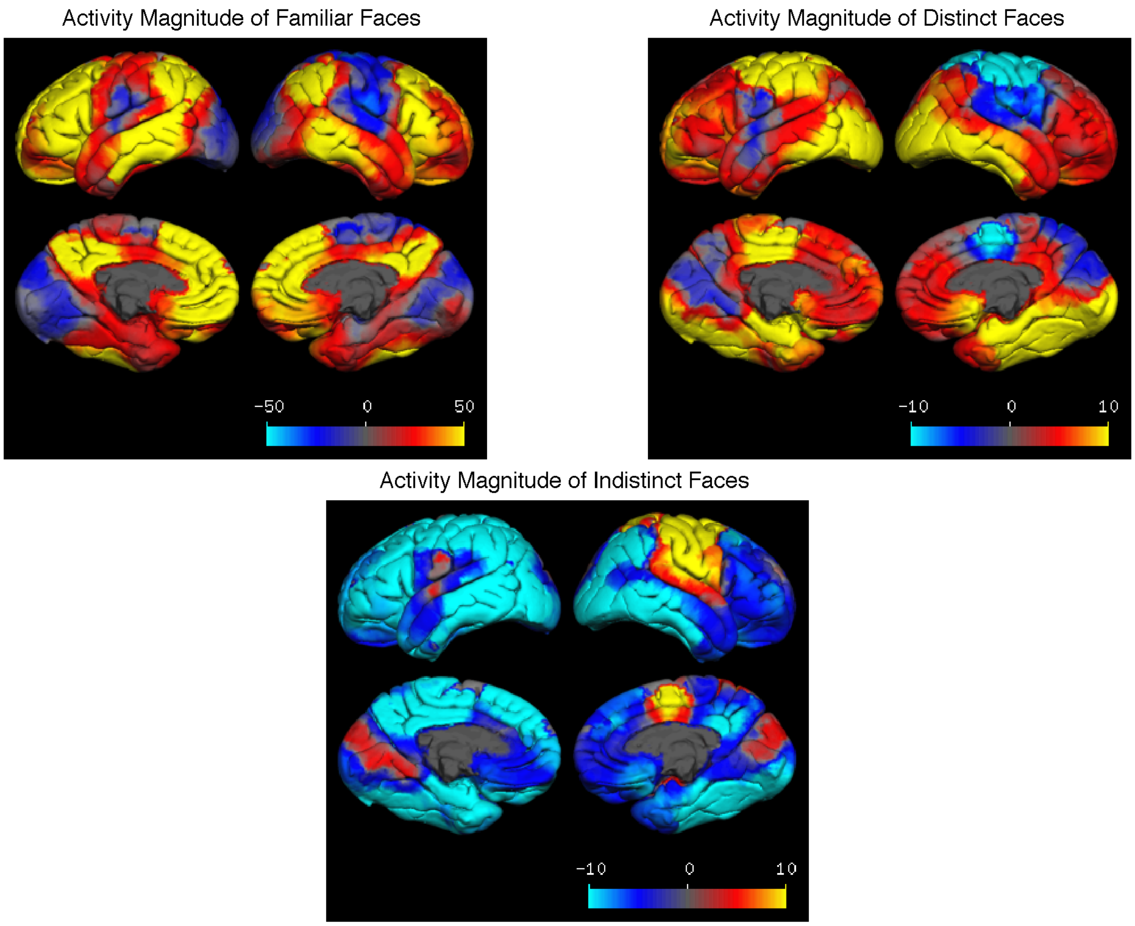
3.8 Supplementary results

As with the supplementary results presented in Chapter 2, these results are less important and are mostly explained by their figure captions. Also, as with Chapter 2, after the supplementary results, there is also an Appendix (Section 3.9), which includes experimental methods.

3.8.1 Main effects: other parameters

3.8.1.1 Magnitude effects mapped across the brain

In the main text of the chapter, we presented map-types results showing how consistency varied across the brain by category of stimulus (familiar, distinct, indistinct). Below, we show how magnitude of response varies with these categories of stimuli.



Activity Magnitude of Familiar Faces

Activity Magnitude of Distinct Faces

Activity Magnitude of Indistinct Faces

Figure 3.14: All values are **Z-score×100**; middle of the range is zero (light blue is negative). *Top left:* Familiar faces activate regions throughout the brain, including inferior parietal, frontal, and middle temporal cortex, and precuneus, but *excluding occipital cortex*; whereas familiar faces were more consistent in left hemisphere FFA, such laterality is not observed for activation magnitude shown above; note scale bar goes to 50 (Z-score 0.5) *Top right:* Distinct faces activate many regions of the brain, including parts of occipital and temporal lobes; note scale bar goes to 8 (Z-score 0.08). *Bottom:* The brain is relatively inactive during presentations of indistinct faces. However (aside from the sensorimotor artifact along the central sulcus), cuneus is slightly activated for indistinct faces, but we find that it fails to reach a statistically significant difference from either familiar or distinct faces by 1-sided t-test.

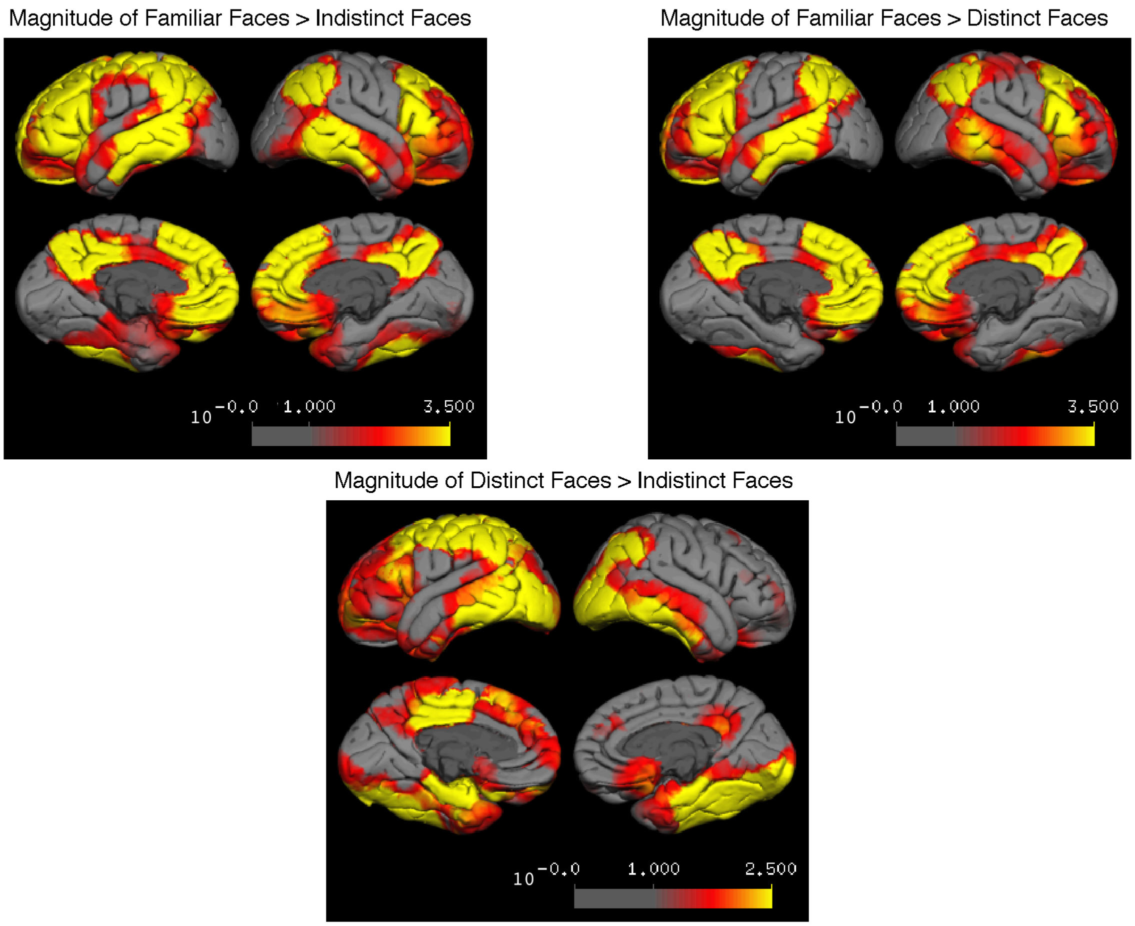

Magnitude of Familiar Faces > Indistinct Faces
10-0.0
1.000
3.500
Magnitude of Familiar Faces > Distinct Faces
10-0.0
1.000
3.500
Magnitude of Distinct Faces > Indistinct Faces
10-0.0
1.000
2.500

Figure 3.15: Regions in which voxel spheres, on average, have significantly different magnitudes of activation for a category of face stimulus (familiar, distinct, or indistinct) are highlighted. Note that familiar faces do ***not*** significantly activate much of occipital lobe relative to distinct or indistinct faces. We map the average empirical statistical significance values using $-log_{10}p$ scale.

3.8.1.2 Consistency mapped with 100-voxel spheres

In the main text of the chapter, we presented map-types results showing how consistency varied across the brain by category of stimulus (familiar, distinct, indistinct) with voxel spheres containing 300 voxels each. Here, we show results with 100-voxel spheres.

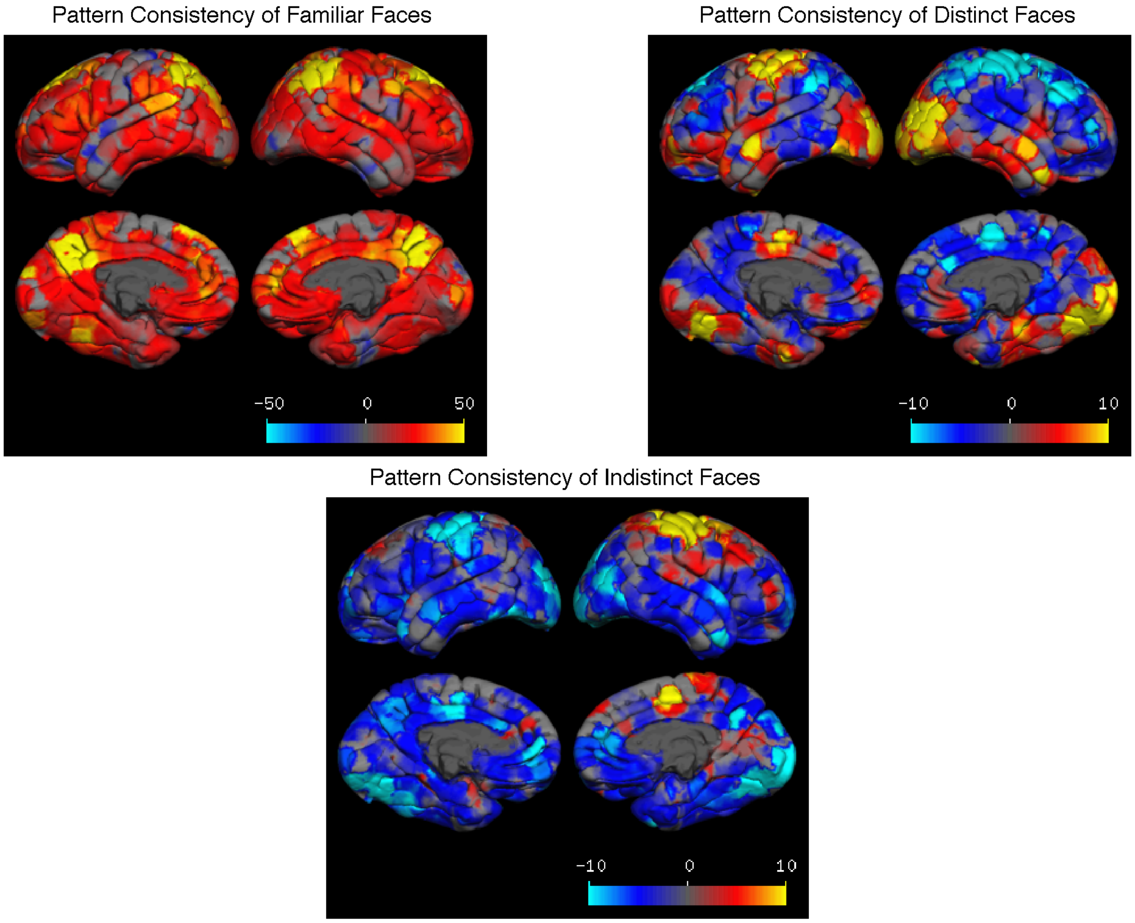
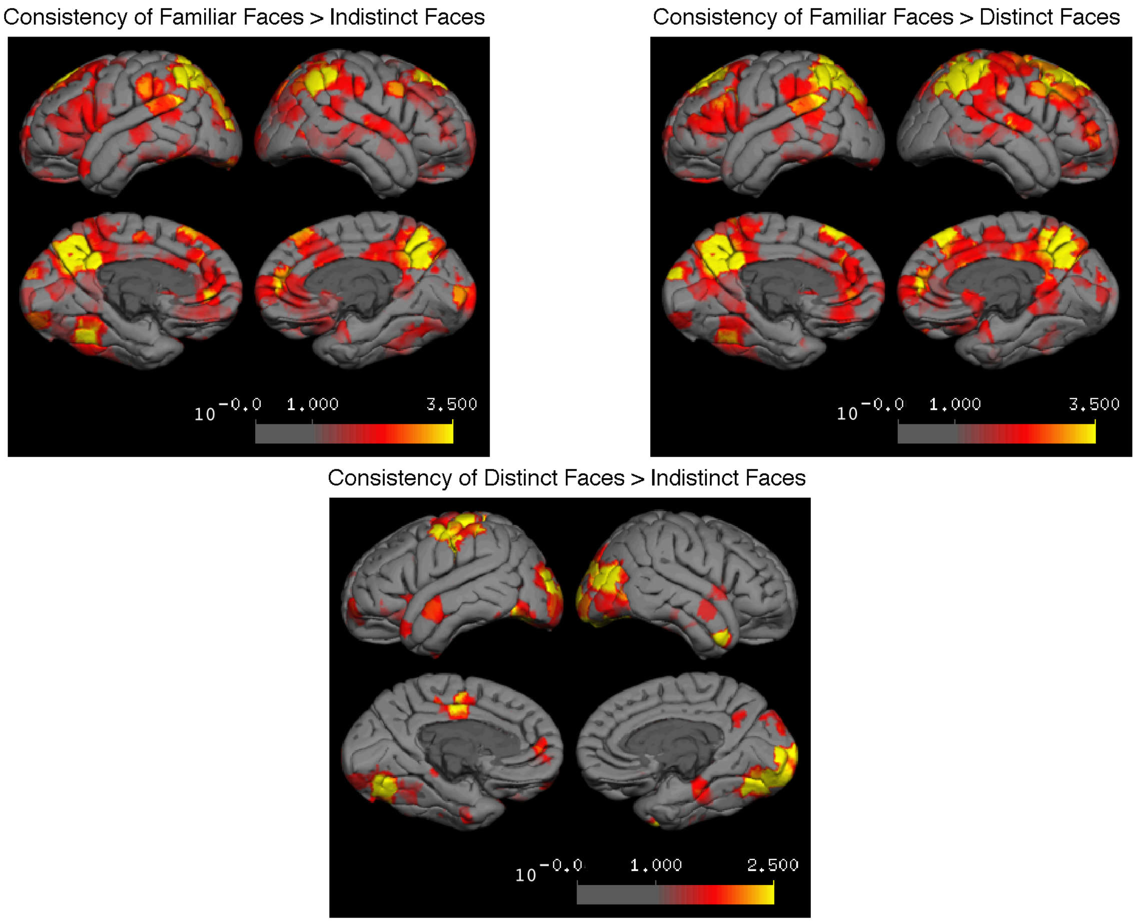

Pattern Consistency of Familiar Faces
-50
0
50
Pattern Consistency of Distinct Faces
-10
0
10
Pattern Consistency of Indistinct Faces
-10
0
10

Figure 3.16: All values are **Z-score×100**; middle of the range is zero. Familiar > Distinct > Indistinct Face consistency for 100-voxel spheres map-type analysis; results are similar to 300-voxel spheres analysis shown in Figure 3.3, but less smooth. *Top left:* Familiar faces are represented very consistently all throughout the brain (especially inferior parietal, precuneus, and fusiform gyrus); note scale bar goes to 50 (Z-score 0.5) *Top right:* Distinct faces are represented more consistently than average in occipital and temporal lobe; note scale bar goes to 8 (Z-score 0.08). *Bottom:* Indistinct faces are less consistent than average in occipital and temporal lobe.



Consistency of Familiar Faces > Indistinct Faces

Consistency of Familiar Faces > Distinct Faces

Consistency of Distinct Faces > Indistinct Faces

Figure 3.17: Map of average empirical significance value ($-log_{10}p$) for 100-voxel spheres map-type analysis. Regions in which voxel spheres, on average, have significantly different neural pattern consistencies for a category of face stimulus (familiar, distinct, or indistinct) are highlighted. As seen with the results for 300 voxels, FFA is more consistent for familiar faces only in left hemisphere.

3.8.1.3 Effects in particular ROIs

In this section, we extend the results shown in Figure 3.5 to more regions, and to in-scanner and pixel-based distinctiveness measures (for which the results are essentially equivalent to the pixel+scanner measure, only less robust).

Here, we will list some of the regions which have significant effects (using pixel+scanner distinctiveness) but which were not shown in Figure 3.5: inferior temporal cortex (IT), part of the ventral stream of visual cortex, shows the familiar > distinct > indistinct trend both for consistency and magnitude. Hippocampus shows a significant increase in both consistency and activation for distinct faces compared to indistinct faces. This is presumably because they are remembered from previous trials and experiments. Familiar faces are significantly more activated and consistent throughout many regions of the brain, including cuneus, precuneus, cingulate cortex, inferior parietal (IPL), and BA44/45 (equivalent to Broca's area in left hemisphere).

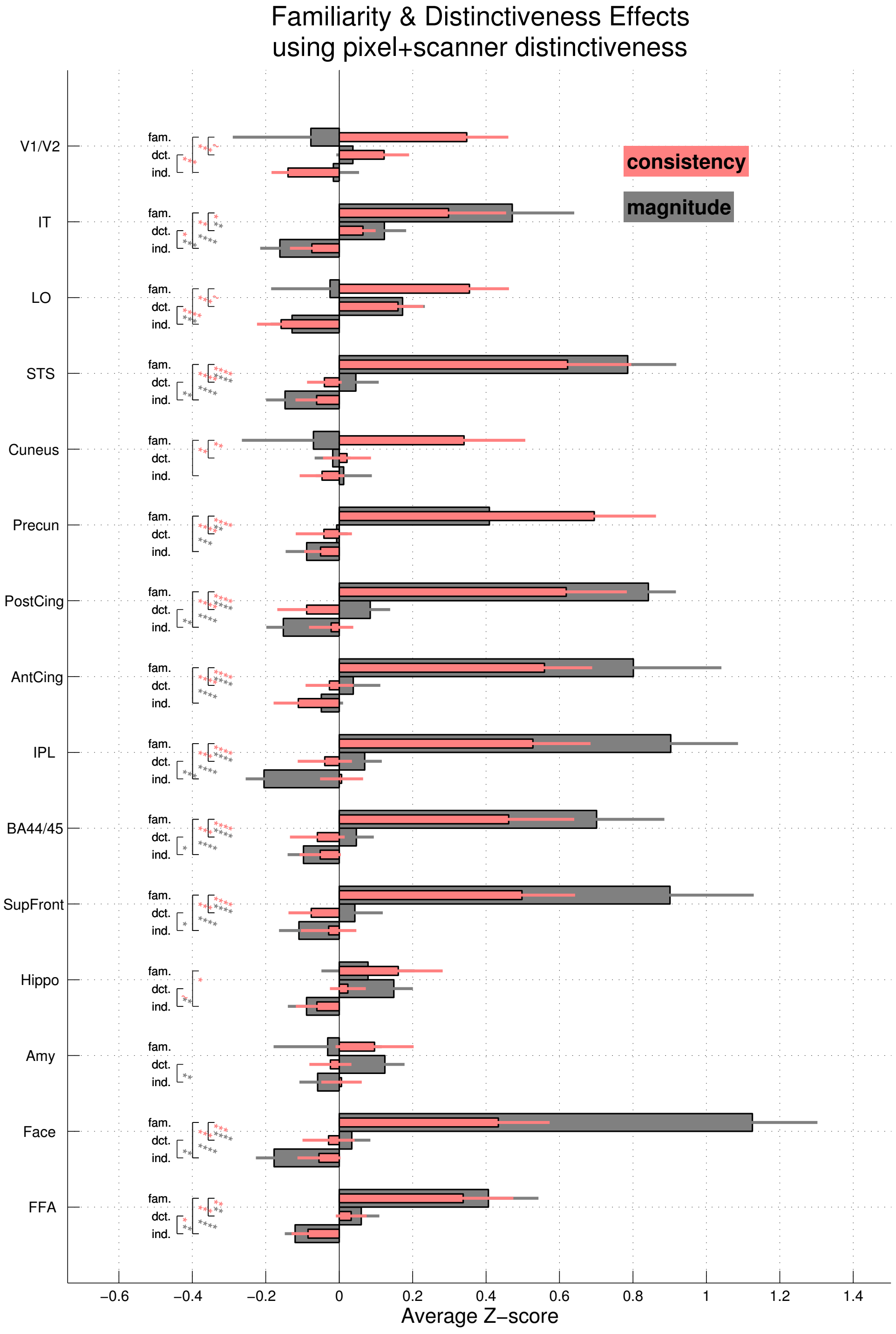


Figure 3.18: Average response magnitude and consistency are shown for an expanded set of regions of interest (expanded form of Figure 3.5). Notice that in V1/V2 and LO familiar faces are not relatively activated (gray bars corresponding to "fam."). Also note a slight effect in the hippocampus for distinct vs. indistinct faces.

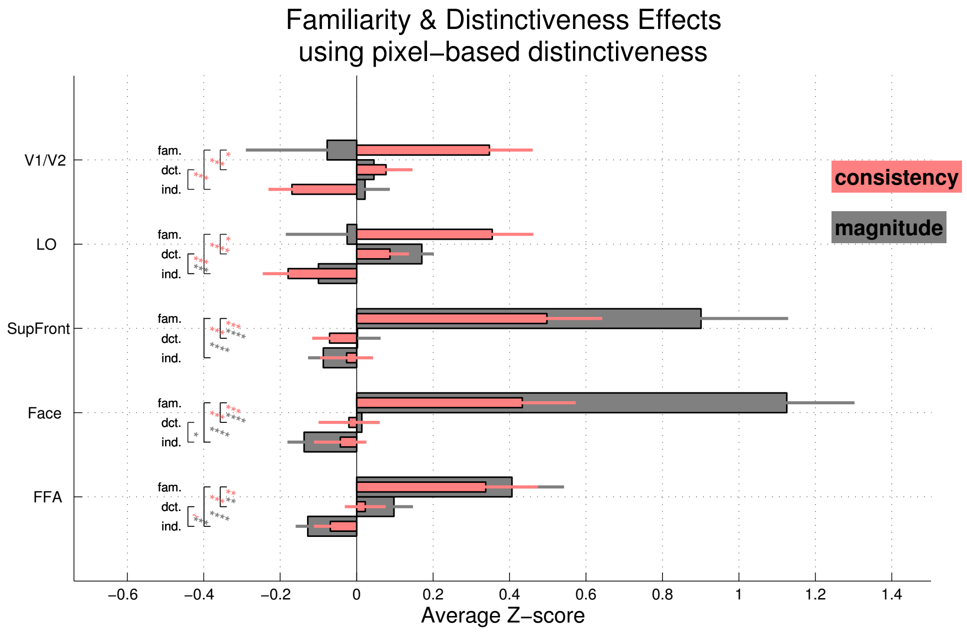


Figure 3.19: Average response magnitude and consistency are shown for several regions of interest (modified form of Figure 3.5), except with distinctiveness modified to be based on only pixel values. Compared to the pixel+scanner distinctiveness results shown in Figure 3.5, the pixel-based distinct faces are slightly less consistent in V1/V2 and LO, but still significantly more so than indistinct faces; however, results are directionally equivalent and meet significance in the same places (except here there is no magnitude difference in SupFront between distinct and indistinct faces).

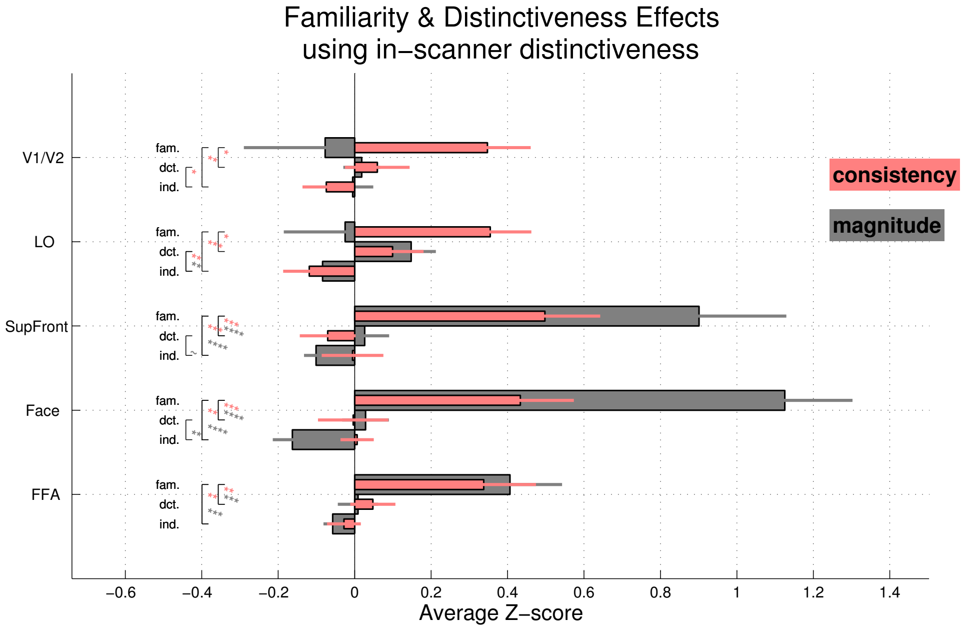


Figure 3.20: Average response magnitude and consistency are shown for several regions of interest (modified form of Figure 3.5), except with distinctiveness modified to be based on only in-scanner ratings of each subject. Compared to the pixel+scanner distinctiveness results shown in Figure 3.5, the scanner-based indistinct faces are slightly more consistent; however, results are directionally equivalent and meet significance in all the same places (except, here, there is no significant difference between distinct and indistinct faces in FFA whereas there is for pixel+scanner based distinctiveness).

3.8.2 Familiarity and distinctiveness in inverted (upside-down) faces

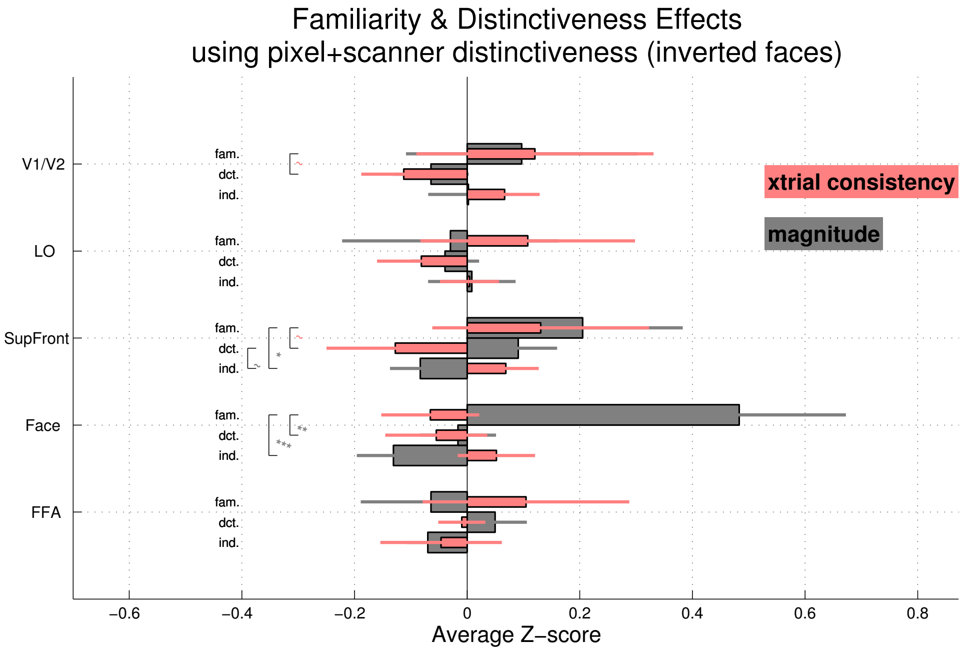


Figure 3.21: In two scans of experiment FaceView, subjects viewed upside-down (inverted) faces. Because each face was only viewed in one such scan, only cross-trial (not cross-session or hybrid) consistency could be calculated. The only highly significant ($p < 0.01$) effect is that familiar faces activate face-selective clusters more than distinct or indistinct faces. The equivalent analysis (base faces, cross-trial), but with upright faces, yields a highly statistically significant effect ($p < 0.01$) of consistency and activation magnitude ($p < 0.001$) for familiar faces in superior frontal lobe (SupFront). We infer that superior frontal is less activated for familiar faces when they are inverted. 3.18.

3.8.3 Non-categorical distinctiveness vs. consistency and magnitude: follow-ups

In this section, we follow up on the results shown in Section 3.2.1 and Figure 3.6. The point of that section was to test whether the greater consistency of distinct faces is a purely categorical effect or one sufficiently reliable to fit in a linear relationship with individual distinctiveness values. We found in that section that consistency linearly increased with distinctiveness with statistical significance in occipital cortex. Below, we show that the same is true in fusiform gyrus, and for activation magnitude in occipital cortex. We also show that, in non-visual-cortex, namely frontal lobe, such trends do not hold.

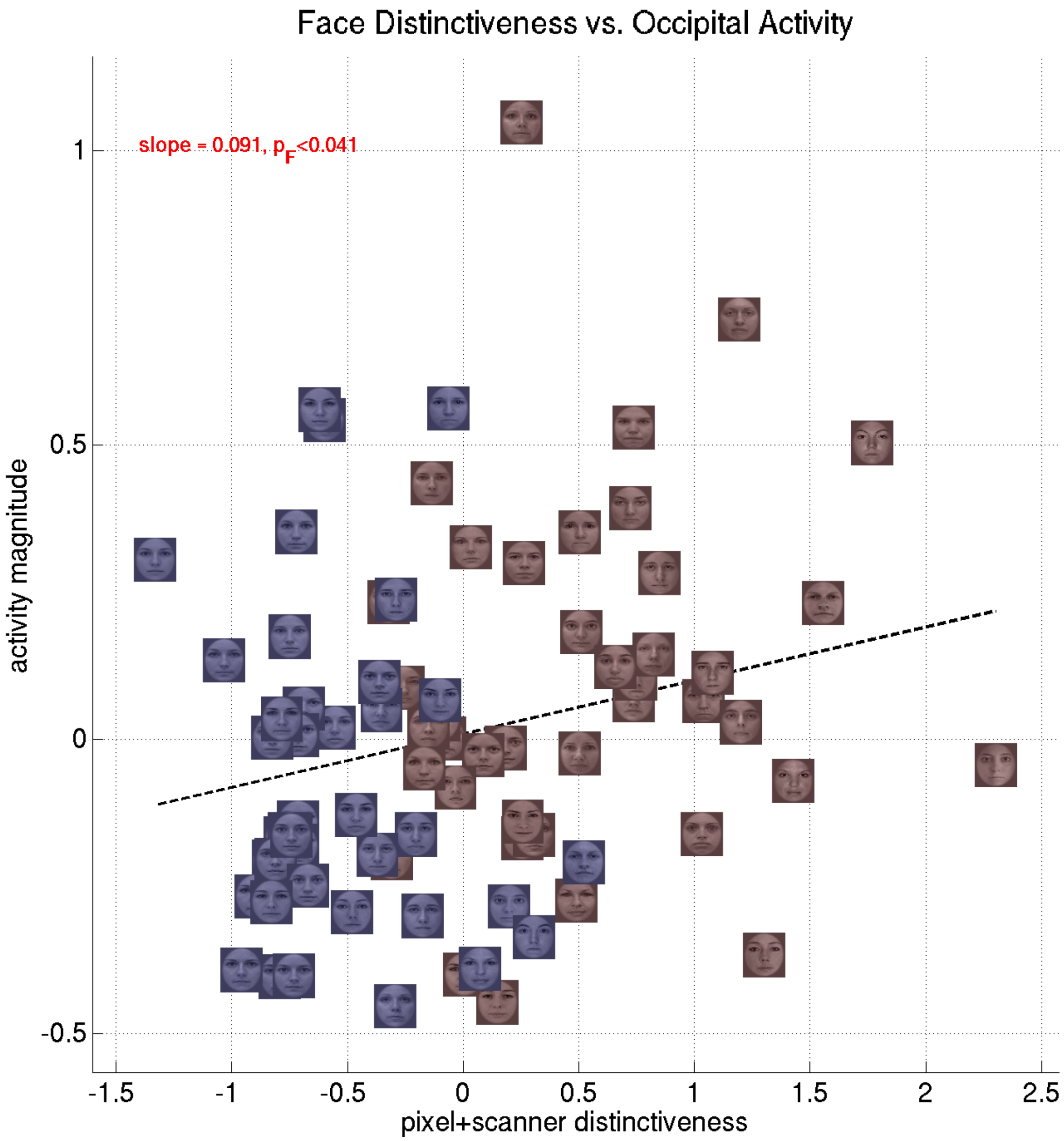


Figure 3.22: The subject-averaged (excluding familiar viewers) activation magnitude in occipital lobe (i.e., averaged across occipital spheres) is plotted against the corresponding subject-averaged distinctiveness for each of the 80 face stimuli. Morphs are shaded blue. Magnitude can be linearly modeled with distinctiveness, yielding a goodness-of-fit F-test significance level of $p_F < 0.041$, which is weaker than what was found with the consistency ($p_F < 0.0059$, see Figure 3.6).

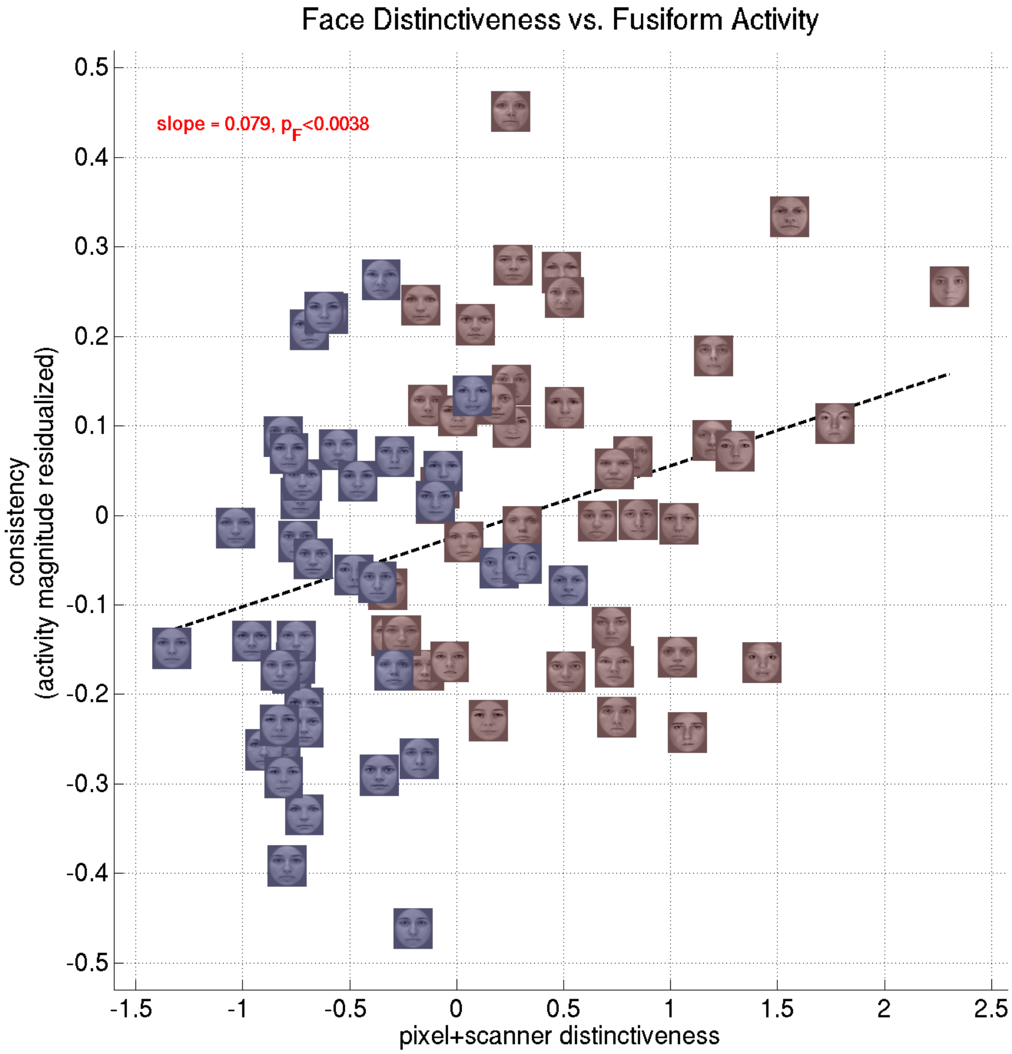


Figure 3.23: The subject-averaged (excluding familiar viewers) neural pattern consistency in fusiform gyrus (i.e., averaged across spheres centered in, or stepped from the center of, fusiform regions) is plotted against the corresponding subject-averaged distinctiveness for each of the 80 face stimuli. Morphs are shaded blue. Neural pattern consistency can be linearly modeled with distinctiveness, yielding a goodness-of-fit F-test significance level of $p_F < 0.0038$. The same analysis, except with magnitude of activation (instead of consistency) in fusiform gyrus, also yields a positive correlation and a fit of $p_F < 0.00028$.

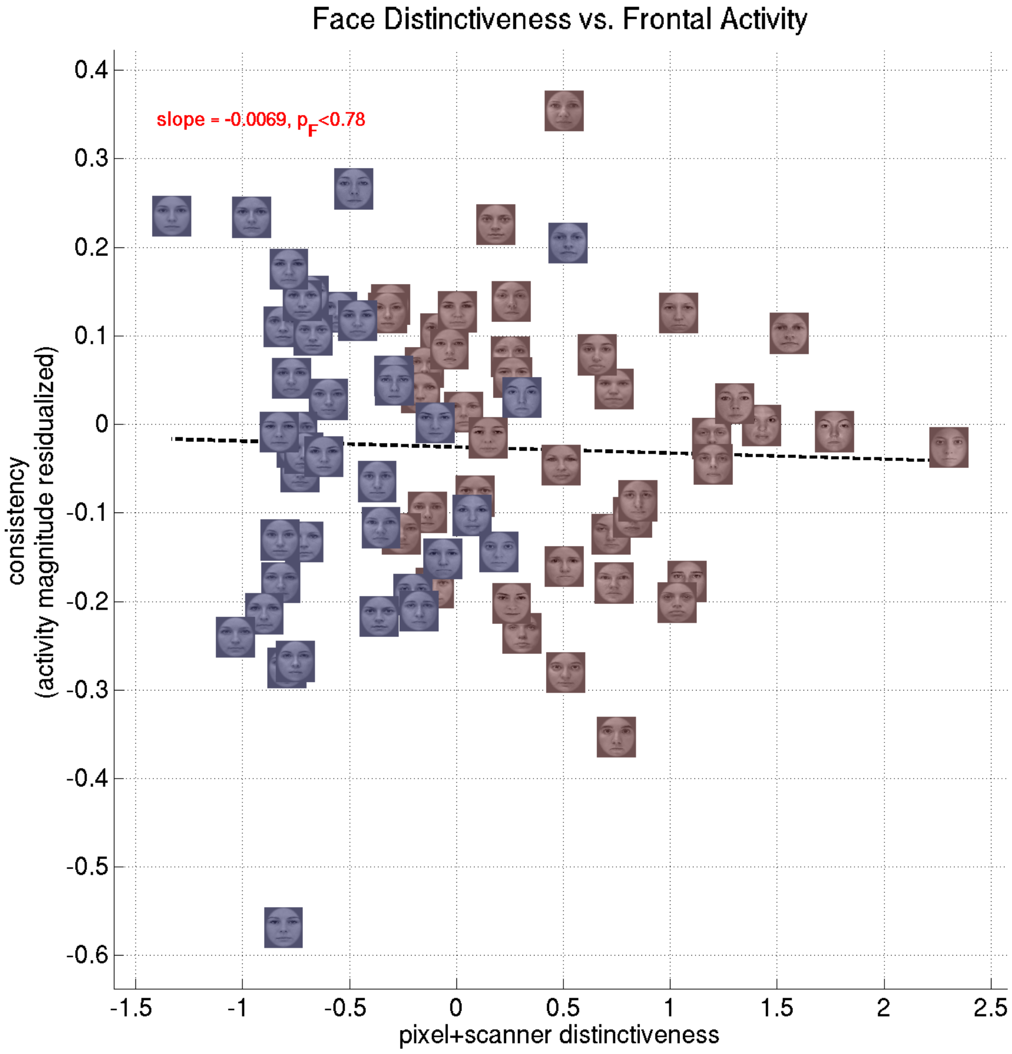


Figure 3.24: The subject-averaged (excluding familiar viewers) neural pattern consistency in frontal lobe (i.e., averaged across spheres centered in, or stepped from the center of, regions in frontal lobe) is plotted against the corresponding subject-averaged distinctiveness for each of the 80 face stimuli. Morphs are shaded blue. In contrast to the results shown above in visual cortex (occipital lobe and fusiform gyrus), there is no significant relationship between the two here. Similarly, distinctiveness cannot be used to model activation magnitude in frontal lobe either ($p_F < 0.6$).

3.8.4 Type of personal familiarity and neural consistency

Refer to Figure 3.25. The self face is highly activated, but inconsistent in visual cortex; however, the *morph* between self and sister is most consistent among familiar face types in visual cortex, including IT, and LO, and almost V1/V2 as well. The self-face is known to be processed differently than other familiar faces [21, 72], so its specialness here is not surprising.

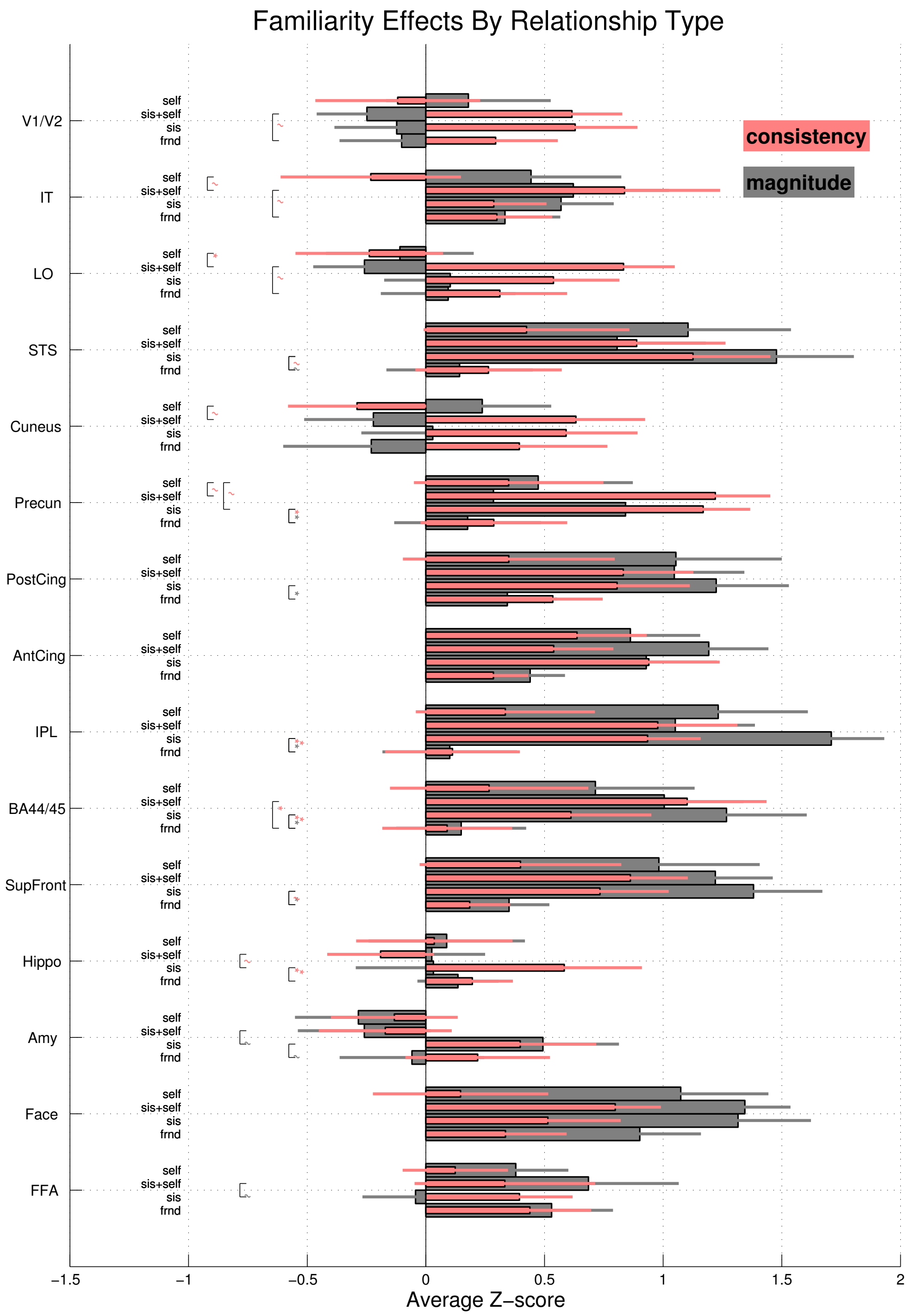


Figure 3.25: Consistency and activation magnitude by type of familiarity relationship. Notice specialness (e.g., relatively lower consistency in V1/V2) of the self-face relative to the other types. Friend here is defined as having familiarity 7 or 8, and sis+self is the morph stimulus which is the 50/50 between the two.

3.8.5 Pairwise relationship between consistency and amplitude across brain regions

For all the figures in this section, consistency and activation magnitude are raw, unresidualized, un-Z-scored. Correlations are computed per subject across face stimuli, then averaged across subjects. Results are also averaged across hemispheres (and only intra-hemispheric correlations are computed). Each matrix has rows and columns indexing each of the 77 regions in the Destrieux Atlas, called "aparc a2009s" by freesurfer. Results are averaged across the 7 spheres per region (though center spheres are only compared to center spheres, posterior to posterior, etc.). The regions are ordered in a smooth, spatially contiguous way, from V1 (calcarine sulcus), to temporal lobe, to parietal lobe, and terminating in the frontal pole.

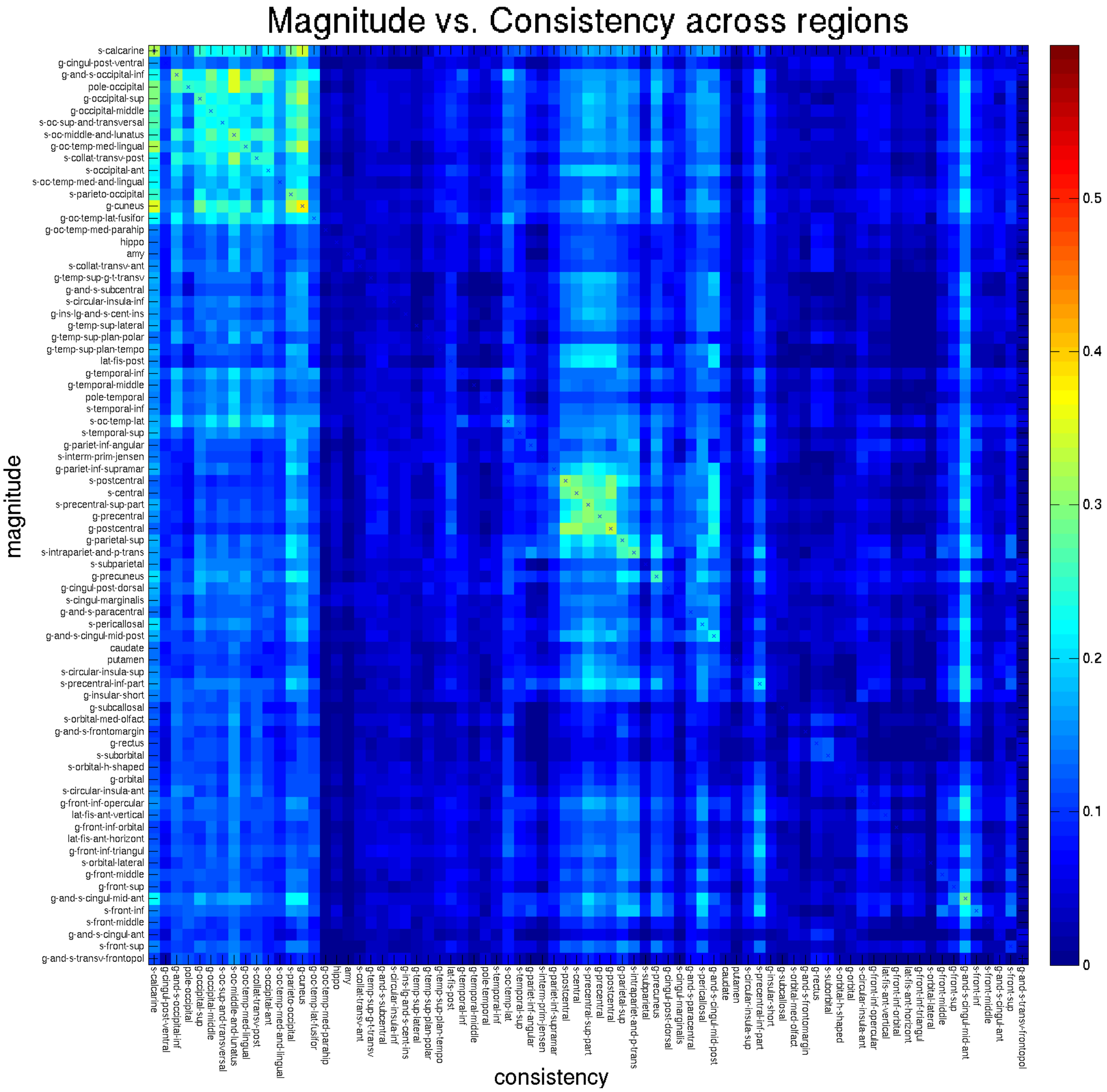


Figure 3.26: The average correlation between the consistency of a region (indexed by columns) and the magnitude (indexed by rows), in the same hemisphere, across all faces. "x"s mark the diagonal entries. The two bright squares of interrelationship between correlation and magnitude correspond to occipital lobe and sensorimotor cortex. The set of leftmost columns (up to the end of occipital lobe) of the matrix correspond approximately to the top *right* panel of Figure 3.7: very little correlation with magnitude outside of occipital areas.

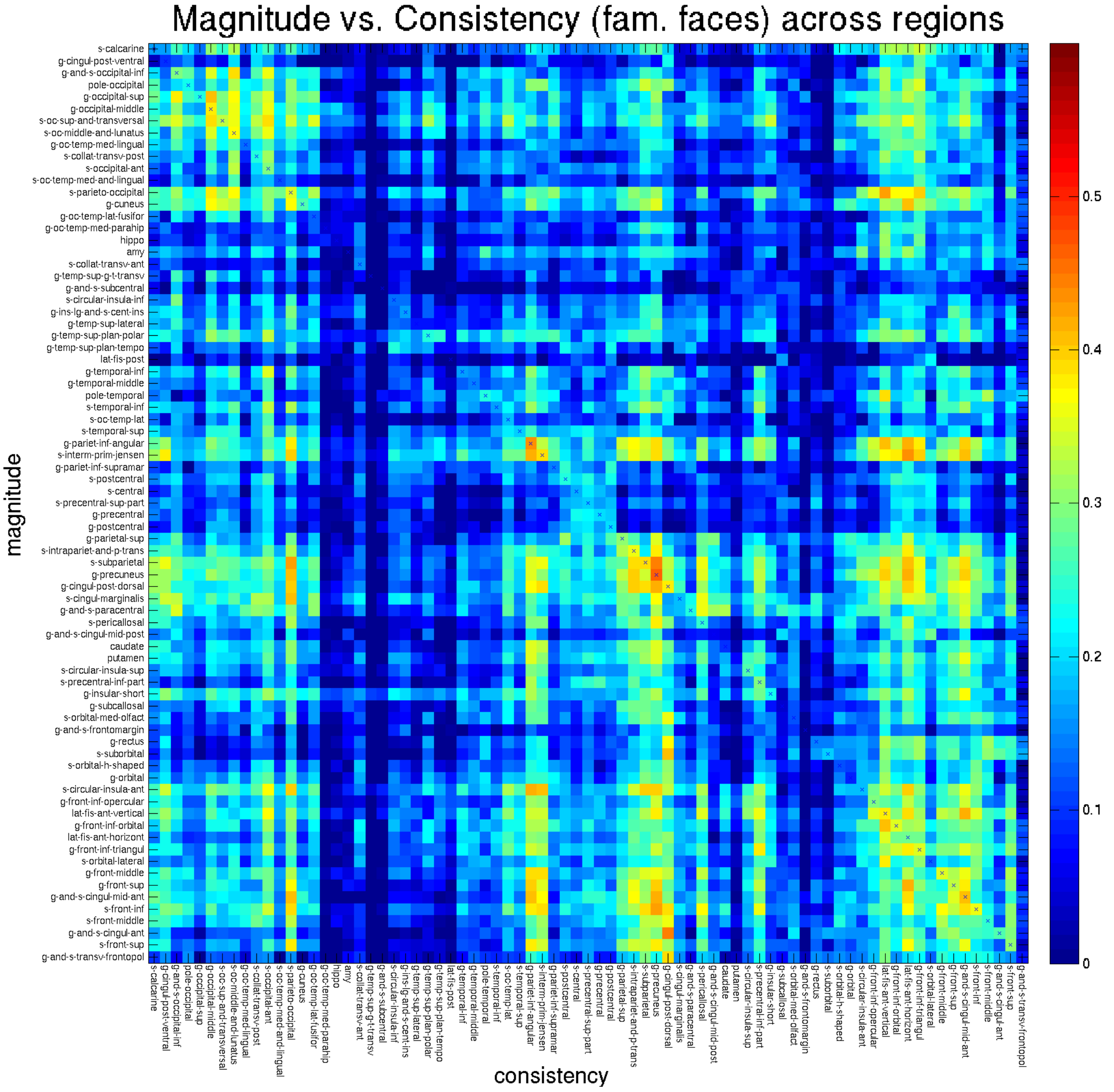


Figure 3.27: The average correlation between the consistency of a region (indexed by columns) and the magnitude (indexed by rows), in the same hemisphere, across only familiar faces. "x"s mark the diagonal entries. The set of leftmost columns (up to the end of occipital lobe) of the matrix correspond approximately to the top *left* panel of Figure 3.7: we find high correlations with magnitude in higher brain regions outside of the occipital lobe.

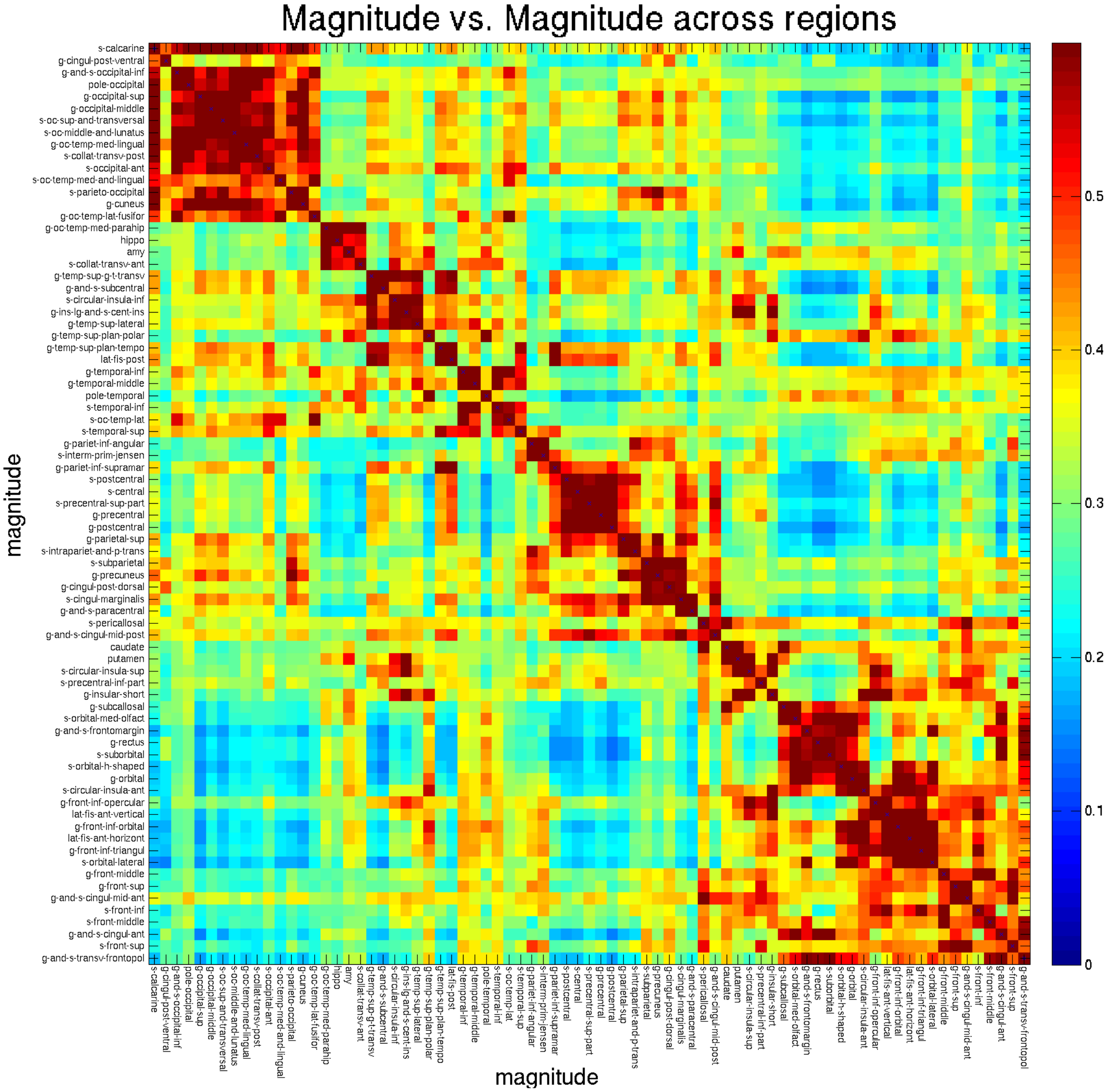


Figure 3.28: The correlation between activation magnitude in one region and activation magnitude in another, across all faces. The activation magnitudes of neighboring regions are very highly correlated, especially within the same cortical lobe.

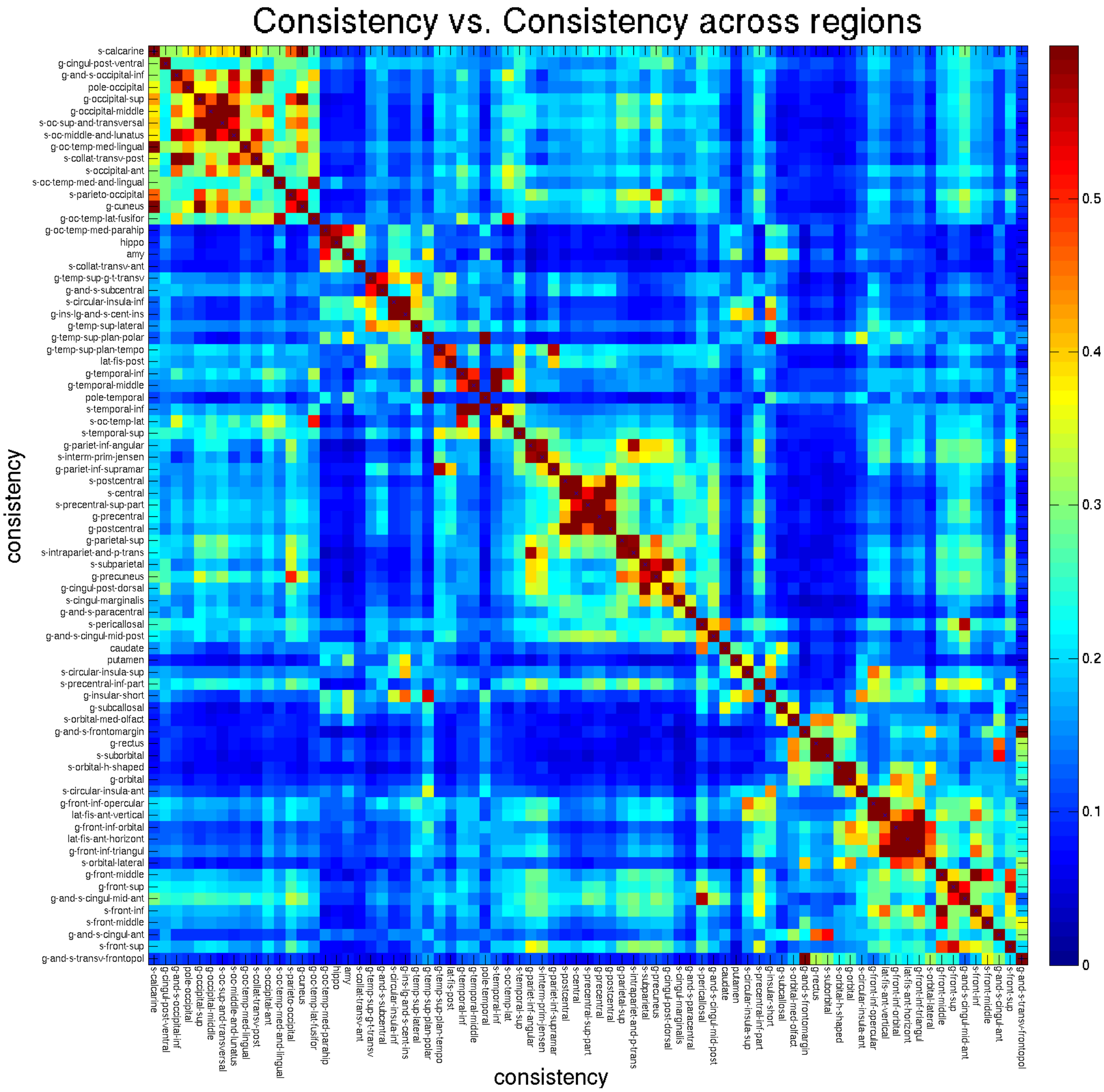


Figure 3.29: The correlation between neural pattern consistency in one region and neural pattern consistency in another, across all faces. The correlation with neighbors is present but to much lesser extent than found in the equivalent analysis of activation magnitude.

3.9 Appendix

Subjects provided informed written consent prior to the experiments. The Caltech Institutional Review Board approved all experimental procedures.

3.9.1 Details of task performed in MRI scanner (Experiment FaceView)

This experiment consisted of showing subjects faces one a time, after which they would enter a number (1=least distinctive to 6=most distinctive) with a keypad controller indicating how distinctive they found the face. Numbers 1-3 could be entered with a controller in the right hand and numbers 4-6 with the left.

Images in the scanner were presented such that faces subtended approximately 5°x6.7° visual degrees. This size was intended to fit an entire face mostly inside foveal and para-foveal vision while maintaining sufficient detail. Guidelines for face "distinctiveness" were provided to the subjects prior to scanning (see above, Section 3.1.2).

Trial Structure: Each face presentation trial consisted of a face that flashed on-off-on, 500 ms each (total 1.5 s), followed by 2.75 seconds during which the subject had to enter a distinctiveness rating, followed by a uniform random delay of 1 to 6.25 seconds. See Figure 3.30 for a graphical representation of this. (The delay and jitter were introduced for robust estimation of the laggy hemodynamic response to each image.) Throughout the entire run between face presentations, a white fixation dot was presented at the center of the screen, and centered between where the eyes would occur on face stimuli. Subjects were instructed to fixate on it. On 7 random **catch trials**, the fixation dot would turn blue 300 ms before stimulus onset, instead of remaining white right up to stimulus onset. Because of the long delay between trials (introduced for better response estimates), such a change was difficult to notice unless attention was maintained. Subjects were instructed to withhold their distinctiveness rating following the face presentation if they noticed this. Failure to notice these catch trials together with excessive head motion was taken as an indication of too much inattention. Scans in which the subject was too determined to be too inattentive were repeated up to several weeks later ("make up scans"); out of 100 individual scans, 9 were repeated for this reason.

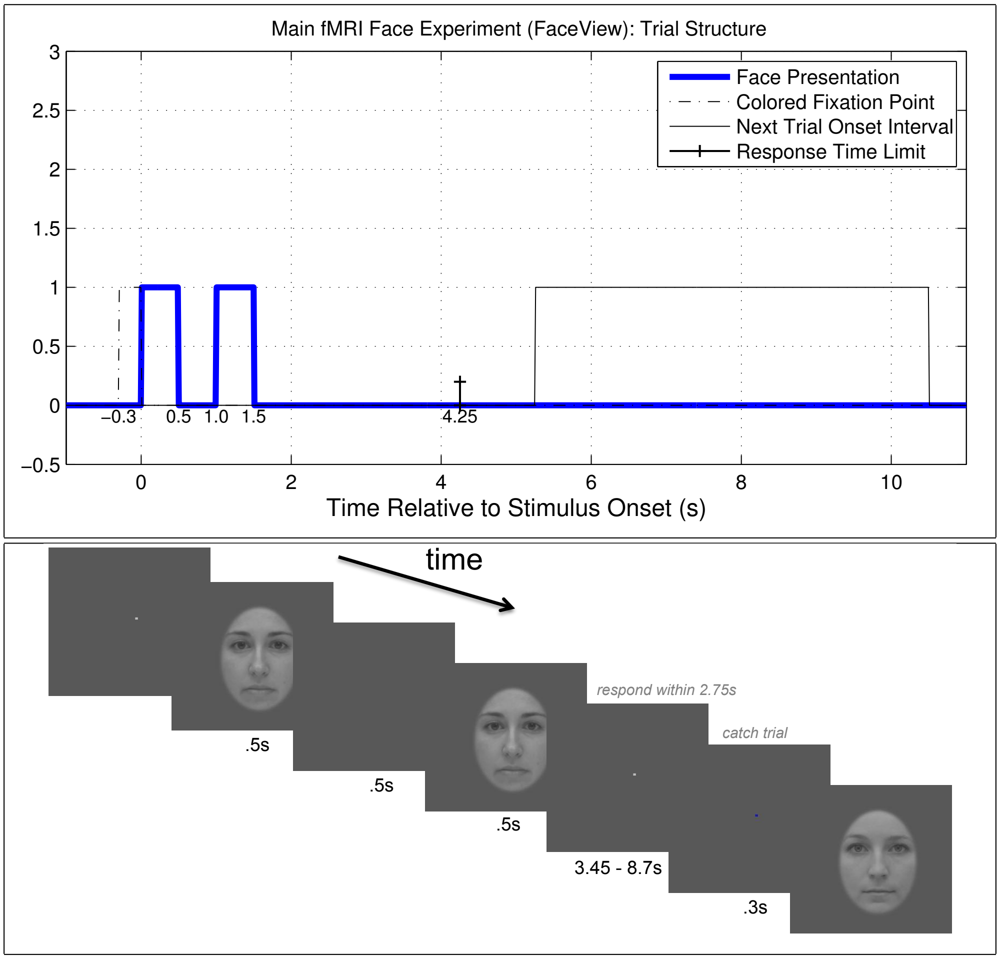


Figure 3.30: The time-course of a single trial in the FaceView experiment (taking place in an fMRI scanner), wherein subjects explicitly rated the distinctiveness of faces viewed one at a time. The bottom panel shows the beginning of a catch trial, indicated by a fixation point color change.

Scan Structure: Each fMRI scan for the FaceView experiment lasted 12 minutes, 11.25 seconds. The first 15 and last 8 seconds were blank except for the fixation point. The jittered delay times after each trial were pseudo-randomly computed once, and then frozen for all subjects and all sessions. Each scan consisted of 20 unique faces, presented an average of 4 times, randomly permuted (different permutation for each subject, for each scan) for a total of 80 trials, not including 7 catch trials. The minimum delay between identical faces ranged from 1 (consecutive) to 81 trials. Averaged across subjects and minimized across stimuli, the closest two identical faces appeared in a single session was 13.03 ± 0.21 trials. The average delay between identical faces was 21.2 ± 0.17 trials. The partitioning/batching of the 80 faces into four disjoint sets of 20 for single scan presentation was identical for all ten subjects.

Repeated Face Batches: Each of the four batches of 20 unique faces was seen in exactly two separate fMRI scans. Thus, there were 8 FaceView scans per subject (so far accounted; see below). The order in which the eight batches were shown was counter-balanced among subjects. Each face was viewed on an average of 8 trials, averaging 4 per scan. The absolute minimum delay among all subjects between two scans consisting of the same face set (same-batch scans – notably, never consisting of trials in the same order) was 3 hours, and the average minimum delay between such scans among subjects was 27 hours. The median delay between same-batch scans was 2.24 days. The single longest delay between two same-batch scans among all subjects was 31 days, due to a make-up scan in which the first was too poor in quality. The median delay between the two most temporally distant same-batch scans among subjects was 4.77 days.

Inverted Face Batches: Finally, subjects viewed the 40 non-morph/base stimuli upside down in two 20 disjoint batches of inverted face runs of experiment FaceView, with the *exactly* the same trial and instructions as before. Unlike the upright faces, each unique face was only seen in one such batch, so neural pattern consistency could only be established across trials, not sessions, for these inverted faces. Together with the 8 upright FaceView runs, this brings the total number of FaceView scans per subject to 10, for an experimental total of 100 scans (10 subjects × 10 scans), not including the functional localizer runs (see below).

3.9.2 Defining distinctiveness measures

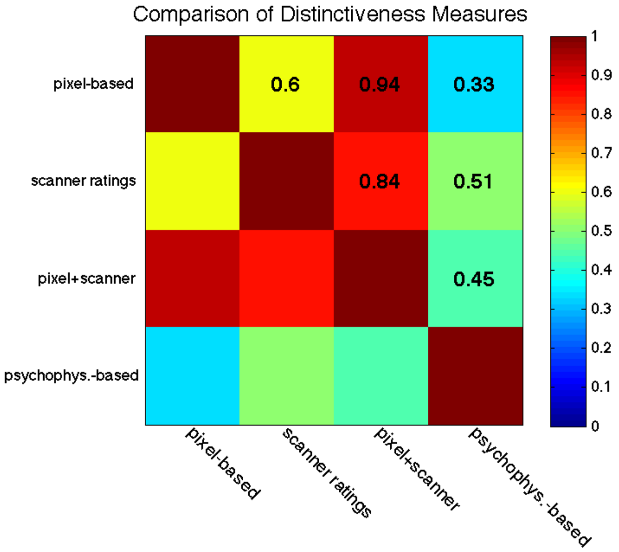


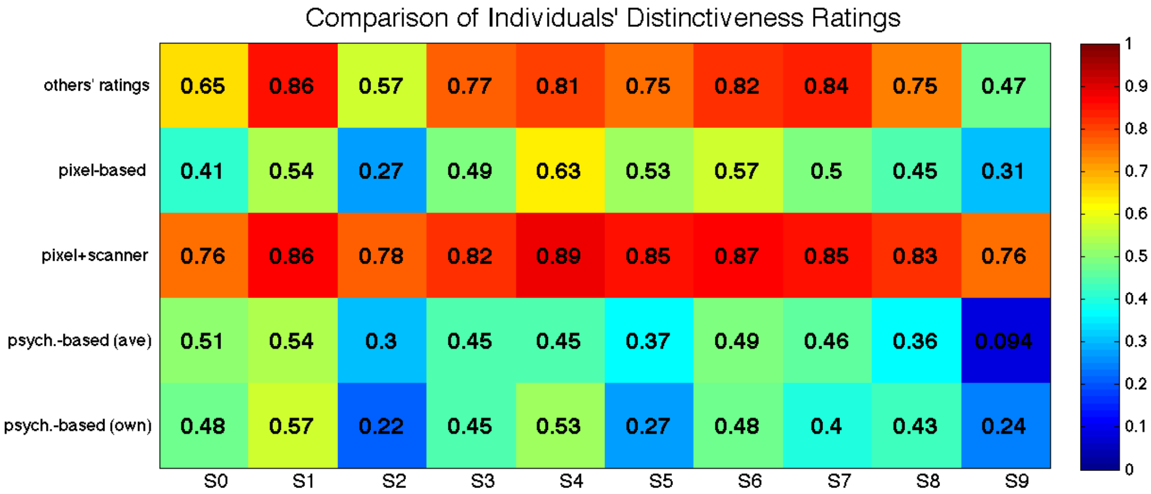
Figure 3.31: The correlations (across 80 face stimuli) between different distinctiveness measures are shown. Scanner ratings are averaged across subjects. The psychophys.-based distinctiveness values are are based on an MDS of the explicit dissimilarity ratings provided in the psychophysics experiments of Chapter 2 (called "dissim. MDS" there). Note that pixel-based distinctiveness correlates more strongly with the in-scanner distinctiveness than the psychophys.-based distinctiveness; one reason for this may be that Gaussian noise was added to stimuli in the experiments of Chapter 2 to discourage subjects from using low-level pixel-based features to compare faces (see Section 2.10.4).

(1) pixel-based distinctiveness This was computed exactly as in Chapter 2 (Section 2.10.6), but extended to all 80 faces (including the 40 morphs), and then Z-scored across stimuli, so that the average pixel-based distinctiveness among all faces was 0. Also see Figure 2.9 for a comparison of stimulus distance metrics among the base 40 faces, and Figure 2.35 for a comparison of stimulus distinctiveness measures among the 40 base faces.

(2) in-scanner distinctiveness 1-6 distinctiveness ratings provided in the scanner by each subject were Z-scored within each scan, so that the average distinctiveness of a face seen by a particular subject in one particular fMRI scan was 0. This value was then averaged together across all unique presentations of the face.

(3) pixel+scanner distinctiveness (default measure used, i.e., unless otherwise stated) In order to simultaneously capture distinctiveness as perceived instantaneously by each subject entering distinctiveness ratings in the scanner, and also distinctiveness as determined solely by the stimulus content, we added together the across the extended set of 80 faces, and the subjects' distinctiveness ratings (which were also Z-scored), into a combined **pixel+scanner** distinctiveness. This combined metric correlated very highly with both the scanner ratings (0.84) and the pixel-based ratings (0.94), so results based on either one of those two constituent metrics alone were very nearly identical to the ones obtained with the combined pixel+scanner metric (See Figure 3.31).

(4) psychophysics-based distinctiveness (a.k.a. "psychophys.-based" or "psych.-based") This distinctiveness measure is based on the measure called Dissimiliarity-ratings-based distinctiveness ("dissim. MDS") in Chapter 2 (Section 2.10.6.2). It is based on the psychophysics experiments in that chapter, wherein subjects explicitly provided similarity (equivalently, dissimilarity) ratings: experiment SimRate. To extend this distinctiveness measure to the morphed faces (which were not compared in Chapter 2), the x- and y- MDS coordinates of each constituent face were summed together, and the resulting face's distance from the origin was taken to be its distinctiveness. Therefore, the average face was given a psychophys.-based distinctiveness of 0. Note that this can be done averaged across subjects (that is, dissimilarity ratings are first averaged then the MDS is performed), or on each individual subject's ratings.



	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
others' ratings	0.65	0.86	0.57	0.77	0.81	0.75	0.82	0.84	0.75	0.47
pixel-based	0.41	0.54	0.27	0.49	0.63	0.53	0.57	0.5	0.45	0.31
pixel+scanner	0.76	0.86	0.78	0.82	0.89	0.85	0.87	0.85	0.83	0.76
psych.-based (ave)	0.51	0.54	0.3	0.45	0.45	0.37	0.49	0.46	0.36	0.094
psych.-based (own)	0.48	0.57	0.22	0.45	0.53	0.27	0.48	0.4	0.43	0.24

Figure 3.32: The correlation between each individual subject's **in-scanner** distinctiveness ratings and another metric is shown. The top row corresponds to the average in-scanner distinctiveness rating among the 9 other subjects.

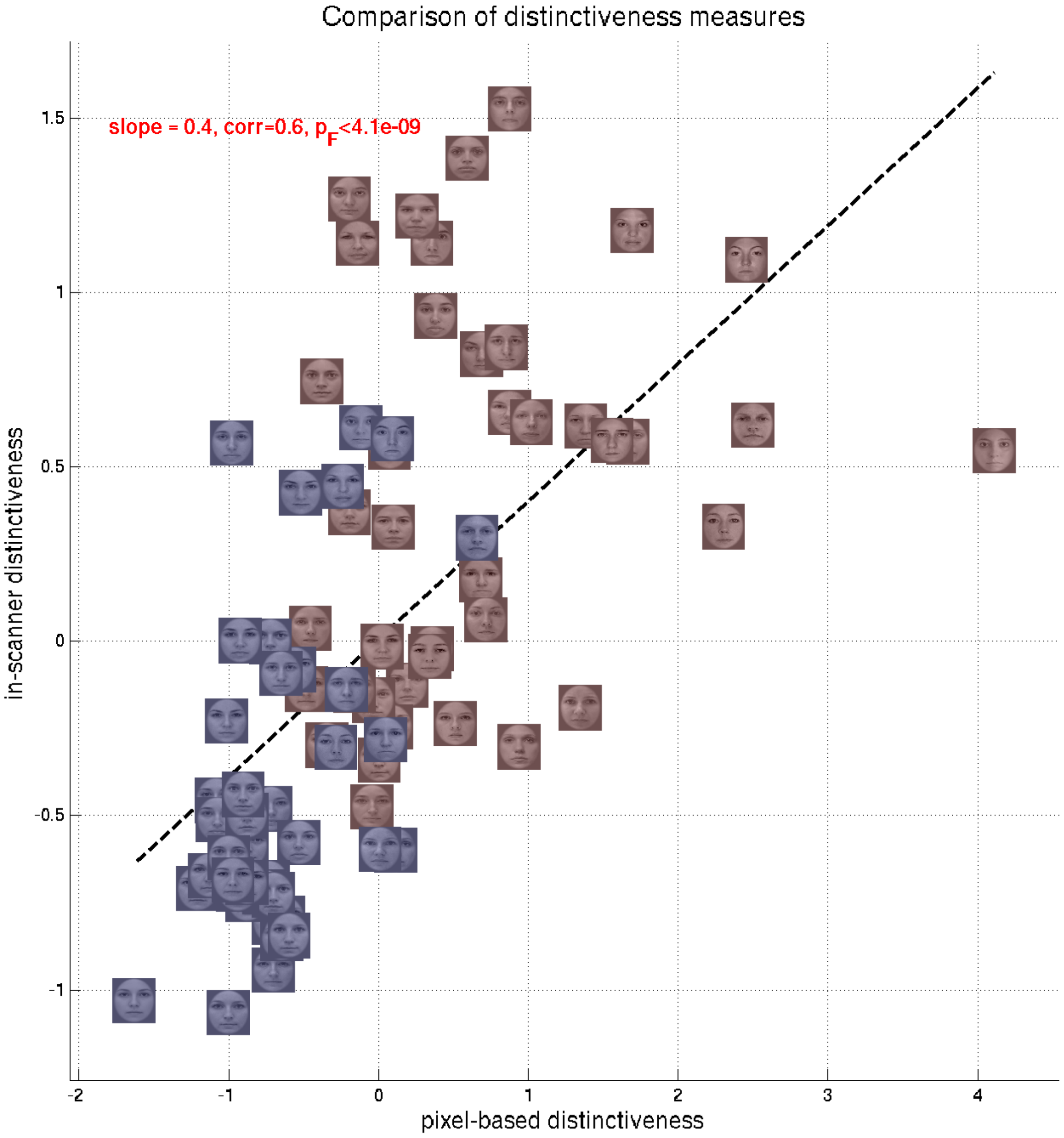


Figure 3.33: The correlation between in-scanner distinctiveness (averaged across subjects) and pixel-based distinctiveness is 0.6 (each is Z-scored). This corresponds to one (symmetric) matrix entry in Figure 3.31. Morphs are shown shaded blue, and base faces are shown shaded red. The morphs have lower distinctiveness values in both metrics.

3.9.3 MRI data acquisition and preprocessing steps

MRI data were collected using a Siemens (Erlang, Germany) 3T Trio. BOLD functional (T2*-weighted) images consisting of 42 slices with 3×3×3 mm voxels were collected with a TR (repetition time) of 2.25 seconds using a 12-channel Head coil (acquisition matrix 64×64, flip angle 80°, echo time 30 ms). Slices were obliquely oriented at 30° for near whole-brain coverage; occipital and temporal lobes were always fully inside the imaged volume. High-resolution (1 mm isotropic) anatomical images were acquired using a T1-weighted MPRAGE (magnetization-prepared rapid gradient echo) sequence. All visual stimuli were projected onto a rear-projection screen visible from within the MRI scanner via an angled mirror.

Functional volumes were slice-time-corrected, self-motion corrected (under rigid-body transformation), and then aligned to a single reference scan for each subject independently. Alignment to reference scans was done in two steps: first using a 12-dimensional affine transformation, then using a highly regularized nonlinear adjustment to slightly improve the fit. These steps were carried out using the FSL software suite (version 4.1.3, http://www.fmrib.ox.ac.uk/fsl/).

3.9.3.1 Siemens Physiologic Monitoring Unit (PMU)

A pulse oxymeter attached to end of one of the subject's fingers and a respiratory belt wrapped around the subject's midsection were employed simultaneously with every fMRI scan. Data from each were recorded at 50 Hz. Because the oxymeter kept moving around on subjects' fingers, the data from it were too unreliable and thus completely discarded.

3.9.3.2 Regressors used by SPM in the general linear model in order to obtain "betas"/response estimates

Beta estimates under the general linear model were calculated using the Statistical Parametric Mapping toolbox for MATLAB (SPM8 version 4010, 21-Jul-10, http://www.fil.ion.ucl.ac.uk/spm/) with standard settings (HRF 32.2 seconds long, convolution order 1). Notably, we did *not* spatially smooth functional values for these estimates. We ran two versions of the general linear model for each fMRI scan volume sequence: one which provided beta (response) estimates for each individual face trial, and one which provided estimates only for each unique face (of which there were multiple trials in a scan, using information from all them). Catch trial betas were estimated, but were subsequently completely ignored in the scientific analysis. In the case of the functional localizer scans, instead of one estimate per *face*, we obtained one beta estimate per *condition* (e.g., "houses" or "faces" or "objects"), of which there were only two per scan.

For each scan, we used the following 15 regressors: (1) two self-motion regressors, based on the head motion parameters estimated within each scan, (2) three slightly time-shifted versions of curve traced out by interpolating the local maxima of the respiratory series (see [7, 8] for a justification of this technique), (3) six time courses computed from within the same brain scan, five of them from CSF (ventricles) or white matter (individually hand-selected locations), and the remaining one was the sum of all of these, (4) two polynomial drift regressors: linear and quadratic Legendre regressors, (5) one "start anomaly" regressor, consisting of 1s for the first two volumes, and 0s thereafter, (6) one "end anomaly" regressor, consisting of 1s for the last two volumes, and 0s before.

3.9.4 Function-based region localization

3.9.4.1 Stimuli and task

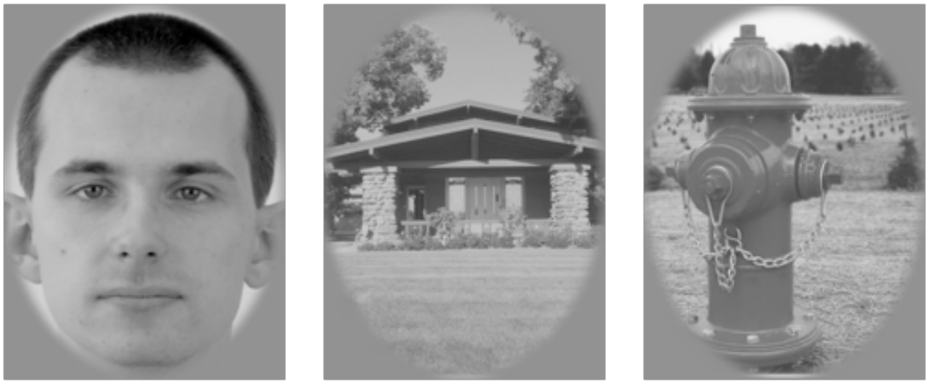


Figure 3.34: Example stimuli from the functional localizers. *Left*: example face from the PUT face database [51]. *Center*: example house from the Pasadena Houses database [41]. *Right*: fire hydrant, from the collection of 40 object images.

Subjects each participated in two functional localizer scans in addition to the FaceView scans: (1) Faces vs. Houses, and (2) Faces vs. Objects. A block-design was employed. Each localizer scan consisted of 12 alternating stimulus blocks, each with either only face stimuli, or only stimuli from the other visual category. Blocks were presented in alternating order starting with faces. Each block consisted of 20 trials, during which a stimulus was presented for 300 ms, followed by a 450 ms fixation-square pause; thus, each block lasted 15 seconds. Before the first block, in between blocks, and after the last block, there was 15 seconds of no-stimulus, fixation-square time to allow for hemodynamic rest. Either 5 or 6 times in each such localizer run, a subject would see a stimulus presented twice consecutively, the only such times this would occur. Subject were instructed to key in a button on their controller when they noticed this happen to ensure that their attention was kept high. Based on their response accuracies, subjects were all sufficiently alert for all localizer scans.

Forty emotionally neutral face images were selected from the PUT Face Database [51], forty house images were selected from the Pasadena Houses 2000 database [41], and 40 images of household objects were collected by the author from Google Images[6]. All images were converted to grayscale, normalized to have equal mean and standard deviation in luminance, then pixel values less than .01 away from black or white were set to .01 away from black or white, respectively, then the image was masked with an ellipse (same for all images), approximating the shape of a face. Luminance values outside the mask were set to 0.5. Images were 300 pixels in width and 400 pixels in height, and centered on a 1024×768 display, with all background pixels colored in at gray 0.5. Images subtended approximately 6°×8° of visual angle as viewed by participants in the scanner.

3.9.4.2 Determining Face-Selective Clusters

After alignment to the reference functional volume, functional images from the localizer scans were smoothed with a spatial Gaussian kernel having FWHM 3.0 mm. SPM was used to obtain a t-statistic and p-value at each voxel associated with the hypothesis that sum of the face responses was equal to the sum of the house(/object) responses at that voxel. Spatially contiguous clusters of voxels, for which the t-statistic was positive (corresponding to a greater response to faces), and with significant p-value, were estimated, such that no cluster contained more than 1000 voxels nor fewer than 12. This strategy reliably identified clusters in regions such as fusiform gyrus, and LOC. FFA was taken to be the union of face-selective clusters overlapping with the fusiform gyrus of either hemisphere. The face-selective clusters for each subject, including those defined as FFA, are shown in Figure 3.35.

3.9.5 Anatomy-based region localization

Each subject's anatomical (MPRAGE) scans (typically 1 per imaging session) were aligned with each other (rigid body) and averaged together to produce a single high-fidelity anatomical image per subject. This image was then parcelled into distinct cortical and subcortical regions, each given a descriptive label, using the FreeSurfer automated surface reconstruction software[7]. Using an affine transformation and regularized nonlinear adjustment (with FSL tools), anatomical images were registered with the reference functional image for each subject. This transformation was used to create masks for FreeSurfer-labeled anatomical regions in the reference-functional space of each subject. An example of FreeSurfer's output is shown in Figure 3.36.

[6] http://images.google.com, Google Inc. (Mountain View, CA)

[7] http://surfer.nmr.mgh.harvard.edu/

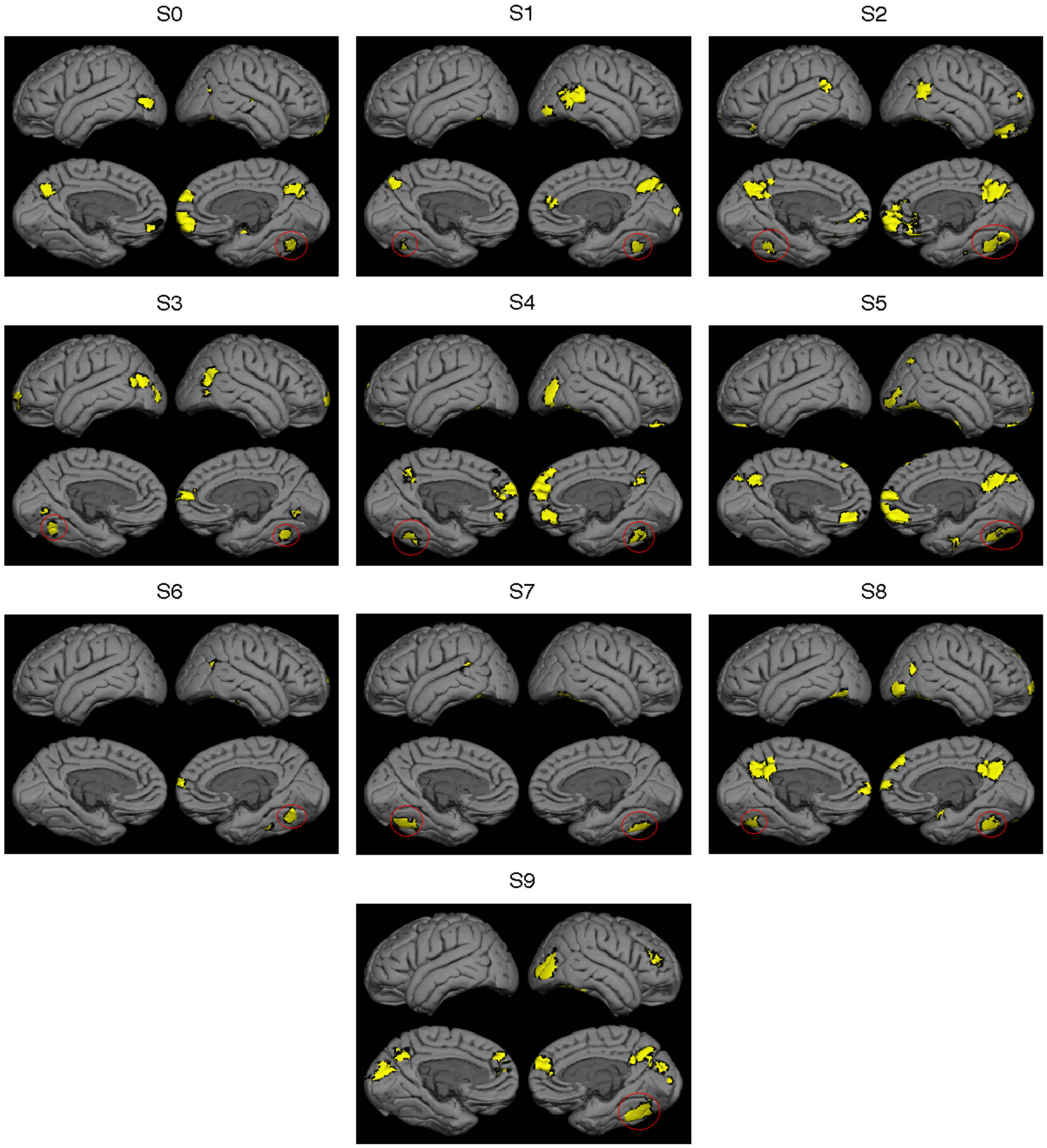


Figure 3.35: The face-selective clusters for each subject are shown (the fusiform face areas, FFAs, are indicated with red circles). Clusters were computed in each subject's native anatomical space. They are shown here projected, via nonlinear warp, onto the MNI standard cortical surface for illustrative purposes only. This warping of subjects' functional data into MNI space was used nowhere else except to make this figure.

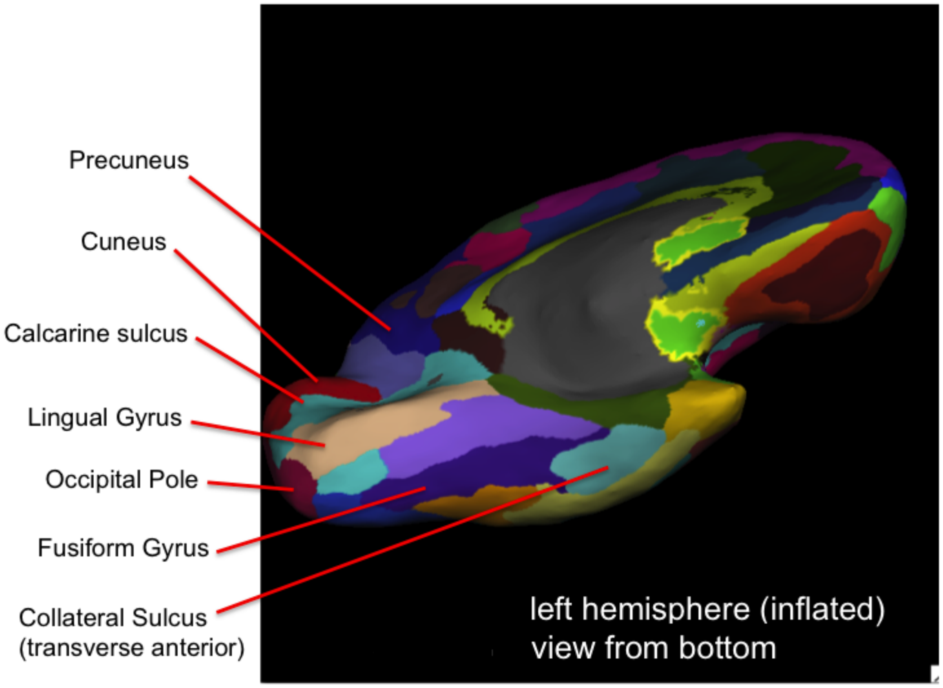


Figure 3.36: A colored-coded output of FreeSurfer's cortical parcellation for one subject, represented on an inflated left hemisphere. V1 is located in the calcarine sulcus and FFA is located in the fusiform gyrus.

3.9.6 How we compute the activity magnitude in a region

We begin with an estimate, at each voxel, of an activity magnitude (which we sometimes refer to as an amplitude, a name due to associations with a hemodynamic response fluctuation) for a particular stimulus. These are the raw "beta" estimates computed by the SPM software. The median value in the voxel set is then computed (to reduce the effect of outliers), and then the series of median values within a single fMRI scan is Z-scored (so the mean activity magnitude, across stimuli or presentations, is set to 0 within a scan for a particular region). The Z-scoring was done to better facilitate comparison across brain regions, which may have different underlying imaging signal-to-noise levels or hemodynamic properties independent of functional selectivity.

3.9.7 How we compute the consistency value in a region

The neural pattern consistency of a particular face stimulus was computed between two sessions (median 2.24 days apart), within which each face was seen on an average of 4 trials, and within each session across these trials. When computing cross-session consistency, the response estimates used were the ones in which there was only one estimate per unique face in a scan. Because such estimates pool information across multiple trials, they are relatively higher in fidelity. When computing cross-trial consistency, the beta estimates used were the ones in which each individual face trial was assigned a unique beta value; these are relatively lower in fidelity.

We represent a brain response pattern to a particular stimulus s at session/trial t as a vector $v_{t,s}$. It has one entry per voxel in the target region, containing the raw beta estimate of the response at that voxel. The unidirectional neural consistency of the response pattern to a face stimulus f between two times A and B was calculated as follows:

$$C_{A \to B}(f) = \frac{1}{|F|} \sum_{g \in F} corr(v_{A,f}, v_{B,f}) \geq corr(v_{A,f}, v_{B,g}) \tag{3.1}$$

where $\geq$ is an operating yielding one if the condition is satisfied and zero otherwise, and $corr(x, y)$ is the Pearson correlation coefficient between vectors x and y. F is a set of "distractor" faces (possibly with repetitions, in the case of cross-trial consistency) shown at some other time B, and including exactly one unique presentation of f. Thus, the consistency $C_{A \to B}(f)$ is the *fraction of distractor faces at B less neurally correlated with the response of f shown at A than f itself is at B.* This is illustrated graphically in Figure 3.37. This raw consistency value can thus take on values in $[0, 1]$. It is logically equivalent to a scaled rank-correlation metric. This calculation is also carried out the in other direction in time, and the consistency of a face f between times A and B is given by:

$$C(f) = \frac{1}{2}\left(C_{A \to B}(f) + C_{B \to A}(f)\right) \tag{3.2}$$

Note on cross-trial consistency: To compute the consistency across single trials within a session, all unique pairs of trials of the same face within each session were used to compute a consistency score as defined in 3.2 (with all single trial responses to other faces as distractors), and finally the single-trial consistency was given as the average across all such pairs, then averaged again across the two sessions. An important technical note is that in computing the single-trial consistency, the set F in Eq. 3.1 was modified to exclude those faces presentations within 5 or fewer trials (out of 87) of$f$. That is because the neural response pattern to faces within a small window of time (roughly +/- 45 s) was self-correlated (due to, e.g., more similar head position, related hemodynamic state, more similar cognitive state, etc.); thus, trials within this window were discarded so that consistency was computed relative to only those distractor trials sufficiently far away in time.

Unified consistency (x-sess+x-trial): Finally, for each face in the stimulus set, and for each subject, we combined the cross-session ("xsess") and cross-trial ("xtrial", within a single session) consistency measures into a single unified pattern consistency as follows:

$$Consistency = \frac{1}{2}(C_{xsess} + C_{xtrial}) \tag{3.3}$$

This final unified consistency metric thus reflected the neural pattern consistency across both days and minutes, with equal weighting for each category of delay. However, due to the empirical distribution of these values, as we discuss below in Section 3.9.10, the cross-session data actually weighed more heavily in practice.

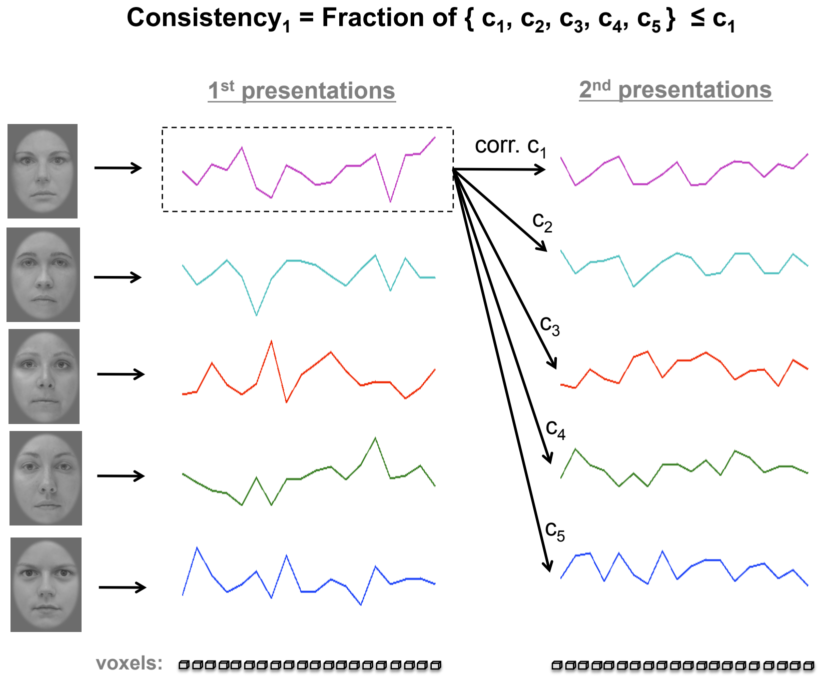

Consistency$_1$ = Fraction of { c$_1$, c$_2$, c$_3$, c$_4$, c$_5$ } ≤ c$_1$
1st presentations
2nd presentations
corr. c$_1$
c$_2$
c$_3$
c$_4$
c$_5$
voxels:

Figure 3.37: An illustration of the unidirectional consistency calculation for Face 1 from first to second presentations, in an example with only 4 other/distractor faces. The response patterns to each face over the voxels are illustrated as colored plots, and the correlation between these across voxels is labeled as $c_1, c_2, ... c_5$.

3.9.8 ROI-wise and Map-type comparisons: comparable regions across subjects and the MNI surface maps

We compared both magnitude and consistency in regions across subjects. We did this in two ways:

1. **ROI-wise comparisons:** Using comparable regions of interest (e.g., as shown in Figure 3.5)

2. **Map-type comparisons:** Using comparable clusters of voxels (e.g., as shown in Figure 3.3)

ROI-wise comparisons These are facilitated simply by using voxel sets defined via anatomical parcellation or functional localization as described above. They are computed in **each hemisphere independently then averaged together**. So, for instance, a consistency value of 0.3 in LO means that the consistency in left occipital cortex, averaged together with the consistency in right occipital cortex, was 0.3 standard deviations above average (as all reported values are Z-scored). This per-hemisphere-then-average strategy is also used for functionally defined ROIs: FFA, and "Face" (the union of face selective clusters).

Map-type comparisons To facilitate local cross-brain comparisons smoothly across the whole brain, while still carrying out all magnitude and consistency computations in the subjects' native, unwarped, unsmoothed functional spaces, we developed a technique for obtaining anatomically similar sets of voxels spanning the entire brain. First, for each brain hemisphere, we define a set of 77×7 "spheres" of N voxels each in each subject's native functional space (by default we used $\mathbf{N = 300}$, though we also tried $N = 100$ and $N = 200$ with similar results). The spheres are not actually spatially "spherical", but have a sphere-like property: they consist of the N nearest *cortical* voxels to a reference center voxel*, restricted to lie in the same hemisphere as the reference.* The first 77 centers, inducing 77 spheres, were defined simply: one per parcellated anatomical region[8]. The x-, y-, and z-coordinates of these centers were taken to be the median in each direction among voxels in the region. The next 6 sets of 77 were defined by, for each region, stepping $\frac{1}{4}$ its diameter (distance between its two most distant constituent voxels) in each of the 6 cardinal brain directions: left, right, inferior, superior, anterior, and posterior. These direction vectors were calculated for each subject's reference functional volume taking into account their exact head position in the scanner. Thus, for each of the 77 parcellated anatomical regions, we obtained 7 overlapping spheres for each hemisphere. An illustration of this overlapping is shown in Figure 3.38. We also parcellated the MNI152 template brain (Montreal Neurological Institute atlas, originally based on the work of Collins [16]), and computed the $77 \times 7 \times 2$ (2 for hemispheres) spheres for it as well.

The $77 \times 7 \times 2$ **(1078) spheres become comparable ROIs:** Just as with ordinary ROIs, we could then, within each one of these spheres, perform calculations across sets of subjects pooled together. For instance, we could compute the average consistency for familiar faces among all subjects in one such sphere (e.g., "calcarine sulcus, step anterior"). This allows us to assign a value to each such sphere which is based on data from all 10 subjects.

Value at voxel is average among spheres which contain it: Given a subject-pooled value

[8] in the Destrieux Atlas, called "aparc a2009s" by freesurfer

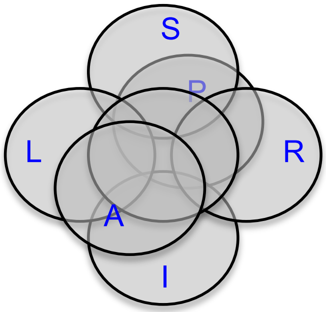


Figure 3.38: Illustration of the overlapping set of 7 spheres per brain region with one in the center and 6 for each of the cardinal brain directions: (L) left, (R) right, (I) inferior, (S) superior, (A) anterior, (P) posterior. Actual voxel "spheres" were not geometrically spherical as in the illustration.

at each of the 77 × 7 × 2 spheres, we could color a whole brain with data. We do this, without exception, in the MNI152 template space. The spheres span the whole brain, covering all cortical voxels, and so each voxel could take on a value. The value at that voxel is the average among the spheres which contain it. Not only are spheres within a region overlapping, but neighboring regions have overlapping spheres, so the data at each point is based on multiple estimates: for $N = 300$, the 25th, 50th, and 75th percentile of number of spheres per voxel are 4, 7, and 10. The maximum number of estimates per voxel among all voxels is 32. For comparison, for $N = 100$, the 25th, 50th, and 75th percentiles are 1, 2, and 4 estimates, and the maximum is 16. Most of our analysis is based on $N = 300$.

p-values are averaged in log space: When the value at each sphere is a p-value, we average together the $-log_{10}p$ values and map those.

MNI surface maps: Having a value for each voxel in the MNI template space, as computed with overlapping spheres, we project that data onto the MNI cortical surface, using freesurfer. This results in images like the those in Figure 3.3. We computed both pial (used in this chapter) and inflated surface maps (shown in Figure 3.36, but not used to show actual data). While pial maps have immediately recognizable locations due to their textbook-like appearance, they technically hide information from deep within sulci. The inflated maps contain more information, because they reveal the sulci, but they are also harder to read due to lack of easily recognizable anatomical landmarks. After examining results in both views, we concluded that the pial surfaces were better suited for illustration, especially when complemented by text and ROI-wise figures.

3.9.9 Residualized and Z-scored magnitude and consistency measures

The magnitude and consistency measures described above suffered from dependence on artifactual experimental factors of no scientific value. Therefore, to minimize the influence of these on the results, these factors were residualized out, as follows:

$$w^* = min_w||[M\,\mathbf{1}]w - v||$$

where v is the column-vector of experimental values in an ROI (for example, one consistency per face, per subject), and M is the matrix of regressors, one per face, per subject. The minimization is carried out via standard linear regression. Then the residualized values were defined as:

$$v' = v - [M\,\mathbf{1}]w^*$$

Finally, the values in v' were Z-scored relative to all the other values in v'.

For map-type comparisons of magnitude, the following 9 regressors were used for each subject for each face: (1) the standard deviation of in-scanner distinctiveness ratings, with one value per session, (2) the standard deviation of the in-scanner distinctiveness ratings, within a single session, then averaged across the two, (3) the number of presentations of the face in the first session, (4) the number of presentations of the face in the second session, (5) the minimum delay between two presentations of the face in the first session, (6) the minimum delay between two presentations of the face in the second session (both in terms of number of trials), (7) the time between the first session and the second session (in days), (8) the average instantaneous head motion (absolute sum of parameters) while the face was being presented in the first session, (9) the average instantaneous head motion during the second session.

For map-type comparisons of consistency, the magnitude of response was also regressed out (see below).

For ROI-wise comparisons, the regressors for magnitude and consistency were the same as above, with the addition of the number of voxels in the region (since that varies for ROI-wise type comparisons, but not for map-type where $N = 300$ for all spheres).

The reasoning behind this selection of regressions was that we wanted to minimize the effect of subjects simply being more consistent in their button presses on the consistency of the brain responses (though in sensorimotor cortex, the neural similarity observed between any two button press movements using the same hand is high independently of this, see Section 3.9.11), and we also wanted to account for the fact that faces which were seen more times or more closely together in time would be more consistent, and also attempt to minimize the influence of changes in neural measures due simply to head motion at the moments the face was being presented.

NOTE – Consistency is magnitude-residualized: Unless otherwise stated, the consistency measures presented everywhere in this chapter are magnitude-residualized, meaning that the boost in or depression of consistency due to simply more or less aggregate activation in a region (due to more or less SNR, assuming constant noise level) is accounted for in this way and subtracted out.

3.9.10 Comparison of consistency measures

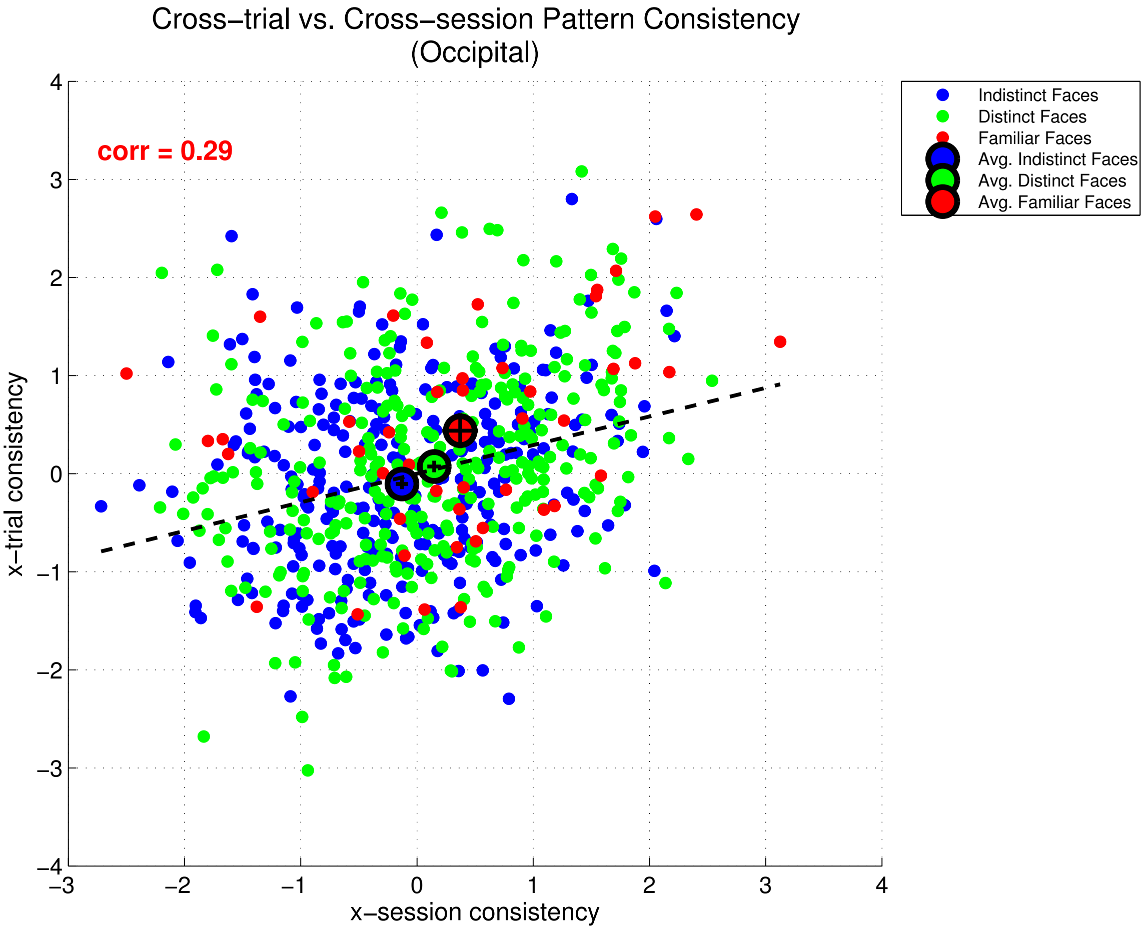


Figure 3.39: Averaged across voxel spheres in occipital cortex, the raw (unresidualized) cross-session consistency values correlate with the raw (unresidualized) cross-trial consistency values at a level of 0.29. They correlate with the raw (unresidualized) combined (x-sess + x-trial) consistency values at a level of 0.94, and 0.59, respectively. We show one data point per face (80), per subject (10), for a total of 800 points. Note that in terms of raw consistency values, on average (bigger circles): Indistinct Faces < Distinct Faces < Familiar Faces.

Because the results using cross-trial and cross-session consistency values were similar, we simply combined them to gain statistical power in our analyses, by averaging them together (before residualization + Z-scoring), see Equation 3.3. Because the cross-session consistency values were on average much higher (due to being based on multiple trials rather than just one, hence much less noise), the sum was dominated by the cross-session consistency: whereas averaged across all voxel spheres (not just occipital), the residualized combined consistencies correlate (across subjects & faces

= 800 points) with the residualized cross-session consistencies at a level of 0.95, they only correlate with the residualized cross-trial consistencies at a level of 0.40. This relatively higher weighting of the cross-session data which emerges naturally is desirable as it is more reliable and based on more data points per consistency calculation.

3.9.11 Neural pattern consistency in sensorimotor cortex

Suppose a face is rated for distinctiveness, across two presentations, by the same hand of a subject (which would occur if they are rated in {1,2,3} or {4,5,6} but not across these). Consider the pattern of activity in sensorimotor cortex in the contralateral hemisphere (the one controlling the hand). Due to the hand sensorimotor neurons being relatively activated compared to other motor neurons (e.g., foot, leg), and in relatively the same way (subject makes a button press – irrespective of which button they press, the motor action is nearly the same relative to the available space of motor actions), the pattern of activity across this sensorimotor cortex will be highly correlated across these two presentations. Now consider the neural pattern correlation in this sensorimotor cortex (same hemisphere), except across two presentations of a face rated by the *other* hand. Because the hand this part of cortex controls is not moving, and the subject is relatively still, this part of cortex is simply in some kind of baseline activity, which is not at all correlated across trials. Thus, if we Z-score the neural pattern consistency across different faces in this hemisphere, we will find a clear trend: faces rated by the hand it controls are above average, faces rated by the hand it does not control are below average. Furthermore, this is more or less independent of the variability in the actual ratings because a hand movement compared to no hand movement is a much greater difference than two different types of hand movements.

Due to the fact that buttons corresponding to distinctiveness ratings {1,2,3} were pressed with the *left* hand, indistinct faces had above-average consistency in sensorimotor cortex in the *right* hemisphere, and because buttons corresponding to distinctiveness ratings {4,5,6} were pressed with the *right* hand, distinct faces had above-average consistency in sensorimotor cortex in the *left* hemisphere.

Part II

Neural pattern similarity across individuals

Chapter 4

Different individuals exhibit similar neural pattern distance between visual objects

In this first part of the thesis, we focused on individual differences in the perception of just one category of visual stimuli, faces. In this chapter, we extend our investigation beyond this one category, and present experimental results aimed at quantifying inter-subject neural pattern similarity in response to a very broad range of visual object categories. Subjects in an fMRI scanner were shown static images, selected from 44 categories (spanning life (including faces), artifacts, food, and places), one at a time. We find that the distances between these categories, induced from activity in cortical visual object areas, correlate highly between subjects, and also to distances inferred from a behavioral clustering task, and that this correlation remains significant even among subsets of closely related categories. We also show that one subject's brain activity can be accurately modeled using another's, and that this allows us to predict which image a subject is viewing based on his/her brain activity. We conclude with a discussion of the possibility of using brain-based features to aid in computer vision, with a small experiment demonstrating an enhancement in recognition performance using these.

4.1 Previous work

The approach we take to comparing subjects' associations between categories, using representational dissimilarity matrices, was also used by Kriegeskorte, in a paper published in the journal *Neuron* [56] in 2008, for a different purpose. He compared object representations in humans to those in monkeys. The most closely related work to our own on modeling subjects' responses to individual images was carried out by Shinkareva et al., reported in *PLoS One* in 2008 [83]. In her study, subjects viewed five distinct images (line-drawings) from each of two high-level categories: tools and dwellings. Responses in each subject's brain were predicted using models built on all the other subjects' responses across the whole brain. In this way, it was possible to decode which image a subject was viewing based on his/her brain activity. The results were provided in rank accuracy: a value of 1 means the predicted image was always the first in the list of candidates, 0.5 means it was halfway down on average. Among 10 subjects, the average performance in this metric ranged from 0.6 to 0.94. We will present results on using a similar approach to decode both image and category identity, and discuss the information content of voxels which are well-modeled across subjects.

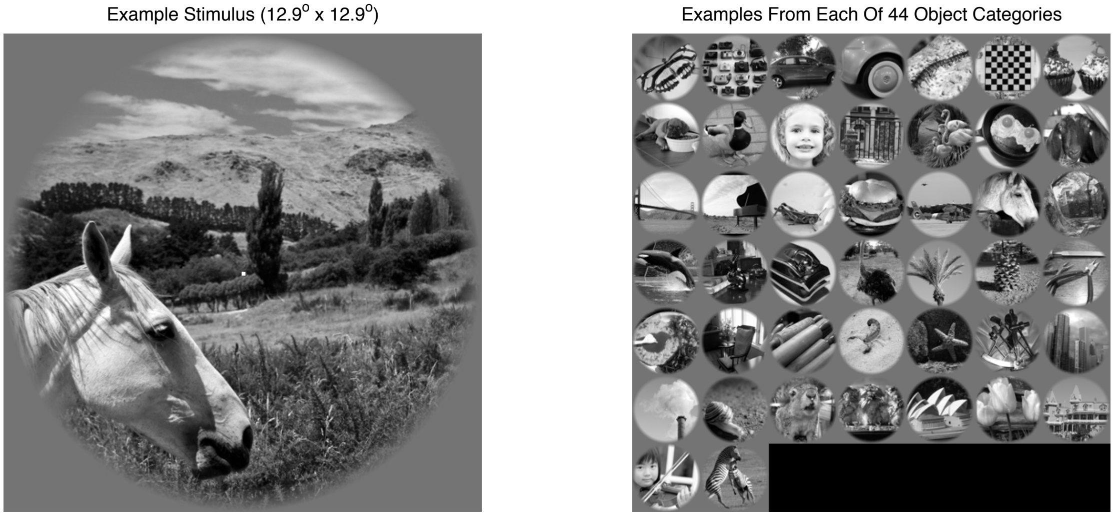
Although decoding information based on cross-brain models is quite novel, decoding fMRI responses based on stimulus features has been a very active area of research in the last decade. Some of the most important studies include Cox and Savoy in 2003 [17], Kamitani and Tong in 2006 [48], Kay et al. in 2008 [53], Mitchell et al. in 2008 [64], and Reddy et al. in 2009 [76]. Cox and Savoy were able to decode which of 10 image categories were being viewed very high accuracy, Kamitani extended this to decoding which stimulus a subject was *attending* to among simultaneous candidates, Kay greatly improved decoding accuracy using a mixed Gabor model of voxels, Mitchell presented decoding results based on a semantic feature approach, and Reddy found above-chance decoding of *imagined* visual stimuli.

4.2 Basic experimental setup

Three healthy adult subjects, KJS, JSB, and XH (1 female, ages 25-31), viewed 440 unique images (repeated non-successively 2 or 4 times), 10 from each of 44 varied categories spanning life forms, artifacts, foods, and places in an fMRI scanner, while instructed to fixate on a central point and key in category information with a controller. See Figures 4.1 and 4.2 for details about the stimuli.

4.2.1 Five regions of interest in the brain: EV1, EV2, LOC, VO, Fusi

Five regions in the occipital and temporal visual cortex of each brain were selected for investigation: (1) "**EV1**", for early visual 1, consisting roughly of primary visual cortex and regions close to it (approximately *V1-V2*), believed to represent elementary visual features such as edges [45], (2) "**EV2**", for early visual 2, regions higher in the visual pathway and immediately adjacent to EV1 (approximately *V3-V4*), associated with intermediate visual features such as shape [70], (3)



Example Stimulus (12.9^o x 12.9^o)

Examples From Each Of 44 Object Categories

Figure 4.1: The image stimuli: the subject of each image was generally not perfectly centered, and set against a natural nonuniform background. No schematics or line drawings were used, only pictures of actual items. Some of the images were selected from the Caltech 256 database [34], and the rest were selected from flickr (http://www.flickr.com). Images occupied roughly 12.9° × 12.9° of visual angle. Each image was 500 px × 500 px, and circularly faded into a uniform gray background as seen in the left panel. The background had gray value equal to the average brightness of all the images. The circular fading consisted of linearly scaling the last 10% radially to background. Images were contrast normalized such that the lowest 0.1st percentile and highest 99.9th percentile of their brightness values were stretched to absolute black and white respectively. *Left*: Example stimulus from "horse" category. *Right*: Example stimuli from each of the 44 categories.

life	artifact	food	place
butterfly	camera	cupcake	golden gate bridge
centipede	car side	fried egg	iron gate
dog	car tire	hamburger	skyscraper
duck	chessboard	pupusa	smokestack
face	fire hydrant	sushi	sydney opera house
flamingo	grand piano		victorian house
goat	helicopter		
grasshopper	microscope		
horse	nail clipper		
killer whale	pliers		
ostrich	rocking chair		
palm tree	rolling pin		
pine cone	sextant		
scorpion	violin		
sea star			
snail			
squirrel			
tulip			
zebra			

Figure 4.2: The 44 categories in the static images experiment could be roughly categorized into four super-categories: life, artifact, food, and place.

"**LOC**" (lateral occipital complex), associated with high-level object category representation [35], (4) "**VO**" (ventral occipital) areas associated with a range of mid-to-high-level representations [10], and (5) "**Fusi**" (fusiform gyrus), associated with high-level category representations such as faces [50]. Additional details concerning the stimulus preparation, MR scanning, the estimation of a response magnitude to each image in each voxel of each brain, and the region of interest (ROI) localization (See Section 4.6.3) are provided in the Appendix to this chapter.

4.3 Inter-subject similarities in object category distances

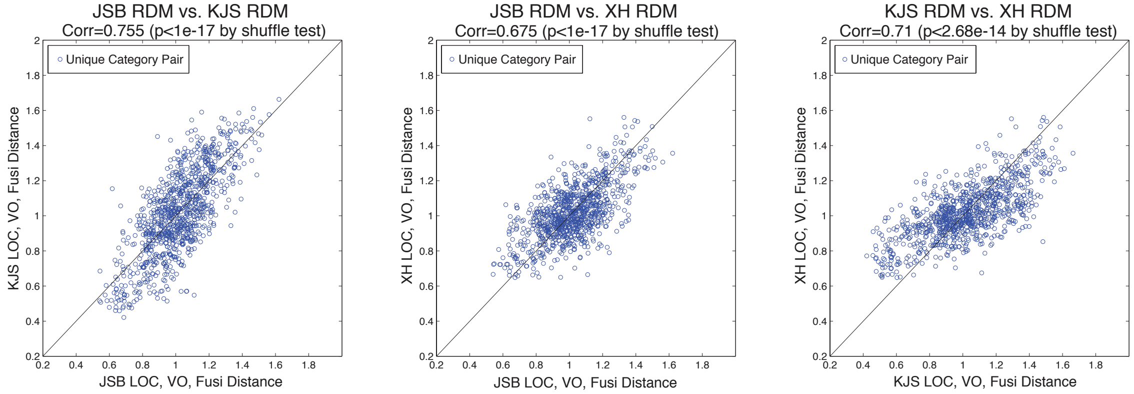


Figure 4.3: Inter-object distances, induced from responses high in the visual pathway (LOC, VO, and Fusi together), are very similar across subjects.

We know that, if directly asked, different people will provide more or less consistent accounts of which objects appear similar to them, and which look more different. It is not obvious how such similarity values are computed. One simple strategy might be that each object type elicits a unique response pattern in the brain, and that somehow it is possible to facilitate a comparison between different response patterns and ultimately report which are more similar. This strategy allows the comparison of two objects never before seen, between which a relationship has never been learned. It also implies that if two people explicitly report similar distances between objects, then somewhere in their brains we should find response patterns which give rise to similar inter-object distances. And, indeed we do.

We perform the same kind of analysis we did in Chapter 3, Section 3.6, comparing the neural distances between individual faces across subjects, except here we compare RDMs among entirely different object categories. Let $x_{s,R}(cat_i)$ be the vector of response amplitudes in the region R of subject s to category i. Each component of the vector corresponds to a unique voxel in region R. We formed a representational dissimilarity matrix (RDM) between object categories i and j in region R of subject s as follows[1]:

$$RDM_{s,R}(i,j) = 1 - corr(x_{s,R}(cat_i), x_{s,R}(cat_j))$$

where $corr()$ the Pearson correlation coefficient between the two vectors. If the two response patterns are perfectly correlated across voxels in R, then the dissimilarity value is 0, and if they are perfectly anti-correlated, the value is maximized at 2. Fig. 4.3 shows that these RDMs were very similar across subjects when formed on the region of visual cortex consisting of all LOC, VO, and Fusi voxels. In fact, their unique entries (RDMs are symmetric) correlated at a level of 0.713, averaged across subject pairs. This number is very highly statistically significant, by an empirical significance test based on shuffling the rows of one of the RDMs. The correlation drops to 0.528 when using regions EV1 and EV2, likely due to less category-specific information represented there, but remains significant.

To rule out the possibility that these inter-category correlations were driven only by the batching of categories into fMRI scans, we considered relationships among category pairs in sets of only 4 categories at a time, comparing only categories shown within the same fMRI scan, and we still found a significantly positive correlation: the mean among these, averaging across category subsets and subject pairs is 0.31, and the 70th percentile is 0.69; the population of such correlations among 4×4 RDMs is significantly to the right of zero with 1 sided t-test (p<0.0011). Of course, we expect these values to be lower not only because they control for stimulus batching but also because they compare a much smaller number of category pairs. We also ran the complementary test, where only one random category from every scan was used, and again found significantly correlated RDMs (mean=0.55, to the right of zero with t-test $p < 9 \times 10^{-8}$).

We note that these RDM correlations are much higher than what we observed among individual faces in Chapter 3 (peak of around 0.1 in visual cortex); several reasons for this may include: 1. the neural distances between different object *categories* are much greater than the neural distances between individual faces due to more distributed representation detectable at the fMRI level; 2. the number of trials per category is higher in this experiment (20 or 40, see Appendix to this chapter) than the number of trials per face in the last experiment (average 8), so the individual responses here have higher fidelity; 3. in this experiment, the individual exemplars in each category are different, possibly leading to more robust estimates, whereas in the last experiment the different trials of a

[1] we actually discard the 30% of voxels in R with lowest SNR

particular face consisted of identical stimuli; 4. here we compare representations using the bilateral set of voxels in a wide swath of visual cortex including LOC, VO, and Fusi, whereas in Chapter 3 we compare only 300 voxels from the same hemisphere at a time; 5. here stimulus order is the same for each subject whereas in Chapter 3 it was not.

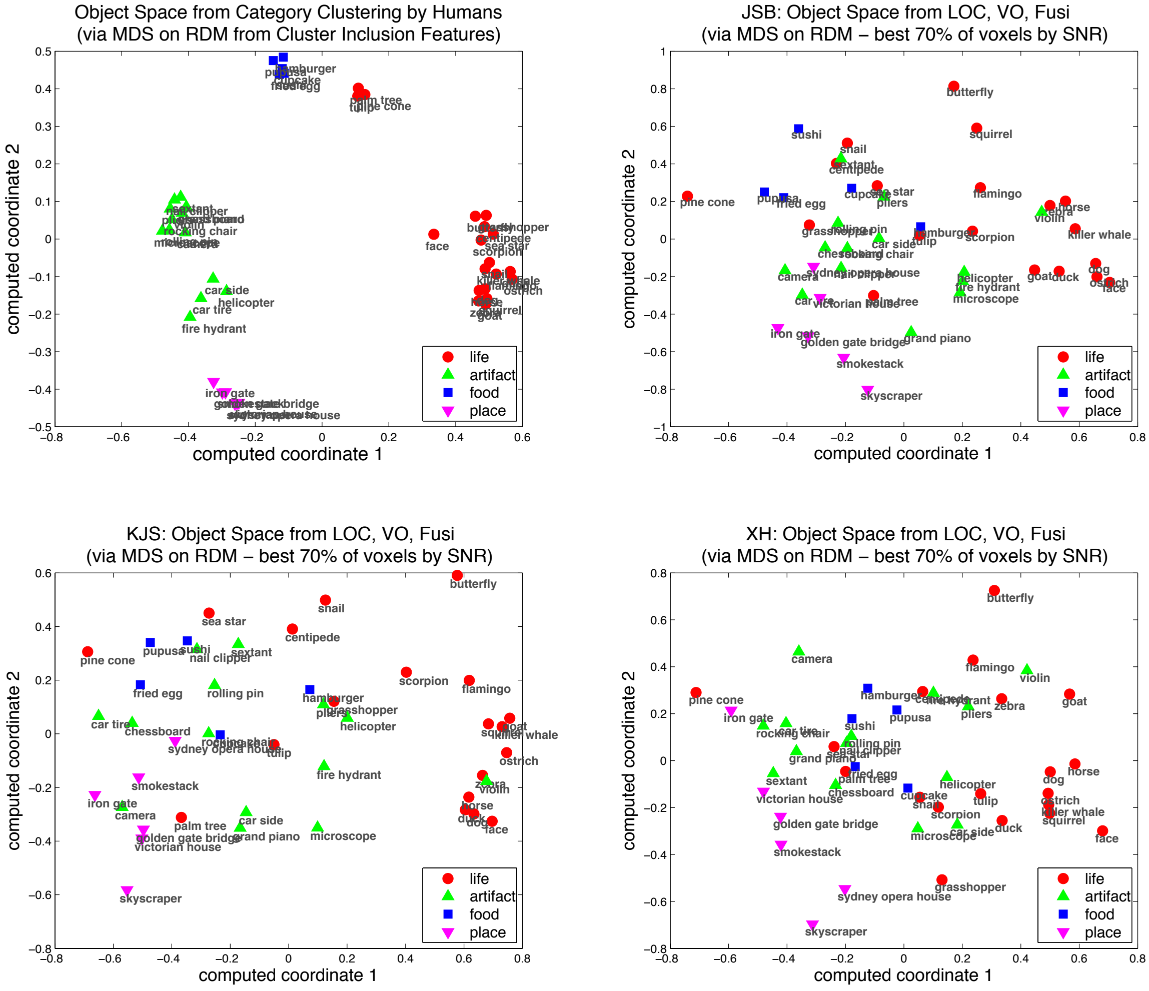


Figure 4.4: Visual relationships between object categories inferred from brain responses are similar. The dissimilarities between categories are approximated in two dimensions using MDS for each subject. *Top-left:* coordinates of objects based on explicit human clustering with a computer interface. *Rest:* coordinates of objects based on similarities between response patterns in higher visual cortex (LOC, VO, and Fusi together), for each of the three subjects. Notice that the four super-categories tend to cluster in roughly the same way, and that within each super-cluster, some structure is similar across subjects (e.g., skyscraper is always extremal).

We compared the RDMs formed from the cortical responses, as above, to an RDM between objects based on explicit human report. The three subjects in the fMRI experiment above, together with four more healthy adults, were asked to cluster the 44 categories into two groups recursively, forming a hierarchical binary tree, until either they felt no reason to separate subsets of categories

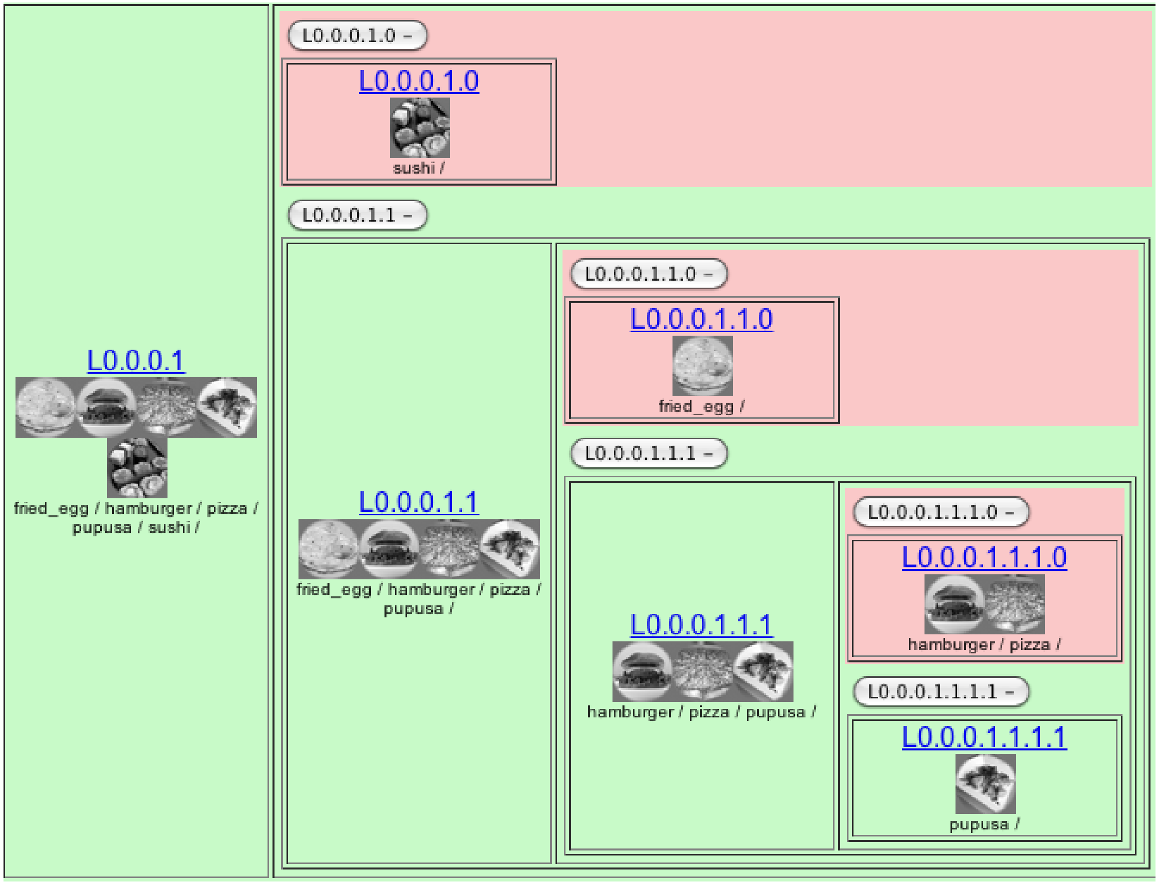
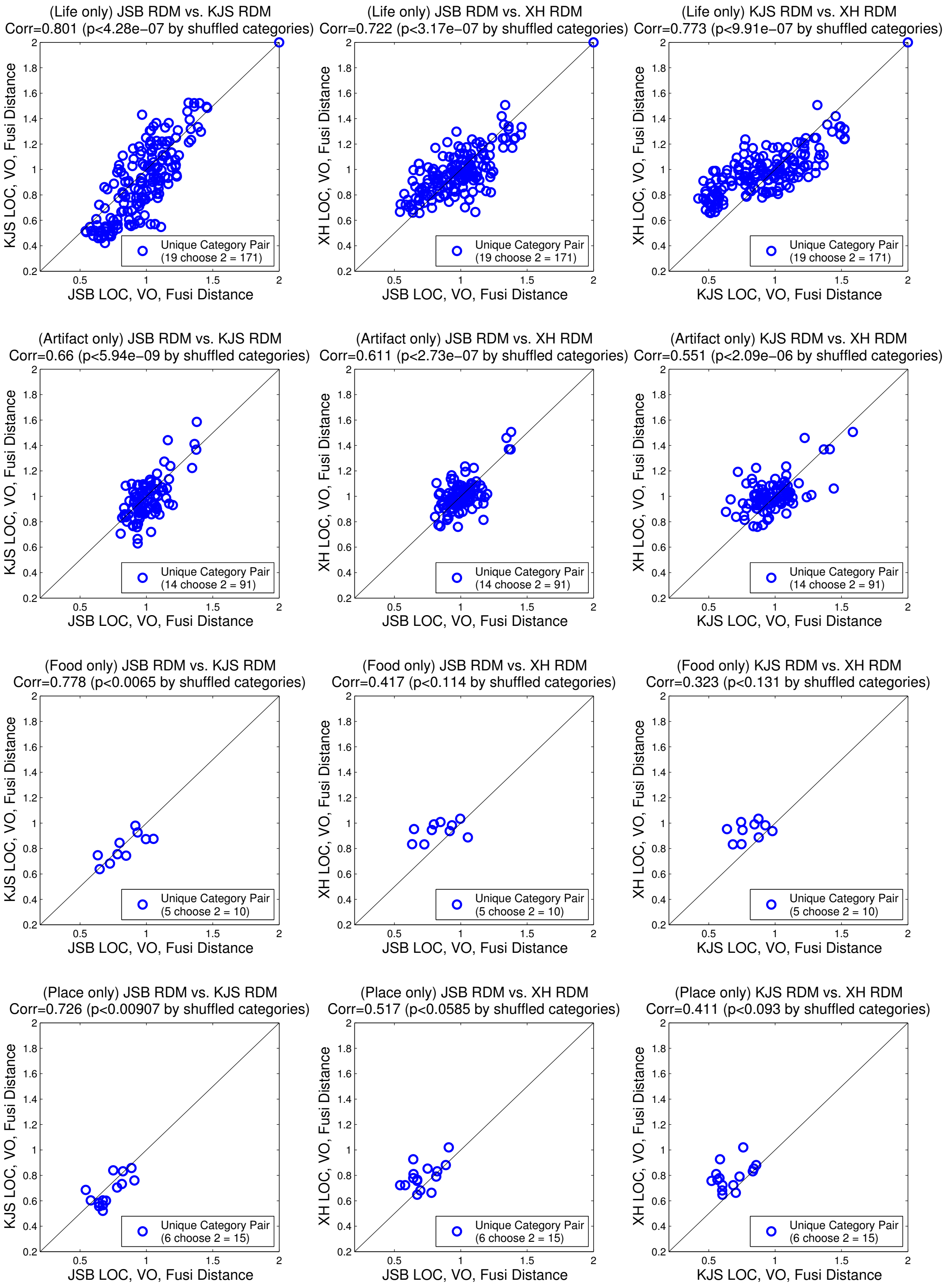


Figure 4.5: A subtree of one subject's hierarchical clustering of object categories using the web interface is shown. The full tree may be browsed at: http://www.klab.caltech.edu/~harel/res/fmri/stimuli/gui/seeallclusters.php?subj=kjs

further, or until they ended up with just a pair of categories. This task was carried out online, at home by the subjects on their own computers, using an interface like that which is shown in Fig. 4.5. They were asked to make similarity judgments based on the appearance of the categories, and could use the interface to explore the 10 exemplars in each category. Although they were asked to do this based on appearance alone, subjects uniformly created hierarchies adhering to logical, semantic relationships (e.g., "a dog and a goat are both four-legged mammals"). Then, each cluster at each level of the hierarchy of each subject was encoded as a binary column vector with 1s at indices corresponding to categories in the cluster, and 0s outside: we call such a vector a *cluster inclusion feature*. The collected set of all these vectors yields a matrix (we actually discard equivalent columns), the rows of which are feature vectors, one per category, which can be used to form an "explicit-report" RDM, based on the collective clustering of all subjects (no meaningful inter-subject differences in the clustering were found). Matlab's multidimensional scaling (MDS) algorithm, *mdscale*, was used with default parameters to obtain a set of 2-dimensional coordinates for each category based on this RDM, approximately preserving the distance relationships. That is shown in Fig. 4.4, top-left.

The correlation between this explicit-report RDM and the RDMs formed on higher visual regions (LOC, VO, Fusi together) was 0.32, averaged across subjects, and highly significant (by shuffle test). This correlation dropped to .05 and was not significant in lower visual areas (EV1 and EV2). However, the lower visual areas did yield RDMs correlating at 0.528 among subjects, meaning that somehow the structure of the similarity between object categories was very similar across subjects in these areas, but was totally unrelated to the high-level semantic relationships revealed in the clustering task: as we move higher up the visual pathway, the relationships between object categories became more consistent across subjects, and much more closely related to semantic relationships.

We investigated whether the inter-subject similarity in RDMs was governed mostly by the dissimilarity among super-categories. We found that even restricted to each super-category alone, the relationships between object categories were highly correlated among subjects, and passed significance tests except where there were very few data points available. Fig. 4.6 illustrates these inter-subject correlations among subsets of objects belonging to just one super-category.



(Life only) JSB RDM vs. KJS RDM
Corr=0.801 (p<4.28e−07 by shuffled categories)

(Life only) JSB RDM vs. XH RDM
Corr=0.722 (p<3.17e−07 by shuffled categories)

(Life only) KJS RDM vs. XH RDM
Corr=0.773 (p<9.91e−07 by shuffled categories)

(Artifact only) JSB RDM vs. KJS RDM
Corr=0.66 (p<5.94e−09 by shuffled categories)

(Artifact only) JSB RDM vs. XH RDM
Corr=0.611 (p<2.73e−07 by shuffled categories)

(Artifact only) KJS RDM vs. XH RDM
Corr=0.551 (p<2.09e−06 by shuffled categories)

(Food only) JSB RDM vs. KJS RDM
Corr=0.778 (p<0.0065 by shuffled categories)

(Food only) JSB RDM vs. XH RDM
Corr=0.417 (p<0.114 by shuffled categories)

(Food only) KJS RDM vs. XH RDM
Corr=0.323 (p<0.131 by shuffled categories)

(Place only) JSB RDM vs. KJS RDM
Corr=0.726 (p<0.00907 by shuffled categories)

(Place only) JSB RDM vs. XH RDM
Corr=0.517 (p<0.0585 by shuffled categories)

(Place only) KJS RDM vs. XH RDM
Corr=0.411 (p<0.093 by shuffled categories)

Figure 4.6: RDMs on subsets of related categories are also similar across subjects. Each row of panels corresponds to a unique super-category, and each column to a unique subject pair.

4.4 Inter-subject similarities at the single-voxel level

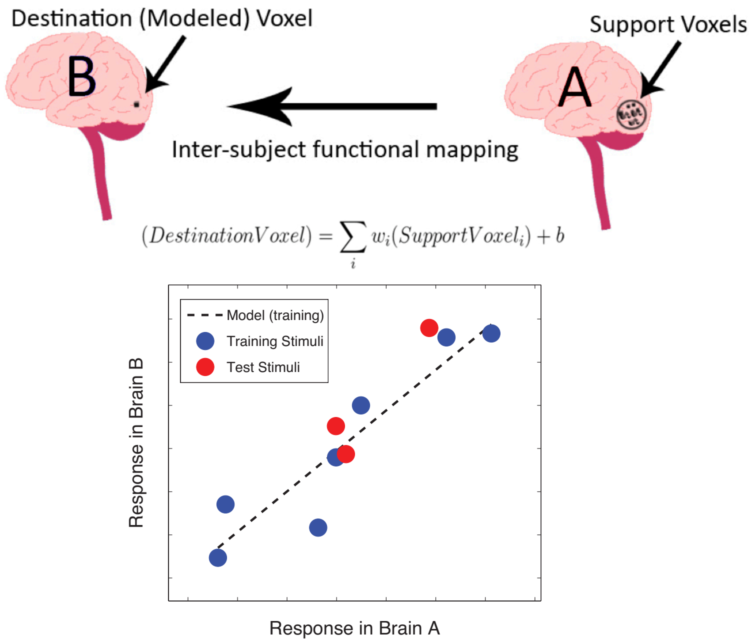

Destination (Modeled) Voxel
B
Support Voxels
A
Inter-subject functional mapping
$(DestinationVoxel) = \sum_i w_i(SupportVoxel_i) + b$
Model (training)
Training Stimuli
Test Stimuli
Response in Brain B
Response in Brain A

Figure 4.7: Illustration of the linear cross-brain model of a single destination voxel based on a combination of support voxels in the same region of another brain. For visual convenience, the response in brain A is shown as a scalar, but in the actual model it is a vector with a component for each support voxel.

We were also interested to find out whether individual voxels had responses to images which were similar to those of voxels in other subjects' brains. To address this question, we built cross-brain models as illustrated in Fig. 4.7. Each voxel in region R (each of the 5 regions considered separately) was linearly modeled based on *all* of the voxels in the same region R of *one* other subject (the "support voxels"). That is, we found a weight vector w relating the image responses of a single voxel i in region R in subject s, $x_{s,i}$, to the matrix of responses $x_{s',R}$ in the same region of another subject s' as follows:

$$x_{s,i}(image_j) = w_0 + \sum_{k \in R(s')} x_{s',k}(image_j)w_k$$

$$x_{s,i} = [1 \; x_{s',R}]w$$

where $x_{s,i}$ is an $n_{train} \times 1$ vector image responses, n_{train} is the number of training images, $x_{s',R}$ is an $n_{train} \times |R(s')|$ matrix of responses, $R(s')$ is the set of voxels in region R of subject s' and $|\cdot|$

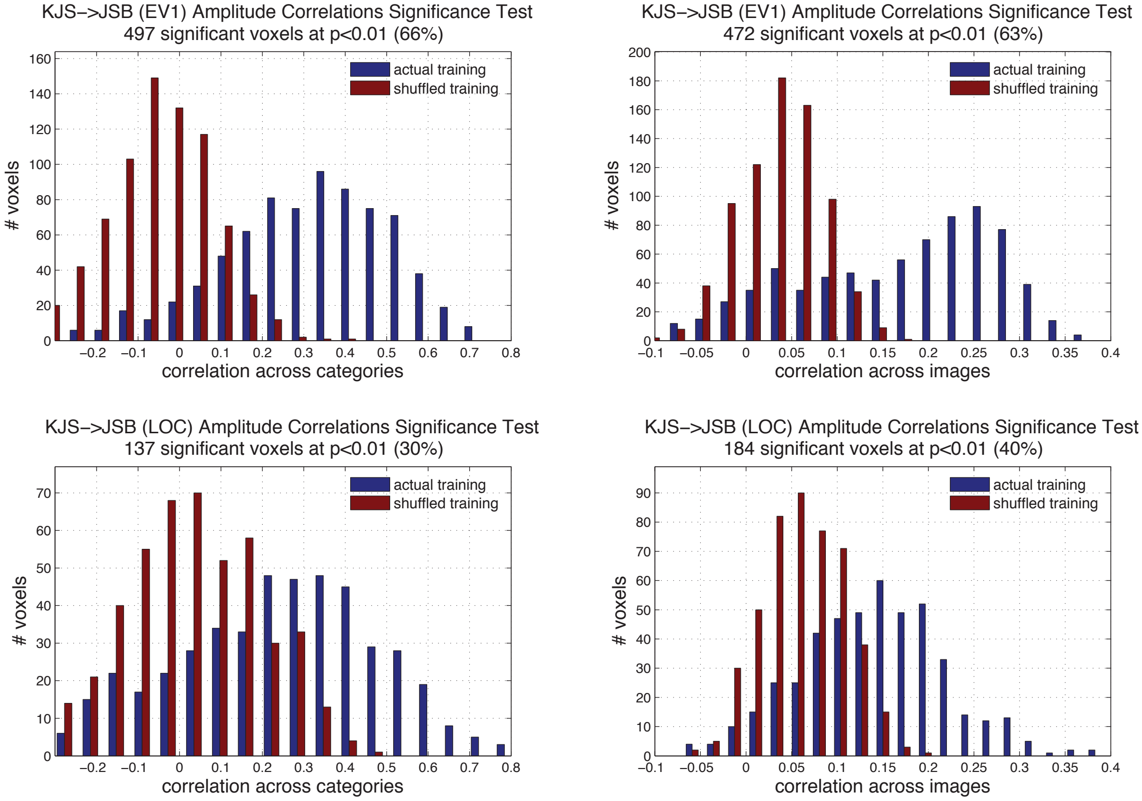


Figure 4.8: The correlation between predicted and observed responses to images and categories in single voxels, based on cross-brain models. All histograms are over voxels in the modeled subject's brain, in a given region. The correlation for each voxel is computed over the entire set of images or categories, but each individual estimate is based on a prediction from a model trained on images shown in different MR scans. Results are only shown for predicting the responses of JSB with the responses of KJS. Left panels correspond to models trained with category responses; right panels correspond to models trained with individual image responses. In order to test which voxels were significantly modeled, training was also performed based on a shuffled image order, and the results from that control model are shown in red. Voxels modeled above the 99^{th} percentile of correlations observed under shuffled training are considered significant. Note that the correlation when modeling category responses (left) is higher than when modeling image responses (right), because the estimates per-category are much higher in fidelity (10 or 20× data per response estimate). However, the shuffled model on categories also spuriously produces voxels at higher correlations, because the correlations are computed over much shorter vectors. Note: it is not surprising that the shuffled distribution on images lies to the right of zero, because a random shuffling in general preserves some nonzero amount of the structure in the original image order.

is the cardinality operator, and w is a $(1 + |R(s')|) \times 1$ vector of weights. To estimate the weight vector w relating one subjects' responses to another's, we set up a linear regression problem with n_{train} training images, selected from all but one scan (e.g., if the left-out scan contained 40 unique images, $n_{train} = 400$). However, the number of voxels over which we wanted to estimate weights was greater than n_{train} in general, and so we employed Tikhonov regularization[2] in order to force the minimization problem to have a unique solution.

We trained the model using both responses to individual object exemplars/images, and entire-categories (for which the number of responses per voxel is decreased an order of magnitude). We then tested how well this model extrapolated to the left-out scan (iterated over each of 12 scans), by comparing the responses predicted by the cross-subject model to those actually observed. Because no two scans featured the same object category, and the amplitude estimations from each scan were computed independently (e.g., they do not share a drift term), the extrapolation required non-trivial generalization.

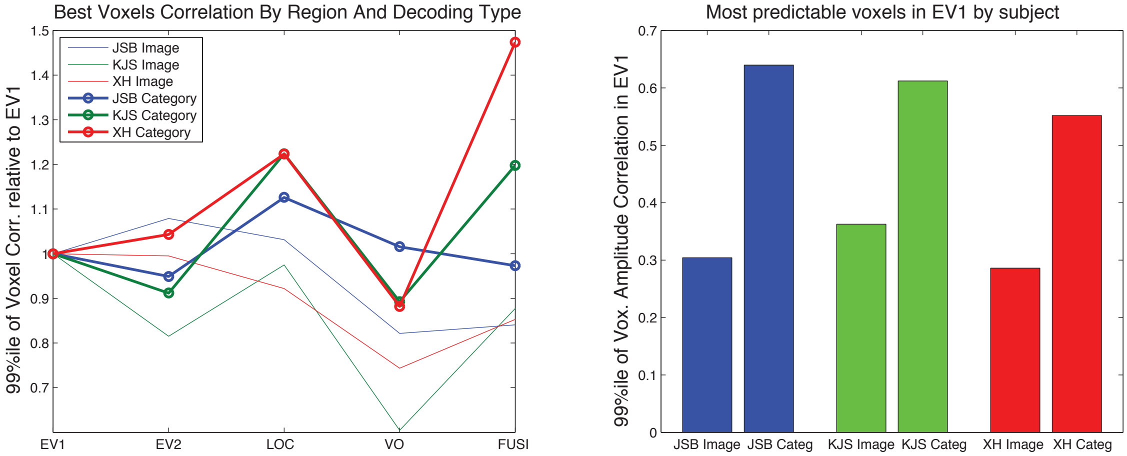


Figure 4.9: *Left:* The correlation between observed and predicted responses is greater in the higher visual pathway (LOC, VO, and Fusi) when training is based on entire categories. When training is based on individual images, the best-modeled voxels tend to fall in early visual cortex (EV1 and EV2), where information distinguishing between them is likely represented. *Right:* The correlation between observed and predicted responses for the 99^{th} percentile of voxels, for each subject, in EV1. The category correlations are higher both due to higher fidelity response estimates (more trials per response), and a higher likelihood of spurious correlations because they are computed over shorter vectors (number of categories $<$ number of images).

The results are that such models do convincingly extrapolate and are able to statistically significantly predict the responses in another brain. Fig. 4.8 shows a histogram over voxels in subject JSB, where each voxel is assigned a value according to how well it is modeled. This value is equal

[2] with parameter $\alpha = 0.1$ when training with entire-category responses, and $\alpha = 50$ when training with individual image responses. The results are not terribly sensitive to these parameters.

to the correlation between the predicted responses and the observed responses, over all images in the stimulus set. The correlations for a single voxel often exceed the 99^{th} percentile of those correlations obtained by shuffled training. Results obtained for other subjects are qualitatively similar. Fig. 4.9 shows that predicting the response to an object category is more easily done in higher visual cortex (esp. LOC and Fusi) than in lower visual cortex (EV1 and EV2). However, predicting the response to an individual image is more easily done in lower visual cortex. This is presumably because neurons in lower visual cortex have response properties which allow them to distinguish between individual images, whereas those higher in the visual pathway may only be able to distinguish between categories.

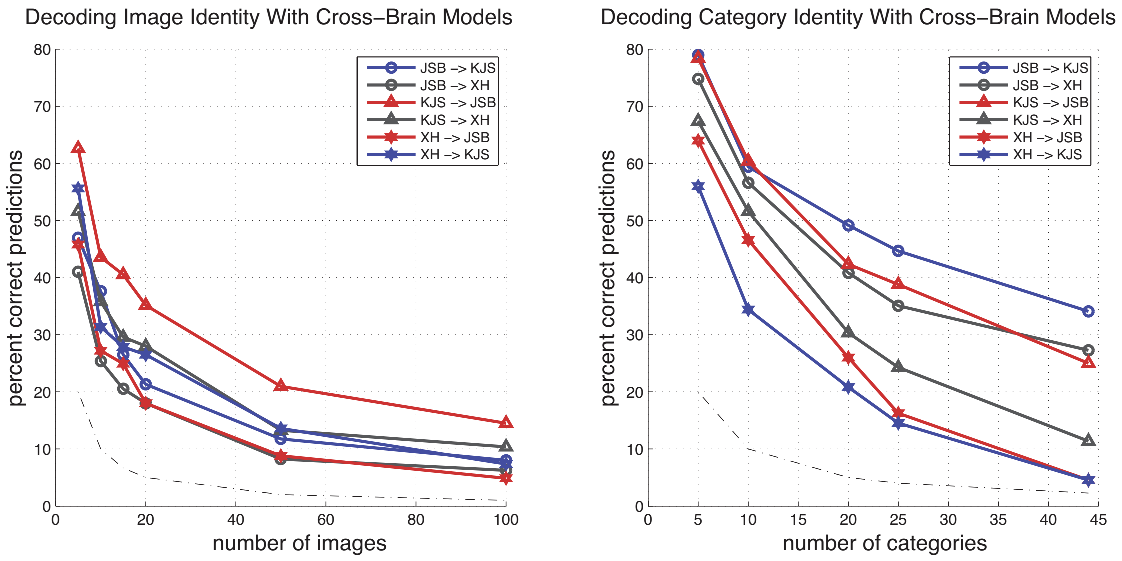


Figure 4.10: It is possible to accurately predict which image in a set a subject was viewing based on his/her brain data. The image predicted is that for which the observed brain activity most closely matches what would be expected given another subject's brain data in response to the same image. All decoding curves lie significantly above chance level. *Left:* decoding based training models to predict responses to individual images. *Right:* decoding based on training models to predict responses to object categories.

The strength of the cross-brain models for all subject pairs and modeling directions is summarized in Fig. 4.10. It is possible to accurately predict which image in a set a subject was viewing based on his/her brain data, and a cross-brain model which predicts what the brain data should be for each stimulus in the set. To get these results, we only used the best 30% of voxels in each region, ranked by the fidelity of the training estimates. Those voxels which were well modeled on the training set of images tended to extrapolate more accurately to the test set. Note that generally the models involving subject XH are worse, probably because the MRI set-up was slightly different for this subject, as discussed in the MRI methods (Section 4.6.2). Strikingly, in the case of the best models, we found it possible to predict exactly which category out of 44 a subject was viewing in roughly 35% of cases (compared to 1/44=2.3% chance level).

4.4.1 Voxels well-modeled across brains are category selective

Given the way our experiment was structured, namely, that subjects all viewed the image stimuli in exactly the same order, it is in principle possible that the cross-subject models are predictive because subjects simply synchronize over the course of the scan in a way which is independent of object category/stimulus information. For instance, perhaps subjects become relatively inattentive and attentive in the same temporal pattern, driving a common pattern of blood flow to visual cortical areas, without any relation to the object category being shown. If that were the case, then the voxels which are best modeled across subjects (based on individual image responses) would not contain any more category information than those which are poorly modeled. For instance, the best modeled voxel could simply be one at a critical point in the vascular network which fluctuates in a very stereotypical way among subjects over the course of each scan, but it is no more likely to respond at one level for one set of categories and at another for the complement than a voxel which is poorly modeled, because its responses are by supposition independent of category information. However, we find that this is in fact not the case, for at least two reasons: (1) Fig. 4.11 shows that those voxels which are best modeled, based on responses to individual images, have a strong preference for images with life forms. They are more active for those images than for images without life forms. (2) We find a strong correlation between the "cluster inclusion features", provided from human annotation, and encoding the identity of subsets of strongly related categories, and the best modeled voxels in LOC and Fusiform gyrus. Assigning each voxels two values: (1) *cross-brain-likeness,* the average correlation between its predicted image responses based on cross-brain models and those observed (indicating how well modeled it is), and (2) *category-likeness,* the average correlation between its responses and the cluster inclusion features (indicating how category-selective it is), we find a mean correlation of .474 between these two values across voxels in LOC, and 0.471 in Fusiform gyrus. The value of *cross-brain-likeness* is significantly explained by value *category-likeness* according to an F-test in all cases (all 3 subjects, both ROIs), with all p-values less than 1.6×10^{-5}.

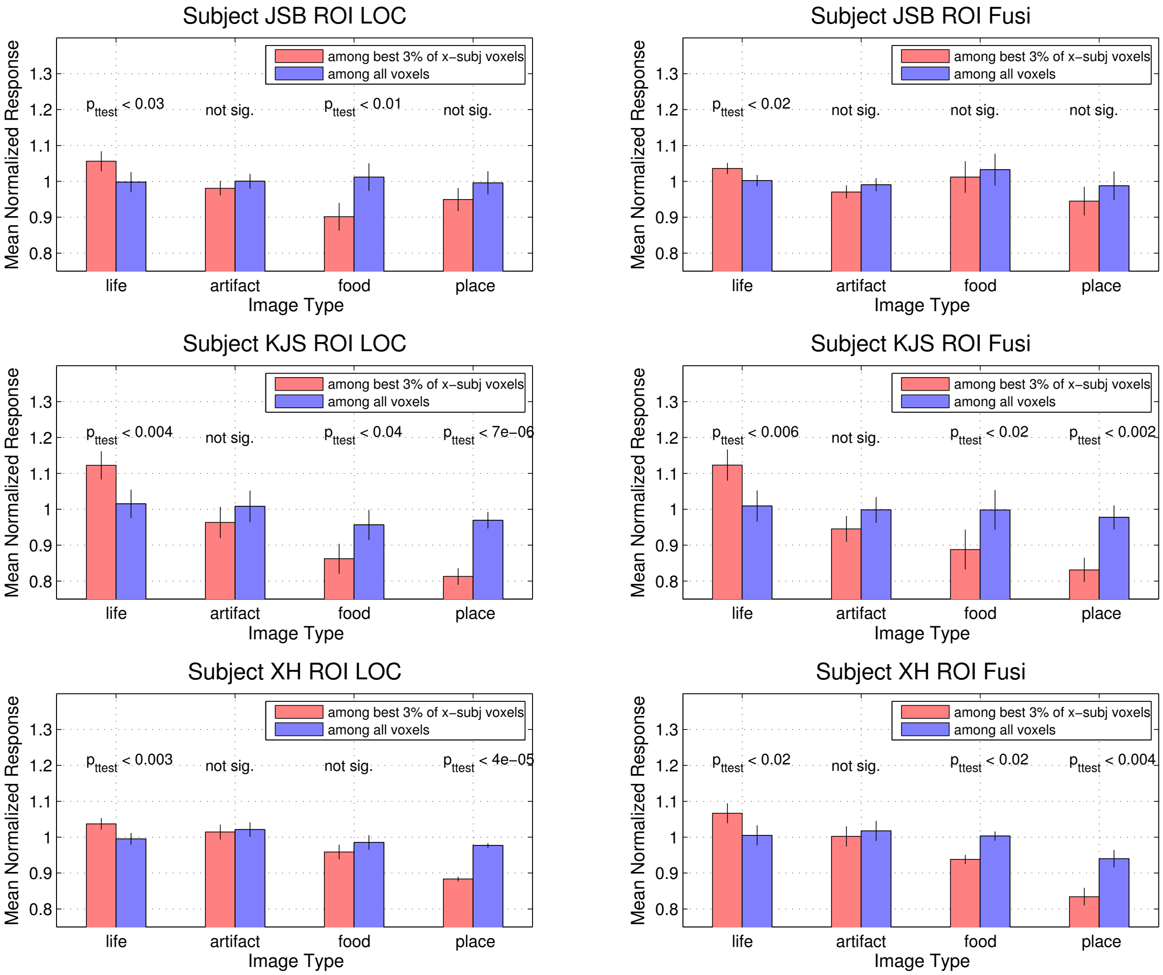


Figure 4.11: The voxels in each brain which are best modeled based on cross-brain models exhibit a preference for images with life forms. Voxels are ranked according to how well their predicted responses to each image (using image-based decoding) are to the observed ones (averaged over the two other subjects used to form the model). Image responses are normalized such that in any single scan, the mean response is 1.0. Among all voxels, the mean response to each category is around 1, however among the 3% which are modeled best across subjects, there is a strong preference for life forms and against places, both in LOC and Fusiform gyrus, in all 3 subjects. p-values are shown in each plot above each pair of corresponding bars. They are computed from a t-test on the list of paired comparisons, one per category in the super-category set.

4.5 Discussion and future directions

First, on the section comparing object category distances and finding high correlation (correlation 0.55 using pairs of categories from different scans): by showing that different subjects have similar inter-category neural pattern distances, we have introduced some quantitative constraints on the extent to which people may differ in visual experience. If the quale associated with the percept of a visual object category is defined or constrained by its similarity at the neural level to other visual percepts, we can say some with degree of certainty that such visual qualia are not totally different among different people.

The section on modeling single-voxel responses also has potential utility. A vital assumption in systems neuroscience is that the receptive field properties of neurons are generally similar across individuals. Very little, however, is understood about how the physical configurations of such functionally similar neurons might correspond across brains, and what patterns govern such mappings if any. The usefulness of having such functional mapping was demonstrated in an experiment by Sabuncu et al. [80], in which standard fMRI analyses could gain significance by applying a smooth "functional" (i.e., based on the function of the brain) mapping, learned beforehand based on a movie presentation, to align subjects' brains. The approach we took in the single-voxel mapping section of this chapter has similar application. One could bring subject brain B into "functional" alignment with brain A by reconstituting every voxel in reference space A using the model learned during training.

Fig. 4.12 suggests another possible application for the kind of analysis presented in this chapter, of the brain's activity under visual stimulation. Brain responses to images are added to features automatically extracted from image content (namely, C2 features from the "hmax" model [78, 67]), and their addition improves object support-vector-machine (SVM-)based object categorization performance without maximizing it to perfect level. This suggests that the information in the brain responses is (1) not identical to object category labels and (2) not redundant relative to hmax features.

These features could potentially just be extremely noisy category labels. However, it is interesting to speculate that, to the extent that they are not perfect object labels, they are meaningful mid-level features computed by the brain in the visual pathway as part of its object recognition processing stream. If so, they could be trained against as part of a computer vision architecture which uses as ground-truth not only image labels, but also these intermediate labels supplied by the brain. Such intermediate features could help bridge the vast gap between low-level features

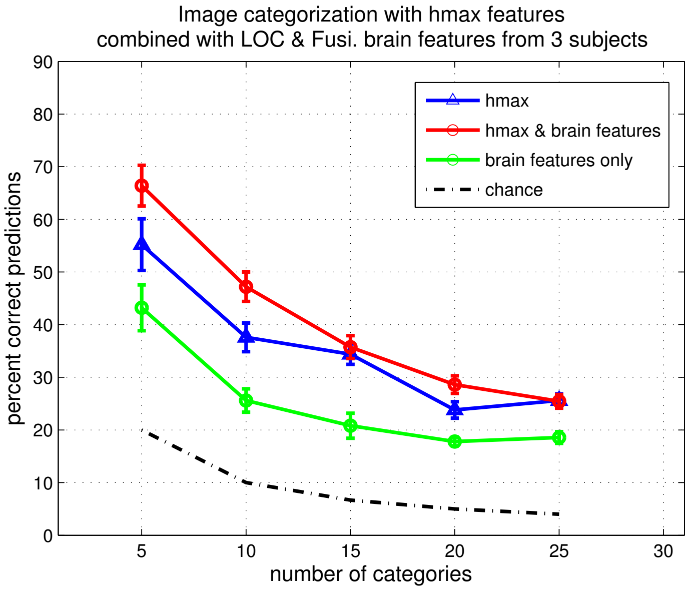


Figure 4.12: Brain-based features enhance image categorization performance. Categorization is based on training a support vector machine with 9 examples per category, and testing it with the example image left out. Error bars are standard errors over 25 trials, each corresponding to a different subset of categories. Features consist of 2000 hmax C2 elements, and 362 voxel responses from LOC and Fusiform gyrus (311 from LOC), corresponding to the 15% of voxels best-modeled in those ROIs across subjects (136 from JSB, 158 from KJS, and 68 from XH).

extracted automatically and high-level labels supplied manually by human annotators. We predict that, performed at a sufficiently large scale (e.g., with 5 monkeys, and thousands of images), mid-level image features could be learned from brain responses. Machine learning techniques could then be applied to automatically construct these for *new* images, hopefully aiding in object recognition.

Technical notes on categorization experiment: For each number of categories (5,10,..., 25), we performed 25 trials. In each trial, a different subset of the 44 categories of visual objects described in this chapter was selected. In each case, the 10 exemplar images per category were partitioned into 9 training and 1 testing image. The features describing each image (hmax and/or brain) were Z-scored relative to the training set, and the mean and standard deviation of each feature in this training set were then used to normalize the test features. A multiclass support vector machine (using highest rank in pool spanning all unique one-vs-one classifiers) was trained using MATLAB's Statistical Pattern Recognition Toolbox ("stprtool"), with default parameters, linear kernel function, and a regularization constant $C = 100$. The 50% of features having the lowest absolute weight in the SVM training were thrown out, and the training was repeated. This was done one more time, so that the final SVM was only trained on 25% of the features (hmax, brain, or hmax+brain). This final SVM was used to predict a category for each test image. This hmax feature creation and pruning techniques were based on a public code library by Jim Mutch[3]. To provide the reader with some idea of how the pruning divided hmax and brain features, in one randomly selected trial in an experiment predicting among 25 categories, of the original 2000 hmax features, 549 were selected (27.5%), and of the original 362 brain-based features, 42 were selected (11.6%), of which 17 were from KJS, 17 from JSB, and 8 from XH.

[3] see http://www.mit.edu/~jmutch/fhlib/, [67]

4.6 Appendix

Subjects provided informed written consent prior to the experiments. The Caltech Institutional Review Board approved all experimental procedures.

4.6.1 Visual stimulation paradigm

The 440 images were shown over a series of 12 individual functional imaging scans, each one lasting 7 minutes, 57 seconds (7:57), and consisting of 80 image presentation trials, from either 2 or 4 categories (2 scans with 2 categories, and 10 with 4 categories). A white fixation dot appeared in the center of the image display throughout the entirety of each scan; subjects were instructed to keep their eyes fixated on the dot the entire time, as images centered on it would appear in sequence. Each of the 10 exemplar images from each of the categories was shown either 4 times (in scans with only 2 categories), or 2 times (in scans with 4 categories), non-successively and distributed non-regularly over the 80 trials. The images were randomly permuted with respect to category, but the exact image order was the same for all three subjects. In each presentation trial, an image flashed at 2.5 Hz for 1 s, followed by a 3.5 or 8 s (exactly once ever 4 images, regularly) delay, during which the subject pressed a button on a controller indicating which object category (among 2 or 4 candidates) he/she thought the just-presented image belonged to. Additionally, there were 13.5 and 21 seconds of lead-time and end-time before and after all the image presentations, respectively, during which the screen was a uniform gray except for the fixation dot. All visual stimuli were projected onto a rear-projection screen visible from within the MRI scanner via an angled mirror.

4.6.2 MRI data acquisition and preprocessing steps

MRI data were collected using a Siemens (Erlangen, Germany) 3T TRIO. All functional images (blood oxygenation level dependent (BOLD) T2*-weighted) were acquired using a gradient-echo echo-planar (EPI) sequence. For two of the subjects (KJS and JSB), functional images consisting of 20 slices with $2.23 \times 2.23 \times 2.5$ mm voxels were acquired with a TR (repetition time) of 1.5 seconds using a Nova Medical occipital coil (acquisition matrix 64×64, flip angle $80°$, echo time 31 ms). For the remaining subject (XH), an 8-channel head coil was used to acquire the functional images, with 20 slices and $3 \times 3 \times 3$ mm voxels (TR of 1.5 seconds, acquisition matrix 64×64, flip angle $70°$, echo time 31 ms). Slices were obliquely oriented for coverage of occipital and temporal visual areas. Anatomical images for all three subjects were acquired using a T1-weighted MPRAGE (magnetization-prepared rapid gradient echo) sequence with the 8-channel head coil. Functional volumes were slice-time-corrected, then registered with a single 12-parameter 3-dimensional affine transformation to a reference scan for each subject (this transformation incorporated self-motion correction). Images were then resampled into a 3 × 3 × 3 mm voxel space, and smoothed with a spatial Gaussian kernel with standard deviation 1.5 mm. Anatomical images were co-registered with functional images using a 12-parameter affine transformation as well.

4.6.3 Defining regions of interest (ROIs) in the brain using Freesurfer

The regions of interest in this chapter were defined relative to the cortical segmentation computed automatically by FreeSurfer (http://surfer.nmr.mgh.harvard.edu/). Because they were defined anatomically, and not based on a functional experiment, these region labels are approximate. EV1 was taken as the cortical matter inside calcarine sulcus and the closest part of the occipital pole, and EV2 was defined as regions adjacent to and immediately dorsal and ventral from that. VO (ventral-occipital) was here taken as a posterior portion of the medial occipito-temporal and lingual sulcus. Here, Fusi (fusiform gyrus) was simply taken as the 60% most posterior portion of all fusiform gyrus, since that's where some visual object areas (e.g., FFA) tend to concentrate. LOC (lateral occipital complex) here simply consisted of lateral voxels in occipital cortex. Fig. 4.13 shows the calcarine sulcus in an anatomical scan of subject KJS.

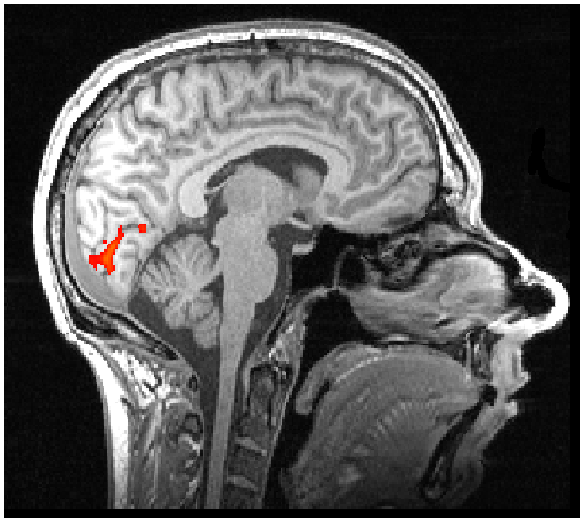


Figure 4.13: The calcarine sulcus, which contains primary visual cortex (V1), is shown highlighted in red in subject KJS in one sagittal slice.

4.6.4 Estimating a response to each image in each voxel of each brain

Fundamentally, we assume that each image presentation trial induces some amount of neural activity in each little chunk of brain tissue, and that this induces an increase in blood flow, which we would like to estimate at each location independently given the imaging data. The approach we took to modeling the response in each discrete voxel (the 3-dimensional analog of a pixel) in the imaging data was based on the linear systems approach first described by Boynton [9] in 1996. At each such voxel, over the course of the scan, we observed a time-course vector t (of dimension $n \times 1$ where n is the number of 3D volume acquisitions, which in our case $n = 318$ for a single 7:57 scan) of image intensities. We modeled the time-course as the sum of three components:

$$t = r + d + n,$$

where r is the "response" component of the signal due to increased blood flow induced by stimulation, d is a "drift" component, which is taken to be very slowly changing (e.g., < 0.02 Hz), and resulting from the baseline intensity of the imaged tissue plus slowly varying changes in the scanner (e.g. heating up), and n is random zero-mean noise resulting from measurement near the physical limits of the system (and independent at each time point). We assume a simple relationship between the vector of stimulus onsets s (a vector of ones at times corresponding to stimulus onset, and zeros elsewhere) and the observed response r, viz.:

$$r = h * a, \text{ i.e.,}$$

$$r(j) = \sum_{i=0}^{n-1} h(j-i)a(i) \text{ for } j = 0, 1, ..., n-1$$

where $a = As$ is the $n \times 1$ vector of response amplitudes (a vector with positive values at times corresponding to stimulus onset, and zeros elsewhere; A is a square matrix appropriately scaling s to the response of each stimulus), and h is a characteristic *hemodynamic response function* (HRF), which usually peaks about 5 seconds after stimulus onset (see Fig. 4.15 bottom for an example). We assume $h(j) = 0$ for $j < 0$ and $j > t_{max}$ where t_{max} is usually around 12 seconds, although in some cases it is as high as 32. Essentially, the model assumes each discrete stimulation event induced a characteristic impulse response of increased blood flow, mostly lasting 10-12 seconds, and that these were additive in sequence.

The challenge was to estimate the response amplitudes a induced by each stimulus given, in

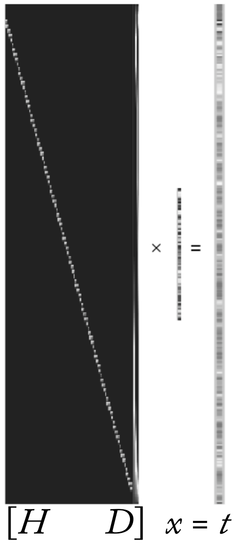


Figure 4.14: The linear model of blood-flow response to stimuli: the matrix on the left consists of HRFs shifted to each event onset, H, concatenated with the matrix of slowly changing drift terms, D. T is the observed time-course.

general, only the slowly-changing nature of d, the times of stimulus onsets s, and the observed output t. To make the problem simpler, we assumed that the HRF h took a known form. HRFs have been studied extensively in visual cortex (see e.g. [52]) and their basic characteristics are well understood. Then, we estimated the non-zero entries in a by obtaining the least-squares solution $\hat{x}$ to the following equation:

$$[H\ D]x = t$$

where $\hat{r} = H\hat{x}_m$ is the estimated response, and $\hat{d} = D\hat{x}_o$ is the estimated drift, and $\hat{y} = \hat{r} + \hat{d}$ is the reconstructed signal. m is the number of stimulus onsets, $\hat{x}_m$ is a vector of the first m components of $\hat{x}$, o is the order of the drift term, and $\hat{x}_o$ is a vector of the last o components of $\hat{x}$. D is an $n \times o$ matrix of slowly-changing basis functions up to order o (we use $o = 9$ and the first 9 Legendre polynomials interpolated between -1 and 1 [52]), and

$$H(j, i) = h(j - onset(i)), \text{ for } i = 1, 2, ..., m$$

where $onset(i)$ is the time of the i^{th} stimulus onset. Fig. 4.14 and Fig. 4.15 illustrate the model and provide an example of its estimations on a real data sample. For better results, we carried this estimation out for 5 different candidate HRFs in each brain region in each subject and, for each one independently, used the estimates corresponding to the HRF where the reconstructed signal was closest to the observed signal in the least-squares sense.

To account for variability between scans, the response amplitudes were normalized by the mean across all image trials in the scan at which they occurred. Finally, the response to each image was taken as the mean of these normalized values over its 2 or 4 presentations, and the response to each image category was taken as the mean of the normalized values over its 20 or 40 presentations. In each voxel, there was some variability in the response to each image across presentation trials, and the mean variance of this value over images was taken as its inverse "SNR": voxels which consistently yielded the same response across trials of the exact same image had very high SNR. Voxels with low SNR were likely to be outside of visual areas.

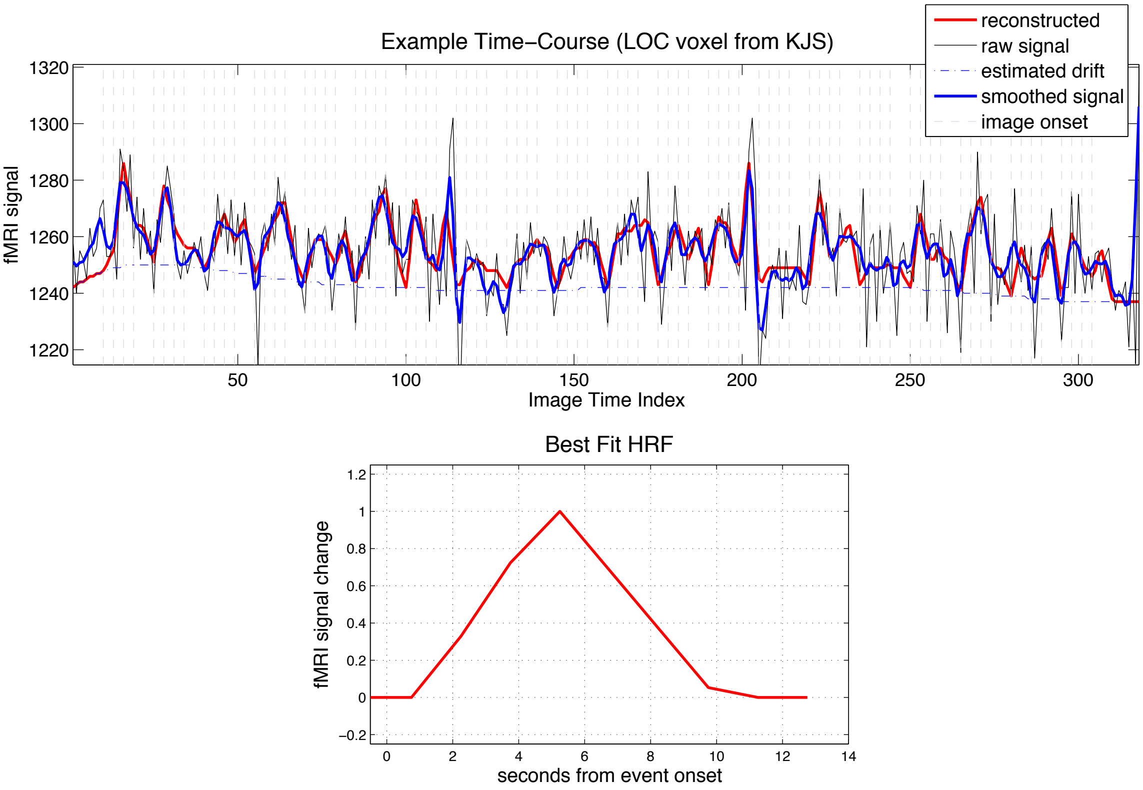


Figure 4.15: Model estimation on a single voxel in the LOC region of subject KJS: *Top:* The observed time-course is shown in black, together with the reconstructed signal in red. Notice that the raw image intensities are in this case around 1240 - 1280; most of this intensity is simply due to the magnetic properties of the imaged tissue, which are not changing over the course of the scan. It is the small fluctuations at the top of the signal which carry information about the extent to which blood flow increased with each image onset. The image onsets are shown as vertical dashed lines. *Bottom:* the shape of the HRF which yielded the best reconstruction for the time-course.

Chapter 5

Different individuals exhibit temporally similar neural patterns under dynamic video stimulation

Whereas in the last chapter we compared patterns of neural activity across voxels in space, in this chapter, we present experimental results aimed at quantifying inter-subject similarities in neural patterns across *time*, as they occur in response to videos both with and without sound. Subjects in an fMRI scanner were shown three categories of videos, each with and without an accompanying audio track. Two of the categories of videos were semantically meaningful whereas the third was more abstract. We find that subjects' brain responses to the same videos correlate in corresponding brain regions to a high degree (71 - 78% compared to correlations with self later in time). We also find evidence that when watching videos *with* sound, visual attention is likely blurred at times and transferred to audition, as subjects relatively decorrelate in visual areas compared to the muted case. Lastly we show that about 34% of the observed variance in inter-subject response similarity can be explained by similarities in the parcellation and/or structure of their brain anatomies.

5.1 Previous work

The most relevant reference for the work in this chapter is an important paper published in *Science* in March 2004 [37], wherein Hasson and colleagues showed that individual subjects exhibited a very strong tendency to go through similar temporal patterns of relative high and low fMRI BOLD activation during natural audiovisual stimulation. Subjects viewed a 30 minute continuous block of the feature film *The Good, The Bad, and the Ugly* (1966, Metro-Goldwyn-Mayer) while in an MR scanner, and then individual points in their brains were compared to corresponding ones in the other subjects. Strong correlations were found in early and intermediate visual areas, including V1, V2, V3, V4, collateral sulcus, the fusiform face area (FFA), the parahippocampal place area (PPA), as well regions associated with auditory processing: A1, superior temporal sulcus (STS), and lateral sulcus (LS), and also the post-central sulcus (PCS), associated with hand movements. Hasson later reported, in the *Journal of Neuroscience* in 2008 [38], that the correlation in early sensory areas (e.g., primary visual cortex) was preserved even for much shorter presentation times, but that the correlation in areas higher in the sensory pathway (e.g., STS, precuneus) depended on information being accumulated over larger time windows. In another paper published in 2008 [46], Jääskeläinen and colleagues showed that it was also possible for subjects to correlate in frontal cortex (especially right frontal cortex), an area in which Hasson et al. did not find significant correlations. They achieved this by having subjects watch the first 72 minutes of the feature film *Crash* (2005, Lions Gate Films) outside of the scanner, then watch the last 36 minutes in the scanner. Presumably, the synchronization of high-level emotional events drove this previously unreported signal. Golland et al. [33] report significant decrease in auditory region correlation for movies without sound, and find a pattern of intra-subject correlation (same subject, later in time) similar to the inter-subject pattern reported originally reported by Hasson in [37]. Finally, in a slightly different experiment, Stephens et al. show that the pattern of neural activity a speaker of a story experiences is significantly correlated with the pattern a listener to that story experiences [87].

5.2 Basic experimental setup

Seven healthy[1] adult subjects (4 female, ages 21-33) freely viewed three videos in each of two conditions: visual stimulation only (i.e., without sound), and auditory & visual stimulation. The videos in the first and second conditions were not identical, but very similar in nature. Three of the subjects were brought back in more than a month after their original scans and repeated the entire experiment. Below, we report similarities between subjects in cortical responses, how they depend on the nature of the stimuli, and how they compare to same-subject similarity across time.

5.2.1 Stimulation paradigm

Subjects participated in two MR scans, lasting 22 minutes, 12 seconds each. In each scan, subjects viewed (1) a 14 minute video, followed by (2) a 4 minute video, followed by (3) another 4 minute

[1] We note that 3 of the subjects (not those brought back for repeat scans) were initially recruited for having self-reported visual-to-auditory synesthesia. However, no significant difference with respect to this was found in the results presented in this chapter.

video, with 4 seconds in between videos and 2 seconds at the very beginning and end of the scan. Subjects were not required to fixate and were only instructed to pay attention. The videos in the first scan had no sound (visual only, “V-only”), and the videos in the second scan all did have sound (auditory and visual stimulation, “AV”). Auditory stimuli were delivered via MRI-compatible stereo headphones. All visual stimuli were projected onto a rear-projection screen visible from within the MRI scanner via an angled mirror. Stimuli subtended roughly $16^\circ \times 12^\circ$ of visual angle, and occupied 640×480 pixels on 800×600 pixel display.

(1) The first video in each condition contained a continuous segment of the feature film *Back to the Future* (1985, Universal Pictures). In the visual-only condition, the clip was from near the beginning of the film, and in the audio & visual condition, the clip was was from near the end of the film. These were intended to fully engage the subjects in a plot, and involve all levels of cognition from simple sensory perception to high-level emotional and theory-of-mind processes.

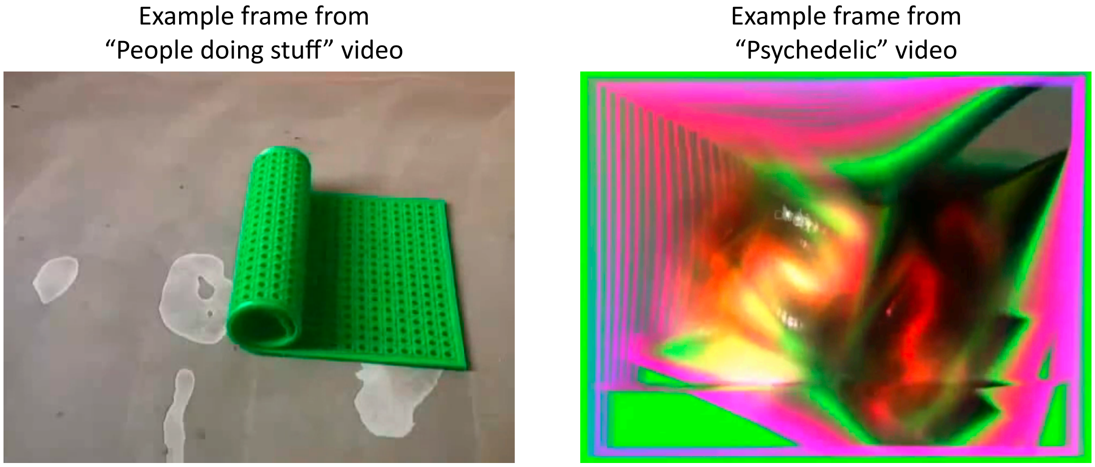


Figure 5.1: Samples from the videos shown in the dynamic images experiment.

(2) The second clip in each condition featured “People Doing Stuff”, taken in part from the artistic film “Everything is Everything” by Koki Tanaka[2]. It consisted of various household objects being manipulated or moved in seemingly meaningless and random manners. Fig. 5.1 shows an example frame from this video. The original video from collegehumor is only about 6 minutes long, so it was divided into two 2:52 segments, each of which was then concatenated with an additional 1:08 of very similar video content, filmed by the author, in order to produce two full 4 minute clips. These videos featured concrete recognizable objects, but did not form a coherent plot.

(3) The third clip in each condition featured “psychedelic” videos found on YouTube[3]. These

[2] http://www.collegehumor.com/video:1924307

[3] http://www.youtube.com, YouTube, LLC (San Bruno, CA)

videos contained very colorful, undulating, abstract shapes and forms. In the video with sound, a trance music soundtrack was played, which did clearly involve human vocals, although only with very vague and difficult-to-decipher words without any clear message.

Thus, a variety of video content was shown ranging from the most concrete to increasingly abstract. All subjects viewed the videos in exactly the same order, and always with the visual-only condition first.

5.2.2 Two kinds of inter-subject correlations

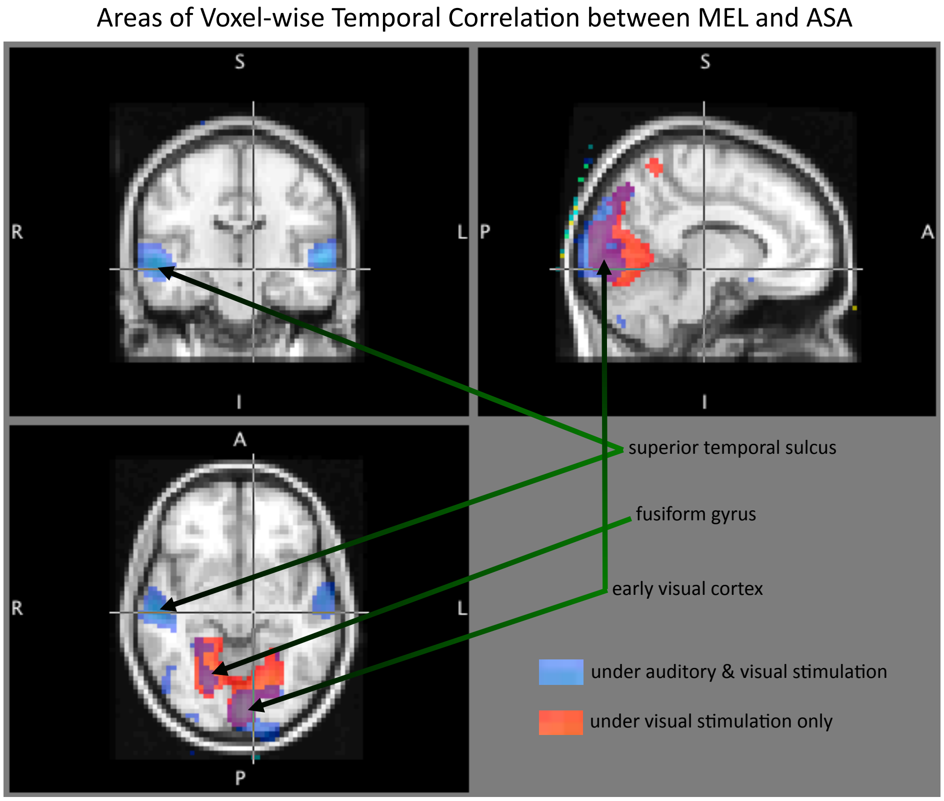

Areas of Voxel-wise Temporal Correlation between MEL and ASA
S
R
L
I
S
P
A
I
A
R
L
P
superior temporal sulcus
fusiform gyrus
early visual cortex
🟦 under auditory & visual stimulation
🟥 under visual stimulation only

Figure 5.2: The spatial extent of inter-subject correlations extends into auditory regions only under auditory & visual stimulation. Correlations are cut off below 0.3 (do not appear as a colored patch), are computed from the entire 3-video presentation, and are voxel-wise on the functional images spatially smoothed with a 5mm Gaussian kernel. The underlying anatomical image is the MNI152 standard brain in the 3.3 mm space.

Below, two kinds of comparisons are made between subjects' functional responses: (1) voxel-wise: in these comparisons, an exact MNI52 coordinate is compared to the same one in another subject. When this kind of comparison is made, an ROI is taken to be only the intersection of all subjects' ROIs in the normalized space, (2) ROI-wise: in these comparisons, the average activity across a whole spatial region of voxels is taken before comparison to another subject is made. When this kind of comparison is made, an ROI is taken to the be the entire ROI for the subject (not only the sub-ROI overlapping with the others'). For voxel-wise comparisons, the imaging data is also spatially smoothed with a Gaussian kernel before drift removal and low-pass filtering. Unless otherwise specified, the smoothing was with a kernel of standard deviation 2 mm. Details concerning the MR scanning and localization of brain regions is provided in the Appendix to this chapter.

5.3 Broad correlations across cortex

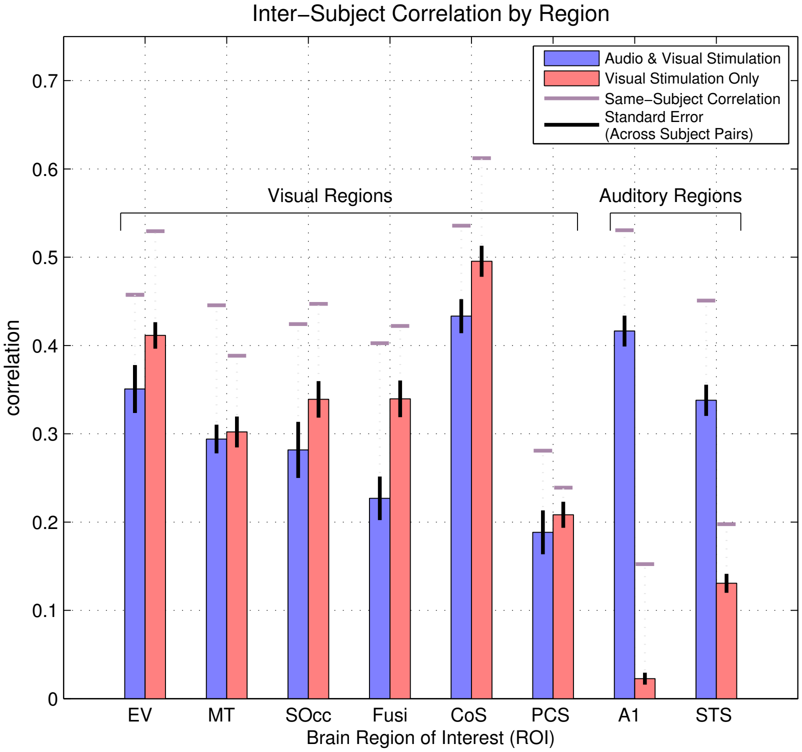


Figure 5.3: ROI-wise functional correlation across subjects is shown for each of the eight regions. For each ROI, the mean same-subject correlation (across the three subjects which came in for a repeat scan more than a month later) is shown as a horizontal line. The inter-subject correlations are about 71% as high as the same-subject correlations in the auditory & visual (AV) case, and 78% in the visual-only (V) case, excluding region A1, which correlates across subjects to only 15% the same-subject extent. All correlations are significantly above zero.

Replicating the finding originally reported by Hasson et al. [37], we find "synchronization" of brain activity among subjects which is extensive, going well beyond just primary sensory cortices. Fig. 5.2 shows the voxel-wise correlation between two subjects (MEL and ASA). Some things are immediately clear: for example, under auditory & visual stimulation, subjects' responses in the superior temporal sulcus (STS) correlated, but much less so when stimuli were soundless. The posterior portion of STS (pSTS) has been implicated in the integration of auditory and visual information into a single unified representation [42]. Fig. 5.3 summarizes the inter-subject correlations found in each of the ROIs: all correlated significantly positively, even region A1 under visual-only stimulation, although only with p-value 0.011 according to a two-sided t-test on the set of subject pairs. Fig. 5.4 shows how the voxel-wise correlations are spread out within each ROI.

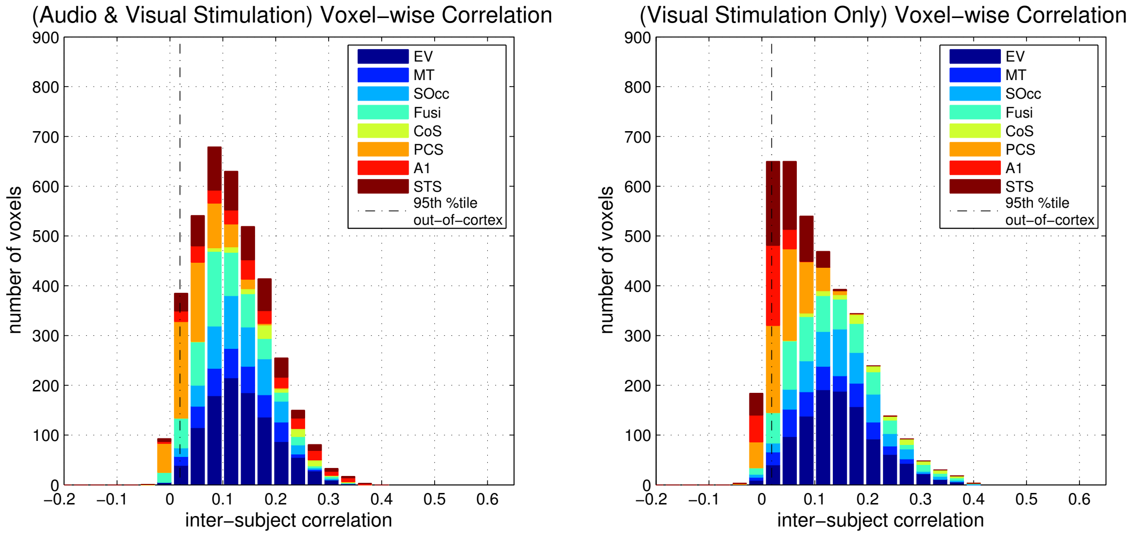


Figure 5.4: The histogram of *inter-subject correlations* across voxels. The dashed line corresponds to the 95th percentile of inter-subject correlations found in voxels outside of cortex.

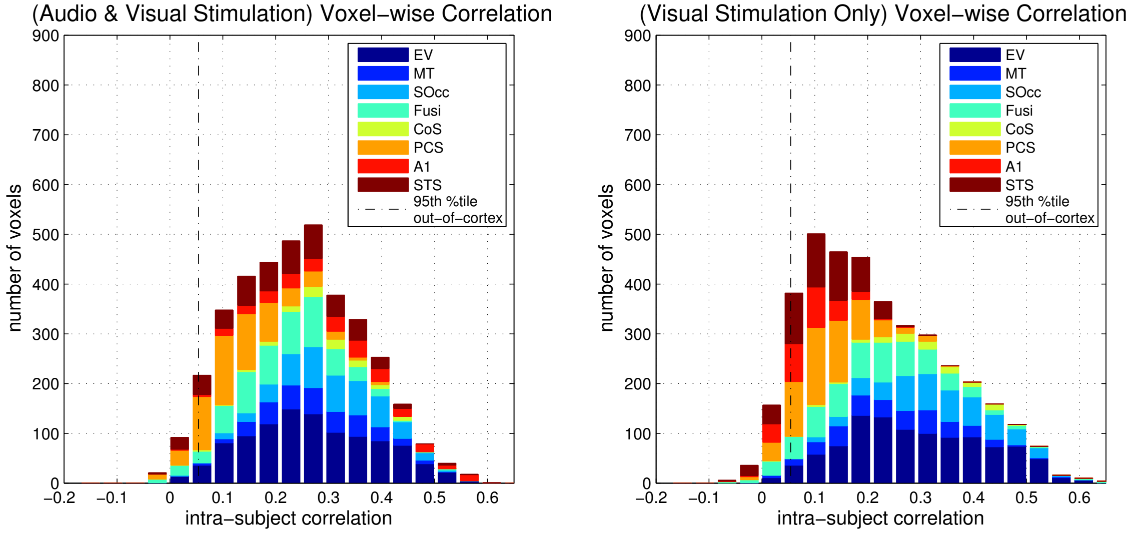


Figure 5.5: The histogram of *same-subject correlations* across voxels. The dashed line corresponds to the 95th percentile of same-subject correlations found in voxels outside of cortex. Correlations are greater than the analogous inter-subject comparisons, shown in Fig. 5.4.

To understand how strong the inter-subject correlations were, we compared them to the correlations between a subject and him/herself later in time. Three subjects (ASA, CW, and JH) came back more than a month after their initial scans for a repeat of the exact same experiment. The resulting voxel-wise same-subject correlations are shown in Fig. 5.5. The mean same-subject correlations (among the three subjects for which we had data) are also indicated in Fig. 5.3, for each ROI. We find that, averaged across ROIs, subjects correlated with others to 71% the same-subject extent in the AV case, with a standard deviation of 8%. In the V-only case, subjects correlated at 78% the same-subject extent, with a standard deviation of 6% among ROIs, if we exclude A1.

A1 under visual-only stimulation correlated much less between subjects than within. Whereas subjects only correlated at a level of 0.023 (standard deviation 0.052) with others, the same subject correlation averaged at 0.15 (standard deviation 0.097). The inter-subject distribution fell nowhere near the same-subject distribution ($p_{t-test} < 7 \times 10^{-6}$). The relative magnitude of this difference is so much greater in A1 than in other regions that it is likely that some additional source of variability, relating to A1 but not the other regions, contributed to degrading the A1 correlations between subjects, but not within subjects. One good explanation relates to proximity to sensory input and functional coupling. The activity in visual regions, close to the main source of sensory input (retinal), mostly followed the temporal structure of the visual stimulus, with STS not far behind. However, the primary auditory cortex, A1, was far from the main source of sensory input (although the MR scanner made noise throughout the scan, it was nearly perfectly uniform averaged a 2 or 3 second window). Thus, it is arguable that the temporal structure of its activity was mostly driven by functional coupling to other regions, including the stimulus-driven visual regions[4], and overall level of attentional arousal. To the extent that functional coupling arises from underlying anatomical connectivity, inter-subject variability in anatomy would relatively degrade inter-subject correlations compared to same-subject correlations. Also, it is expected that subjects would be more consistent with themselves (later in time) than with others about which visual stimuli were most suggestive of sound and also which were most attention-grabbing. Consistency in both of these dimensions could have relatively enhanced same-subject synchronization in A1 more than in other regions, where activity was less dependent on these factors.

[4] even in visual-only stimulation, there is weak functional coupling between early visual (EV) cortex and A1. A spectral coherence analysis revealed many EV-A1 voxel pairs with coherence above chance level for all subjects.

5.4 Effect of stimulus on inter-subject correlation

We found two interesting trends relating the stimulus type to the inter-subject correlations: (1) a decrease in visual region correlations under AV stimulation compared to V-only, (2) higher correlations in primary auditory cortex under V-only stimulation for the "People Doing Stuff" video.

5.4.1 Less attention to vision likely when audio is introduced

Fig. 5.6 summarizes the trend of generally lower visual correlation under AV stimulation compared to V-only. In early visual cortex, superior occipital cortex, fusiform gyrus, and collateral sulcus, inter-subject correlations were significantly lower. Averaged across these regions, a subject pair correlated 29% less when stimuli were audiovisual. At first, we found this result to be counter-intuitive. Subjects informally reported feeling more alert and interested in the videos when they included sound. Thus, it might be expected that subjects' brain activity would be more locked to the stimuli globally, and that inter-subject correlations would not only appear in auditory regions (compared to the V-only case), but also be enhanced in visual regions. Instead, the opposite was found.

Because subjects always viewed the silent videos first, in a nearly hour-long stretch in the MR scanner, it could be argued that they were simply more tired during the AV videos, and that is why synchronization then was relatively attenuated. However, several observations suggest otherwise. First, as already mentioned above, audiovisual stimuli are simply more engaging to subjects, and thus increase attention, possibly enough to compensate for AV videos being presented after V-only. Second, early visual cortex, fusiform gyrus, and collateral sulcus were significantly less correlated between subjects in the AV case even just in the first two minutes (all $p < 8 \times 10^{-4}$ by two-sided t-test) and just the first five minutes (all $p < 4 \times 10^{-3}$) of stimulus presentation, when subjects were most alert in both cases. Third, if we compare inter-subject correlations during the second 7 minutes of the *Back to the Future* video in the V-only case to inter-subject correlations during the first half in the AV case, we might expect the trend to reverse or disappear, since correlations right after stimulus onset might be higher than those 7 minutes in. However, we find that the direction of the trend remained the same, both in terms of number of subject pairs for which correlation was lower in AV case, and mean effect size, with AV correlations lower in early visual cortex, MT, fusiform gyrus and collateral sulcus; in this most strict test, however, statistical significance was only achieved in fusiform gyrus ($p < 2 \times 10^{-3}$).

The simplest interpretation of subjects' decreased synchronization under AV stimulation is that,

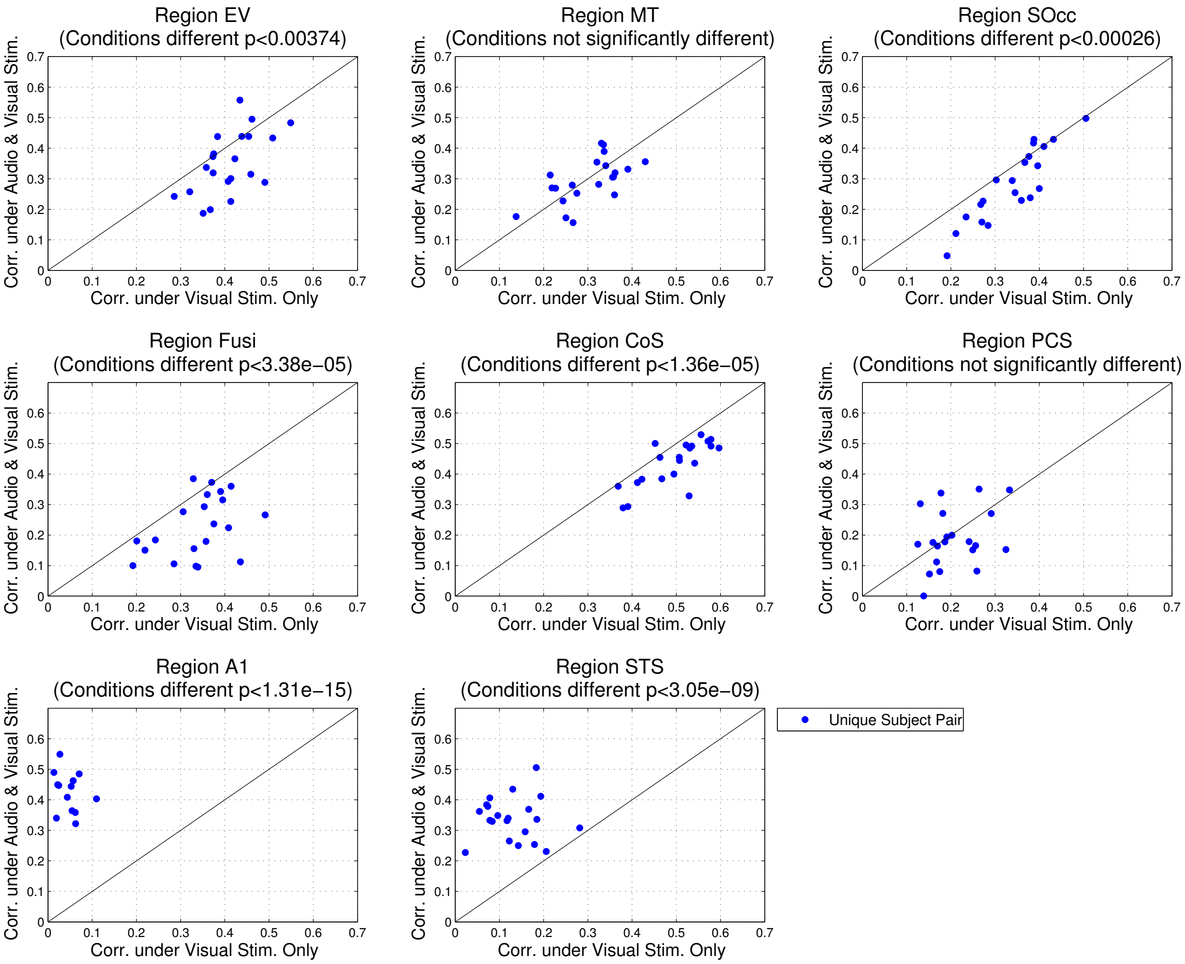


Figure 5.6: The ROI-wise inter-subject correlations are observed to be higher in visual cortex in the visual-only stimulation condition than in the auditory & visual stimulation condition. Correlations in A1 and STS are clearly higher under auditory & visual stimulation. All p-values are computed from a two-sided t-test, with one data point per subject pair.

although they felt more engaged, visual attention itself might have been blurred as subjects could occasionally gain relevant information (e.g. dialogue) from listening alone. Put another way, shutting down auditory processing heightened subjects' sense of visual perception.

5.4.2 Synchronization in A1 under visual-only stimulation

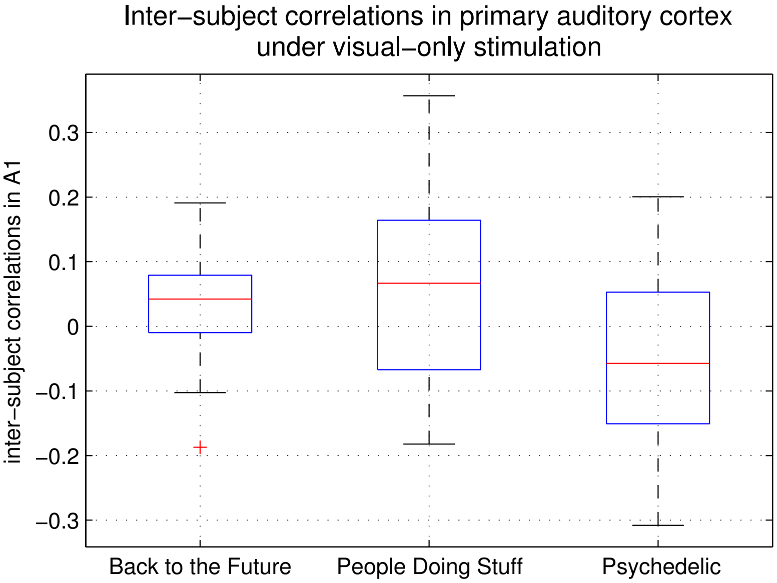


Figure 5.7: Median inter-subject correlations (ROI-wise) in A1 under visual-only stimulation are above zero for the first two videos. Box plots indicate median, 25th, 75th percentiles, minimum and maximum values over the set consisting of both brain sides in each of 21 subject pairs. *Back to the Future* and "People Doing Stuff" are both higher than the psychedelic video ($p < 2 \times 10^{-4}$, one-sided t-test) and "People Doing Stuff" is slightly higher than *Back to the Future* ($p < .15$, one-sided t-test).

Fig. 5.7 shows the distribution of inter-subject correlations in A1 under visual-only stimulation, for each of the 3 videos separately. Subjects correlated during *Back to the Future* and during "People Doing Stuff", significantly above zero ($p < 7 \times 10^{-3}$, one-sided t-test), but not during the psychedelic video. Furthermore, the mean inter-subject correlations were slightly higher during "People Doing Stuff" than during *Back to the Future* (.069 and .047 in left and right hemispheres, averaged across subject pairs, compared to 0.036 and 0.027), although the difference was not enough to establish significance given the variance of the data.

We looked for times in the "Stuff" video at which group activity (ROI-averaged in each hemisphere of each subject, then averaged across subjects) was most above baseline, and significantly so according to a one-sided t-test ($p < 10^{-3}$) on the set of unique time-courses (one per hemisphere per subject). The two times at which subjects were most synchronized in this way with positive activation closely coincided with the two longest segments in the video (of 57): the first involving toilet paper unraveling by the wind of a box fan (lasting about 20 seconds), and the second involving water filling a glass container with plastic bottle caps inside (lasting about 16 seconds). That is, as these clips were being shown, activity in A1 ramped up across subjects, peaking at 151 and 219 seconds into the video (total length: 240 seconds), thus contributing positively to inter-subject correlation.

Again, because subjects all viewed videos in the same order, there is the possibility that correlations in A1 were higher in the first two videos only because they came first, and subjects became bored later, thus decorrelating relative to each other. However, this possibility is weakened by the observation that the correlations during "Stuff", which came second, were stronger than those during *Back to the Future*, despite the latter being shown first, and lasting longer, both of which increase opportunity for synchronization.

Although it is surprising that subjects do correlate even in A1 without auditory stimulation, it is not surprising that among the three videos shown, correlations during "Stuff" were highest. Compared to the psychedelic video, "Stuff" had clear temporal structure; it was comprised of dozens of individual segments, each with a clear beginning and end. Furthermore, many of the segments featured visual stimuli, such as objects dropping and colliding, which, in the natural world, would be accompanied by very salient and identifiable sounds. Viewing these clips without sound elicits auditory imagery, which can modulate the input to A1 and contribute to an increased fMRI signal, even in the absence of spiking activity there [62]. Similarly, "Stuff" had more clear temporal structure and featured more aurally suggestive content than *Back to the Future,* the constituent scenes of which lasted considerably longer than the segments in "Stuff". However, *Back to the Future* did contain some scenes which would be expected to elicit auditory imagery (e.g., a car accelerating, people talking), just not as many as "Stuff", and so it is makes sense that synchronization was intermediate for this video.

5.5 Individual differences in cortical parcellation of anatomy

We investigated whether individual differences in brain anatomy, as determined by automatic parcellation, could explain any of the variability in inter-subject functional correlations. To address this question, we first assessed whether inter-subject correlations were at all stable over time or stimulus conditions. The results of this analysis are shown in Fig. 5.8. We find that the more a pair of subject synchronized under AV stimulation, the more they generally did under V-only stimulation. The correlation between these two values was measured at 0.71, and was significant by a non-parametric shuffle test.

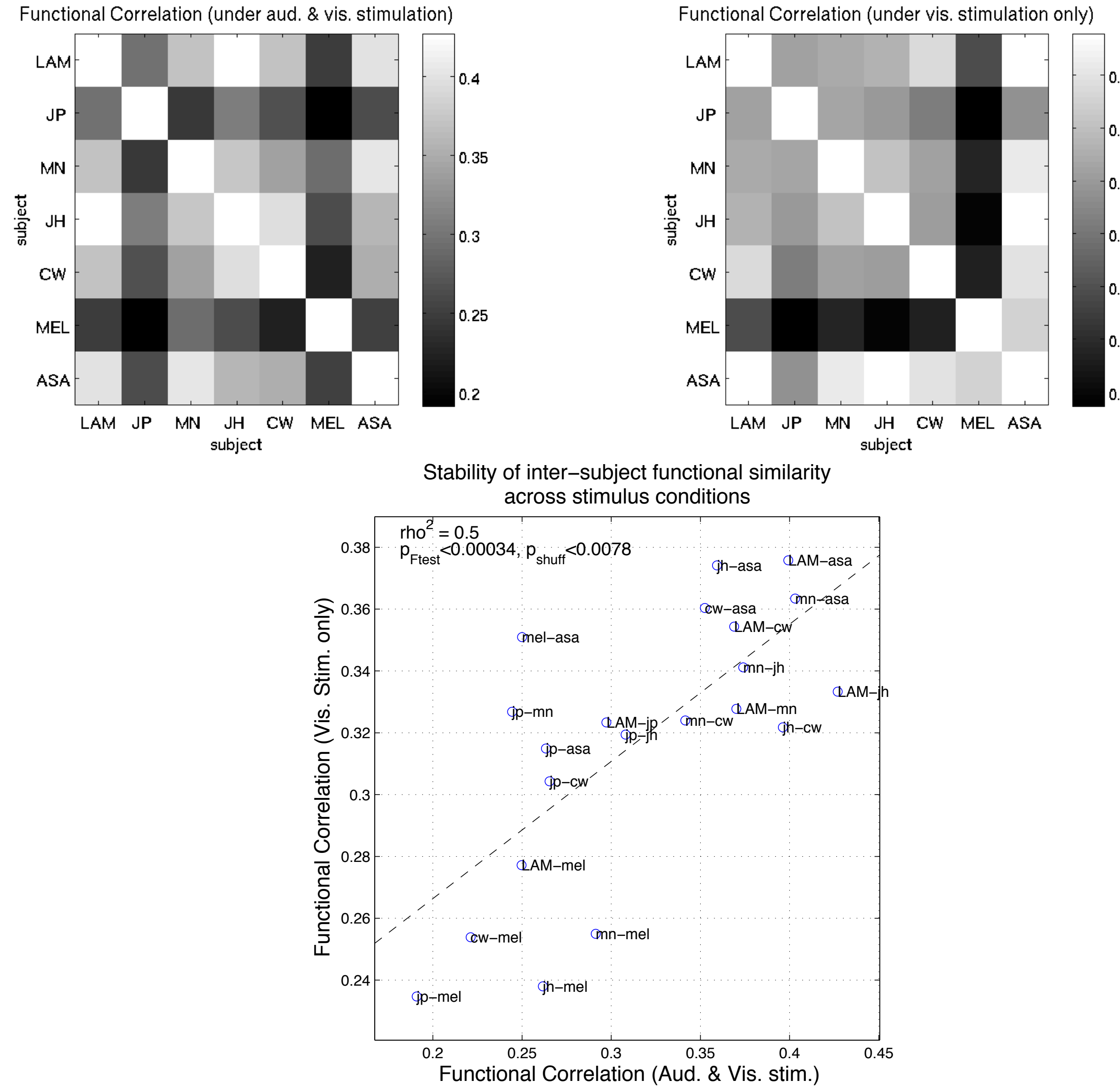


Figure 5.8: Functional similarity between a particular pair of subjects remains consistent across stimulus conditions. Correlations in both stimulus conditions are taken as the average ROI-wise correlations across regions, excluding A1 under V-only stimulation. Shuffling the identity of the subject pairs under AV stimulation results in a correlation as high as the one actually observed in only 0.78% of trials. We also performed a standard F-test to test whether the variability in AV correlations significantly explained the variance in V correlations: the result was that it explained 50% of the variance ($p < 3.4 \times 10^{-4}$).

Next, we developed a measure of anatomical similarity between subjects, based on the relative volumes of certain cortical areas, as provided by FreeSurfer. Fig. 5.9 illustrates the kind of variability we find among different individuals. For instance, whereas subject JP has only 18% of cortex in exclusively frontal regions (inferior frontal gyrus and sulcus, middle frontal gyrus and sulcus, superior frontal gyrus and sulcus, frontomarginal gyrus and sulcus, transverse frontopolar gyrus and sulcus), this fraction is 21% in subject MEL.

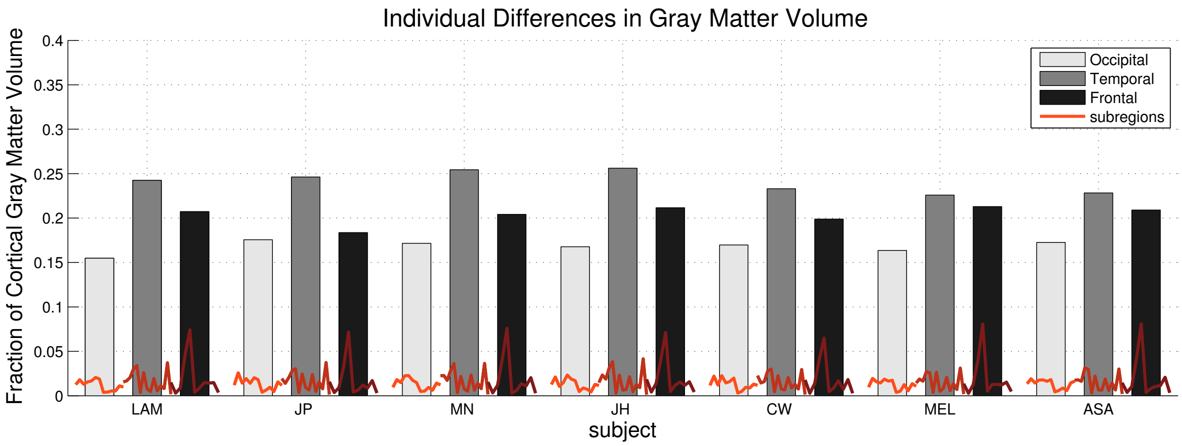


Figure 5.9: Small differences in the distribution of gray matter to different cortical regions are observed between subjects. The red lines in front of each bar show the volume profile over the corresponding sub-regions for each subject. The order of the indices (from left to right in front of each bar) is not meaningful, but the order is the same for all subjects in each region.

We tried a number of methods to encode the anatomical properties of each subject. Every method involved encoding each subject's anatomy as a single anatomy descriptor vector, with an entry for each cortical region, with value equal to the volume that region filled expressed as a fraction of total cortical volume. We also tried absolute volumes, but the results predicted functional similarity slightly worse than the normalized volumes. The anatomical similarity between subjects was defined as the correlation between their anatomy descriptors.

The method which worked best is shown in Fig. 5.10. In that method, the anatomy descriptor had a component for each sub-region in temporal, occipital, and frontal cortices. The correlation between anatomical similarity and functional similarity (defined as the maximum inter-subject functional correlation over the two stimulus conditions, V and AV) was 0.586 ($\rho^2 = 0.34$). Shuffling the true order of the subjects and computing their faked anatomical similarity, then comparing to the unshuffled functional similarity, yielded an anatomical-functional correlation as high as the true one observed in only 0.1% of trials, suggesting that we did not arrive at this high a correlation by chance. Something about the cortical allocation did truly relate to the inter-subject functional similarity.

A radically simpler method for computing anatomical similarity yielded similar results: we used descriptor vectors with only three values, one for each of occipital, temporal, and frontal cortex (volumes summed across sub-regions). In that case, we found a correlation of 0.64 between anatomical

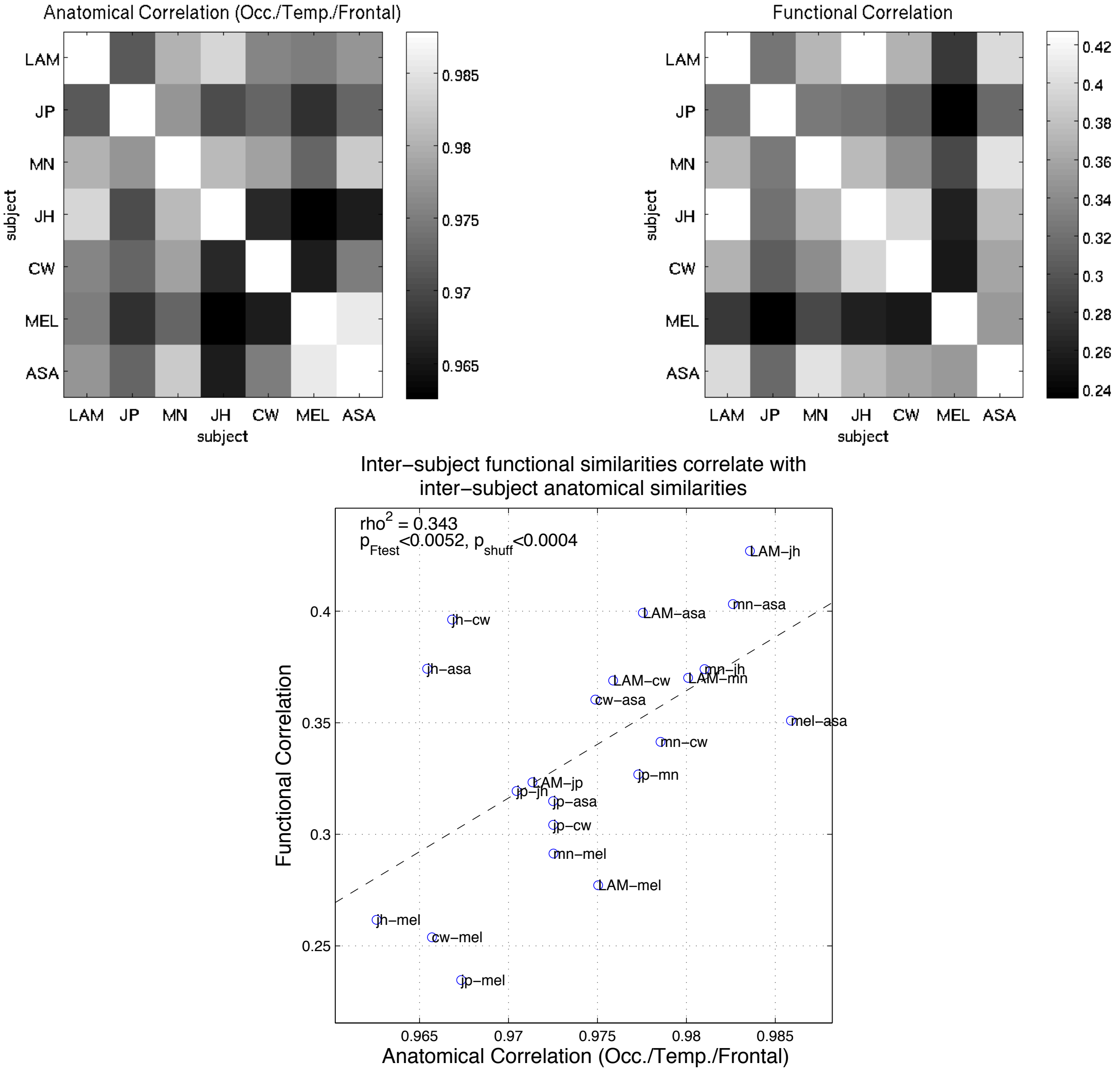


Figure 5.10: Some of the variance in inter-subject similarity is explained by anatomical differences.

and functional similarity, although the shuffle test revealed a significance of only 0.032 (compared with 0.001 for the descriptor based on sub-regional volumes).

Notably, some anatomy descriptors did not yield similarities close to the functional ones. For instance, a descriptor based on all sub-regions *outside* of frontal, temporal, and occipital cortices yielded an anatomical-functional correlation of 0.079, with $p < 0.38$ by the shuffle test. Perhaps more intriguingly, a descriptor based on sub-regions only in occipital cortex was also not significant ($p < 0.56$). This is somewhat surprising since several of the regions in which we measured functional similarity were in occipital cortex. A descriptor based on sub-regions in frontal cortex alone was slightly better ($p < 0.17$), but still not significant. Sub-regions in temporal cortex alone did somewhat significantly correlate with functional similarity, however (correlation=0.35, $p < 0.036$ by shuffle test).

The best explanation for the observed relationship between anatomical and functional similarity is that, because ROIs were defined anatomically, subjects with more similar anatomical parcellation were more likely to have more similar functional specialization within the ROIs relative to those with less similar anatomy. For instance, early visual cortex may have consisted of 95% of V1 and V2 for one subject, and 90% for another. This explanation agrees with the observation that temporal regions were most predictive of functional similarity, since most of the ROIs were in or directly adjacent to temporal cortex. The fact that frontal cortex was slightly more predictive than occipital cortex could be meaningless, as neither individually achieved significance, or, most speculatively, it might mean that synchronization between a pair of subjects is actually related to similarities in high-level cognitive functions arising from similarity in frontal cortex anatomy.

5.6 Discussion and future directions

In this chapter, we further quantitatively constrained the extent to which people differ in visual experience. The temporal patterns elicited in visual cortex in response to a video stimulus correlate across subjects at an average of 78% the same-subject level; the fact that subjects exhibit such a high degree of temporal synchronization at the neural level during visual experience provides additional evidence that the cortical manifestations of visual experience are not divergent among different people; rather, they are surprisingly similar, suggesting that the associated qualia may be as well, if we take the view that more similar physical processes likely give rise to more similar qualia.

We found that anatomical similarity (or anatomical label similarity) could significantly model neural pattern similarity. This result suggests that identical twins should have very similar neural patterns in time, and possibly distributed in space as well, even if their life experiences are different, for instance due to being separated at birth, though this last condition would be hard to meet experimentally at a large scale. Also, based on our experimental procedures, it is impossible to know whether functional similarities correlated with anatomical parcellation or anatomy itself (though these two *should* be closely related); to address this, one would require an expert anatomist to annotate each brain.

We also found evidence that subjects neurally diverge in visual cortex when an audio track is introduced to a video. One possible cause for this is that eye movement patterns also diverge relative to watching a clip without sound: a possibility which could be confirmed with a simple and elegant follow-up experiment. The result of gaze decorrelation in the presence of audio may be especially surprising if one supposes that an audio track introduces high-level, semantic information which would not only engage the overall alertness of the viewer but also direct their spatial attention towards topic-relevant visual locations.

In Chapter 3, we found that familiarity with a face enhanced its consistency: out in the real world, most of visual experience is full of motion. Therefore, we predict this result should generalize to moving stimuli as well. In particular, we would predict that the temporal neural pattern consistency of a video stimulus increases with its extent of familiarity. One could test this by measuring whether people who have seen a video more times are more self-correlated than those who have not. Further, one could test whether familiarity, not with the video itself, but rather the elements it contains – such as faces, voices, places, etc. – also increases self-correlation/consistency.

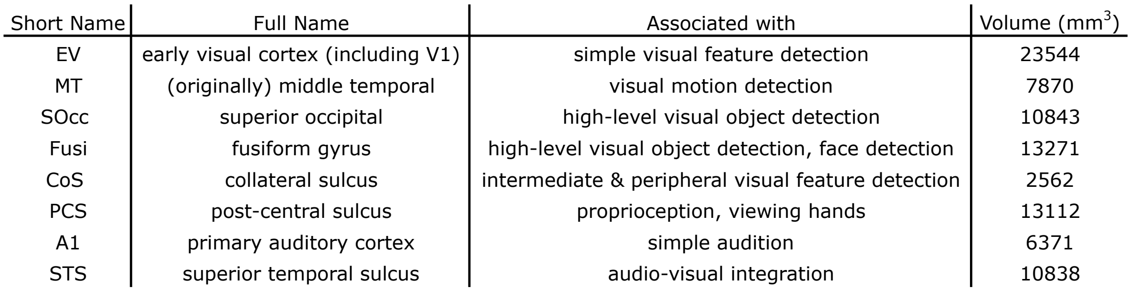
5.7 Appendix

Subjects provided informed written consent prior to the experiments. The Caltech Institutional Review Board approved all experimental procedures.

5.7.1 MRI data acquisition and preprocessing steps

MRI data were collected using a Siemens (Erlangen, Germany) MAGNETOM Trio, A Tim System 3T, and a 12-channel head coil. All functional images (blood oxygenation level dependent (BOLD) T2*-weighted) were acquired using a gradient-echo echo-planar (EPI) sequence. Functional images consisting of 30 slices with $3 \times 3 \times 4$mm voxels were acquired with a TR of 2 seconds (acquisition matrix 64x64, flip angle 80°, echo time 30 ms). Anatomical images were acquired using at T1-weighted MPRAGE sequence with the same 12-channel head coil. Functional volumes were slice-time corrected, then normalized to the standard MNI152 (Montreal Neurological Institute atlas) space using a single 12-parameter 3-dimensional affine transformation (this transformation incorporated self-motion correction), and resampled to $3.3 \times 3.3 \times 3.3$ mm. Anatomical images were co-registered with the functional images using a 12-parameter affine transformation as well. Finally, to reduce contributions to the imaging data not due to functional activation, from the time-course at each voxel was subtracted the best fit to it from the first 21 Legendre polynomials (interpolated between -1 and 1), comprising a "drift" estimate, and then a low-pass filter (cut-off frequency of 0.12 Hz) was applied.

5.7.2 Defining regions of interest (ROIs) in the brain



Short Name	Full Name	Associated with	Volume (mm^3)
EV	early visual cortex (including V1)	simple visual feature detection	23544
MT	(originally) middle temporal	visual motion detection	7870
SOcc	superior occipital	high-level visual object detection	10843
Fusi	fusiform gyrus	high-level visual object detection, face detection	13271
CoS	collateral sulcus	intermediate & peripheral visual feature detection	2562
PCS	post-central sulcus	proprioception, viewing hands	13112
A1	primary auditory cortex	simple audition	6371
STS	superior temporal sulcus	audio-visual integration	10838

Figure 5.11: Brain regions of interest for the dynamic image experiment. Volumes are bilateral, and based on averages across subjects, in the MNI152 space.

Based on the published experiments similar to this one [37, 38], eight regions of interest were selected for investigation. They are summarized in Fig. 5.11. Regions were all defined relative to the cortical segmentation computed automatically by FreeSurfer (http://surfer.nmr.mgh.harvard.edu/). Because they were defined anatomically, and not based on a functional experiment, these region labels are approximate. Each subject's anatomical-based ROI labels were projected into the MNI152 space separately, landing in sightly different locations for each subject.

Chapter 6

Conclusions

In Part I of the thesis, we showed that distinctive and familiar faces have more consistent representation in early visual cortex, even without any enhanced activation magnitude there: that is, the perception of such faces at the low neural level is relatively fixed compared to the perception of unfamiliar faces. Based on our results, we proposed a framework within which to understand the learning of new faces: faces initially distinctive are more consistent in neural representation; such consistency may preferentially cause memories to form, the presence of which creates feedback into early visual cortex which further consolidates representational consistency, thus completing the cycle. Simultaneously, at the perceptual level, as a face becomes more familiar, all those around it become more dissimilar looking from each other, possibly due to a process in which neurons learn to a likeness function to the face, in what can be termed an exemplar-based coding system.

Whereas in Part I we focused on what makes a visual stimulus appear to look different to different individuals, in Part II we studied the extent of similarity in the perception of general visual stimuli. We showed that different subjects have similar inter-category neural pattern distances (mean correlation 0.55 among category pairs selected from different scans), and that they exhibit similar patterns of neural activity in time in response to videos (in visual cortex, 78% the same-subject level). One may think of these as introducing some quantitative constraints on the extent to which people may differ in visual experience, assuming for practical reasons that more similar physical processes give rise to more similar qualia.

Throughout Part I of this thesis, we developed a connection between memory/familiarity and distinctiveness. However, such memory was assessed by a single interview, and not updated nor longitudinally varied over the course of the experiment. A follow-up study to both the psychophysical experiments in Chapter 2 and the fMRI experiments of Chapter 3 could investigate how distinctiveness and neural pattern consistency coevolve during the gradual learning of a new face.

Finally, to further strengthen the findings we reported on neural pattern consistency in response to familiar and distinctive faces, it would be desirable to investigate such patterns using other modalities of neuroscientific investigation, such as EEG, or intracranial recordings. We hypothesize that, in V1, one would find that familiarity and distinctiveness correlate better with measures of consistency among populations of neurons than with their median spiking rate or other measures of sheer activation magnitude. Furthermore, if this consistency were to be found, it would be possible to test whether it is something which arises immediately following stimulus onset, or only following some latency, after which signals from higher areas have time to feed back. If it is discovered that feedback is in fact *not* necessary for neural pattern consistency in V1, that might evidence of a form of sensory memory, in which the same cells within redundant subpopulations fire in response to a stimulus.

Bibliography

[1] R Adolphs, D Tranel, and H Damasio. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, 1994.

[2] R Adolphs, D Tranel, and H Damasio. Fear and the human amygdala. *The Journal of Social Neuroscience*, 1995.

[3] JY Baudouin and M Gallay. Is face distinctiveness gender based? *Journal of Experimental Psychology: Human Perception and Performance*, 2006.

[4] J M Beale and FC Keil. Categorical effects in the perception of faces. *Cognition*, 57:217–239, 1995.

[5] MS Beauchamp, KE Lee, BD Argall, and A Martin. Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41(5):809–823, 2004.

[6] P J Benson and D I Perrett. Visual processing of facial distinctiveness. *Perception*, 23(1):75–93, 1994.

[7] RM Birn, JB Diamond, MA Smith, and PA Bandettini. Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *NeuroImage*, 31(4):1536–1548, 2006.

[8] RM Birn, K Murphy, and PA Bandettini. The effect of respiration variations on independent component analysis results of resting state functional connectivity. *Human brain mapping*, 29(7):740–750, 2008.

[9] GM Boynton, SA Engel, GH Glover, and DJ Heeger. Linear systems analysis of functional magnetic resonance imaging in human V1. *The Journal of neuroscience*, 1996.

[10] Alyssa A Brewer, Junjie Liu, Alex R Wade, and Brian A Wandell. Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nature Neuroscience*, 8(8):1102–1109, August 2005.

[11] JB Brewer, Z Zhao, JE Desmond, GH Glover, and JDE Gabrieli. Making memories: brain activity that predicts how well visual experience will be remembered. *Science*, 281(1185), 1998.

[12] George Bush, Phan Luu, and Michael I Posner. Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6):215–222, June 2000.

[13] R Caldara and H Abdi. Simulating the 'other-race' effect with autoassociative neural networks: further evidence in favor of the face-space model. *Perception-London*, 2006.

[14] A.E. Cavanna and M.R. Trimble. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583, January 2006.

[15] D.J. Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 1995.

[16] D.L. Collins. 3D model-based segmentation of individual brain structures from magnetic resonance imaging data. *Thesis, McGill Univ., Canada*, 1994.

[17] DD Cox and RL Savoy. fMRI Brain Reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 2003.

[18] T. N. Wiesel D H Hubel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215, March 1968.

[19] AR Damasio, H Damasio, and GW Van Hoesen. Prosopagnosia. *Neurology*, 32(4), 1982.

[20] Adélaïde de Heering, Sarah Houthuys, and Bruno Rossion. Holistic face processing is mature at 4 years of age: Evidence from the composite face effect. *Journal of Experimental Child Psychology*, 96(1):57–70, January 2007.

[21] C Devue, F Collette, E Balteau, C Degueldre, A Luxen, P Maquet, and S Brédart. Here I am: the cortical correlates of visual self-recognition. *Brain Research*, 1143:169–182, 2007.

[22] RJ Douglas and KAC Martin. Recurrent neuronal circuits in the neocortex. *Current biology*, 17(13), 2007.

[23] B Duchaine and K Nakayama. The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44:576–585, 2006.

[24] J Duncan and GW Humphreys. Visual search and stimulus similarity. *Psychological review*, 96:433–458, 1989.

[25] E Eger, S Schweinberger, RJ Dolan, and RN Henson. Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *NeuroImage*, 2005.

[26] I Fried and KA MacDonald. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron*, 1997.

[27] JL Gallant, J Braun, and DC VANESSEN. Selectivity for Polar, Hyperbolic, and Cartesian Gratings in Macaque Visual-Cortex. *Science*, 259(5091):100–103, 1993.

[28] JL Gallant, CE Connor, S Rakshit, JW Lewis, and DC VANESSEN. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of Neurophysiology*, 76(4):2718–2739, 1996.

[29] I Gauthier, P Skudlarski, JC Gore, and AW Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2):191–197, 2000.

[30] Michael Goard and Yang Dan. Basal forebrain activation enhances cortical coding of natural scenes. *Nature Neuroscience*, 12(11):1444–1449, October 2009.

[31] MI Gobbini and JV Haxby. Neural systems for recognition of familiar faces. *Neuropsychologia*, 45(1):32–41, 2007.

[32] AJ Golby, JDE Gabrieli, JY Chiao, and JL Eberhardt. Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, 4(8):845–850, 2001.

[33] Y Golland, S Bentin, H Gelbard, Y Benjamini, R Heller, Y Nir, U Hasson, and R Malach. Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. *Cerebral Cortex*, 17(4):766, 2007.

[34] G. Griffin, A.D. Holub, and P. Perona. The Caltech-256. *Caltech Technical Report*, 2007.

[35] K Grill-Spector, Z Kourtzi, and N Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10-11):1409–1422, 2001.

[36] M Haldane, G Cunningham, C Androutsos, and S Frangou. Structural brain correlates of response inhibition in Bipolar Disorder I. *Journal of Psychopharmacology*, 22(2):138–143, January 2008.

[37] U Hasson, Y Nir, I Levy, G Fuhrmann, and R Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634, 2004.

[38] U Hasson, E Yang, I Vallines, DJ Heeger, and N Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539, 2008.

[39] JV Haxby, LG Ungerleider, B Horwitz, JM Maisog, SI Rapoport, and CL Grady. Face encoding and recognition in the human brain. In *Proceedings of the National Aacademy of Sciences*, pages 922–927, 1996.

[40] JJ Heisz and DI Shore. More efficient scanning for familiar faces. *Journal of Vision*, 8(1), 2008.

[41] C. Helle and P. Perona. Pasadena Houses 2000, http://www.vision.caltech.edu/html-files/archive.html.

[42] J Hocking and CJ Price. The role of the posterior superior temporal sulcus in audiovisual processing. *Cerebral Cortex*, 18(10):2439, 2008.

[43] KL Hoffman, AA Ghazanfar, I Gauthier, and NK Logothetis. Category-specific responses to faces and objects in primate auditory cortex. *Frontiers in Systems Neuroscience*, 1, 2007.

[44] KL Hoffman, KM Gothard, and MC Schmid. Facial-expression and gaze-selective responses in the monkey amygdala. *Current biology*, 2007.

[45] DH Hubel and TN Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 19:215–243, 1968.

[46] IP Jääskeläinen, K Koskentalo, MH Balk, T Autti, J Kauramäki, C Pomren, and M Sams. Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. *The Open Neuroimaging Journal*, 2:14, 2008.

[47] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[48] Y Kamitani and F Tong. Decoding seen and attended motion directions from activity in the human visual cortex. *Current biology*, 16:1096–1102, 2006.

[49] N Kanwisher and J Liu. Perception of Face Parts and Face Configurations: An fMRI Study. *Journal of Cognitive Neuroscience*, 2010.

[50] N Kanwisher, J McDermott, and MM Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, 1997.

[51] Andrzej Kasinski, Andrzej Florek, and Adam Schmidt. The PUT Face Database. *Image Processing & Communications*, 13(3-4):59–64, 2008.

[52] KN Kay, SV David, RJ Prenger, KA Hansen, and JL Gallant. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. *Human brain mapping*, 2008.

[53] KN Kay, T Naselaris, RJ Prenger, and JL Gallant. Identifying natural images from human brain activity. *Nature*, 2008.

[54] Gabriel Kreiman, Christof Koch, and Itzhak Fried. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3(9):946–953, September 2000.

[55] N Kriegeskorte, E Formisano, B Sorger, and R Goebel. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104(51):20600, 2007.

[56] N Kriegeskorte, M Mur, DA Ruff, R Kiani, J Bodurka, and et al. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 2008.

[57] K Lee, G Byatt, and G Rhodes. Caricature effects, distinctiveness, and identification: testing the face-space framework. *Psychological science*, 11(5):379–385, September 2000.

[58] CM Leonard, ET Rolls, and FAW Wilson. Neurons in the amygdala of the monkey with responses selective for faces. *Behavioural brain research*, 1985.

[59] DA Leopold, IV Bondar, and MA Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–575, 2006.

[60] G Loffler, G Yourganov, F Wilkinson, and HR Wilson. fMRI evidence for the neural representation of faces. *Nature Neuroscience*, 8(10):1386–1391, 2005.

[61] N K Logothetis, J Pauls, M Augath, T Trinath, and A Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, July 2001.

[62] NK Logothetis and BA Wandell. Interpreting the BOLD signal. *Annual review of physiology*, 2004.

[63] Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7(7):293–299, July 2003.

[64] TM Mitchell, SV Shinkareva, A Carlson, and KM Chang. Predicting human brain activity associated with the meanings of nouns. *Science*, 2008.

[65] Sebastian Moeller, Winrich A Freiwald, and Doris Y Tsao. Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*, 320(5881):1355–1359, June 2008.

[66] CJ Mondloch, TL Lewis, DR Budreau, D Maurer, JL Dannemiller, BR Stephens, and KA Kleiner-Gathercoal. Face perception during early infancy. *Psychological science*, 10(5):419–422, 1999.

[67] J Mutch and D Lowe. Multiclass Object Recognition with Sparse, Localized Features. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:11–18, 2006.

[68] Vaidehi Natu, David Raboy, and Alice J O'toole. Neural correlates of own- and other-race face perception: Spatial and temporal response differences. *NeuroImage*, pages 1–9, November 2010.

[69] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996.

[70] Anitha Pasupathy and Charles E Connor. Population coding of shape in area V4. *Nature Neuroscience*, 5(12):1332–1338, December 2002.

[71] KS Pilz, IM Thornton, and HH Bulthoff. A search advantage for faces learned in motion. *Experimental Brain Research*, 171(4):436–447, 2006.

[72] SM Platek, JW Loughead, RC Gur, S Busch, K Ruparel, N Phend, IS Panyavin, and DD Langleben. Neural substrates for functionally discriminating self-face from personally familiar faces. *Human brain mapping*, 27(2):91–98, 2006.

[73] Joaquim Radua, Mary L Phillips, Tamara Russell, Natalia Lawrence, Nicolette Marshall, Sridevi Kalidindi, Wissam El-Hage, Colm McDonald, Vincent Giampietro, Michael J Brammer, Anthony S David, and Simon A Surguladze. Neural response to specific components of fearful faces in healthy and schizophrenic adults. *NeuroImage*, 49(1):939–946, January 2010.

[74] Rajeev D S Raizada and Nikolaus Kriegeskorte. Pattern-information fMRI: New questions which it opens up and challenges which face it. *International Journal of Imaging Systems Technology*, 20(1):31–41, February 2010.

[75] M Ramon, S Caharel, and B Rossion. The speed of recognition of personally familiar faces. *Perception*, 2011.

[76] Leila Reddy, Naotsugu Tsuchiya, and Thomas Serre. Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, pages 1–8, December 2009.

[77] Gillian Rhodes, Susan Brennan, and Susan Carey. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, 19(4):473–497, October 1987.

[78] M Riesenhuber and T Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, November 1999.

[79] JJ Ryu. Representations of familiar and unfamiliar faces as revealed by viewpoint-aftereffects. *Vision Research*, 2006.

[80] M R Sabuncu, B D Singer, B Conroy, R E Bryan, P J Ramadge, and J V Haxby. Function-based Intersubject Alignment of Human Cortical Anatomy. *Cerebral Cortex*, 20(1):130–140, 2010.

[81] GE Schafe and JE LeDoux. Memory consolidation of auditory pavlovian fear conditioning requires protein synthesis and protein kinase A in the amygdala. *The Journal of neuroscience*, 20(18), 2000.

[82] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–6429, April 2007.

[83] SV Shinkareva, RA Mason, VL Malave, W Wang, TM Mitchell, and MA Just. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS One*, 3(1):1394, 2008.

[84] S Shipp, JDG Watson, and RSJ Frackowiak. Retinotopic maps in human prestriate visual cortex: the demarcation of areas V2 and V3. *NeuroImage*, 1995.

[85] A Slater and PC Quinn. Face recognition in the newborn infant. *Infant and Child Development*, 10, 2001.

[86] LR Squire. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological review*, 1992.

[87] Greg J Stephens, Lauren J Silbert, and Uri Hasson. Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the United States of America*, 107(32):14425–14430, August 2010.

[88] M Sugiura, NJ Shah, K Zilles, and GR Fink. Cortical representations of personally familiar objects and places: functional organization of the human posterior cingulate cortex. *Journal of Cognitive Neuroscience*, 17(2):183–198, 2005.

[89] M Taylor, M Arsalidou, S Bayless, D Morris, JW Evans, and EJ Barbeau. Neural correlates of personally familiar faces: Parents, partner and own faces. *Human brain mapping*, 30:2008–2020, 2009.

[90] S Thorpe, D Fize, and C Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, June 1996.

[91] G Tononi. An information integration theory of consciousness. *BMC neuroscience*, 2004.

[92] RB Tootell, JB Reppas, KK Kwong, R Malach, RT Born, TJ Brady, BR Rosen, and JW Belliveau. Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *The Journal of neuroscience*, 15(4):3215–3230, 1995.

[93] Doris Y Tsao, Winrich A Freiwald, Roger B H Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, February 2006.

[94] Doris Y Tsao, Sebastian Moeller, and Winrich A Freiwald. Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19514–19519, December 2008.

[95] Doris Y Tsao, Nicole Schweers, Sebastian Moeller, and Winrich A Freiwald. Patches of face-selective cortex in the macaque frontal lobe. *Nature Neuroscience*, 11(8):877–879, August 2008.

[96] T Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, 43(2):161–204, May 1991.

[97] Tim Valentine and Vicki Bruce. Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 40(3):300–305, 1986.

[98] G Van Belle, M Ramon, P Lefèvre, and B Rossion. Fixation patterns during recognition of personally familiar and unfamiliar faces. *Frontiers in Cognitive Science*, 1(20), 2010.

[99] J R Vokey and J D Read. Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory and Cognition*, 20(3):291–302, May 1992.

[100] AD Wagner, DL Schacter, M Rotte, W Koutstaal, A Maril, AM Dale, BR Rosen, and RL Buckner. Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, 281(5380):1188–1191, 1998.

[101] GE Walton and TGR Bower. Newborns form "prototypes" in less than 1 minute. *Psychological science*, 4(3):203–205, 1993.

[102] G Xue, Q Dong, C Chen, Z Lu, J A Mumford, and R A Poldrack. Greater Neural Pattern Similarity Across Repetitions Is Associated with Better Memory. *Science*, 330(6000):97–101, October 2010.