Neuron 50, 159–172, April 6, 2006 ©2006 Elsevier Inc. DOI 10.1016/j.neuron.2006.03.012

Evaluation of a Shape-Based Model of Human Face Discrimination Using fMRI and Behavioral Techniques

Xiong Jiang,[1] Ezra Rosen,[1] Thomas Zeffiro,[2] John VanMeter,[2] Volker Blanz,[3] and Maximilian Riesenhuber[1,*]
[1] Department of Neuroscience
Georgetown University Medical Center
Research Building Room WP-12
3970 Reservoir Road NW
Washington, District of Columbia 20007
[2] Center for Functional and Molecular Imaging
Georgetown University Medical Center
Washington, District of Columbia 20007
[3] Max-Planck-Institut für Informatik
Saarbrücken
Germany

Summary

Understanding the neural mechanisms underlying object recognition is one of the fundamental challenges of visual neuroscience. While neurophysiology experiments have provided evidence for a "simple-to-complex" processing model based on a hierarchy of increasingly complex image features, behavioral and fMRI studies of face processing have been interpreted as incompatible with this account. We present a neurophysiologically plausible, feature-based model that quantitatively accounts for face discrimination characteristics, including face inversion and "configural" effects. The model predicts that face discrimination is based on a sparse representation of units selective for face shapes, without the need to postulate additional, "face-specific" mechanisms. We derive and test predictions that quantitatively link model FFA face neuron tuning, neural adaptation measured in an fMRI rapid adaptation paradigm, and face discrimination performance. The experimental data are in excellent agreement with the model prediction that discrimination performance should asymptote as faces become dissimilar enough to activate different neuronal populations.

Introduction

Understanding how the human brain performs complex cognitive tasks requires understanding the mechanistic relationship of stimuli, neural activity, and behavior. Due to the fundamental importance of object recognition for cognition, significant research effort has focused on elucidating these relationships in this domain. In visual physiology, there is now broad support for a general class of models of cortical visual processing based on a hierarchical bottom-up organization with succeeding stages sensitive to image features of increasing specificity and tolerance to stimulus transformations such as scaling or translation (Hubel and Wiesel, 1962; Kobatake and Tanaka, 1994; Logothetis and Sheinberg, 1996; Ungerleider and Haxby, 1994).

However, this class of models is widely thought to be insufficient to explain human discrimination performance for the important object class of faces. In particular, the so-called "face inversion effect" (FIE), referring to the observation that inversion appears to disproportionately affect the discrimination of faces compared to other objects (Yin, 1969), has given rise to theories postulating that face discrimination is based on face-specific mechanisms, such as configural coding (Carey and Diamond, 1986), that are different from the shape-based mechanisms used to discriminate nonface objects. Regarding the neural substrate of face perception, numerous imaging studies have suggested that the "fusiform face area" (FFA) plays a pivotal role in human face recognition (Gauthier et al., 2000; Grill-Spector et al., 2004; Haxby et al., 2000; Kanwisher et al., 1997; Loffler et al., 2005; Rotshtein et al., 2005; Yovel and Kanwisher, 2004). Using faces of famous individuals, a recent study (Rotshtein et al., 2005) has argued for an identity-based representation in the FFA, with neurons showing categorical tuning for different individuals. Another study has posited a model of face representation in the FFA in which faces are encoded in a global face space by their direction and distance from a "mean" face (Loffler et al., 2005). In contrast, electrophysiological studies have suggested that the face representation in monkey cortex is *sparse*, i.e., a given face is represented by the activation of a small subset of neurons, part of a larger population, whose preferred face stimuli are similar to the currently viewed face, and whose joint activation pattern is sufficient to encode the identify of a particular face (Young and Yamane, 1992, but see Rolls and Tovee, 1995). Sparse coding has distinct computational advantages over other coding schemes (e.g., "grandmother codes" and distributed codes), in terms of energy efficiency and ease of learning, and has been posited to be a general principle underlying sensory processing in cortex (Olshausen and Field, 2004).

Using a combination of computational modeling, fMRI, and behavioral techniques, we have found that both human face discrimination performance and FFA activation can be quantitatively explained by a simple shape-based model in which human face discrimination is based on a sparse code of tightly tuned face neurons.

Results

Model Face Neurons

Our hypothesis is that human face discrimination is mediated by neurons tuned to face shapes that are located in the FFA, similar to the "face neurons" observed in the monkey (Baylis et al., 1985; Desimone, 1991; Perrett et al., 1982; Young and Yamane, 1992). We model face neuron properties using our previously published computational model of object recognition in cortex (Riesenhuber and Poggio, 1999a, 2002). The model (see Figure 1) simulates processing in the cortical ventral

*Correspondence: mr287@georgetown.edu

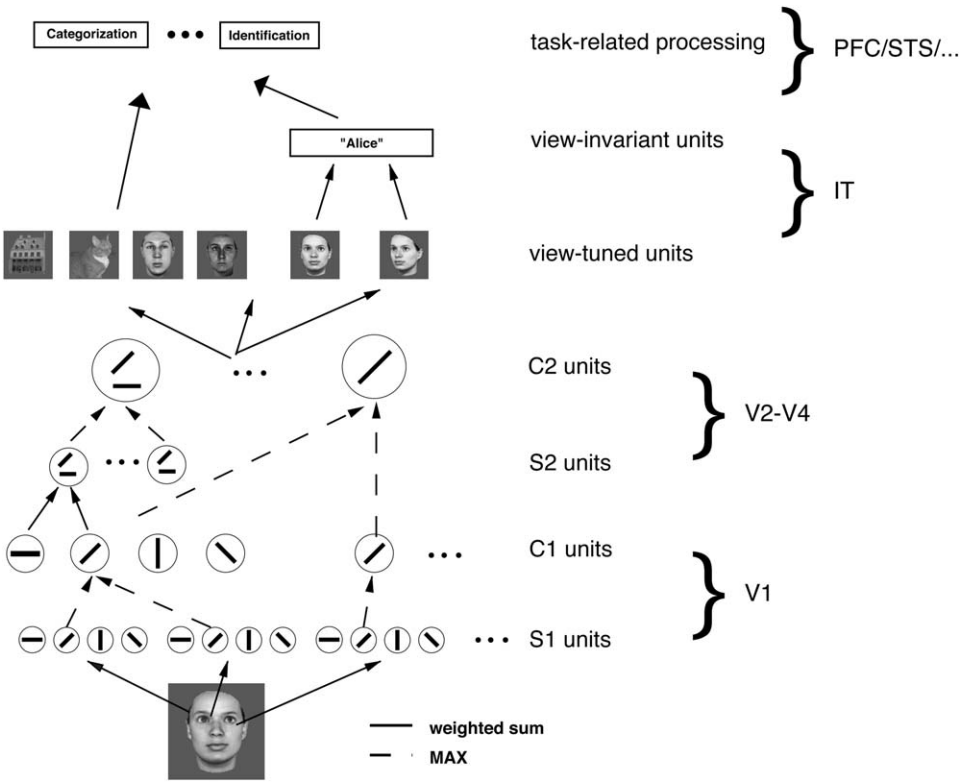

Figure 1. Scheme of Our Model of Object Recognition in Cortex

Feature specificity and invariance to translation and scale are gradually built up by a hierarchy of "S" (using a weighted sum operation, solid lines) and "C" layers (using a MAX pooling operation, dashed lines), respectively (see Experimental Procedures), leading to view-tuned units (VTUs) that show shape tuning and invariance properties in quantitative agreement with physiological data from monkey IT. These units can the provide input to task-specific circuits located in higher areas, e.g., prefrontal cortex.

visual stream thought to mediate object recognition in humans and other primates (Ungerleider and Haxby, 1994). This stream extends from primary visual cortex, V1, to inferotemporal cortex, IT, forming a processing hierarchy in which the complexity of neurons' preferred stimuli and receptive field sizes progressively increase. At the top of the ventral stream, in IT, so-called "view-tuned" cells are differentially sensitive to views of complex objects while being tolerant to changes in stimulus size and position (Logothetis and Sheinberg, 1996). View-tuned units (VTUs, see Figure 1) in the model have been able to quantitatively account for the shape tuning and invariance properties of IT neurons (Freedman et al., 2003; Riesenhuber and Poggio, 1999b).

As Figure 1 illustrates, the tuning and stimulus specificity of a VTU is defined by its connectivity to afferent units, called C2, which are the model counterpart of neurons in V4 or posterior IT. In particular, VTUs can differ in the number and identity of their C2 afferents, their tightness of tuning (σ), and their response variability (see Equation 1 in Experimental Procedures). Varying these parameters allows us to create face-tuned units of varying specificity (Figure 2). Note that a unit's response is based on the shape difference between the unit's preferred face and the current stimulus, without any explicit (categorical) identity tuning. We modeled the FFA as a population of 180 such face units, each tuned to a different face.

Selecting Realistic Parameter Sets

Given our goal of obtaining a realistic model of human face discrimination, we constrained model parameters by simulating a recent behavioral face discrimination study (Riesenhuber et al., 2004), employing the same stimuli and conditions as used in the experiment (Figure 3). In that study, subjects had to detect "featural" or "configural" changes in face pairs that were presented either upright or inverted. Figure 4 (black points with error bars) shows the experimental results for the different trial types.

It is important to note that this face discrimination experiment involved two stimulus manipulations that have been widely used to investigate face recognition mechanisms: (1) stimulus inversion and (2) "featural" versus "configural" stimulus changes. The latter manipulations are motivated by a conceptual model of face processing in which the visual system is thought to process faces by first recognizing "face features," i.e., the eyes, mouth, and nose, and then computing their "configuration" as a function of the spatial position of the "features" (Carey and Diamond, 1986; Farah et al., 1995; Mondloch et al., 2002). The FIE is then hypothesized to arise from a failure to calculate face configuration for inverted faces, forcing the visual system to rely on "featural processing" alone. It is thus of special interest to determine whether our model, with its generic shape-based processing and no explicit coding of either "face features" or

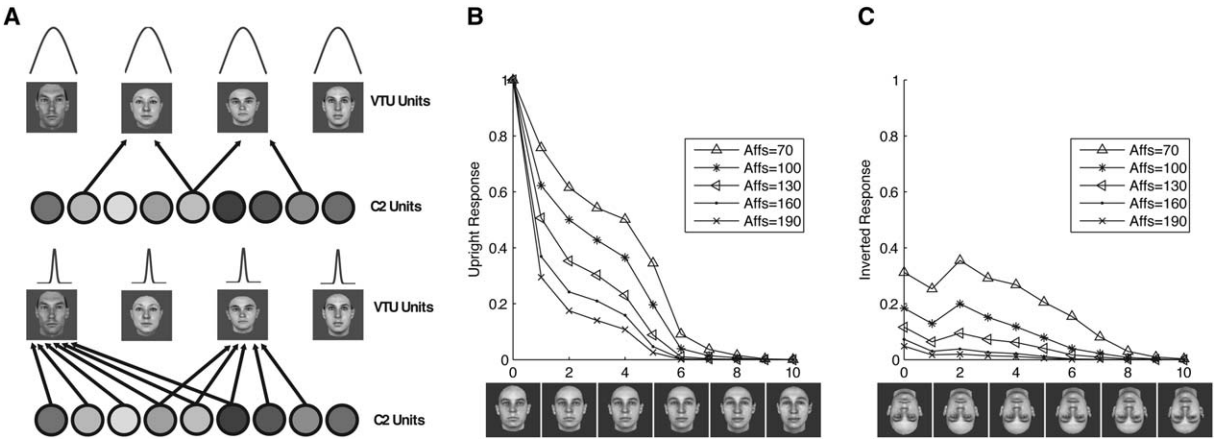

Figure 2. Modeling Face Units of Varying Specificity

(A) shows a cartoon of how tuning specificity can be varied from broadly tuned (upper figure) to more sharply tuned (lower figure) by varying tuning width σ and number of afferents *d* to each unit. For clarity, only the C2 to VTU connections for two face units each are shown. (B) and (C) show the responses of example face units (with σ = 0.1 and varying *d)* to their preferred face (set to be the leftmost face shown in [B]), and increasingly dissimilar faces along one example morph line for upright (B) and inverted (C) orientations.

"configuration," can account for these qualitative effects and also for the quantitative behavioral performance.

Simulating the same/different paradigm used in Riesenhuber et al. (2004) requires modeling the mechanism by which face unit activity gives rise to behavioral decisions. Following the aforementioned electrophysiological results that argue for a sparse face representation in cortex, we also varied the number of face units whose activation entered into the "same/different" decision, in addition to varying face unit tuning parameters. In particular, we assumed that in the same/different paradigm, subjects remembered the activation levels of the n_M most activated face units and then compared the activation pattern over these units to that resulting from the second stimulus. This is also a computationally efficient strategy for object recognition in scenes with more than one simultaneously presented object (Riesenhuber and Poggio, 1999a). Subjects were assumed to choose a "different" response if the difference between the two activation patterns exceeded a threshold τ, with τ selected by a brute-force search procedure for each parameter set to obtain the best fit to the experimental data.

We found that, out of the 6804 parameter sets investigated (see Experimental Procedures), 35 produced fits with $p > 0.05$ (i.e., no significant difference between the experimental data and the model fit), demonstrating that although the fitting problem is nontrivial, the model is nevertheless able to quantitatively account for the

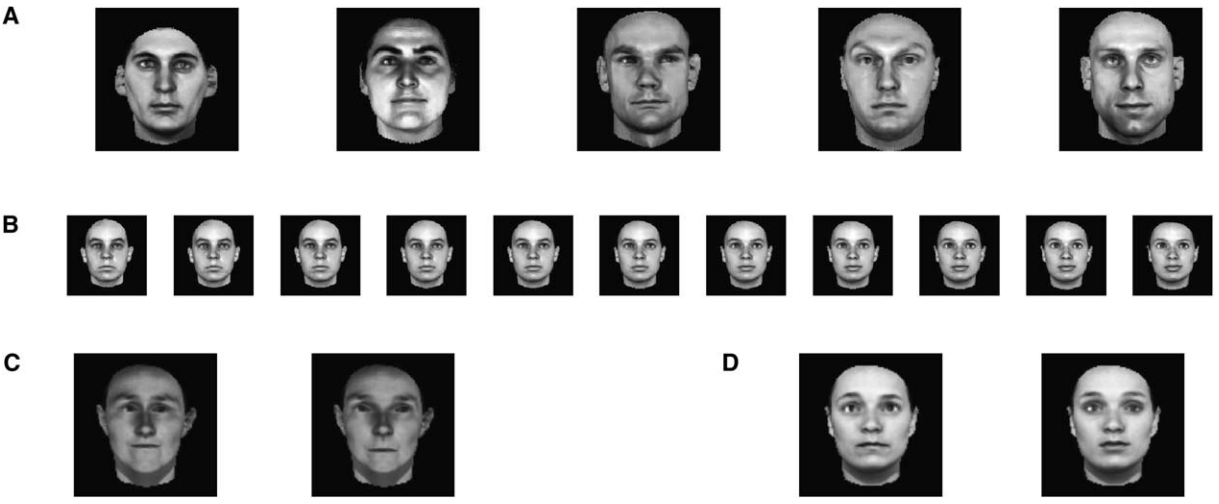

Figure 3. Examples of Stimuli Used in the Simulations

(A) shows some examples of face prototypes. (B) shows an example "morph line" between a pair of face prototypes (shown at the far left and right, respectively), created using the photorealistic morphing system (Blanz and Vetter, 1999). The relative contribution of each face prototype changes smoothly along the line. For example, the fourth face from the left (the third morph) is a mixture of 70% of the face on the far left and 30% of the face on the far right. (C) and (D) give examples of configural and featural change pairs, respectively (see Experimental Procedures).

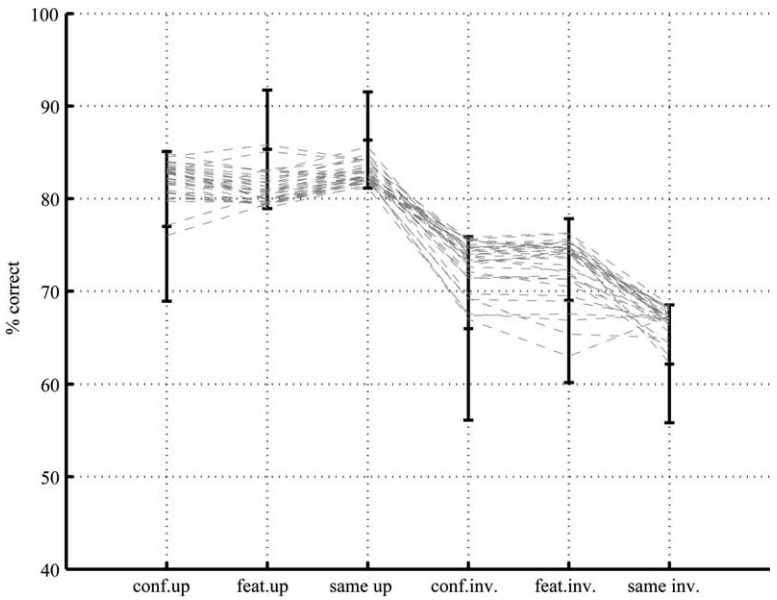

Figure 4. Comparison of the Experimental "Featural/Configural" Face Discrimination Results to Model Fits

Experimental data are in black, with error bars showing the 95% confidence intervals for the six different trial types (all combinations of upright/inverted orientations × same/different trials × "feature"/"configuration" change trials). The fits given by the 35 parameter sets with $p > 0.05$ are shown in gray, with dashed lines connecting fits belonging to the same parameter set drawn for illustration.

experimental data. Figure 4 shows the experimental data and the "good" fits to these data (i.e., all fits with $p > 0.05$). The model captures both "featural" and "configural" aspects of face perception, despite the fact that neither facial "features" nor "configuration" were explicitly coded by the face units. This is both an interesting and surprising result because it demonstrates that experiments that show either "configural" or "featural" effects do not necessitate explicit coding of either "features" or "configuration" in the underlying neural representation. Supporting this model prediction, a recent fMRI study failed to find support for differential processing of "featural" and "configural" changes in the FFA (Yovel and Kanwisher, 2004).

A closer examination of our simulation results reveals that "good" fits share the following qualities: (1) relatively sharply tuned units—for example, for $\sigma = 0.1$, only units with between 100 and 130 C2 afferents produce solutions with $p > 0.05$ (see Figures 2B and 2C); (2) a small n_M (of the "good" parameter sets, all have $n_M \leq 10$). These constraints on face unit tuning arising from simulations are a reflection of the constraints imposed by the behavioral data: to obtain an FIE in the model, neurons have to show relatively tight tuning to their preferred (upright) faces so that they respond well to upright, but less to inverted, faces. By the same token, tuning cannot be so sharp that the resulting face representation cannot generalize, i.e., also respond to the unfamiliar face stimuli to be discriminated in the experiment, which were different from the 180 faces used in the face representation. The simulations further show that, in the model, the comparable performance on "featural" and "configural" trials in the experiment is found for a wide range of parameters and appears to be due to the underlying representation at the C2 level that does not discriminate between face "features" and "configuration" but rather treats both as changes of stimulus shape. Further, the comparable performance on the "same" and "different" trials found in the experiment serves to constrain noise parameters and threshold τ to produce the appropriate variability and overlap of response distributions for "same" and "different" trials. Finally, the high degree of face specificity of model units that causes only a small subset of face units to strongly respond to a given face stimulus favors parameter sets with small n_M, producing a read-out model well-matched to the sparseness of the stimulus representation.

Figure 5 shows the relationship between face unit tuning width and inversion effect in more detail. Both decreasing the number of afferents, d, or increasing σ broadens model unit tuning, while increasing d or decreasing σ tightens model face unit tuning. Figure 5A shows that for very tightly tuned units (high d and/or low σ), modeled face discrimination performance is at chance, as model face units do not respond significantly to the unfamiliar faces used in the experiment. In contrast, more broadly tuned units can support performance on upright faces as found in the experiment. Importantly, however, very broadly tuned units (low d and high σ) produce an inversion effect lower than found in the experiment (Figure 5C), as face units now show increasing responses also to inverted faces (cf. Figure 2). Thus, only a small range of tuning broadness values (shown by the ridge in the d/σ space) produces units that can support the experimentally observed high performance on upright faces and low performance on inverted faces (Figure 5D).

In our model, face discrimination is mediated by neurons tuned to specific face shapes, putatively located in the FFA, whereas the recognition of other objects, such as dogs, cars, or shoes, would be mediated by neurons tuned to representatives from these object classes (see Figure 1). To show that the model can also account for the recognition of nonface objects, we performed an additional set of simulations using morphed cars (Figure 6). We modeled two hypothetical cases: (1) performance on cars at a level comparable to faces (i.e., showing an inversion effect, similar to the putatively expertise-related inversion effect for dogs exhibited by the dog show judges of Carey and Diamond (1986) and (2) lower performance with no inversion effect (the generic case expected for nonface objects that subjects are not experts

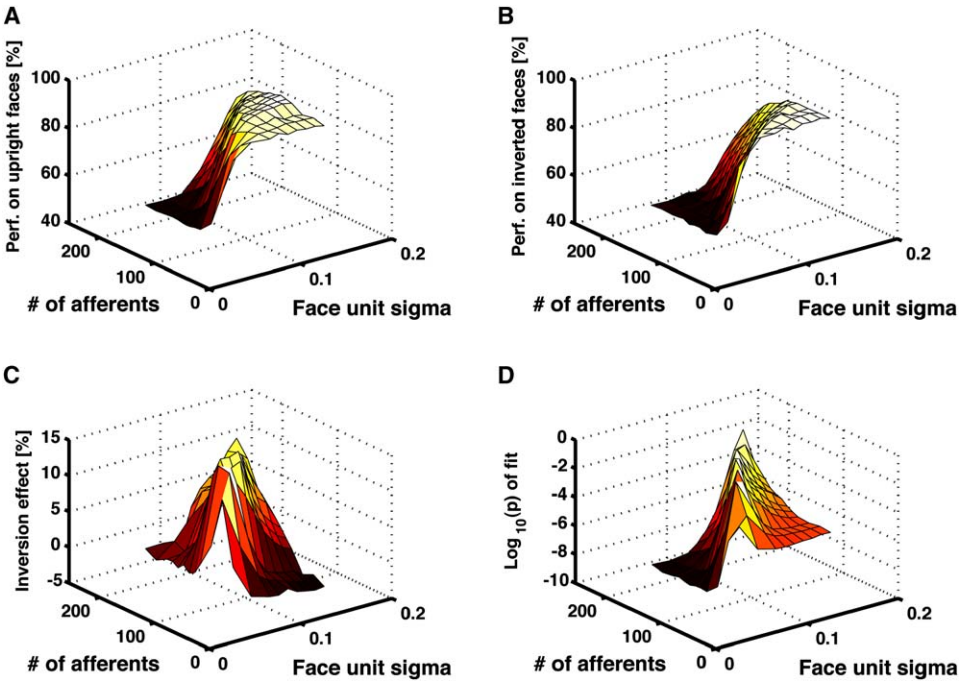

Figure 5. Illustration of the Dependence of the Inversion Effect on Model Face Unit Tuning Width

Based on the best model fit (p = 0.14, with d = 130, $\sigma_n = 0.01$, $n_M = 4$, $\sigma = 0.1$) to the behavioral data of Riesenhuber et al. (2004), the plots show the effect of varying tuning width of the model face units (by varying σ and d while leaving σ_n and n_M constant, and fitting τ to obtain the best match to the behavioral data, as described in the text) on (A) the mean performance on upright face stimuli for each parameter set (i.e., the mean of the "featural," "configural," and "same" trials using upright stimuli), (B) the performance on inverted faces (i.e., the mean of the "featural," "configural," and "same" trials using inverted stimuli), (C) the resulting inversion effect, i.e., the difference between (A) and (B), and (D) the goodness of the fit.

in). As Figure 6 shows, the model predicts, as in the case for faces, that high performance with an inversion effect is based on tightly tuned model units—in this case tuned to cars—(Figure 6C), while lower performance with no inversion effect is based on more broadly tuned units (Figure 6D). Very interestingly, recent experimental data (Yovel and Kanwisher, 2005) that have shown a correlation between subjects' behavioral FIE and inversion effect in the FFA (defined as the difference in response to upright versus inverted faces) directly support this model prediction, as an increase in face unit tuning broadness leads to an increase in face unit responses to inverted faces (Figure 2C) and thus a decreased neuronal inversion effect, along with a decreased behavioral inversion effect (Figures 5C and 6D).

Using the Model of Face Neurons to Predict FFA Responses

As described in the previous section, the model predicts that human face discrimination is based on a sparse code of face-selective neurons. We next set out to test the model's predictions using both fMRI and behavioral techniques.

fMRI studies of face processing have commonly focused on the average BOLD-contrast response to stimuli (e.g., faces versus objects). However, because of the limited spatial resolution of fMRI, relating BOLD-contrast signal change, behavioral performance, and neural population activity is complicated by the fact that both the density of selective neurons as well as the broadness of their tuning contribute to the average activity level in a voxel: a given BOLD-contrast signal change could arise from a small number of strongly activated units or a large number of unspecific neurons with low activity. By contrast, fMRI rapid adaptation (fMRI-RA) paradigms have been suggested to be capable of probing neuronal tuning in fMRI more directly (for a recent review see Grill-Spector et al., 2006). The fMRI-RA approach is motivated by findings from IT monkey electrophysiology that found that when pairs of stimuli were presented sequentially, the neural response to the second stimulus was suppressed relative to presentation as the first stimulus (Lueschow et al., 1994; Miller et al., 1993). In particular, the degree of adaptation has been suggested to depend on the degree of similarity between the two stimuli, with repetitions of the same stimulus causing the greatest amount of suppression. In the fMRI version of this experiment, the BOLD-contrast response to a pair of stimuli presented in rapid succession is measured for pairs differing in certain aspects (e.g., viewpoint or shape), and the combined response level is taken to correlate with dissimilarity of stimulus representation at the neural level (Grill-Spector et al., 2006; Murray and Wojciulik, 2004).

The model predicts a direct link between face neuron tuning specificity and behavioral discrimination performance. In particular, for tasks requiring the discrimination of a target face (T) from a distractor face (D), behavioral performance should increase with dissimilarity between target and distractor faces, as the corresponding activity patterns get increasingly dissimilar. Crucially, due to the tight tuning of face neurons, for some

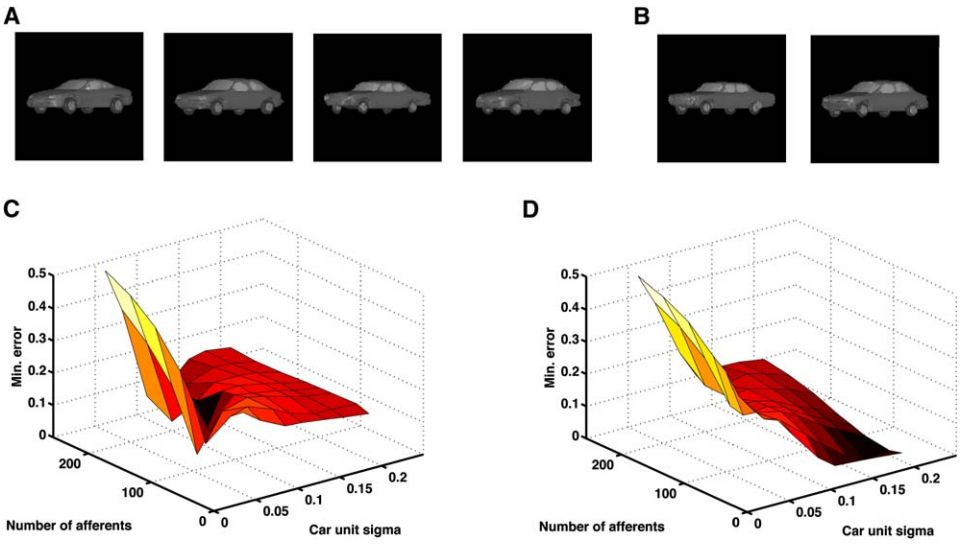

Figure 6. Demonstration of the Generality of the Model

This figure shows simulation results for the same model as in Figure 1, but now using a population of 100 "car units" instead of the 200 face units at the VTU level. These car units are tuned to different cars generated with a computer graphics morphing system (Shelton, 2000), some examples of which are shown in (A). The stimuli to be discriminated were 90 pairs of cars, different from the ones forming the car representation (but belonging to the same car morph space), an example of which is shown in (B). We then modeled a same/different car discrimination task in the same way as the face discrimination task of Figure 4, with the difference that the optimization criterion for selecting the same/different threshold τ was minimizing the Euclidean distance to a hypothetical mean subject performance (since there were no individual subject data). In particular, we investigated two cases: (1) average subject performance identical to the one found for faces in Riesenhuber et al. (2004), i.e., 83% correct on upright cars (for simplicity, performance was assumed to be equal for "same" and "different" trials), and 66% correct on inverted cars, i.e., showing a robust inversion effect, and (2) average performance equal to the grand mean in the Riesenhuber et al. (2004) case, i.e., 75% for upright and inverted cars. We explored all combinations of the following parameters σ_n from {0.05 0.1 0.15 0.2}, d from {20 40 80 100 120 160 200 240}, σ from {0.05 0.07 0.1 0.12 0.15 0.17 0.2}, and n_M from {1 2 4 6 8 10 16}, for a total of 1568 parameter sets. (C) and (D) illustrate how the fitting error varies with broadness of tuning (for each combination of σ and d, the value shown is the lowest error for all combinations of σ_n and n_M tested) for cases (1) and (2), respectively.

T-D dissimilarity level, both stimuli will activate disjoint subpopulations and performance should asymptote, as further increasing the T-D dissimilarity would not increase the dissimilarity of face neuron activation patterns. Likewise, in an fMRI-RA paradigm, adaptation of FFA face neurons stimulated with pairs of faces of increasing dissimilarity should decrease and asymptote when the faces activate different subpopulations.

We used the face neuron model to quantitatively predict the physical T-D dissimilarity for which BOLD-contrast and behavioral performance were expected to asymptote. In particular, we calculated the Euclidean distance between the face unit activity patterns corresponding to two faces of varying similarity chosen to lie on a "morph line" created by interpolating between two face prototypes using a photorealistic face morphing program (Blanz and Vetter, 1999). These distances were calculated for pairs of faces along ten morph lines, spanned by 20 face prototypes that were chosen as comparably discriminable in pilot psychophysical experiments. Note that these faces were different from the 180 faces used in the face representation, as well as different from the faces used in Riesenhuber et al. (2004) on which the parameter fitting was based, making this a true prediction of the model. Each face pair consisted of a prototype face, F, and a second face that was a morph of F and another prototype, F′, ranging from 100% F to 0% F′ to 0% F to 100% F′, in ten "morph steps" of 10% (see Figure 3B). This produced 11 distinct face stimuli between any two prototype faces, m0–m10. The conditions, M0–M10, were defined by the distance in morph steps between the faces in the pairs used in each condition (see Figures 3B, 8A, and 9A). We then calculated an average activation pattern distance over the face units for the pairs of faces at each M step.

Figure 7A shows the predicted average activation pattern distance as a function of dissimilarity for each of the 35 "good" parameter sets. While there is some variability across parameter sets for smaller shape differences, we find that the predicted activation pattern distances for the different parameter sets all predict an initial steep rise of activation pattern distance followed by an asymptote around M6. This close agreement is not trivial, as is demonstrated in Figure 7B, which shows a much broader range of slopes and asymptotes for other groups of parameters that did not fit the psychophysical data ($p < 0.0001$).

Thus, for an fMRI-RA experiment using the same stimuli, the model predicts that BOLD-contrast signal modulations should increase steadily with increasing pairwise dissimilarity and then asymptote. Moreover, the model makes the quantitative prediction that this effect should be observed around M6, coinciding with an asymptote of behavioral performance in a face discrimination paradigm. Note that this focus on the asymptote of activation pattern distances does not require us to make additional quantitative assumptions about how neural activity overlap and BOLD signal adaptation are related (which is still poorly understood [Grill-Spector et al., 2006]), leading to a more robust prediction.

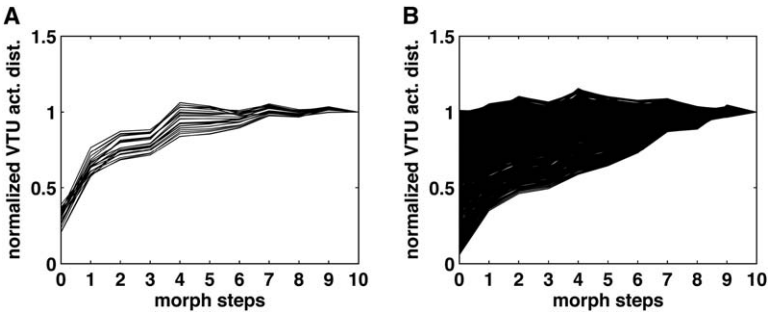

Figure 7. Predicted Distance of Model Face Unit Activation Patterns between a Face Prototype and Progressively Dissimilar Faces, Normalized by the Value at M10

Values shown are averages over the ten morph lines and 100 repetitions each. (A) shows the activation pattern distances for the 35 "good" parameter sets (with fits $p > 0.05$). For comparison, (B) shows the corresponding distance values for all fits with $p < 0.0001$.

fMRI Results

To test the model predictions concerning the asymptotic behavior of the fMRI BOLD-contrast signal, two fMRI experiments were conducted using an fMRI-RA paradigm. Experiment 1 was conducted to test the prediction of a BOLD-contrast signal asymptote around M6, and experiment 2 was conducted to address whether FFA neurons might represent a categorical identity code instead of the shape-based representation predicted by the model.

As performing a discrimination task on the face pairs in the scanner might introduce a confounding task modulation of the BOLD-contrast signal due to the different task difficulty levels among the M0, M3, M6, and M9 conditions (Grady et al., 1996; Sunaert et al., 2000), a facial target detection task was used to make certain that subjects paid attention to the face stimuli (Jiang et al., 2000; Rotshtein et al., 2005). Other than the target face, the same faces used in the model simulations were used for the experiments.

Experiment 1

Three different levels of dissimilarity, morph steps M3, M6, and M9, were tested in experiment 1. Figure 8 shows both the paradigm and the average hemodynamic response to the morph conditions of interest in the right FFA (size: 653 ± 51 mm^3) of 13 subjects. Since the hemodynamic response peaked in the 4–6 s time window, statistical analyses were carried out on these peak MR signal values (Kourtzi and Kanwisher, 2001). Paired t tests revealed that there was no significant difference between the M6 and M9 conditions ($t = 1.138$, $p = 0.277$).

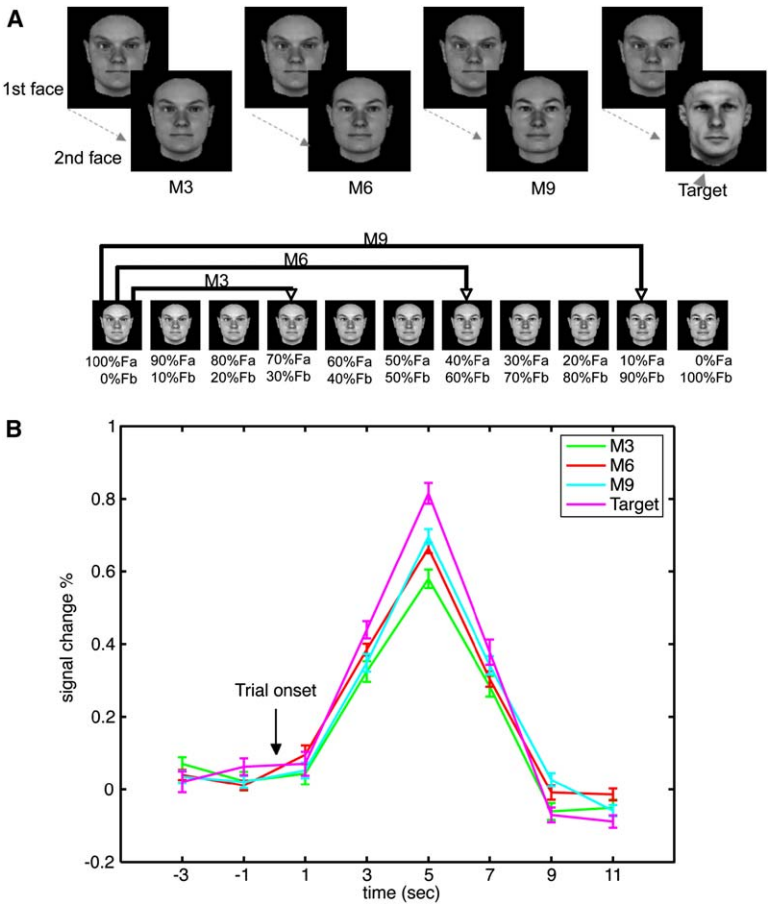

Figure 8. fMRI Experiment 1

(A) Illustrates the four types of non-null trials tested in experiment 1, and the stimulus selection for the M3, M6, and M9 conditions along one sample morph line. Subjects were asked to identify the target face that could appear either first or second in a pair of faces. (B) The mean (n = 13) hemodynamic response in the right FFA to the pair of faces of the M3, M6, M9, and target trials. Error bars show within-subject SEM.

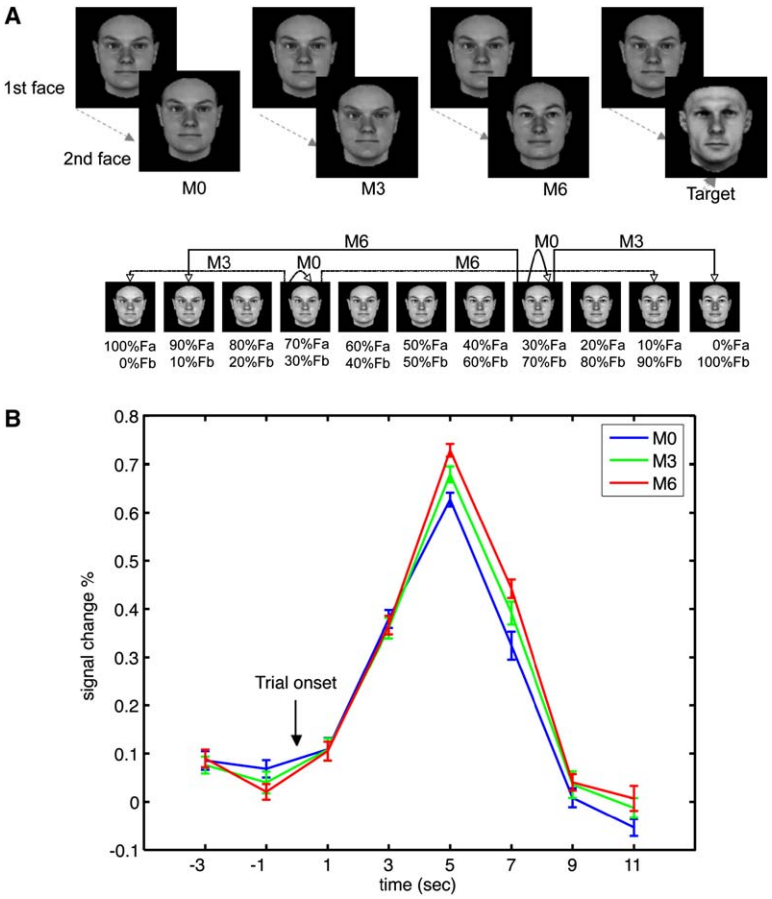

Figure 9. fMRI Experiment 2

(A) The four types of non-null trials tested in experiment 2, and description of stimulus selection for the M0, M3, and M6 conditions along one sample morph line in experiment 2 (see text). (B) The mean hemodynamic response (n = 13) in the right FFA to the pair of faces of the M0, M3, and M6 trials. Error bars show within-subject SEM.

In contrast, the BOLD-contrast signals related to both M6 and M9 were significantly higher than that related to M3 (M6 versus M3, $t = 3.254$, $p < 0.01$; M9 versus M3, $t = 3.209$, $p < 0.01$). Additionally, similar to previous reports (Jiang et al., 2000), the BOLD-contrast signal of target trials was significantly higher than that seen during the other three non-null trials ($p < 0.01$ for paired t test), demonstrating that the saturation at M9 was not due to a possible saturation of the BOLD-contrast response itself.

The results of experiment 1 thus agreed well with the model prediction. There was a sharp increase of BOLD-contrast signal from M3 to M6, but not from M6 to M9, even though the change of shape difference between the pair of faces in M3 and between the pair of faces in M6 was the same as the change of shape difference between those in M6 and in M9 (t test on change in pixel difference between image pairs in M6 relative to image pairs in M3 versus change in pixel difference between M6 and M9, n.s.).

Experiment 2

Experiment 2 addressed a possible concern regarding the signal changes observed in experiment 1, that is, whether the asymptote we observed was the effect of a perceived categorical face identity change instead of being due to an asymptoting of response differences for face shape-tuned units that, while responding preferentially to a specific face, do not show identity-based tuning (see Figure 2). In particular, while "face identity" units would respond as long as a certain face is categorized as belonging to a certain person and not for other people, the shape-tuned units in our model show a response that steadily decreases with dissimilarity between a unit's preferred face and the current stimulus, based on low-level (C2) feature differences. In the identity-based interpretation, the difference between M3 and M6 would arise because the two faces of M3 trials (0% and 30% morphs) would be judged to be of the same identity, while the two faces of M6 trials (0% and 60% morphs) would be judged to be of different identity (as would the two faces of M9 trials) (Rotshtein et al., 2005). Considering that our subjects had never before seen the faces used in the experiment, we considered such an explanation, based on the existence of face units categorizing the identity of the experimental faces, to be unlikely. Moreover, the behavioral discrimination performance (see Figure 10 and below) did not show signs of an abrupt categorical discrimination behavior, compatible with the results of a previous psychophysical study (Beale and Keil, 1995).

To explain the results of experiment 1 based on face identity, the assumption has to be made that the two faces of M3 share the same identity (Rotshtein et al., 2005). Based on this assumption, there should be no difference between M0 and M3. However, based on the predictions of our model, there should be a sizeable difference between M0 and M3, since the 0% and 30% morphs already cause substantially differing face unit

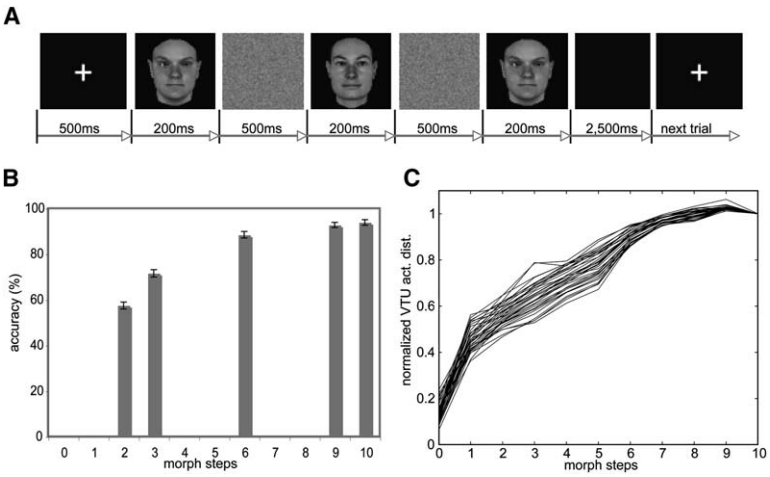

Figure 10. The Parametric Face Discrimination Experiment

(A) The behavioral paradigm. Images were shown sequentially and subjects were asked to judge whether the second or third face was the same as the first face (target). (B) The mean accuracy at the five tested morph steps. Error bars show SEM. (C) shows the predicted distance in face unit activation patterns for the n_M most activated units for the 35 "good" parameter sets.

activation patterns (cf. Figure 7A). We thus tested three different levels of dissimilarity, M0, M3, and M6, in experiment 2 (Figure 9). The paradigm used in experiment 2 was slightly modified from experiment 1 to correspond to the one used in Rotshtein et al. (2005). Figure 9B shows the average hemodynamic response in the right FFA (size: 645 ± 62 mm^3, n = 13). For the fMRI BOLD signal at peak (4–6 s), paired t tests indicated that there was a significant difference between M0 and M3 ($t = 2.639$, $p < 0.05$), between M3 and M6 ($t = 2.374$, $p < 0.05$), and between M0 and M6 ($t = 4.804$, $p < 0.001$). Thus, the results further support the shape-based, but not an identity-based, model.

Parametric Face Discrimination Performance

As explained above, our shape-based model predicts that behavioral performance in a discrimination paradigm should correlate with face unit activation pattern distance. We tested this prediction using a two-alternative-forced-choice paradigm, expecting this paradigm to more directly correlate with the predicted neural activation pattern distance by avoiding possible subject bias issues associated with a same/different paradigm, in particular for small shape differences.

After finishing the fMRI experiments, subjects (n = 22) were tested at five different levels of face similarity (Figure 10). Repeated-measures ANOVA revealed a significant effect across the five different morph steps [$F(2.29, 48.09) = 220.23$, $p < 0.001$]. Again, behavioral performance can be seen to asymptote around M6, with the increase in accuracy from M6 to M9 being less than 4% compared to over 17% for the increase from M3 to M6.

While the BOLD-contrast response in the FFA ROI is assumed to reflect the aggregate activation over all neurons in the FFA, the model predicts that only a subset of these neurons is actively engaged in support of the discrimination decision. We thus calculated the average activation pattern distance for the same stimuli as before, now averaging only over the n_M (as appropriate for each of the "good" parameter sets) most active units for each face (Figure 10C), and compared the correlation of predicted activation pattern distance and behavior for the two cases. We find that both distance measures correlate well with behavior ($r > 0.94$), supporting the model's prediction that behavioral performance is based on a comparison of activation patterns over FFA neurons. However, the correlation is significantly better ($p < 10^{-7}$, t test) for the correlation of behavior with activation pattern distance calculated only over the n_M most strongly activated units (Figure 10C) than with the activation pattern distance calculated over all units (Figure 7A): the average over all 35 "good" parameter sets was $r = 0.958$ (all $p < 0.016$) for the correlation with activation pattern distance over all units, and $r = 0.973$ (all $p < 0.013$) for the n_M most active units, compatible with the sparse coding prediction of the model. This effect was even stronger when including the results from an additional psychophysical experiment in which subjects were tested on conditions M2, M4, M6, M8, and M10, see Figure S5 in the Supplemental Data available online.

Discussion

The neural mechanisms underlying the representation and recognition of faces in the human brain have recently been the subject of much controversy and debate. Resolving these differences is of significant importance for our understanding of the general principles underlying object representation and recognition in cortex.

In this paper, we have shown that a computational implementation of a physiologically plausible neural model of face processing can quantitatively account for key data, leading to the prediction that human face discrimination is based on a sparse population code of sharply tuned face neurons. In particular, we have shown that a shape-based model can account for the face inversion effect, can produce "configural" effects without explicit configural coding, and can quantitatively account for the experimental data. The model thus constitutes a computational counterexample to theories that posit that human face discrimination necessarily relies on face-specific processes. We have further shown that the model can be used to derive quantitative predictions for fMRI and behavioral experiments, integrating the different levels of description in one framework. This study directly establishes a quantitative and mechanistic

relationship between neural adaptation level, neuronal stimulus selectivity, and behavioral performance.

Which features of the shape-based model are critical for its success? The model consists of a simple hierarchy composed of just two different kinds of neural operations: MAX pooling and feature combination. While the former is crucial for obtaining invariance to scaling and translation and robustness to clutter, the experiments in this paper did not focus on these aspects. Therefore, the experimental data are likely compatible with a number of computational models in which face recognition is based on a population of sharply tuned face units. Nevertheless, by using our model "out of the box," we were able to show that the same model that is able to quantitatively account for tuning properties of IT neurons can also contribute to our understanding of human face discrimination.

An asymptotic course of the FFA BOLD-contrast response with increasing face dissimilarity has also been found in a recent block-design fMRI study (Loffler et al., 2005). That study used face stimuli generated with a computer graphics system to vary "identity" and "distinctiveness." The observed BOLD-contrast signal saturation observed in the block with faces showing the greatest variation in "distinctiveness" was interpreted to suggest a face representation based on a polar coordinate system centered on a "mean" face, with neurons in the FFA showing sigmoidal tuning for variations in radial "distinctiveness" away from the "mean" face for a given "identity" angle. Given that face pairs of increasing "distinctiveness" also have increasingly different shape, our model offers a parsimonious explanation for these data as well, suggesting that the increased response in the "distinctive" blocks was due to decreased levels of adaptation rather than an increase in the response of face units. However, as our stimuli were created by morphing between faces of different identity but comparable "distinctiveness," as shown by the comparable subject performance on different morph lines, our imaging data are not compatible with the "identity/distinctiveness" hypothesis that would predict an identity-based code with different neurons coding for different "identities." Interestingly, a behavioral study previously conducted with the same stimuli (Wilson et al., 2002) reported that subjects' metric of face space was "locally Euclidean," i.e., not polar, appearing to likewise argue against the sigmoidal tuning and polar face space hypotheses. Moreover, subjects in Wilson et al. (2002) were better at discriminating stimulus variations around the "mean" face than around a face distant from the mean. However, putative sigmoidal face neurons would respond least and show the lowest sensitivity for stimulus changes around the mean face, making it difficult to understand how performance should be best for these stimuli. In contrast, our shape-based model offers a way to easily reconcile the behavioral and imaging data by assuming that there are more face neurons tuned to "typical" faces that are close to the mean than to less common faces that are farther away from the mean. This would also explain why Loffler et al. (2005) observed less BOLD-contrast signal for faces chosen around a center away from the "mean" face, which by definition would be less common. Interestingly, a similar argument might also provide an explanation for recent fMRI results investigating the "other race effect" in which less FFA activation and poorer discrimination performance was found for "other race" faces than for faces belonging to one's own race (Golby et al., 2001).

In our experiments, we combined results from behavioral and functional neuroimaging experiments that used the same stimuli to link the different levels of description. To minimize possibly confounding task-induced modulations of FFA activity, we had subjects perform a task in the scanner unrelated to the variable of interest (intra-pair similarity) and collected the parametric face discrimination data in a separate behavioral experiment performed after the imaging experiment. The possibility that subjects covertly performed an identification task might explain the differences between our results and those of Rotshtein et al. (2005), who found identity-related FFA activation in subjects viewing morphs of famous faces after performing a familiarity rating task on the same faces. As that and other studies (Winston et al., 2004) have shown identity-related signals in anterior temporal regions, it appears possible that the identity-related FFA activation found in those studies is due to top-down identity-based modulation of the shape-based bottom-up FFA response we found, compatible with recent results from monkey neurophysiology experiments (Freedman et al., 2003) that have found a delayed category-related modulation of shape-tuned responses in IT in monkeys performing a categorization task, most likely due to feedback from category-related circuits in prefrontal cortex, in line with the model of Figure 1 (see also Tomita et al., 1999). Consistent with this interpretation, a recent study (G. Avidan and M. Behrmann, 2005, abstract, Journal of Vision *5*, 633a, http://journalofvisionorg/5/8/633/, doi:101167/58633) failed to find identity-based adaptation for familiar faces in the FFA.

While the shape-based model is able to account for human face discrimination performance in our experiments, it is of course possible that additional, face-specific neuronal mechanisms are needed for other aspects of face perception. Clearly, given the computational complexity of visual object recognition, a close interaction of model and experiment in future studies will be helpful in further elucidating the underlying neural mechanisms.

Experimental Procedures

The HMAX Model of Object Recognition in Cortex

The HMAX model of object recognition in the ventral visual stream of primates has been described in detail elsewhere (Riesenhuber and Poggio, 1999b). We here only briefly describe the model; source code and implementation details can be found at http://maxlab.neuro.georgetown.edu/hmax/.

Figure 1 shows a sketch of the model. Input images (here, gray-scale images 128 × 128 pixels in size) are densely sampled by arrays of two-dimensional derivative-of-Gaussian filters, the so-called S1 units, each responding preferentially to a bar of a certain orientation, spatial frequency, and spatial location, thus roughly resembling properties of simple cells in striate cortex. In the next step, C1 cells, roughly resembling complex cells in primary visual cortex, pool over sets of S1 cells of the same preferred orientation (using a maximum, "MAX," pooling function [Lampl et al., 2004]) with similar preferred spatial frequencies and receptive field locations, to increase receptive field size and broaden spatial frequency tuning. Thus, a C1 unit

responds best to a bar of the same orientation as the S1 units that feed into it, but over a range of positions and scales. To increase feature complexity, neighboring C1 units (in a 2 × 2 arrangement) of similar spatial frequency tuning are then grouped to provide input to an S2 unit, roughly corresponding to neurons tuned to more complex features, as found in V2 or V4 (Cadieu et al., 2004). In the "standard" version of the model used here (Riesenhuber and Poggio, 1999b), there are 256 different types of S2 units in each filter band, corresponding to the 4^4 possible arrangements of four C1 units of each of four orientations. The S2 unit response function is a Gaussian with a center of 1 and standard deviation of 1 in each dimension, meaning that an S2 unit has a maximal firing rate of 1 that is attained if each of its four afferents responds at a rate of 1 as well.

To finally achieve size invariance over all filter sizes and position invariance over the modeled visual field, the S2 units are again pooled by a MAX operation to yield C2 units, the output units of the HMAX core system, designed to correspond to neurons in extrastriate visual area V4 or posterior IT (PIT). There are 256 C2 units, each of which pools over all S2 units of one type at all positions and scales. Consequently, a C2 unit will respond at the same level as the most active S2 unit with the same preferred feature, but regardless of its scale or position. C2 units provide input to the view-tuned units (VTUs), named after their property of responding well to a certain view of a complex object.

Face-Tuned Units

Face units are modeled as VTUs (see Figure 1). Mathematically, given a stimulus s and a corresponding d-dimensional activation pattern $v_i(s)$ over the d C2 units the VTU i is connected to, the response r_i of a VTU i with preferred C2 activation vector w_i (corresponding to, e.g., a specific face) is a Gaussian with tuning width σ,

$$r_i(s) = \exp\left(\frac{-\|w_i - v_i(s)\|^2}{2\sigma^2}\right) + n; \quad (1)$$

where n is additive Gaussian noise with standard deviation σ_n. The responses are constrained to be positive, i.e., if adding the noise would have caused a negative activation, activation is set to 0.

As each C2 unit is tuned to a different complex feature, a VTU's tuning specificity is directly determined by which C2 units it is connected to. For instance, a "face" VTU can be connected to all 256 C2 afferents with weights set to correspond to the C2 activation caused by a particular face, constraining its tuning in a 256-dimensional feature space. The VTU could also be connected to a smaller number of C2 afferents, for instance just the $d < 256$ most strongly activated ones (Riesenhuber and Poggio, 1999a), with the same connecting weights to those afferents as before. This VTU would still respond maximally to the same face, but now a greater number of other objects (those that cause similar activation patterns over the face unit's d C2 afferents, but possibly different activation patterns over the remaining C2 units) would cause similarly high activation. The lower d, the less specific the tuning would be. The specificity of a VTU, i.e., how much its response changes for different stimuli, is thus a function of the dimensionality d of its w, its tuning width σ (with higher σ producing broader tuning), and the amount of noise n (with higher noise amplitudes producing more variable responses).

The preferred stimulus of each face unit was defined by setting the unit's w_i equal to the $v_i(s)$ caused by the face stimulus the unit was to be tuned to, thus producing maximal response (see Equation 1). All other connections (i.e., up to and including the C2 layer) were as in the original 1999 model (Riesenhuber and Poggio, 1999b). This mechanism of creating a population of face-tuned units was chosen for simplicity, as the development of the face representation was not the focus of this paper.

Stimuli

Faces: Prototypes and Morphs

Face stimuli were based on 200 original faces (100 male, 100 female, grayscale, 128 × 128 pixels, Figure 3A). Of those, 180 were used for the face representation, and 20 different faces were used as stimuli in the discrimination task (see below). From these 20 faces, pairwise morphs were constructed using a photorealistic morphing system (Blanz and Vetter, 1999) (see Figure 3B), allowing us to finely parameterize shape similarity.

"Featural" and "Configural" Changes

To investigate potential differences in "featural" and "configural" processing and fit model parameters, we used the face stimuli of Riesenhuber et al. (2004), consisting of pairs of faces differing either by a "configural" (Figure 3C) or a "featural" (Figure 3D) change using a custom-built morphing system that allowed us to freely move and exchange face "parts" (i.e., eyes, mouth, and nose) of prototype face images. In particular, following the experiment of Riesenhuber et al. (2004), there were 80 image pairs, which could be presented either upright or inverted. Forty face pairs were associated with "featural trials": 20 face pairs with the faces in each pair differing by a "feature" change (replacement of eyes and mouth regions with those from other faces prototypes) and 20 "same" face pairs composed of the same outlines and face component positions as the corresponding "different" faces, with both faces having the same eyes/mouth regions. Another 40 face pairs were used in the "configural change" trials, consisting of 20 face pairs differing by a "configural" change (displacement of the eyes and/or mouth regions), plus 20 "same" face pairs composed of faces with the same face outlines and parts as the corresponding "different" pairs, with both faces in the pair having the same configuration. In the psychophysical experiment, a face pair (randomly selected to consist either of faces differing in a "featural" or a "configural" change, or one of the "same" face pairs, and in randomly chosen upright or inverted orientation) were presented to subjects, who needed to make same/difference judgments. For details, see Riesenhuber et al. (2004).

Modeling Behavior

In the same/different face discrimination task, two face images, s and s', are presented sequentially, and the subject's task is to decide whether the images were identical or different. We modeled this task by assuming that subjects retained information about the first stimulus by remembering the face unit activation values of a subset M of face units, notably the n_M units most strongly activated by the first stimulus (see Results). Subjects were then assumed to compare the activation over this subset of units caused by the second stimulus to the stored activation pattern and decide "same" or "different" depending on whether the difference fell below or above, respectively, a threshold τ. That is, the subject was assumed to compare

$$\sqrt{\sum_{j \in M} \left(r_s^j - r_{s'}^j\right)^2} < \tau$$

Thus, τ corresponds to a subject's "just noticeable difference" threshold. Simulated performance levels on the same/different task of Riesenhuber et al. (2004) were then determined by using the same stimuli as in the experiment and averaging performance over 100 repetitions for each pair of images to decrease variability of the performance estimates.

Fitting the model parameters to the experimental data of Riesenhuber et al. (2004) thus required choosing values for four free parameters (cf. Equation 1): the noise amplitude σ_n, the number of afferents d to each face unit, the face neuron tuning width σ, and the number of most strongly activated face units on which the same/different decision was based, n_M. In particular, we performed an exhaustive search through parameter space to find parameter sets that could fit the experimental data, testing parameter sets made up of all combinations of choosing σ_n from {0.05 0.01 0.02 0.03 0.04 0.05 0.06}, d from {40 60 80 100 110 120 130 140 150 160 180 200}, σ from {0.05 0.07 0.09 0.1 0.11 0.12 0.13 0.14 0.15}, and n_M from {1 2 3 4 6 8 10 12 14} for a total of 7 × 12 × 9 × 9 = 6804 parameter sets. These ranges were chosen based on pilot simulations to be most likely to produce face unit tuning and behavior compatible with the experimental data. Then, for each of these parameter sets, τ was optimized automatically to produce the best fit (in terms of statistical significance) to the experimental data.

Psychophysical and fMRI Experiments

Participants

Sixteen (ten female), and 15 (nine female) right-handed normal adults (aged 18–40) took part in fMRI experiments 1 and 2, respectively, with three of them participating in both experiments. Experimental procedures were approved by Georgetown University's

Institutional Review Board, and written informed consent was obtained from all subjects prior to the experiment. The data from five participants were excluded because of excessive head motion (two in experiment 1, one in experiment 2), sleep onset inside the scanner (one in experiment 1), or failure to detect the target face (one in experiment 2). Following the MRI scanning sessions, all participants (except for one who was unavailable) completed the face discrimination experiment.

Face Discrimination Experiment

Based on the results of several pilot studies, morphed faces (200 × 200 pixels) along ten within-gender morphing lines of 20 individual prototype faces (ten females) were used to test participants' face discrimination ability with a two-alternative forced-choice paradigm. Each trial started with a 500 ms fixation period, followed by three sequentially presented faces, each presented for 200 ms and separated by a 500 ms static random mask (see Figure 10A). The participants' task was to judge whether the second or the third face was the same as the first face by pressing one of two buttons. The next trial would automatically start 2500 ms after the offset of the third face, regardless of participant's response. One of the choice faces was always the same as the sample face, while the other choice face differed from the first one by one of five possible different levels of similarity, and could be the face at 2, 3, 6, 9, or 10 morphing steps away from the first face along the morph line, corresponding to the M2, M3, M6, M9, and M10 conditions, respectively. Among the three faces within each trial, at least one of them was always one of the 20 prototype faces. Stimuli were presented to participants on a liquid crystal display monitor on a dark background, 1024 × 768 resolution, 60 Hz refresh rate, at a distance of 60 cm. An in-house software package was used to present the stimuli and to record the responses. Participants completed a total of 400 trials (80 per condition) in four blocks.

Functional Localizer Scans

To locate the FFA regions, a block design was used to collect MRI images from two localizer scans for each subject (Haxby et al., 1999; Kanwisher et al., 1997). During each run, following an initial 10 s fixation period, 50 grayscale images of faces, houses, and scrambled faces were presented to participants in blocks of 30 s (each image was displayed for 500 ms and followed by a 100 ms blank screen), and were separated by a 20 s fixation block. Each block was repeated twice in each run that lasted for 310 s, and participants were asked to passively view these images while keeping their fixation at the center of the screen. The face and house images used in the localizer scans were purchased from http://www.hemera.com and postprocessed using programs written in MATLAB (The Mathworks, MA) to eliminate background variations and to adjust image size, luminance, and contrast. The final size of all images was scaled to 200 × 200 pixels, and half of the faces were scrambled using a grid of 20 × 20 pixel elements while the outlines of the faces were kept intact.

Rapid Event-Related Scans

To probe the effects of varying shape similarity on BOLD-contrast response in the FFA, MRI images from four ER scans were collected in experiments 1 and 2. Each run lasted 528 s and had two 10 s fixation periods, one at the beginning and the other at the end. Between the two fixation periods, a total of 127 trials were presented to participants at a rate of one every 4 s. During each trial (except the null trial), two faces were displayed sequentially (300 ms each with a 400 ms blank screen in-between), and followed by a 3000 ms blank screen (Kourtzi and Kanwisher, 2001). For each run, the data from the first two trials were discarded, and analyses were performed on the data of the other 125 trials—25 each of the five different conditions: three conditions of interest of varying intra-pair stimulus similarity (M3/M6/M9 in experiment 1 and M0/M3/M6 in experiment 2), the task trials, in which a target face, which participants needed to identify, could appear as either the first or the second one of the pair of faces, and the null trials (Figures 8A and 9A). Trial order was randomized and counterbalanced using M-sequences (Buracas and Boynton, 2002). While inside the scanner, participants were asked to watch all the faces but only respond to the target face by pressing a button with the right hand. Except for the target face, all face stimuli were the same as the ones used in the psychophysical experiment (with slightly decreased luminance levels to adjust for the darkness inside the scanner). The stimuli of both localizer and ER scans were presented on black background using E-Prime software (http://www.pstnet.com/products/e-prime/), back-projected on a translucent screen located at the rear of the scanner, and viewed by participants through a mirror mounted on the head coil. Resolution and refresh rate were the same as those used in the psychophysical experiment.

MRI Acquisition

MRI data were acquired at Georgetown University's Center for Functional and Molecular Imaging using an echo-planar imaging (EPI) sequence on a 3.0 Tesla Siemens Trio scanner (flip angle = 90°, TR = 2 s, TE = 30 ms, FOV = 205, 64 × 64 matrix) with a single-channel head coil. Forty-four interleaved axial slices (thickness = 3.2 mm, no gap; in-plane resolution = 3.2 × 3.2 mm^2) were acquired for the two functional localizer and four functional runs. At the end, three dimensional T1-weighted MPRAGE images (resolution 1 × 1 × 1 mm^3) were acquired from each subject.

MRI Data Analysis

After discarding the images acquired during the first 10 s of each run, the EPI images were temporally corrected to the middle slice, spatially realigned and unwarped together with the images from the localizer scans using the SPM2 software package (http://www.fil.ion.ucl.ac.uk/spm/software/spm2/). Images from the localizer scans were then smoothed with an isotropic 6 mm Gaussian kernel without normalization (Kiebel et al., 1999), while the images from the functional runs were neither smoothed nor normalized since the data analyses were conducted on independently defined ROIs (FFA). For comparison to previous studies, in a secondary analysis, the images from experiment 1 were resliced to 2 × 2 × 2 mm^3 and normalized to a standard MNI reference brain in Talairach space using SPM2. The same analysis procedures were conducted on the thus defined right FFA (44 ± 1.1, −56 ± 2.0, −21 ± 1.4, mean ± SEM), and the final results were very similar (see Figure S2).

The FFA regions were identified for each individual subject independently with the data from the localizer scans. We first modeled the hemodynamic activity for each condition (face, scrambled face, and house) in the localizer scans with the standard canonical hemodynamic response function (Friston et al., 1995), then identified the FFA ROI with the contrast of face versus house masked by the contrast of face versus baseline ($p < 0.0001$) (see Figure S1 for the results from a representative subject). As in previous studies (Loffler et al., 2005; Rotshtein et al., 2005), we focused our analysis on the right FFA, which was reliably identified in all subjects. For comparison, a left FFA was identified in 12/13 subjects in experiment 1 and 11/13 subjects in experiment 2, in line with previous studies (Haxby et al., 1999; Kanwisher et al., 1997) that likewise found stronger activation in the right than in the left FFA. Furthermore, supporting a dominant role of the right FFA in face perception (Barton, 2003), we found that left FFA activation was not sensitive to the fine shape differences used in our study (Figures S3 and S4). To obtain comparably sized FFAs across subjects, we defined the FFA ROI by choosing an approximately equal number of contiguous voxels with the highest statistical threshold for each subject (Murray and Wojciulik, 2004). To probe the robustness of the results, four additional data sets were extracted in experiment 1 from the right FFA for each subject at different sizes (and with or without normalization), ranging from 189 ± 18 mm^3 to 1244 ± 78 mm^3. As for the 653 ± 51 mm^3 set discussed in the Results section, we found that across all data sets, the peak BOLD signal (within the 4–6 s window) for the M6 and M9 conditions was always significantly higher than that for M3, whereas there was no significant difference between M6 and M9 (paired t test).

After removing the low-frequency temporal noise from the EPI runs with a high-pass filter (1/128 Hz), fMRI responses were modeled with a design matrix comprising the onset of each non-null trial and movement parameters as regressors using SPM2, followed by the extraction of the hemodynamic response for each subject in the right FFA using a finite impulse response (FIR) model with the Mars-Bar toolbox (M. Brett et al., 2002, abstract, presented at the 8th International Conference on Functional Mapping of the Human Brain (Sendai, Japan: NeuroImage *16*) and in-house software written in Matlab. To exclude the possibility that differences across conditions were caused by systematic baseline differences, the data from two scans prior to the onset of each trial were also extracted and shown in the figures.

Supplemental Data

The Supplemental Data for this article can be found online at http://www.neuron.org/cgi/content/full/50/1/159/DC1/.

Acknowledgments

We thank Thomas Vetter for initial face stimuli; Christoph Zrenner for the featural/configural face morphing software; Galit Yovel and especially Zoe Kourtzi for advice on the fMRI design; and Tomaso Poggio, Christine Schiltz, and Pawan Sinha for comments on earlier versions of this manuscript. This research was supported in part by NIMH grants 1P20MH66239-01A1, 1R01MH076281-01, and an NSF CAREER Award (#0449743).

Received: October 11, 2005
Revised: December 8, 2005
Accepted: March 1, 2006
Published: April 5, 2006

References

Barton, J.J. (2003). Disorders of face perception and recognition. Neurol. Clin. *21*, 521–548.

Baylis, G.C., Rolls, E.T., and Leonard, C.M. (1985). Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. Brain Res. *342*, 91–102.

Beale, J.M., and Keil, F.C. (1995). Categorical effects in the perception of faces. Cognition *57*, 217–239.

Blanz, V., and Vetter, T.(1999). A morphable model for the synthesis of 3D faces. Paper presented at SIGGRAPH '99 (ACM Computer Soc. Press; Los Angeles, California).

Buracas, G.T., and Boynton, G.M. (2002). Efficient design of event-related fMRI experiments using M-sequences. Neuroimage *16*, 801–813.

Cadieu, C., Kouh, M., Riesenhuber, M., and Poggio, T. (2004). Shape Representation in V4: Investigating Position-Specific Tuning for Boundary Conformation with the Standard Model of Object Recognition (Cambridge, MA: MIT AI Lab and CBCL).

Carey, S., and Diamond, R. (1986). Why faces are and are not special: An effect of expertise. J. Exp. Psychol. Gen. *115*, 107–117.

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. J. Cogn. Neurosci. *3*, 1–8.

Farah, M.J., Tanaka, J.W., and Drain, H.M. (1995). What causes the face inversion effect? J. Exp. Psychol. Hum. Percept. Perform. *21*, 628–634.

Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2003). Comparison of primate prefrontal and inferior temporal cortices during visual categorization. J. Neurosci. *23*, 5235–5246.

Friston, K.J., Frith, C.D., Turner, R., and Frackowiak, R.S. (1995). Characterizing evoked hemodynamics with fMRI. Neuroimage *2*, 157–165.

Gauthier, I., Tarr, M.J., Moylan, J., Skudlarski, P., Gore, J.C., and Anderson, A.W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. J. Cogn. Neurosci. *12*, 495–504.

Golby, A.J., Gabrieli, J.D., Chiao, J.Y., and Eberhardt, J.L. (2001). Differential responses in the fusiform region to same-race and other-race faces. Nat. Neurosci. *4*, 845–850.

Grady, C.L., Horwitz, B., Pietrini, P., Mentis, M.J., Ungerleider, L.G., Rapoport, S.I., and Haxby, J.V. (1996). Effects of task difficulty on cerebral blood flow during perceptual matching of faces. Hum. Brain Mapp. *4*, 227–239.

Grill-Spector, K., Knouf, N., and Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. Nat. Neurosci. *7*, 555–562.

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. *10*, 14–23.

Haxby, J.V., Ungerleider, L.G., Clark, V.P., Schouten, J.L., Hoffman, E.A., and Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. Neuron *22*, 189–199.

Haxby, J.V., Hoffman, E.A., and Gobbini, M.I. (2000). The distributed human neural system for face perception. Trends Cogn. Sci. *4*, 223–233.

Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. *160*, 106–154.

Jiang, Y., Haxby, J.V., Martin, A., Ungerleider, L.G., and Parasuraman, R. (2000). Complementary neural mechanisms for tracking items in human working memory. Science *287*, 643–646.

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

Kiebel, S.J., Poline, J.B., Friston, K.J., Holmes, A.P., and Worsley, K.J. (1999). Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. Neuroimage *10*, 756–766.

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

Kourtzi, Z., and Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. Science *293*, 1506–1509.

Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. J. Neurophysiol. *92*, 2704–2713.

Loffler, G., Yourganov, G., Wilkinson, F., and Wilson, H.R. (2005). fMRI evidence for the neural representation of faces. Nat. Neurosci. *8*, 1386–1391.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual Object Recognition. Annu. Rev. Neurosci. *19*, 577–621.

Lueschow, A., Miller, E.K., and Desimone, R. (1994). Inferior temporal mechanisms for invariant object recognition. Cereb. Cortex *4*, 523–531.

Miller, E.K., Li, L., and Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. J. Neurosci. *13*, 1460–1478.

Mondloch, C.J., Le Grand, R., and Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. Perception *31*, 553–566.

Murray, S.O., and Wojciulik, E. (2004). Attention increases neural selectivity in the human lateral occipital complex. Nat. Neurosci. *7*, 70–74.

Olshausen, B.A., and Field, D.J. (2004). Sparse coding of sensory inputs. Curr. Opin. Neurobiol. *14*, 481–487.

Perrett, D.I., Rolls, E.T., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. Exp. Brain Res. *47*, 329–342.

Riesenhuber, M., and Poggio, T. (1999a). Are cortical models really bound by the "Binding Problem?". Neuron *24*, 87–93.

Riesenhuber, M., and Poggio, T. (1999b). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

Riesenhuber, M., and Poggio, T. (2002). Neural Mechanisms of Object Recognition. Curr. Opin. Neurobiol. *12*, 162–168.

Riesenhuber, M., Jarudi, I., Gilad, S., and Sinha, P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. Proc. Biol. Sci. (Suppl) *6*, S448–S450.

Rolls, E.T., and Tovee, M.J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. Exp. Brain Res. *103*, 409–420.

Rotshtein, P., Henson, R.N., Treves, A., Driver, J., and Dolan, R.J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat. Neurosci. *8*, 107–113.

Shelton, C. (2000). Morphable surface models. Int. J. Comp. Vis. *38*, 75–91.

Sunaert, S., Van Hecke, P., Marchal, G., and Orban, G.A. (2000). Attention to speed of motion, speed discrimination, and task difficulty: an fMRI study. Neuroimage *11*, 612–623.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., and Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. Nature *401*, 699–703.

Ungerleider, L.G., and Haxby, J.V. (1994). 'What' and 'where' in the human brain. Curr. Opin. Neurobiol. *4*, 157–165.

Wilson, H.R., Loffler, G., and Wilkinson, F. (2002). Synthetic faces, face cubes, and the geometry of face space. Vision Res. *42*, 2909–2923.

Winston, J.S., Henson, R.N., Fine-Goulden, M.R., and Dolan, R.J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. J. Neurophysiol. *92*, 1830–1839.

Yin, R.K. (1969). Looking at upside-down faces. J. Exp. Psychol. *81*, 141–145.

Young, M.P., and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. Science *256*, 1327–1331.

Yovel, G., and Kanwisher, N. (2004). Face perception: domain specific, not process specific. Neuron *44*, 889–898.

Yovel, G., and Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. Curr. Biol. *15*, 2256–2262.

Evaluation of a Shape-Based Model of Human Face Discrimination Using fMRI and Behavioral Techniques

Supplementary Material

Xiong Jiang[1], Ezra Rosen[1] , Thomas Zeffiro[2], John VanMeter[2], Volker Blanz[3], and Maximilian Riesenhuber[1]

[1]Department of Neuroscience, Georgetown University Medical Center, Research Building Room WP-12, 3970 Reservoir Rd. NW, Washington, DC 20007, USA

[2]Center for Functional and Molecular Imaging, Georgetown University Medical Center, Washington, DC 20007, USA

[3] Max-Planck-Institut für Informatik, Saarbrücken, Germany

Correspondence should be addressed to M.R. (mr287@georgetown.edu).

Supplementary Figures

Figure S1

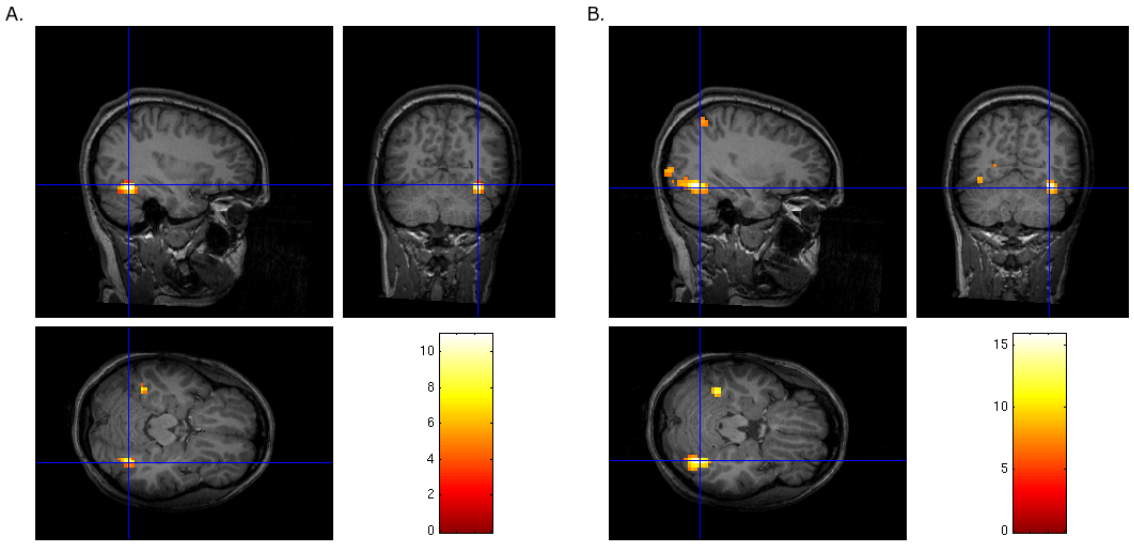

Figure S1. The activations of the right FFA from one representative subject. **(A)** From localizer scans, the comparison of face versus houses ($p < 0.001$ uncorrected for illustration). **(B)** From the ER scans of the same subject, the comparison of face versus baseline ($p < 10^{-10}$ uncorrected).

Figure S2

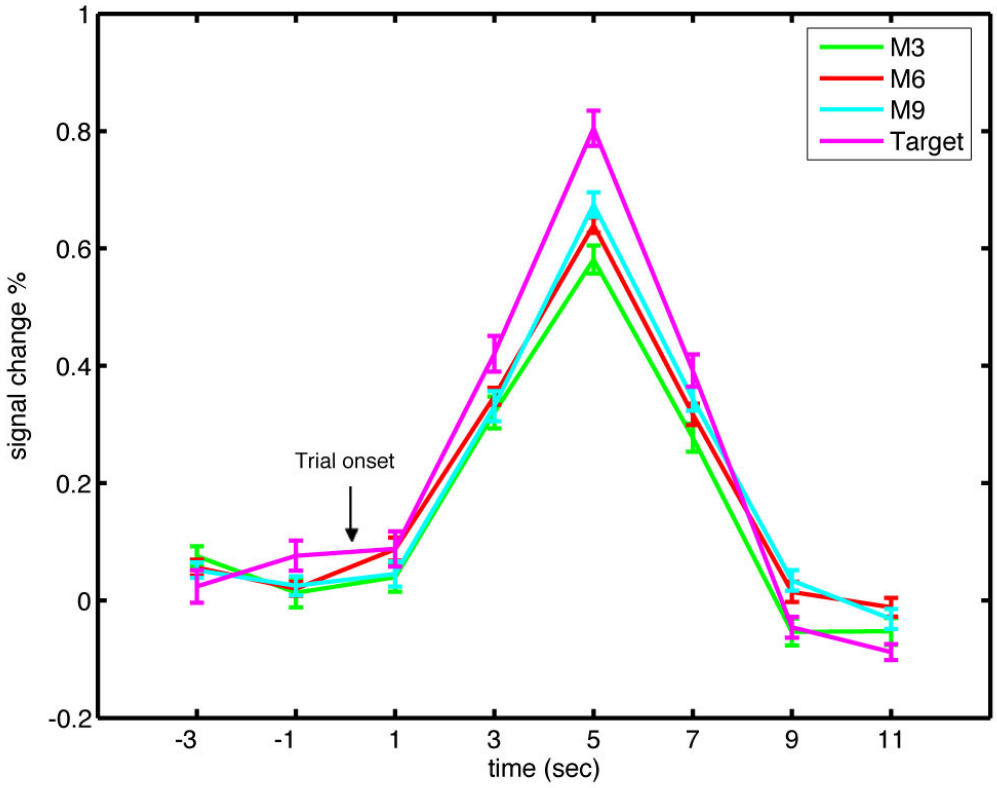

Figure S2. The mean hemodynamic response in the right FFA (location in MNI space: 44±1.1, -56±2.0, -21±1.4; size: 1244±78mm^3) for the M3, M6, M9, and target trials from normalized images in Experiment 1. Paired t-test revealed a significant difference between M6 and M3 (p<0.05), but not between M6 and M9 (p>0.2). Error bars show within-subject s.e.m.

Figure S3

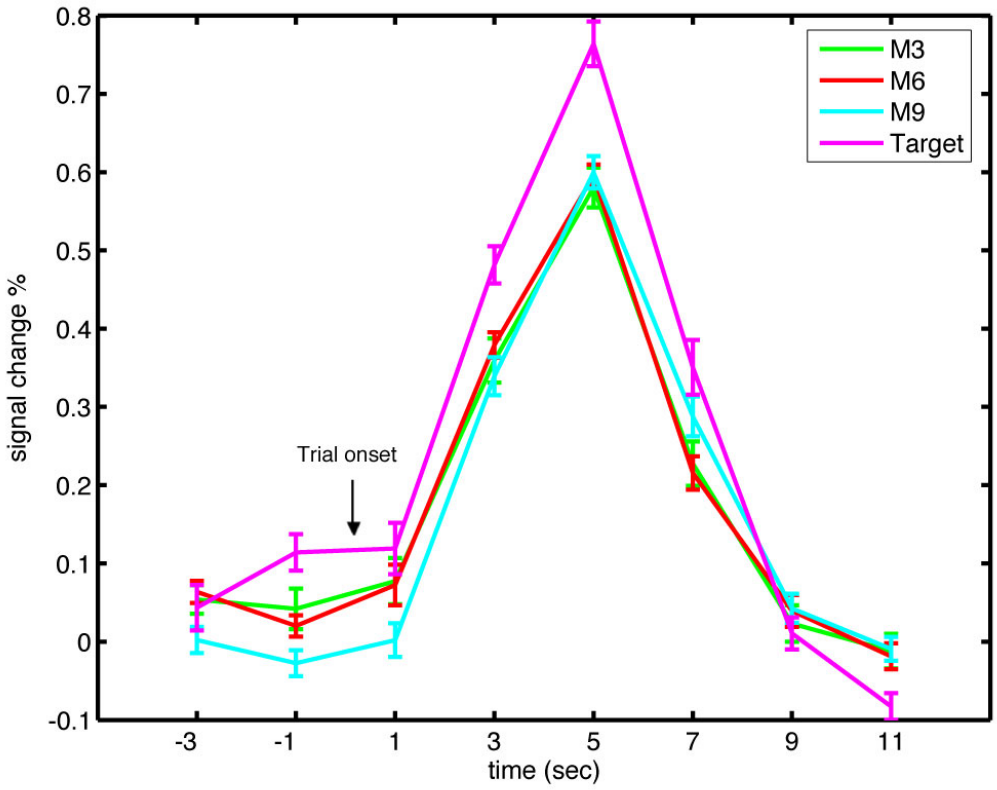

Figure S3. The mean hemodynamic response in the left FFA (size: 522±43mm^3) to the M3, M6, M9, and target trials from non-normalized images in Experiment 1. Paired t-test revealed that there was no significant difference between M3 and M6 ($p>0.5$), or between M3 and M9 ($p>0.5$). Error bars show within-subject s.e.m.

Figure S4

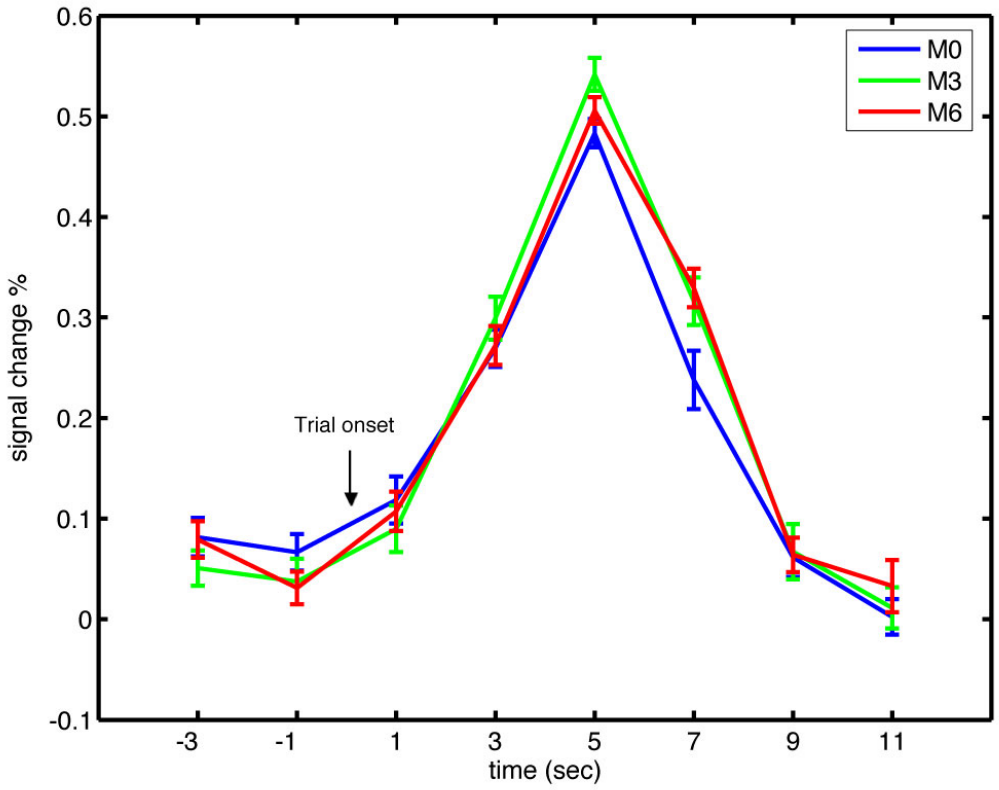

Figure S4. The mean hemodynamic response in the left FFA (size: 499±51mm^3) for the M0, M3, and M6 trials from non-normalized images in Experiment 2. Paired t-test revealed that there was no significant difference between M0 and M3 ($p>0.3$), or between M3 and M6 ($p>0.4$). Error bars show within-subject s.e.m.

Figure S5

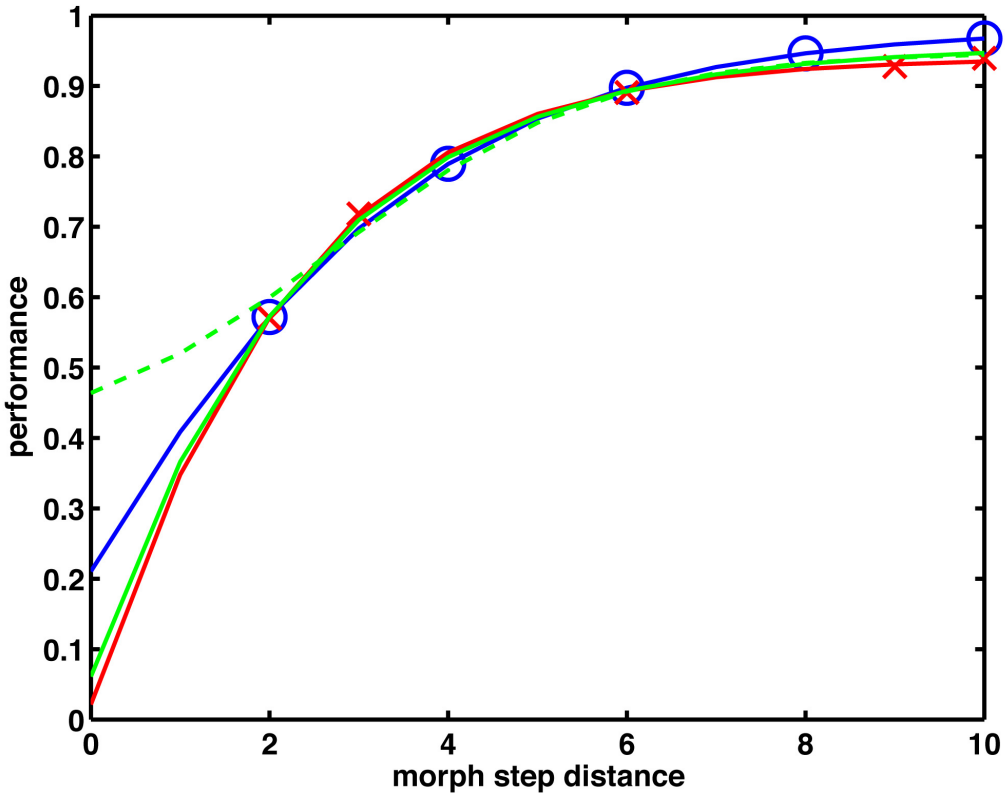

Figure S5. Additional psychophysical data and analyses. To more quantitatively analyze the asymptoting of face unit activation distance and behavior, we fit the simulation and experimental data to sigmoid functions,

$$f(x) = \frac{a}{1 + \exp\left(\frac{-(b-x)}{t}\right)} + e$$

with x being the morph step difference variable (ranging from 0 to 10), and a,b,t,e parameters of the sigmoid. Optimum values for these four parameters were determined independently for each data set using the simplex search method implemented in MATLAB's fminsearch function. We defined the point at which the function reached its

asymptote ("asymptotic point", a_P) as the value of x (discretized in steps of 0.1) where f(x) reached 95% of its maximum in the [0..10] interval.

Fitting f(x) to the face unit activation distance data over all face units, d_{all} , of Figure 7A in the paper reveals that 30/35 "good" parameter sets predict an asymptotic point on or before M6 (median at 51% morph), with 35/35 sets predicting >90% of maximum response by M6, yielding the model prediction of an asymptote of BOLD contrast signal in the face adaptation ER paradigm by M6.

Turning to Figure 10, fitting the sigmoids to the data of Fig. 10C, *i.e.,* the distance d_{top} over the n_M most strongly activated units (the "top units") for each parameter set, we find that, interestingly, the model in general predicts a later asymptotic point (median of 80% morph, range 65%-93% morph) for d_{top} shown in 10C than for d_{all} in Figure 7A. The reason for this later asymptote is that face units strongly activated by a particular face (see, *e.g.,* Figure 2 in the paper) show an appreciable response drop-off over more morph steps away from that face than face units that are only weakly activated to start with.

This yields the very general prediction (that does not depend on any additional assumptions about how neural activation is translated into behavioral decisions in the 2AFC task) that the asymptote for d_{top} constitutes an *upper bound* for the behavioral asymptote: If the behavioral decision is indeed based on the activation pattern difference over the top units, then we would not expect an increase in behavioral performance after this difference asymptotes. On the other hand, behavior could asymptote already for smaller morph differences, if the activation pattern distance required for perfect discrimination is lower than the maximum value of d_{top} for the morphed faces. Thus, the model predicts that the behavioral performance asymptote must be reached *before* d_{top} asymptotes.

Fitting the sigmoid to the behavioral performance in the M2, M3, M6, M9, M10 experiment (red line) yielded a_P = 59% morph, below any of values predicted for the asymptote of d_{top}. To address the possible criticism that the M2, M3, M6, M9, M10 conditions in the experiment are not well matched to the asymptotic range, we ran an additional 12 subjects on the same 2AFC paradigm and morph lines, but using distance conditions M2, M4, M6, M8, and M10 (blue line). The fit produced a_P = 67, similar to the previous fit, and still lower than 34/35 of the a_P of d_{top}. Fitting to the combined behavioral dataset (seven points: M2, M3, M4, M6, M8, M9, M10) gives a_P = 63 (solid green line)[1]. The behavioral asymptote is thus compatible with all parameter sets under the "sparse coding" assumption, but at the investigated resolution also with the asymptotes of d_{all} for most parameter sets. More clear-cut support for the former comes from the correlational analysis that gives equal weight to all data points. Similar as for the correlation analysis of behavior in the M236910 experiment described in the main text, we here find that d_{top} in the M246810 experiment shows significantly higher correlation with behavior than d_{all} (average r_{top} = 0.974 vs. r_{all} = 0.910, $p < 6*10^{-4}$). This also holds true for the joint dataset, $p < 2*10^{-5}$ (average r_{top} = 0.965 vs. r_{all} = 0.9). We should add that all correlations are significant for d_{all} (all $p < 0.05$) and highly significant for d_{top} (all $p \leq 0.002$). Thus, d_{top} appears to provide a better fit to the behavioral data.

---

[1] Note that the low extrapolated values below M2 predicted by the fits are numerical artefacts of the fitting procedure, which do not significantly affect the prediction of the asymptote, however. To illustrate this, the dashed green line shows the best fit for the combined dataset to which a hypothetical point at M0 with the expected chance performance (50%) was added. This produces a more reasonable fit for M values <2, while producing a practically indistinguishable asymptote (a_p = 62).