Adjudicating between face-coding models with individual-face fMRI responses

Short title: Adjudicating between face-coding models with fMRI

Johan D. Carlin* & Nikolaus Kriegeskorte

* to whom correspondence should be addressed:
Johan D. Carlin
MRC Cognition and Brain Sciences Unit
15 Chaucer Road
Cambridge CB2 7EF
UK
johan.carlin@mrc-cbu.cam.ac.uk
+44 (0) 1223 355294

Revised manuscript submitted to PLoS Computational Biology on 13 January 2017.

PRE-PUBLICATION MANUSCRIPT

Abstract

The perceptual representation of individual faces is often explained with reference to a norm-based face space. In such spaces, individuals are encoded as vectors where identity is primarily conveyed by direction and distinctiveness by eccentricity. Here we measured human fMRI responses and psychophysical similarity judgments of individual face exemplars, which were generated as realistic 3D animations using a computer-graphics model. We developed and evaluated multiple neurobiologically plausible computational models, each of which predicts a representational distance matrix and a regional-mean activation profile for 24 face stimuli. In the fusiform face area, a face-space coding model with sigmoidal ramp tuning provided a better account of the data than one based on exemplar tuning. However, an image-processing model with weighted banks of Gabor filters performed similarly. Accounting for the data required the inclusion of a measurement-level population averaging mechanism that approximates how fMRI voxels locally average distinct neuronal tunings. Our study demonstrates the importance of comparing multiple models and of modeling the measurement process in computational neuroimaging.

Author Summary

Humans recognize conspecifics by their faces. Understanding how faces are recognized is an open computational problem with relevance to theories of perception, social cognition, and the engineering of computer vision systems. Here we measured brain activity with functional MRI while human participants viewed individual faces. We developed multiple computational models inspired by known response preferences of single neurons in the primate visual cortex. We then compared these neuronal models to patterns of brain activity corresponding to individual faces. The data were consistent with a model where neurons respond to directions in a high-dimensional space of faces. It also proved essential to model how functional MRI voxels locally average the responses of tens of thousands of neurons. The study highlights the challenges in adjudicating between alternative computational theories of visual information processing.

Introduction

Humans are expert at recognizing individual faces, but the mechanisms that support this ability are poorly understood. Multiple areas in human occipital and temporal cortex exhibit representations that distinguish individual faces, as indicated by successful decoding of face identity from functional MRI (fMRI) response patterns (1–10). Decoding can reveal the presence of face-identity information as well as invariances. However, the nature of these representations remains obscure because individual faces differ along many stimulus dimensions, each of which could plausibly support decoding. To understand the representational space, we need to formulate models of how individual faces might be encoded and test these models with responses to sufficiently large sets of face exemplars. Here we use representational similarity analysis (RSA) (11) to test face-coding models at the level of the representational distance matrices they predict. Comparing models to data in the common currency of the distance matrix enables us to pool the evidence over many voxels within a region, obviating the need to fit models separately to noisy individual fMRI voxels.

Many cognitive and neuroscientific models of face processing do not make quantitative predictions about the representation of particular faces (12,13). However, such predictions can be obtained from models based on the notion that faces are encoded as vectors in a space (14). Most face-space implementations apply principal components analysis (PCA) to face images or laser scans in order to obtain a space, where each component is a dimension and the average face for the training sample is located at the origin (15,16). In such PCA face spaces, eccentricity is associated with judgments of distinctiveness, while vector direction is associated with perceived identity (17–19). Initial evidence from macaque single-unit recordings and human fMRI suggests that brain responses to faces are strongly modulated by face-space eccentricity, with most studies finding increasing responses with distinctiveness (20–23). However,

74 there has been no attempt to develop a unified account for how a single underlying face-
75 space representation can support both sensitivity to face-space direction at the level of
76 multivariate response patterns and sensitivity to eccentricity at the level of regional-
77 mean fMRI activations. Here we develop and evaluate several face-space coding models,
78 which differ with respect to the proposed shape of the neuronal tuning functions across
79 face space and with respect to the distribution of preferred face-space locations over the
80 simulated neuronal population.

81 Face-space coding models define high-level representational spaces, which are
82 assumed to arise through unspecified low-level featural processing. An alternative
83 possibility is that some cortical face-space representations can be explained directly by
84 low-level visual features, which typically covary with position in PCA-derived face
85 spaces. To explore this possibility, we evaluated a Gabor-filter model, which receives
86 stimulus images rather than face-space coordinates as input, and has previously been
87 used to model response preferences of individual voxels in early visual cortex (24).

88 We found that cortical face responses measured with fMRI strongly reflect face-
89 space eccentricity: a step along the radial axis in face space results in a much larger
90 pattern change than an equal step along the tangential axis. These effects were
91 consistent with either a sigmoidal-ramp-tuning or a Gabor-filter model. The
92 performance of these winning models depended on the inclusion of a measurement-
93 level population-averaging mechanism, which accounted for local averaging of neuronal
94 tunings in fMRI voxel measurements.

Results

Sampling face space with photorealistic but physically-controlled animations

In order to elicit strong percepts of the 3D shape of each individual face, we generated a set of photorealistic animations of face exemplars. Each 2s animation in the main experiment featured a face exemplar in left or right half profile, which rotated outward continuously (S2 Movie, Experimental Procedures). The animations were based on a PCA model of 3D face shape and texture (25). Each frame of each animation was cropped with a feathered aperture and processed to equate low-level image properties across the stimulus set (Experimental Procedures). We generated 12 faces from a slice through face space (Figure 1b). Euclidean distances between the Cartesian coordinates for each face were summarized in a distance matrix (Figure 1a), which served as the reference for comparisons against distances in the perceptual and cortical face spaces (Figure 1c-h). We generated a physically distinct stimulus set with the same underlying similarity structure for each participant by randomizing the orientation of the slice through the high-dimensional PCA space (for examples, see S1 Figure). This served to improve generalizability by ensuring that group-level effects were not strongly influenced by idiosyncrasies of face exemplars drawn from a particular face-space slice. In formal terms, group-level inference therefore treats stimulus as a random effect (26).

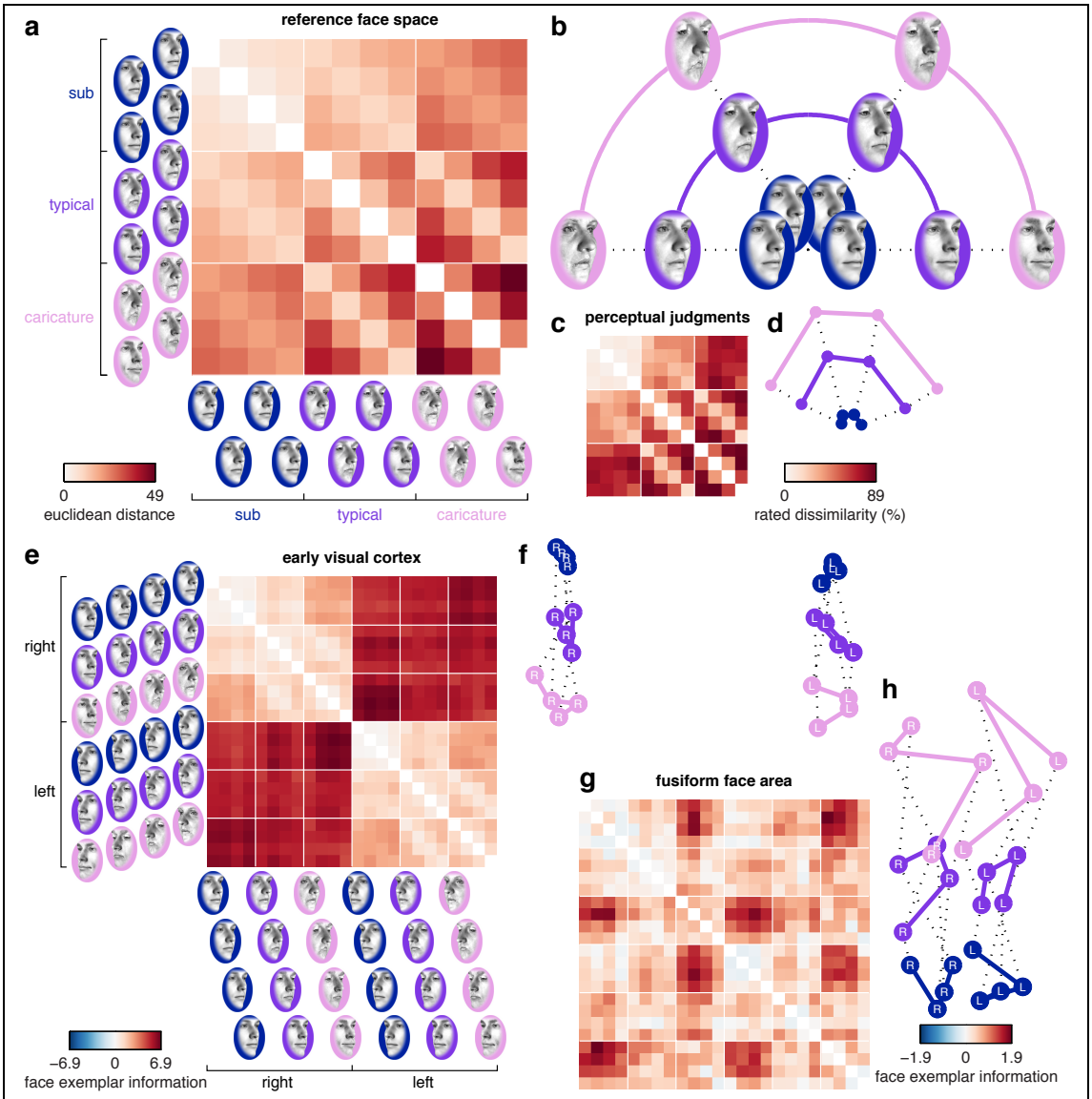


Figure 1. Face spaces obtained from a reference PCA model, perceptual similarity judgments and fMRI response patterns in human visual cortex. (**a-b**) We generated 12 faces in a polar grid arrangement on a 2D slice through the reference PCA space (3 eccentricity levels and 4 directions). The grid was centered on the stimulus space norm (not shown). Euclidean distances between the 12 faces are illustrated in a distance matrix (a) and in a 2D visualization of these distances (b, multidimensional scaling, metric-stress criterion). (**c-d**) Group-average perceptual face space (N=10) estimated from psychophysical similarity judgments. The distance matrix depicts the percentage of trials on which each pair of faces was rated as relatively more dissimilar. (**e-h**) Group-average cortical face-space (N=10) estimated from fMRI response patterns in human visual cortex. The distance matrices depict a cross-validated estimate of multivariate discriminability where 0 corresponds to chance-level performance (Experimental Procedures). Each cortical region was defined in individual participants using independent data. The distance matrices use separate color scales since we expect variable effect sizes across cortical regions. For data from other regions of interest, see S2 Figure.

Cortical face spaces are warped relative to the reference PCA space

Human participants (N=10) participated in a perceptual judgment task (S1 Movie, 2145 trials over 4 sessions) followed by fMRI scans (S2 Movie, 2496 trials over 4 recording days). Brain responses were analyzed separately in multiple independently localized visual regions of interest. We compared representational distance matrices estimated from these data sources to distances predicted according to different models using the Pearson correlation coefficient. These distance-matrix similarities were estimated in single participants and the resulting coefficients were Z-transformed and entered into a summary-statistic group analysis for random-effects inference generalizing across participants and stimuli (Experimental Procedures).

We observed a strong group-average correlation between distance matrices estimated from perceptual dissimilarity judgments and Euclidean distances in the reference PCA space (mean(r)=0.83, mean(Z(r))=1.20, standard error=0.05, $p<0.001$, Figure 1c, Figure 5a, S1 Table). Correlations between the reference PCA space and cortical face spaces were generally statistically significant, but smaller in magnitude (Figure 5b-c, S5 Figure, S1 Table). Distances estimated from the fusiform face area were weakly, but highly significantly correlated with the reference PCA space (mean(r)=0.17, mean(Z(r))=0.17, standard error=0.04, $p<0.001$, Figure 5c). Distances estimated from the early visual cortex were even less, though still significantly, correlated with the reference PCA space (mean(r)=0.07, mean(Z(r))=0.07, standard error=0.04, $p=0.044$, Figure 5b). These smaller correlations in cortical compared to perceptual face spaces could not be attributed solely to lower functional contrast-to-noise ratios in fMRI data, because the effects generally did not approach the noise-ceiling estimate for the sample (shaded region in Figure 5). The noise ceiling was based on the reproducibility of distance matrices between participants (Experimental Procedures). Instead, these findings indicate that the reference PCA space could not capture all the explainable dissimilarity variance in cortical face spaces.

Cortical face spaces over-represent eccentricity

We quantified the apparent warps in the cortical face spaces by constructing a multiple-regression RSA model, with separate distance-matrix predictors for eccentricity and direction, and for within and across face viewpoints (Figure 2a, Experimental Procedures). These predictors were scaled such that differences between the eccentricity and direction parameters could be interpreted as warping relative to a veridical encoding of distances in the reference PCA face space. We also observed strong viewpoint effects in multiple regions, including the early visual cortex (Figure 1e-f). Such effects were modeled by separate constant terms for distances within and across viewpoint.

Eccentricity changes had a consistently greater effect on representational patterns than direction changes, suggesting that a step change along the radial axis resulted in a larger pattern change than an equivalent step along the tangential axis (Figure 2c-d). The overrepresentation of eccentricity relative to direction was observed in each participant, and in both cortical and perceptual face spaces, although the effect was considerably larger in cortical face spaces. A repeated-measures two-factor ANOVA (eccentricity versus direction, within versus across viewpoint) on the single-participant parameter estimates from the multiple-regression RSA model was consistent with these apparent differences, with a statistically significant main effect of eccentricity versus direction for perceptual and cortical face spaces (all $p<0.012$, S2 Table). Thus, compared to the encoding in the reference PCA space, cortical face spaces over-represented the radial, distinctiveness-related axis compared to the tangential, identity-related axis.

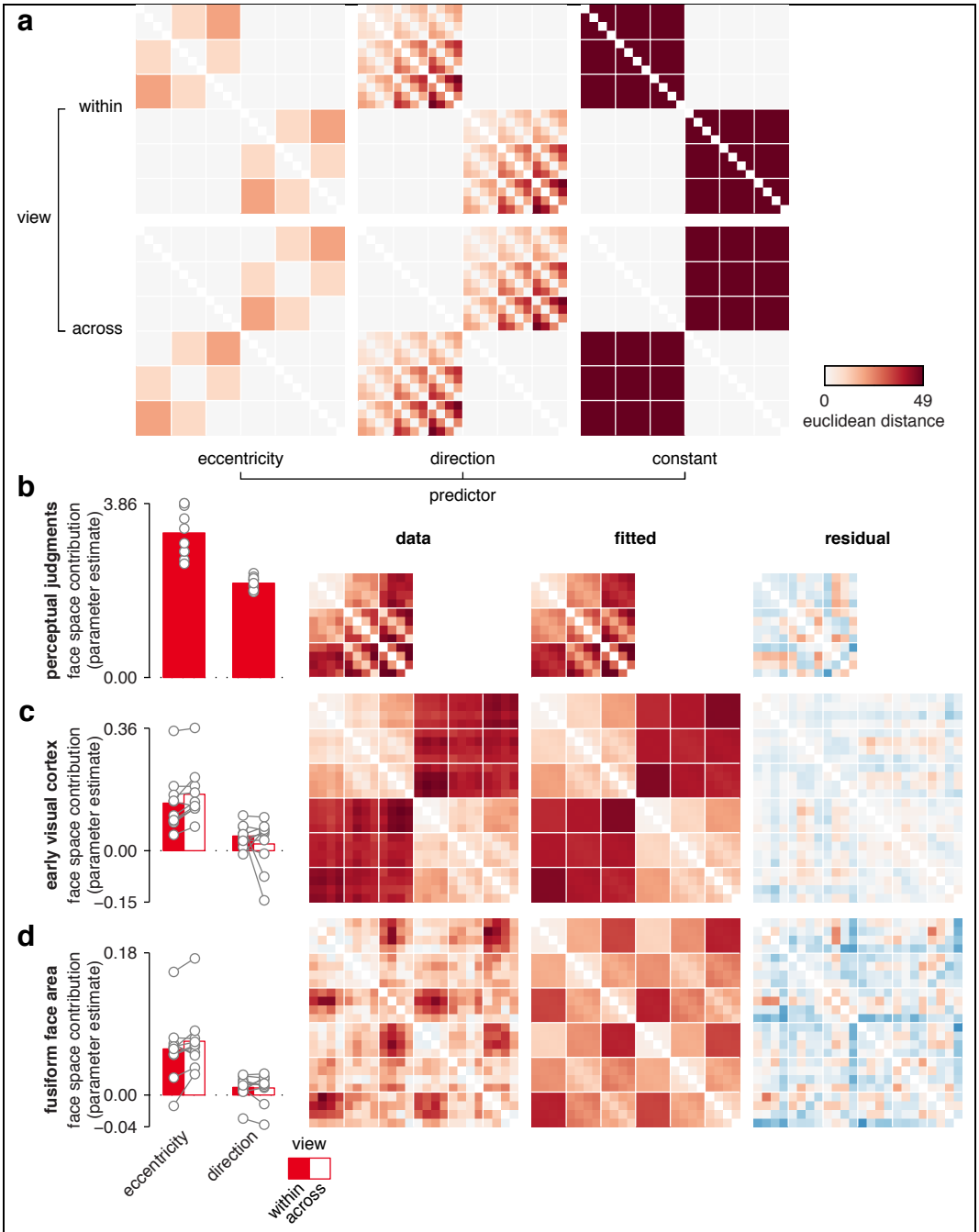


Figure 2. Face-space warping reflects an over-representation of eccentricity over direction information. (**a**) Squared distances in the reference PCA space were parameterized into predictors coding all 4 combinations of face-space metric (direction, eccentricity) and viewpoint (within, across). This multiple regression RSA model also included separate constant terms for each viewpoint, and was fitted separately to each face space using ordinary least squares (Experimental Procedures). Importantly, the scaling of the eccentricity and direction predictors ensures that equal parameter estimates corresponds to a preserved reference PCA space. (**b**) Multiple regression fit to the perceptual face space with group-average parameter estimates, distance matrix, fitted distances and residuals. Gray lines reflect single participant parameter estimates (**c-d**) Multiple regression fit to the cortical face spaces, plotted as in b. See S2 Table for inferential statistics.

164

Although these findings suggest a larger contribution of face-space eccentricity than direction in visual cortex, we also observed clear evidence for greater-than-chance discrimination performance among faces that differed only in face-space direction. Group-average cross-validated discriminant distances for faces that differed in direction but not eccentricity exceeded chance-level performance ($p<0.05$) for typical and caricatured faces in both the early visual cortex and the fusiform face area (S3 Figure). Sub-caricatured faces were less consistently discriminable. Indeed, direction discrimination increased with eccentricity both within (mean=0.013, standard error=0.007, p=0.036) and across (mean=0.012, standard error=0.005, p=0.016) viewpoint in the fusiform face area (linear effect of sub>typical>caricature), suggesting that direction discrimination increased with face-space eccentricity in a dose-dependent manner (S3 Table). By contrast, within viewpoint discrimination performance in the early visual cortex scaled with eccentricity (mean=0.039, standard error=0.008, $p<0.001$), but distances that spanned a viewpoint change did not vary with eccentricity (mean=0.009, standard error=0.019, p=0.297). This is consistent with a view-dependent representation in early visual cortex. Thus, cortical regions discriminate identity-related direction information even in the absence of a difference in distinctiveness-related eccentricity information, suggesting that cortical face representations cannot be reduced to a one-dimensional code based on distinctiveness alone. In summary, cortical coding of face-space position is systematically warped relative to the reference PCA space, with a substantial overrepresentation of eccentricity and a smaller, but reliable contribution of face-space direction.

Regional-mean fMRI activation increases with face-space eccentricity, but removing such effects does not substantially alter cortical face-space warping

Cortical face-space warping could not be explained by regional-mean activation preferences for caricatures. We performed a regional-mean analysis of responses in each cortical area, which confirmed previous reports that fMRI responses increase with distinctiveness across much of visual cortex (Figure 6, S6 Figure) (23). In order to test the influence of such regional-mean activation effects on representational distances, we adapted our discriminant distance metric to remove additive and multiplicative overall activation effects (Experimental Procedures). Distance matrices estimated using this alternative method were highly similar to ones estimated without removal of overall activation effects (all r=0.9 or greater for the Pearson correlation between group-average distance matrices with and without mean removal, S4 Figure). Thus, although eccentricity affected the overall activation in all visual areas, the warping of the cortical face spaces could not be attributed to overall activation effects alone.

Accounting for cortical and perceptual face spaces with PCA face-space-tuning models and image-computable models

We developed multiple computational models, each of which predicts a representational distance matrix and a regional-mean activation profile. These models can be divided into three classes: the sigmoidal-ramp tuning and exemplar tuning models receive face-space coordinates as input, while the Gabor-filter model receives gray-scale pixel intensities from the stimulus images as input. We evaluated each of these three model classes with and without a measurement-level population-averaging mechanism, which approximates how fMRI voxels locally average underlying neural activity.

The sigmoidal-ramp-tuning model proposes that the representational space is covered with randomly oriented ramps, each of which exhibits a monotonically increasing response along its preferred direction in face space (Figure 3a). This model is inspired by known preferences for extreme feature values in single units recorded from area V4 and from face-selective patches in the macaque visual cortex (20,27,28). We modeled the response along each model neuron's preferred direction using a sigmoidal function with two free parameters, which control the horizontal offset and the saturation of the response function (Experimental Procedures). A third parameter controlled the strength of measurement-level population averaging by translating each individual model unit's response toward the population-mean response. The way this accounts for local averaging by voxels is illustrated for the fusiform face area in Figure 3a-b. It can be seen that measurement-level population averaging introduces a substantial U-shape in the individual response functions, with only a minor deflection in favor of a preferred face-space direction. At the level of Euclidean distances between population response vectors evoked by each face, this leads to exaggerated distances for radial relative to tangential face differences (Figure 3c). In summary, measurement-level population averaging provides a simple means to interpolate between two extreme cases: A value of 0 corresponds to the case where the model's response is perfectly preserved in the fMRI voxels, whereas a value of 1 corresponds to the case where the model's response to a given stimulus is reduced to the arithmetic mean over the model units.

In the exemplar model, each unit prefers a location in face space, rather than a direction, and its tuning is described by a Gaussian centered on the preferred location. The representational space is covered by a population of units whose preferred locations are sampled from a Gaussian centered on the norm face (Figure 3d, Experimental Procedures). We fitted the Gaussian exemplar-tuning model similarly to the sigmoidal ramp-tuning model, using two parameters that controlled the width of the

239 Gaussian tuning function and the width of the Gaussian distribution from which
240 preferred faces were sampled. We also evaluated a variant of the exemplar model where
241 the distribution of preferred faces followed an inverted-Gaussian distribution (Figure
242 3e). Population averaging was modeled in the same way as for the sigmoidal-ramp-
243 tuning model using a third parameter.

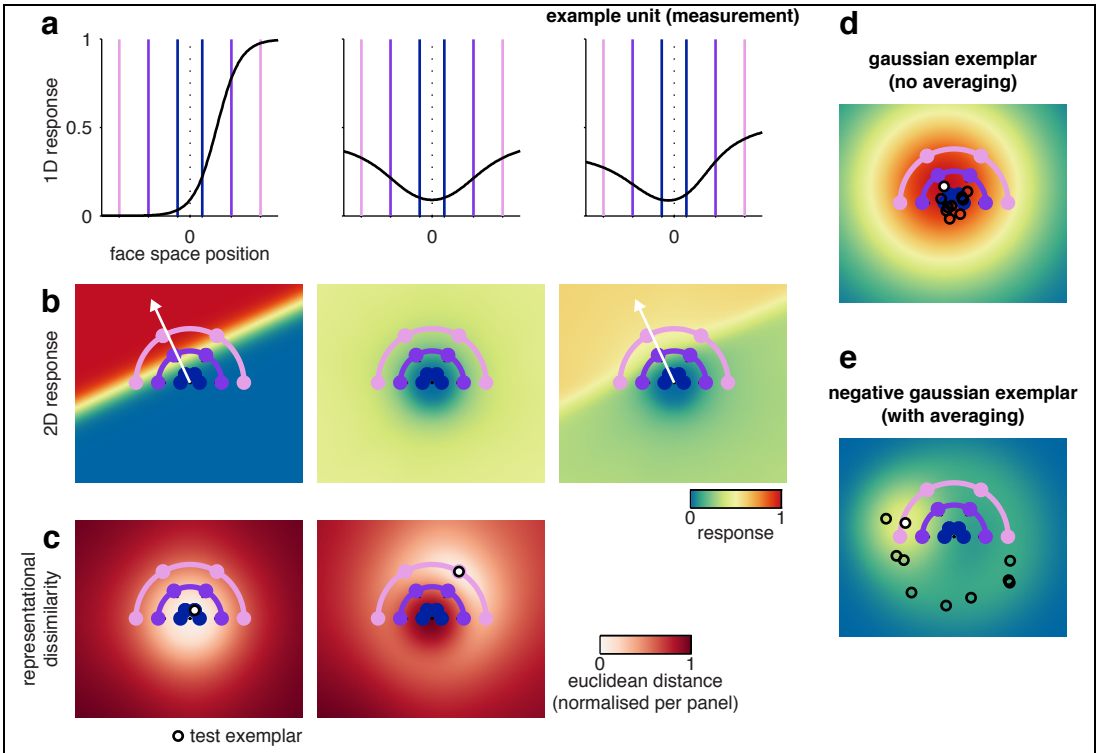


Figure 3. Computational models for face-space coding based on sigmoidal ramp or exemplar tuning coupled with measurement-level population averaging. The visualization uses the model parameters that were optimal for predicting the face space in the fusiform face area. (**a**) Response function from the sigmoidal ramp model's internal representation (left panel) and its measurement (right panel) following translation toward the population-average response function (middle panel). (**b**) Two-dimensional generalization of the sigmoidal ramp tuning function to encode a direction in the face-space slice. An example unit is plotted in the left and right panels with the population-average response in the middle panel. (**c**) The model's representational dissimilarity structure was estimated as the Euclidean distance between the population response vectors elicited by a coordinate in the face-space slice (white circle) and every other coordinate on the face space slice. Two example coordinates are plotted in the left and right panels. It can be seen that dissimilarity increases more rapidly with radial (eccentricity) than with tangential (direction) face-space distance. (**d**) Two-dimensional response function for an example unit from the Gaussian exemplar model (white marker). A sub-set of other units is overlaid in black markers to illustrate the width of the Gaussian distribution of tuning centers. (**e**) Two-dimensional response function for an example unit from the negative Gaussian exemplar model, plotted as in panel d.

244

The Gabor-filter model differs from the previous model classes in that it receives gray-scale image intensities as input, rather than face-space position (Figure 4a). Such models have previously been used to account for response preferences of individual voxels in early visual cortex (24). The model comprises Gabor filters varying in orientation, spatial frequency and phase, and spatial position. The filters are organized into banks, each corresponding to a spatial frequency and comprising a different number of spatial positions (coarser for lower spatial frequencies; Figure 4b, Experimental Procedures). We assumed that all orientations and spatial positions are equally represented. For the spatial frequencies, however, we let the data determine the weighting. We fitted a weighted representational model with one weight for each spatial-frequency bank (5 free parameters, Figure 4c). Local averaging in fMRI voxels was modeled using two stages of measurement-level population averaging: First, filters with tuning centers on either side of the vertical meridian were translated separately toward their respective hemifield-specific population averages. Second, a global-pool averaging was performed similarly to the other models. The contribution of these two population-average signals to the measured responses was modeled by 2 additional parameters (Figure 4d). The additional hemifield-specific averaging stage was necessary to account for strong view-specific effects in early visual cortex, but did not materially contribute to the fit in ventral temporal regions.

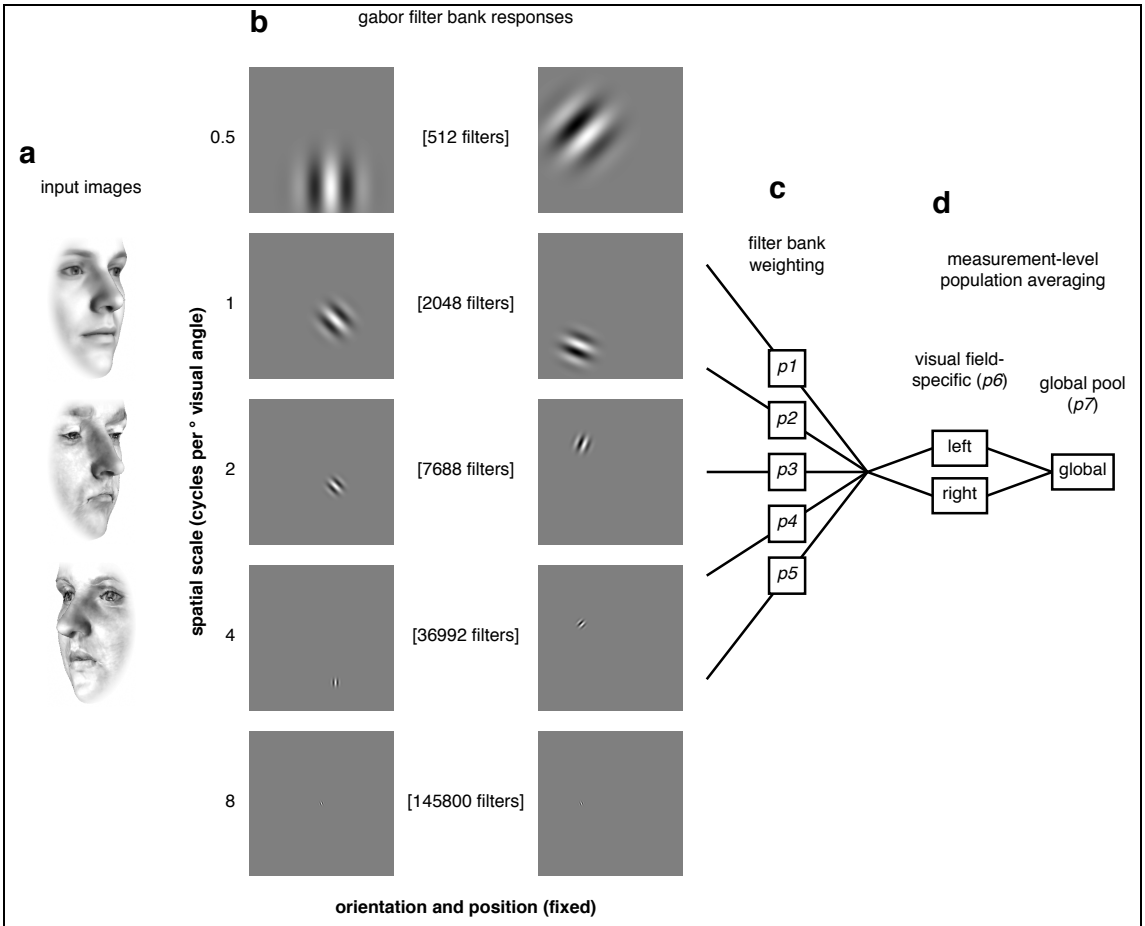


Figure 4. Schematic illustration of processing stages for the Gabor filter model. (**a**) The model receives gray-scale image intensities for each face exemplar. Three examples input images are illustrated in rows. (**b**) Image intensities are passed through banks of Gabor filters. The banks vary in spatial scale (filter standard deviation and grid spacing). The rows illustrate example filters from each bank. (**c**) The output of each filter bank is weighted and measurement effects are modeled using subsequent hemifield-specific and global-pool population averaging stages (**d**). The final output of the model is a Euclidean distance matrix estimated from the activation vectors in response to each face.

Multiple models can explain cortical and perceptual face spaces

We fitted each of the computational models so as to best predict the representational distance matrices from cortical regions and perceptual judgments. We used a leave-one-participant-out cross-validation approach, in which model performance was evaluated on participants and face identities not used in fitting the parameters (Experimental Procedures). Model performance was summarized as the Fisher-Z-transformed Pearson correlation coefficient between the model distances and the data distances. We performed statistical inference on the average Fisher-Z-

transformed correlations over all train-test splits of the data using t tests. Our cross-validation scheme tests for generalization across participants and face identities and ensures that models that differ in complexity (number of free parameters) can be compared. In order to investigate the effect of local averaging in fMRI voxels on representational similarity, we fitted two variants of each model: the full model and a variant that excluded measurement-level population averaging.

We found that all evaluated models explained almost all the explainable variance for the perceptual face space (Figure 5a), with only negligible differences in cross-validated generalization performance (for all pairwise model comparisons, see S4 Table). The inclusion of measurement-level population averaging had little effect on performance. This is expected because perceptual judgments, unlike fMRI voxels, are not affected by local averaging across representational units. Thus, the behavioral data was ambiguous with regard to the proposed models, which motivates model selection by comparison to the functional imaging data.

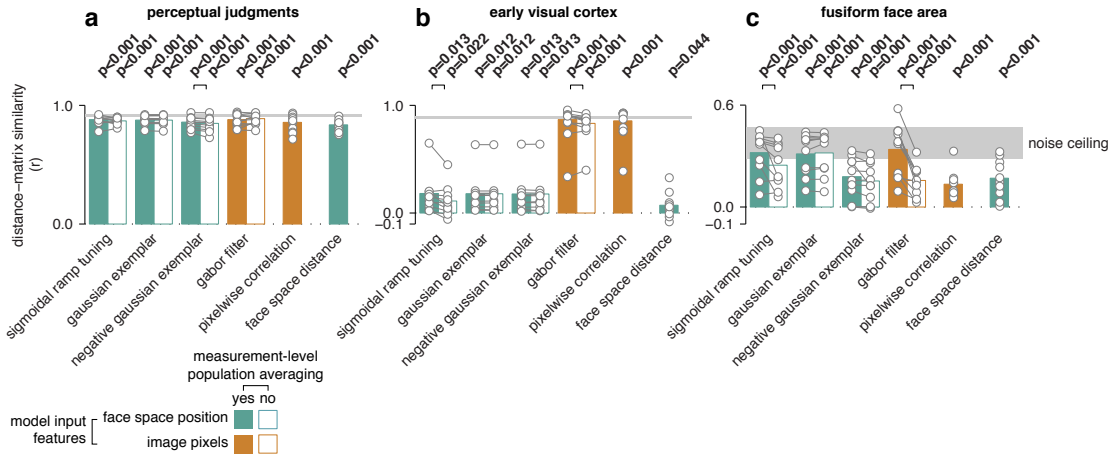


Figure 5. Cross-validated generalization performance for computational face-space (green bars) and image-based (orange bars) models, as well as mean performance for fixed predictors (final two bars in each panel). The bars provide group-averaged Pearson correlation coefficients, while individual participants are overlaid in gray markers. Filled bars indicate full model fits, while outlined bars indicate model fits excluding measurement-level population averaging. Statistically significant differences between model variants with and without population averaging are illustrated with black connection lines ($p<0.05$, see S4 Table for all pairwise comparisons). An estimate of the maximal performance expected given signal-to-noise levels in the sample is illustrated as a shaded noise ceiling (Experimental Procedures). Performance is plotted in separate panels for perceptual judgments (**a**), early visual cortex (**b**), and the fusiform face area (**c**). For other regions of interest, see S5 Figure.

Unlike the perceptual judgments data, the cortical face spaces exhibited substantial differences between the model fits, with a robust advantage for measurement-level population averaging in most cases. In the following, we focus on generalization performance for fits to the early visual cortex and the fusiform face area (for fits to other regions, see S5 Figure).

The early visual cortex was best explained by the Gabor-filter model, which beat the alternative computational models ($p<0.001$ for all pairwise model comparisons, S4 Table) and came close to explaining all explainable variance given noise levels in the data. Generalization performance for this model was slightly, but significantly better ($p=0.009$, Figure 5b) when population averaging was enabled. As a control, we also tested raw pixel intensities as the representational units. We found no significant

299 difference in performance between the 0-parameter pixel-intensity model and the fitted
300 Gabor-filter model with population averaging (p=0.380).

301 The fusiform face area was also well explained by the Gabor-filter model, but in this
302 region we observed similar performance for the sigmoidal-ramp-tuning model. Both
303 models, with population averaging, reached the lower bound of the noise ceiling
304 (Experimental Procedures; Figure 5c, S4 Table), suggesting that these models were able
305 to explain the variance in the dataset that was consistent between participants.
306 Measurement-level population averaging improved generalization performance in the
307 fusiform face area for both the Gabor-filter (p<0.001) and sigmoidal-ramp-tuning
308 models (p=0.032), but did not improve either of the exemplar models (p=0.263 for
309 Gaussian exemplar, p=0.096 for negative Gaussian exemplar). In summary, the
310 representation in the fusiform face area could be explained by multiple models, and in
311 most cases measurement-level population averaging improved the quality of the fit.

312 ***Regional-mean activation profiles are consistent with sigmoidal-ramp***
313 ***and Gabor-filter, but not Gaussian-exemplar models***

314 Multiple computational models provided qualitatively similar fits to our cortical data
315 at the distance-matrix level. However, we might still be able to adjudicate between them
316 at the level of regional-mean activation profiles. To this end, we obtained activation-
317 profile predictions from each model by averaging over all model units. We then
318 correlated the predicted population-mean activation profile with the regional-mean
319 fMRI activation profile for each participant. Even though these models were fitted to
320 distance matrices rather than to regional-mean fMRI responses, we found that they
321 exhibited systematic population-mean response preferences as a function of face-space
322 eccentricity (Figure 6, for other regions see S6 Figure). In particular, the sigmoidal-ramp
323 and Gabor-filter models both predicted increasing population-mean responses with
324 face-space eccentricity, while the Gaussian-exemplar model predicted decreasing

responses with eccentricity (S5 Table, S6 Table for pairwise comparisons). This constitutes evidence against the Gaussian-exemplar model, under the assumption that neuronal activity is positively associated with regional-mean fMRI response in visual regions (29–33). The preference for faces closer to the face-space origin (sub-caricatures) in the Gaussian-exemplar model arises as a necessary consequence of the Gaussian distribution of preferred faces, which is centered on the average face. We also tested a Gaussian exemplar-tuning model with an inverted Gaussian distribution of preferred faces. In this model, more units prefer faces far from the norm (caricatures) than faces close to the norm (sub-caricatures). However, the inverted-Gaussian exemplar model's generalization performance was considerably worse than the standard-Gaussian exemplar model's (Figure 5).

In summary, exemplar models accurately predicted representational distances for cortical face spaces when the preferred-face distribution was Gaussian, but such distributions led to inaccurate predictions of regional-mean fMRI activation profiles. Thus, analysis of regional-mean fMRI responses enabled us to adjudicate between models that made similar predictions at the distance-matrix level, and specifically indicated that the Gaussian exemplar model is unlikely to be the correct model for cortical face-space representation, despite a good fit at the distance-matrix level.

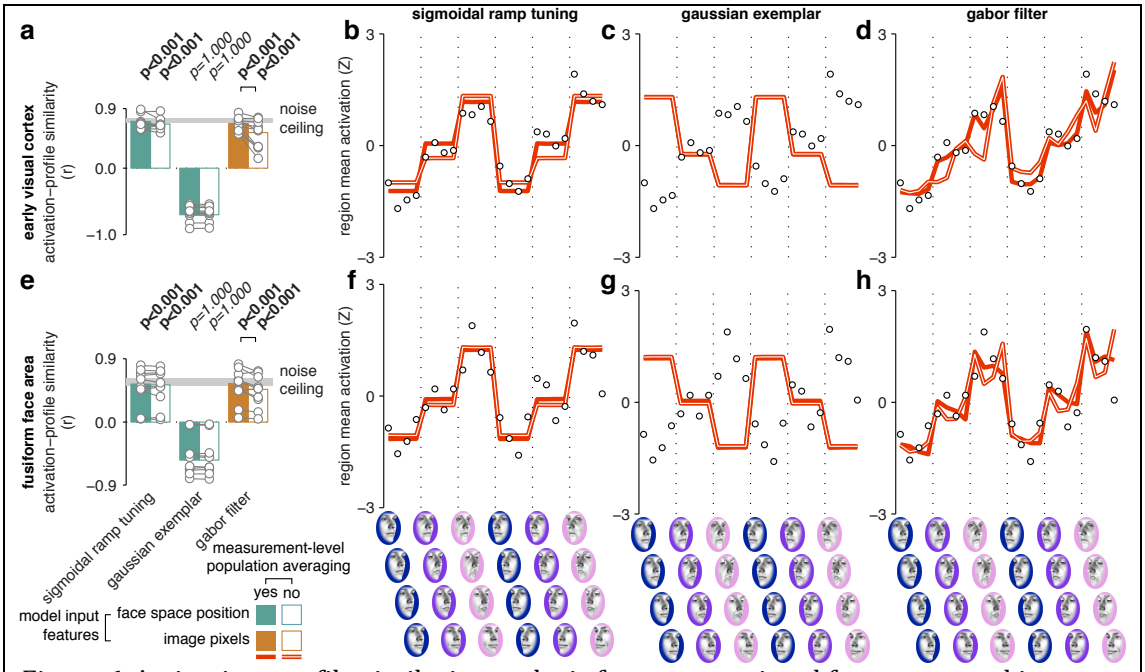


Figure 6. Activation-profile similarity analysis for computational face-space and image processing models. Bars (panels **a**, **e**) are plotted similarly as in Figure 5, with group-average effects in bars and single-participant estimates in markers. Markers in the line graphs (panels **b-d, f-h**) indicate group-average fMRI responses evoked by each face exemplar, while lines indicate the computational model predictions (with measurement-level population averaging in filled lines, without in outlined lines). Data and model predictions have been Z-scored to better illustrate the correlation-based fit metric we use here. The top panels illustrate effects in the early visual cortex, while the bottom panels illustrate effects in the fusiform face area. For other regions of interest, see S6 Figure.

Measurement-level population averaging is necessary to account for symmetric-view tolerance and over-representation of eccentricity

We found that models that included measurement-level population averaging generally outperformed models that did not. This advantage appeared to originate in how models with population averaging captured two effects in the cortical face spaces: symmetric view-tolerance and over-representation of face-space eccentricity relative to direction.

First, population averaging enabled the image-based Gabor-filter model to exhibit symmetric-view-tolerant responses to the face exemplars. Two mirror-symmetric views will drive a mirror symmetric set of Gabor features. Thus, while the pattern of activity

differs, the population-mean activity is similar. Measurement-level pooling over filters centered on distinct visual field locations therefore renders symmetric views more similar in the representation. For instance, the face space in the fusiform face area exhibited little sensitivity to viewpoint, and the Gabor-filter model fit to this region was greatly improved by the inclusion of measurement-level population averaging (Figure 5c, S2 Figure). Indeed, with population averaging, generalization performance for the image-based Gabor-filter model was similar to the sigmoidal-ramp-tuning and exemplar models, for which view tolerance is assumed at the input stage. General view tolerance, beyond the symmetric views we used here, is computationally more challenging. However, for our stimulus set, it was not necessary to posit any intrinsic view-invariant computations in the fusiform face area to explain how its face spaces come to exhibit symmetric view tolerance.

Second, measurement-level population averaging increased the degree to which both the sigmoidal-ramp and the Gabor-filter model over-represented face-space eccentricity relative to direction, which improved the fit for multiple cortical regions. To isolate this smaller effect from the larger symmetric-view-tolerance effect, we collapsed viewpoint in the first-level single-participant fMRI linear model and re-estimated the cortical face spaces and all model fits for the resulting simplified 12-condition design matrix, where each predictor coded appearances of a given face identity regardless of its viewpoint (S7 Figure). Even after collapsing across viewpoints at the first level in this way, measurement-level population averaging still improved the generalization performance of the sigmoidal-ramp and the Gabor-filter model in nearly all cases ($p<0.05$, see S7 Table for descriptive statistics and S8 Table for all pairwise comparisons), including the early visual cortex ($p<0.001$ for sigmoidal ramp tuning, $p=0.004$ for Gabor filter) and the fusiform face area ($p=0.001$ for sigmoidal ramp tuning, $p=0.007$ for Gabor filter). Thus, the advantage for measurement-level population averaging could not be accounted for by the fact that it helps explain symmetric view

381 tolerance. In sum, the addition of population averaging to the model improved model
382 generalization performance, and this advantage appeared to originate in accounts for
383 two distinct observed phenomena.

Discussion

This study investigated human face processing by measuring how a face space of individual exemplars was encoded in visual cortical responses measured with fMRI and in perceptual judgments. Relative to a reference PCA model of the 3D shape and texture of faces, cortical face spaces from all targeted regions systematically over-represented eccentricity relative to direction (i.e., the radial relative to the tangential axis). Cortical regions varied in their sensitivity to face viewpoint. We fitted multiple computational models to the data. Considered collectively, the cortical face spaces in the fusiform face area were most consistent with a face-space-based sigmoidal-ramp-tuning model and an image-based Gabor-filter model, and less consistent with models based on exemplar coding. As expected, effects in the early visual cortex were consistent primarily with the Gabor-filter model. In all cases, the winning models' performance depended on the inclusion of a measurement-level population-averaging mechanism, which approximates how individual model units are locally averaged in functional imaging measurements.

Functional MRI responses in the fusiform face area are best explained by sigmoidal-ramp-tuning and Gabor-filter models

Out of the models we considered, the best accounts for the fusiform face area were a face-space-based sigmoidal-ramp-tuning model and an image-based Gabor-filter model. Exemplar-coding models exhibited relatively lower generalization performance, or made inaccurate predictions for regional-mean fMRI activation profiles. Importantly, the advantage for both the sigmoidal-ramp and Gabor-filter model depended on the measurement-level population averaging mechanism. The key contribution of our modeling effort is to narrow the set of plausible representational models for the fusiform face area to two models that can explain both representational distances and the regional-mean activation profile.

It may appear surprising that a face-space coding model based on sigmoidal-ramp tuning and an image-based model based on Gabor filters should perform so similarly when fitted to face spaces in the fusiform face area. However, the sigmoidal-ramp-tuning model captures continuous variation in face shape and texture, which covaries with low-level image similarity. For instance, local curvature likely increases with face space eccentricity and is encoded in a ramp-like manner at intermediate stages of visual processing in macaque V4 (27). Conversely, the Gabor-filter model likely possesses sensitivity to face-space direction because the contrast of local orientation content varies with major face features such as eyebrow or lip thickness. There are multiple ways to parameterize face space, not all of which require domain-specific face features. This might also clarify why scene-selective areas such as the parahippocampal place area exhibited somewhat similar representational spaces as face-selective regions in the current study. Such widely-distributed face-exemplar effects are consistent with previous decoding studies (2,4). The Gabor-filter model provides one simple account for how such widely distributed face-exemplar effects can arise. It is likely that the face-space effects we report are driven at least in part by a general mechanism for object individuation in visual cortex rather than the engagement of specialized processing for face recognition.

Symmetric-view tolerance and over-representation of eccentricity in representational distances can be modeled as an fMRI measurement effect

The models we evaluate here raise the provocative possibility that in some cases, fMRI effects that might conventionally be attributed to high-level featural coding could instead arise from the neuroimaging measurement process. Such an explanation appears possible for two effects in our data. First, even though the Gabor-filter model is a single-layer network with limited representational flexibility, this model nevertheless

436 exhibited near-complete tolerance to mirror-symmetric viewpoint changes, when
437 coupled with measurement-level population averaging. Second, both this model and the
438 sigmoidal-ramp-tuning model showed greater over-representation of eccentricity when
439 measurement-level population averaging was enabled, suggesting that this over-
440 representation in the fMRI data might also plausibly arise through local averaging in
441 voxels.

442 Previous studies have tended to interpret view-tolerant fMRI effects in terms of
443 cortical processing to support invariant object recognition (2,34,35). The Gabor-filter
444 model suggests a mechanism by which functional imaging measures can exaggerate
445 apparent view-tolerance through spatial pooling over neuronal responses. This result
446 does not contradict previous reports of view-tolerant coding for faces in neuronal
447 population codes measured with single-unit recording (36–39), but rather
448 demonstrates that the type of tolerance to symmetric viewpoint changes that we
449 observed in the current study can be explained without resorting to such intrinsic view-
450 tolerant mechanisms (see also Ramirez et al. (40)). Such findings may go some way
451 toward reconciling apparent discrepancies between single-unit and functional imaging
452 data. For instance, tolerance to symmetrical viewpoints is widespread in human visual
453 cortex when measured with fMRI (34), but appears specific to a subset of regions in the
454 macaque face-patch system when measured with single-unit recordings (36). These
455 results are only contradictory if the measurement process is not considered. In
456 summary, we demonstrate that measurement effects can produce apparent view
457 tolerance in fMRI data. This finding does not suggest that fMRI cannot detect view-
458 tolerant coding (see also 41). For example, population averaging may not account for all
459 cases of non-symmetric view tolerance. However, our results do suggest that modeling
460 of the measurement process is important to correctly infer the presence of such
461 mechanisms from neuroimaging measurements.

Greater regional-mean activation for distinctive faces can arise from local averaging of neuronal responses

The winning models in this study exemplify how sensitivity to face-space eccentricity at the regional-mean activation level can arise as an artifact of averaging, with no individual neuron encoding distinctiveness or an associated psychological construct. Previous functional imaging studies often interpreted response modulations with face eccentricity as evidence for coding of distinctiveness or related social perception attributes (21–23,43). However, both the face-space-based sigmoidal-ramp-tuning model and the image-based Gabor-filter model exhibited increasing population-average responses with eccentricity, even though neither model encodes eccentricity at the level of its units. Although one could, of course, construct a competing model that explicitly codes eccentricity, the models used here are more consistent with single-unit recording studies, where cells generally are tuned to particular features, with a preference for extreme values, rather than responding to eccentricity regardless of direction (20,27,28). Here we demonstrate that when the local averaging of such biologically plausible neuronal tunings is modeled, eccentricity sensitivity emerges without specialized encoding of this particular variable. Related effects have been reported in attention research, where response-gain and contrast-modulation effects at the single-neuron level may sum to similar additive-offset effects at the fMRI-response level (44). In summary, direct interpretation of regional-mean fMRI activations in terms of neuronal tuning can be misleading when the underlying neuronal populations are heterogeneous.

Modeling of measurement-level population averaging is important for computational studies of cortical representation

A simple model of measurement-level population averaging was sufficient here to substantially improve the generalization performance of multiple computational models

for multiple cortical regions. The precise way that fMRI voxels sample neuronal activity patterns remains a topic of debate (30,31,33,45). A mechanistic model of voxel-sampling is unlikely to be robustly identifiable at the single-voxel level. However, under the simple assumption that voxels average random subsets of neurons with nonnegative weights, the effect on the apparent representational geometry will be a uniform stretching along the all-one vector (representing the population average). Here we approximated this effect for models with nonlinear parameters by mixing the population average into the predicted representational feature space. Despite the simplicity of this method, our noise-ceiling estimates indicate that the winning models captured nearly all the explainable variance in the current dataset. For model representations without nonlinear parameters, this measurement model can be implemented more easily by linearly combining the model's original distance matrix and the distance matrix obtained for the population average dimension of the space (using squared Euclidean distances estimates, see also 46–48). Thus, the measurement-level population averaging mechanism we propose here is widely and easily applicable to any case where a computational model is compared to neuroimaging data at the distance-matrix level.

The distance-matrix effects of local pooling of neuronal responses in fMRI voxels is correctly accounted for by our measurement model under the assumption that neurons are randomly intermixed in cortex (i.e. voxels sample random subsets of neurons). This simplifying assumption is problematic for early visual areas, where there is a well-established retinotopic organization with a strong contralateral response preference. For the retinotopic Gabor filter model, we therefore added a hemifield-specific pooling stage, which helped account for strong view-sensitivity in occipital regions of interest. Accounting for measurement effects in the presence of topographic organization is likely to prove more challenging for naturalistic stimulus sets. One solution is to account for local averaging in fMRI voxels by local averaging of the model's internal

representational map (42). This local-pooling approach can be thought of as providing a further constraint on the comparison between model and data, because smoothing the model representation is only expected to improve the fit if the model response topography resembles the cortical topography. This may provide a means of adjudicating between topographically organized models, even when the models predict similar distance matrices in the absence of measurement-effect modeling. In summary, the global population average is a special dimension of the representational space, which is overrepresented in voxels that pool random subsets of neurons. This effect accounts for much variance in the representational distances in the current study, and is easy to model. Modeling the overrepresentation of the global average, as we did here, is suitable for models that do not predict a spatial organization (e.g. face-space coding models). For models that do predict a spatial organization, it may be more appropriate to simulate fMRI voxels by local averaging of the model's representational map (42).

Model comparison is essential for computational neuroimaging

This study demonstrates the importance of considering multiple alternative models to guide progress in computational neuroimaging. In particular, the finding that practically every model we evaluated exhibited significantly greater-than-zero generalization performance strongly suggests how studies that only evaluate a limited set of candidate models can arrive at misleading conclusions (see e.g. 49).

Although the central goal of model comparison is to select the best account of the data, the finding that some models are not dissociable under the current experimental context also has important implications for the design of future studies. Here we demonstrated that Gaussian-distributed exemplar-coding models are less likely to account for human face coding, while accounts based on sigmoidal ramp tuning and Gabor filter outputs perform very similarly. This suggests the need to design stimulus sets that generate distinct predictions from these winning models. For example,

541 presenting face stimuli on naturalistic textured backgrounds may be sufficient to
542 adjudicate between the two models, because the Gabor-filter model lacks a mechanism
543 for figure-ground separation. In conclusion, our study exemplifies the need to test and
544 compare multiple models and suggests routes by which the sigmoidal-ramp-tuning
545 model of face-space coding could be further evaluated.

Materials and Methods

Data and software availability

The distance matrices we estimated for cortical and perceptual face spaces are available, along with software to re-generate all computational model fits (separate copies deposited on https://osf.io/5g9rv; https://doi.org/10.5281/zenodo.242666).

Sampling the reference PCA face space

We generated faces using a norm-based model of 3D face shape and texture, which has been described in detail previously (15,25). Briefly, the model comprises two PCA solutions (each trained on 200 faces), one based on 3D shape estimated from laser scans and another based on texture estimated from digital photographs. The components of each PCA solution are considered dimensions in a space that describes natural variation in facial appearance. All stimulus generation was performed using the PCA solution offered by previous investigators, and no further fitting was performed for this study. We yoked the shape and texture solutions in all subsequent analyses since we did not have distinct hypotheses for these.

We developed a method for sampling faces from the reference PCA space in a manner that would maximize dissimilarity variance. This is related to the concept of design efficiency in univariate general linear modeling (51), and involves maximizing the variance of hypothesized distances over the stimulus set. Because randomly sampled distances in high dimensional spaces tend to fall in a narrow range of distances relative to the norm (52), we reduced each participant's effective face space to 2D by specifying a plane which was centered on the norm of the space and extended at a random orientation. The face exemplars constituted a polar grid on this plane, with 4 directions at 60 degrees separation and 3 eccentricity levels (scaled at 30%, 100% and 170% of the mean eccentricity in the training face set). The resulting half-circle grid on a

plane through the high-dimensional space is adequate for addressing our hypotheses concerning the relative role of direction and eccentricity coding under the assumption that the high-dimensional space is isotropic. The orientation of the face-space slice was randomized between participants and model fits were based on cross-validation over participants. Under these conditions, any non-isotropicity is only expected to impair generalization performance. In preliminary tests we observed that this method yielded substantially greater dissimilarity variance estimates than methods based on Gaussian or uniform sampling of the space.

Face animation preparation

We used Matlab software to generate a 3D face mesh for each exemplar. This mesh was rendered at each of the orientations of interest in the study in a manner that centered the axis of rotation on the bridge of the nose for each face. This procedure ensured that the eye region remained centered on the fixation point throughout each animation in order to discourage eye movements. Renders were performed at sufficient increments to enable 24 frames per second temporal resolution in the resulting animations. Frames were converted to gray-scale and cropped with a feathered oval aperture to standardize the outline of each face and to remove high-contrast mesh edges from the stimulus set. Finally, we performed a frame-by-frame histogram equalization procedure where the average histogram for each frame was imposed on each individual face. Thus, the histogram was allowed to vary across time but not across faces. Note that histogram matching implies that the animations also have identical mean gray-scale intensity and root-mean-square contrast.

A potential concern with these matching procedures is that they could affect the validity of the comparison to the reference PCA space. However, we found that the opposite appeared to be true: distances in the reference PCA space were more predictive of pixelwise correlation distances in the matched images than in the original

597 images. Thus, the matching procedure did not remove features that were encoded in the
598 PCA space and may in fact have acted to emphasize such features.

599 ## *Participants*

600 10 healthy human participants participated in a similarity judgment task and fMRI
601 scans. The psychophysical task comprised 4 separate days of data collection which were
602 completed prior to 4 separate days of fMRI scans. All procedures were performed under
603 a protocol approved by the Cambridge Psychology Research Ethics Committee (CPREC).
604 Participants were recruited from the local area (Cambridge, UK) and were naïve with
605 regard to the purposes of the study. Five additional participants participated in initial
606 data collection but were not invited to complete the study due to difficulties with
607 vigilance, fixation stability, claustrophobia and/or head movements inside the scanner.
608 The analyses reported here include all complete datasets that were collected for the
609 study.

610 ## *Perceptual similarity judgment experiment*

611 We used a pair-of-pairs task to characterize perceptual similarity (S1 Movie).
612 Participants were presented with two vertically offset pairs of faces on a standard LCD
613 monitor under free viewing conditions, and judged which pair was relatively more
614 dissimilar with a button press on a USB keyboard (two-alternative force choice). Each
615 face rotated continuously between a leftward and a rightward orientation (45 degrees
616 left to 45 degrees right of a frontal view over 3 seconds). Ratings across all possible
617 pairings of face pairs (2145 trials: all pairings of the 66 possible pairs of the 12 faces)
618 were combined into a distance matrix for each participant, where each entry reflects the
619 percentage of trials on which that face pair was rated as relatively more dissimilar.

Functional MRI experiment

We measured brain response patterns evoked by faces in a rapid event-related fMRI experiment (S2 Movie). Participants fixated on a central point of the screen where a pseudo-random sequence of face animations appeared (7 degrees visual angle in height, 2s on, 1s fixation interval). We verified fixation accuracy online and offline using an infrared eye tracking system (Sensomotoric Instruments, 50Hz monocular acquisition). The faces rotated outward in leftward and rightward directions on separate trials (18 to 45 degrees rotation left or right of a frontal view), and participants responded with a button press to occasional face repetitions regardless of rotation (one-back task). This served to encourage attention to facial identity rather than to incidental low-level physical features. Consistent with a task strategy based on identity recognition rather than image matching, participants were sensitive to exemplar repetitions within viewpoint (mean d'+-1 standard deviation 2.68+-0.62) and to exemplar repetitions where the viewpoint changed (2.39+-0.52).

The experiment was divided into 16 runs where each run comprised 156 trials bookended by 10s fixation intervals. Each scanner run comprised two experimental runs, which were modeled independently in all subsequent analyses. The trial order in each run was first-order counterbalanced over the 12 faces using a De Bruijn sequence (53) with 1 additional repetition (diagonal entries in transfer matrix) added to each face in order to make the one-back repetition task more engaging and to increase design efficiency (51). The rotation direction in which each face appeared was randomized separately, since a full 24-stimulus De Bruijn sequence would have been over-long (576 trials). Although the resulting 24-stimulus sequences were not fully counter-balanced, we used an iterative procedure to minimize any inhomogeneity by rejecting rotation direction randomizations that generated off-diagonal values other than 0 and 1 in the 24-condition transfer matrix (that is, each possible stimulus-to-stimulus transfer in the sequence could appear once or not at all). These homogeneous trial sequences served to

enhance cross-validation performance by minimizing over-fitting to idiosyncratic trial sequence biases in particular runs. We modeled the data from each run with one predictor per face exemplar and viewpoint.

Magnetic resonance imaging acquisition

Functional and structural images were collected at the MRC Cognition and Brain Sciences Unit (Cambridge, UK) using a 3T Siemens Tim Trio system and a 32-channel head coil. Functional runs used a 3D echoplanar imaging sequence (2mm isotropic voxels, 30 axial slices, 192 x 192mm field of view, 128 x 128 matrix, TR=53ms, TE=30ms, 15° flip angle, effective acquisition time 1.06s per volume) with GRAPPA acceleration (acceleration factor 2 x 2, 40 x 40 PE lines). Each participant's functional dataset (7376 volumes over 8 scanner runs for the main experiment) was converted to NIFTI format and realigned to the mean of the first session's first experimental run using standard functionality in SPM8 (fil.ion.ucl.ac.uk/spm/software/spm8/). A structural T1-weighted volume was collected in the first session using a multi-echo MPRAGE sequence (1mm isotropic voxels)(54). The structural image was de-noised using previously described methods (55), and the realigned functional dataset's header was co-registered with the header of the structural volume using SPM8 functionality. The structural image was then skull-stripped using the FSL brain extraction tool (fmrib.ox.ac.uk/fsl), and a re-sliced version of the resulting brain mask was applied to the fMRI dataset to remove artifacts from non-brain tissue. We constructed design matrices for each run of the experiment by convolving the onsets of experimental events with the SPM8 canonical hemodynamic response function. Slow temporal drifts in MR signal were removed by projecting out the contribution of a set of nuisance trend regressors (polynomials of degrees 0-4) from the design matrix and the fMRI data in each run.

Cross-validated discriminant analysis

We estimated the neural discriminability of each face pair for each region of interest using a cross-validated version of the Mahalanobis distance (56). This analysis improves on the related Fisher's linear discriminant classifier by providing a continuous metric of discriminability without ceiling effects. Similarly to the linear discriminant, classifier weights were estimated as the contrast between each condition pair multiplied by the inverse of the covariance matrix of the residual time courses, which was estimated using a sparse prior (57). This discriminant was estimated separately for the concatenated design matrix and fMRI data in each possible leave-one-run-out training split, and the resulting weights were transformed to unit length and projected onto the contrast estimates from each training split's corresponding test run (16 estimates per contrast). The 16 run-specific distance estimates were averaged to obtain the final neural discriminability estimate for that participant and region. When the same data is used to estimate the discriminant and evaluate its performance, this algorithm returns the Mahalanobis distance, provided that a full rather than sparse covariance estimator is used (56). However, unlike a true distance measure, the cross-validated version that we use here is centered on 0 under the null hypothesis. This motivates summary-statistic random-effects inference for above-chance performance using conventional T tests.

We developed a variant of this discriminant analysis where effects that might be broadly described as region-mean-related are removed (S4 Figure). This control analysis involved two modifications to how contrasts were calculated at the level of forming the discriminant and at the level of evaluating the discriminant on independent data. First, each parameter estimate was set to a mean of zero in order to remove any additive offsets in response levels between the conditions. Second, for each pair of mean-subtracted parameter estimates, the linear contribution of the mean estimate over the pair was removed from each estimate before calculating the contrast. This corrects for the case where a single response pattern is multiplicatively scaled by the

conditions. The resulting control analysis is insensitive to effects driven by additive or multiplicative scaling offsets between the conditions.

Multiple regression RSA

We used a multiple regression model to estimate the relative contribution of eccentricity and direction to cortical and perceptual face-space representations. Multiple regression fits to distance estimates can be performed after a square transform, since squared distances sum according to the Pythagorean theorem. We partitioned the squared distances in the reference PCA space into variance associated with eccentricity changes by creating a distance matrix where each entry reflected the minimum distance for its eccentricity group in the squared reference PCA matrix (that is, cases along the group's diagonal where there was no direction change). The direction matrix was then constructed as the difference between the squared reference PCA matrix and the eccentricity matrix (Figure 2a). These predictors were vectorized and entered into a multiple regression model together with a constant term. This partitioning of the dissimilarity variance in the reference PCA matrix is complete in the sense that a multiple regression RSA model where the reference PCA matrix is used as the dependent variable yields parameter estimates of [1,1,0] for eccentricity, direction and constant, respectively, with no residual error. These three predictors were then split according to viewpoint, with separate sets of predictors for distances within and across viewpoint. The absolute values of the cortical and perceptual distance matrices were squared and then transformed back to their original sign before being regressed on the predictor matrix using ordinary least squares. Finally, the absolute values of the resulting parameter estimates were square-root transformed and returned to their original signs.

Functional regions of interest

We used a conventional block-based functional localizer experiment to identify category-selective and visually-responsive regions of interest in human visual cortex. Participants fixated a central cross on the screen while blocks of full-color images were presented (36 images per block presented with 222ms on, 222ms off, 16 s fixation). Participants were instructed to respond to exact image repetitions within the block. Each run comprised 3 blocks each of faces, scenes, objects and phase-scrambled versions of the scene images. Each participant's data (8 runs of 380 volumes) was smoothed with a Gaussian kernel (6mm full width at half maximum) and responses to each condition were estimated using a standard SPM8 first-level model. Regions of interest were identified using a region-growing approach, where a peak coordinate for each region was identified in individual participants, and a region of interest was grown as a contiguous set of the most selective 100 voxels extending from this coordinate. We defined the face-selective occipital and fusiform face areas with the minimum-statistic conjunction contrast of faces over objects and faces over baseline, and the scene-selective parahippocampal place area and transverse occipital sulcus as the minimum-statistic conjunction contrast of scenes over objects and scenes over baseline, and the early visual cortex as the contrast of scrambled stimuli over the fixation baseline. We also attempted to localize a face-selective region in the posterior superior temporal sulcus, a face-selective region in anterior inferotemporal cortex and a scene-selective region in retrosplenial cortex, but do not report results for these regions here since they could only be identified in a minority of the participants. All regions of interest were combined into bilateral versions before further analysis since we did not have distinct predictions concerning functional lateralization.

Sigmoidal ramp tuning model

The sigmoidal ramp model comprises 1000 model units, each of which exhibits a monotonically increasing response in a random direction extending from the origin of the face space (Figure 3). The response *y* at position *x* along the preferred direction is described by the sigmoid

y[raw] = 1 / (1+exp((-x+o)/s));

where the free parameters are *o*, which specifies the horizontal offset of the response function (zero places the midpoint of the response function at the norm of the space, values greater than zero corresponds to responses shifted away from the norm), and *s*, which defines response function saturation (4 corresponds to a near-linear response in the domain of the face exemplars used here, while values near zero correspond to a step-like increase in response). The raw output of each model unit is then translated toward the population-mean response

y[final] = (y[raw]-y[mean]) * (1-p) + y[mean]

where *p* is a free parameter that defines the strength of measurement-level population averaging (0 corresponds to no averaging, 1 corresponds to each model unit returning the population-mean response).

Exemplar model

The exemplar model comprises 1000 model units, each of which prefers a Cartesian coordinate in the face space with response fall-off captured by an isotropic Gaussian. The free parameters are *w*, which controls the full width at half-maximum tuning width of the Gaussian response function, and *d*, which controls the width of the Gaussian distribution of tuning centers (0.1 places Z=2.32 at 10% of the eccentricity of the caricatures while 3 places this tail at 300% of the eccentricity of the caricatures).

We also constructed an inverted-Gaussian variant of this model where the distribution of distances was inverted at Z=2.32 and negative distances truncated to zero (1% of exemplars). This model was fitted with similar parameters as the original Gaussian exemplar model.

Gabor filter model

The Gabor filter model is an adaptation of a neuroscientifically-inspired model that has previously been used to successfully predict single-voxel responses in the early visual cortex (24, github.com/kendrickkay/knkutils/tree/master/imageprocessing). The model is composed of 5 banks of Gabor filters varying in spatial position, phase (2 values) and orientation (8 directions, Figure 4). We measured each filter's response to the last frame of each animation, and corrected for phase shifts. The resulting rectified response vectors were weighted according to filter bank membership (5 free parameters). We estimated measurement-level population averaging using two pooling stages: a hemifield-specific pool, where filters were pooled according to whether their centers fell left or right of the vertical meridian, followed by a global pool.

Pixelwise correlation predictor

We used a fixed control predictor to estimate whether coding based on pixelwise features would produce the same face space warping we observed in our data (Figure 4). The pixelwise correlation predictor was generated by stacking all the pixels in each of the face animations into vectors and estimating the correlation distance between these intensity values.

Estimating the noise ceiling

We estimated the noise ceiling for Z-transformed Pearson correlation coefficients based on methods described previously (56). This method estimates the explained variance that is expected for the true model given noise levels in the data. Although the

796 true noise level of the data cannot be estimated, it is possible to approximate its upper
797 and lower bounds in order to produce a range within which the true noise ceiling is
798 expected to reside. The lower bound estimate is obtained by a leave-one-participant-out
799 cross-validation procedure where the mean distance estimates of the training split are
800 correlated against the left-out-participant's distances, while the upper bound is obtained
801 by performing the same procedure without splitting the data. These estimates were
802 visualized as a shaded region in figures after reversing the Z-transform (Figure 4).

803 ***Statistical inference***

804 All statistical inference was performed using T-tests at the group-average level
805 (N=10 in all cases except the occipital face area and transverse occipital sulcus, N=9).
806 Correlation coefficients were Z-transformed prior to statistical testing. Average Z
807 statistics were reverse-transformed before visualization for illustrative purposes.

808 Fold-wise generalization performance estimates are partially dependent, which can
809 lead to sample variance underestimates (58,59) and greater than intended false positive
810 rates when conventional parametric statistics are used. However, we simulated the
811 effects of this potential bias and found no consistent inflation in false-positive rates for
812 simulations of the parameters used in the current study (S1 Code, S8 Figure, S9 Figure).
813 Thus, the inferential statistics reported in the current study appear to be robust to this
814 slight dependence and maintain their intended frequentist properties.

Acknowledgments

We are grateful to Jenna Parker for assistance with data collection and to Marta Correia for assistance with 3D EPI protocols. This work was funded by a British Academy Postdoctoral fellowship (J.D.C) and by a grant from ERC (N.K., code 261352).

References

1. Anzellotti S, Caramazza A. From parts to identity: Invariance and sensitivity of face representations to different face halves. Cereb Cortex. :1–10.
2. Anzellotti S, Fairhall SL, Caramazza A. Decoding representations of face identity that are tolerant to rotation. Cereb Cortex [Internet]. 2014 Mar 5 [cited 2013 Mar 6];24:1988–95. Available from: http://www.cercor.oxfordjournals.org/cgi/doi/10.1093/cercor/bht046
3. Axelrod V, Yovel G. Successful decoding of famous faces in the fusiform face area. PLoS One [Internet]. 2015;10:e0117126. Available from: http://dx.plos.org/10.1371/journal.pone.0117126
4. Goesaert E, Op de Beeck HP. Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. J Neurosci [Internet]. 2013 May 8 [cited 2013 May 8];33(19):8549–58. Available from: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1829-12.2013
5. Kriegeskorte N, Formisano E, Sorger B, Goebel R. Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci. 2007;104:20600–5.
6. Natu VS, Jiang F, Narvekar A, Keshvari S, Blanz V, O'Toole AJ. Dissociable neural patterns of facial identity across changes in viewpoint. J Cogn Neurosci. 2010;22(7):1570–82.
7. Nestor A, Plaut DC, Behrmann M. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proc Natl Acad Sci [Internet]. 2011 May 31 [cited 2011 Jun 1];108:9998–10003. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1102433108
8. Nestor A, Behrmann M, Plaut DC. The neural basis of visual word form processing: A multivariate investigation. Cereb Cortex [Internet]. 2013 Jun 12 [cited 2013 Jun 5];23:1673–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22693338
9. Gao X, Wilson HR. The neural representation of face space dimensions. Neuropsychologia [Internet]. Elsevier; 2013 Jul 10 [cited 2013 Aug 11];51(10):1787–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23850598
10. Verosky SC, Todorov A, Turk-Browne NB. Representations of individuals in ventral temporal cortex defined by faces and biographies. Neuropsychologia [Internet]. Elsevier; 2013 Jul 16 [cited 2013 Jul 23];51:2100–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23871881
11. Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci. 2008;2:1–28.
12. Haxby J V, Hoffman E, Gobbini M. The distributed human neural system for face perception. Trends Cogn Sci. 2000;4:223–33.
13. Bruce V, Young AW. Understanding face recognition. Br J Psychol [Internet]. 1986 Aug [cited 2010 Sep 23];77 (Pt 3):305–27. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3756376
14. Valentine T. A unified account of the effects of distinctiveness, inversion, and race in face recognition. Q J Exp Psychol [Internet]. 1991 [cited 2011 Jun

14];43A(2):161–204. Available from: http://www.informaworld.com/index/776369317.pdf

15. Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. Proc 26th Annu Conf Comput Graph Interact Tech - SIGGRAPH '99 [Internet]. New York, New York, USA: ACM Press; 1999;187–94. Available from: http://portal.acm.org/citation.cfm?doid=311535.311556

16. O'Toole AJ, Abdi H, Deffenbacher KA, Valentin D. Low-dimensional representation of faces in higher dimensions of the face space. J Opt Soc Am A [Internet]. 1993 [cited 2012 Jun 14];10(3):405–15. Available from: http://www.opticsinfobase.org/abstract.cfm?id=4552

17. Ross DA, Hancock PJB, Lewis MB. Changing faces: Direction is important. Vis cogn [Internet]. 2010 Jan [cited 2012 May 30];18(1):67–81. Available from: http://www.tandfonline.com/doi/abs/10.1080/13506280802536656

18. Schulz C, Kaufmann JM, Walther L, Schweinberger SR. Effects of anticaricaturing vs. caricaturing and their neural correlates elucidate a role of shape for face learning. Neuropsychologia [Internet]. 2012 Jun 27 [cited 2012 Aug 1];50:2426–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22750120

19. Wilson HR, Loffler G, Wilkinson F. Synthetic faces, face cubes, and the geometry of face space. Vision Res [Internet]. 2002 Dec;42(27):2909–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12450502

20. Leopold DA, Bondar I V, Giese MA. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature [Internet]. 2006 Aug 3;442(7102):572–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16862123

21. Loffler G, Yourganov G, Wilkinson F, Wilson HR. fMRI evidence for the neural representation of faces. Nat Neurosci. 2005;10:1386–90.

22. Davidenko N, Remus D a., Grill-Spector K. Face-likeness and image variability drive responses in human face-selective ventral regions. Hum Brain Mapp [Internet]. 2012 Aug 5 [cited 2011 Aug 7];33:2334–49. Available from: http://doi.wiley.com/10.1002/hbm.21367

23. Said CP, Dotsch R, Todorov A. The amygdala and FFA track both social and non-social face dimensions. Neuropsychologia [Internet]. Elsevier Ltd; 2010 Aug [cited 2010 Sep 7];48(12):3596–605. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20727365

24. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008;452:352–5.

25. Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T. A 3D Face Model for Pose and Illumination Invariant Face Recognition. 2009 Sixth IEEE Int Conf Adv Video Signal Based Surveill [Internet]. Ieee; 2009 Sep;296–301. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5279762

26. Westfall J, Nichols TE, Yarkoni T. Fixing the stimulus-as-fixed-effect fallacy in task fMRI. BioRxiv. 2016;

27. Pasupathy A, Connor CE. Shape Representation in Area V4 : Position-Specific Tuning for Boundary Conformation. J Neurophysiol. 2001;86:2505–19.

28. Freiwald WA, Tsao DY, Livingstone M. A face feature space in the macaque temporal lobe. Nat Neurosci. 2009;12:1187.

29. Boynton GM. Spikes, BOLD, attention, and awareness: a comparison of electrophysiological and fMRI signals in V1. J Vis [Internet]. 2011 Jan;11(5):12.

Available from:
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4124818&tool=pmcentrez&rendertype=abstract

30. Goense J, Logothetis N. Neurophysiology of the BOLD fMRI signal in awake monkeys. Curr Biol. 2008;18:631–40.

31. Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann a. Neurophysiological investigation of the basis of the fMRI signal. Nature [Internet]. 2001 Jul 12;412(6843):150–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11449264

32. Sirotin Y, Das A. Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. Nature. 2009;457:475–9.

33. Cardoso MMB, Sirotin YB, Lima B, Glushenkova E, Das A. The neuroimaging signal is a linear sum of neurally distinct stimulus- and task-related components. Nat Neurosci [Internet]. Nature Publishing Group; 2012;15(9):1298–306. Available from: http://dx.doi.org/10.1038/nn.3170

34. Kietzmann TC, Swisher JD, Konig P, Tong F. Prevalence of Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Visual Pathways. J Neurosci [Internet]. 2012 Aug 22 [cited 2012 Aug 22];32(34):11763–72. Available from: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0126-12.2012

35. Axelrod V, Yovel G. Hierarchical Processing of Face Viewpoint in Human Visual Cortex. J Neurosci [Internet]. 2012 Feb 14 [cited 2012 Feb 14];32(7):2442–52. Available from: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4770-11.2012

36. Freiwald WA, Tsao DY. Functional compartmentalization and viewpo--int generalization within the macaque face-processing system. Science (80-) [Internet]. 2010 Nov [cited 2010 Nov 5];330(6005):845–51. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1194908

37. Hasselmo M, Rolls E, Baylis G, Nalwa V. Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. Exp Brain Res. 1989;75:417–29.

38. Wachsmuth E, Oram M, Perrett DI. Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. Cereb Cortex. 1994;4:509–22.

39. Perrett DI, Rolls E, Caan W. Visual neurones responsive to faces in the monkey temporal cortex. Exp Brain Res. 1982;47:329–42.

40. Ramirez FM, Cichy RM, Allefeld C, Haynes J-D. The neural code for face orientation in the human fusiform face area. J Neurosci. 2014;34:12155–67.

41. Carlin JD. Decoding Face Exemplars from fMRI Responses: What Works, What Doesn't? J Neurosci. 2015;35(25):9252–4.

42. Kriegeskorte N, Diedrichsen J. Inferring brain-computational mechanisms with models of activity measurements. Philos Trans R Soc B Biol Sci. 2016;1–19.

43. Mattavelli G, Andrews TJ, Asghar AUR, Towler JR, Young AW. Response of face-selective brain regions to trustworthiness and gender of faces. Neuropsychologia [Internet]. Elsevier; 2012 May 30 [cited 2012 Jul 17];50(9):2205–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22659107

44. Hara Y, Pestilli F, Gardner JL. Differing effects of attention in single-units and populations are well predicted by heterogeneous tuning and the normalization model of attention. Front Comput … [Internet]. 2014 [cited 2014 Feb 28];8:1–13.

Available from:
http://www.frontiersin.org/Journal/10.3389/fncom.2014.00012/abstract

45. Kriegeskorte N, Bandettini PA, Cusack R. How does an fMRI voxel sample the neuronal activity pattern: Compact-kernel or complex-spatiotemporal filter? Neuroimage. 2009;49(3):1965–76.

46. Khaligh-Razavi S-M, Henriksson L, Kay K, Kriegeskorte N. Fixed versus mixed RSA : Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. BiorXiv. 2016;1–32.

47. Khaligh-Razavi S-M, Kriegeskorte N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol. 2014;10(November):1–29.

48. Jozwik KM, Kriegeskorte N, Mur M. Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. Neuropsychologia [Internet]. Elsevier; 2016;83:201–26. Available from: http://dx.doi.org/10.1016/j.neuropsychologia.2015.10.023

49. Carlin JD, Kriegeskorte N. Ramp coding with population averaging predicts human cortical face-space representations and perception [Internet]. BiorXiv. 2015 Oct. Available from: http://biorxiv.org/lookup/doi/10.1101/029603

50. Cusack R, Vicente-Grabovetsky A, Mitchell DJ, Wild CJ, Auer T, Linke AC, et al. Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. Front Neuroinform [Internet]. 2014;8(January):90. Available from: http://journal.frontiersin.org/article/10.3389/fninf.2014.00090/abstract

51. Henson RNA. Analysis of fMRI timeseries: Linear time-invariant models, event-related fMRI and optimal experimental design. In: Frackowiak RSJ, Friston KJ, Frith CD, Dolan RJ, Price CJ, editors. Human Brain Function. New York: Academic Press; 2003. p. 793–822.

52. Burton AM, Vokey JR. The face-space typicality paradox: Understanding the face-space metaphor. Q J Exp Psychol [Internet]. 1998 [cited 2012 May 31];3:475–83. Available from: http://www.tandfonline.com/doi/abs/10.1080/713755768

53. Aguirre GK, Mattar MG, Magis-Weinberg L. de Bruijn cycles for neural decoding. Neuroimage. Elsevier Inc.; 2011 Mar 24;56(3):1293–300.

54. van der Kouwe a. JW, Benner T, Salat DH, Fischl B. Brain morphometry with multiecho MPRAGE. Neuroimage. 2008;40(2):559–69.

55. Manjón J V., Coupé P, Martí-Bonmatí L, Collins DL, Robles M. Adaptive non-local means denoising of MR images with spatially varying noise levels. J Magn Reson Imaging. 2010;31(1):192–203.

56. Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Comput Biol. 2014;10:e1003553.

57. Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage [Internet]. 2010 May [cited 2010 Aug 2];53:103–18. Available from: http://dx.doi.org/10.1016/j.neuroimage.2010.05.051

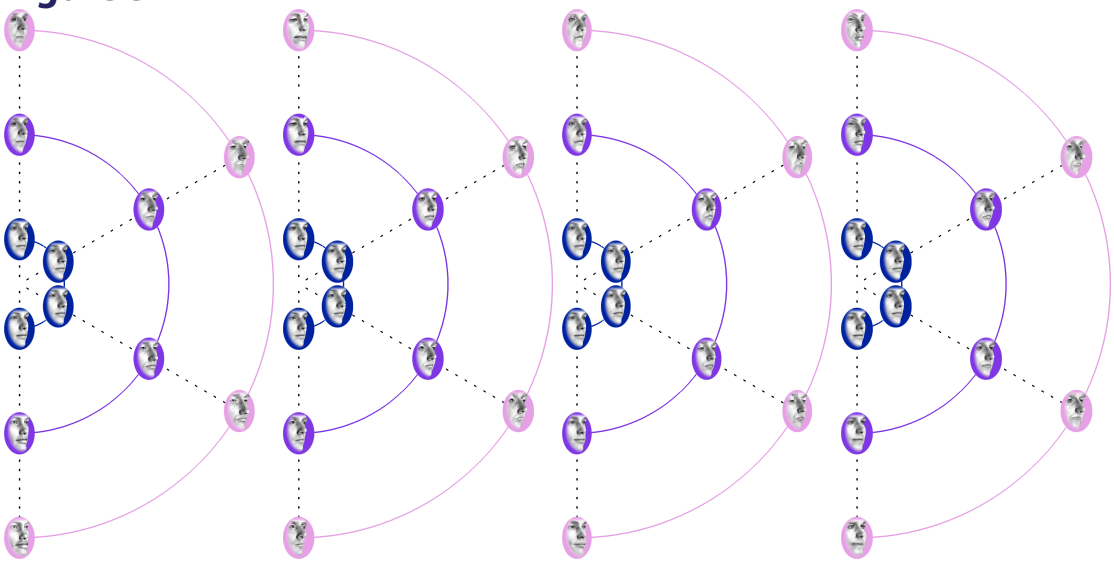
58. Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn. 2003;52(3):239–81.

59. Bengio Y, Grandvalet Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. J Mach Learn Res [Internet]. 2004;5:1089–105. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12646250

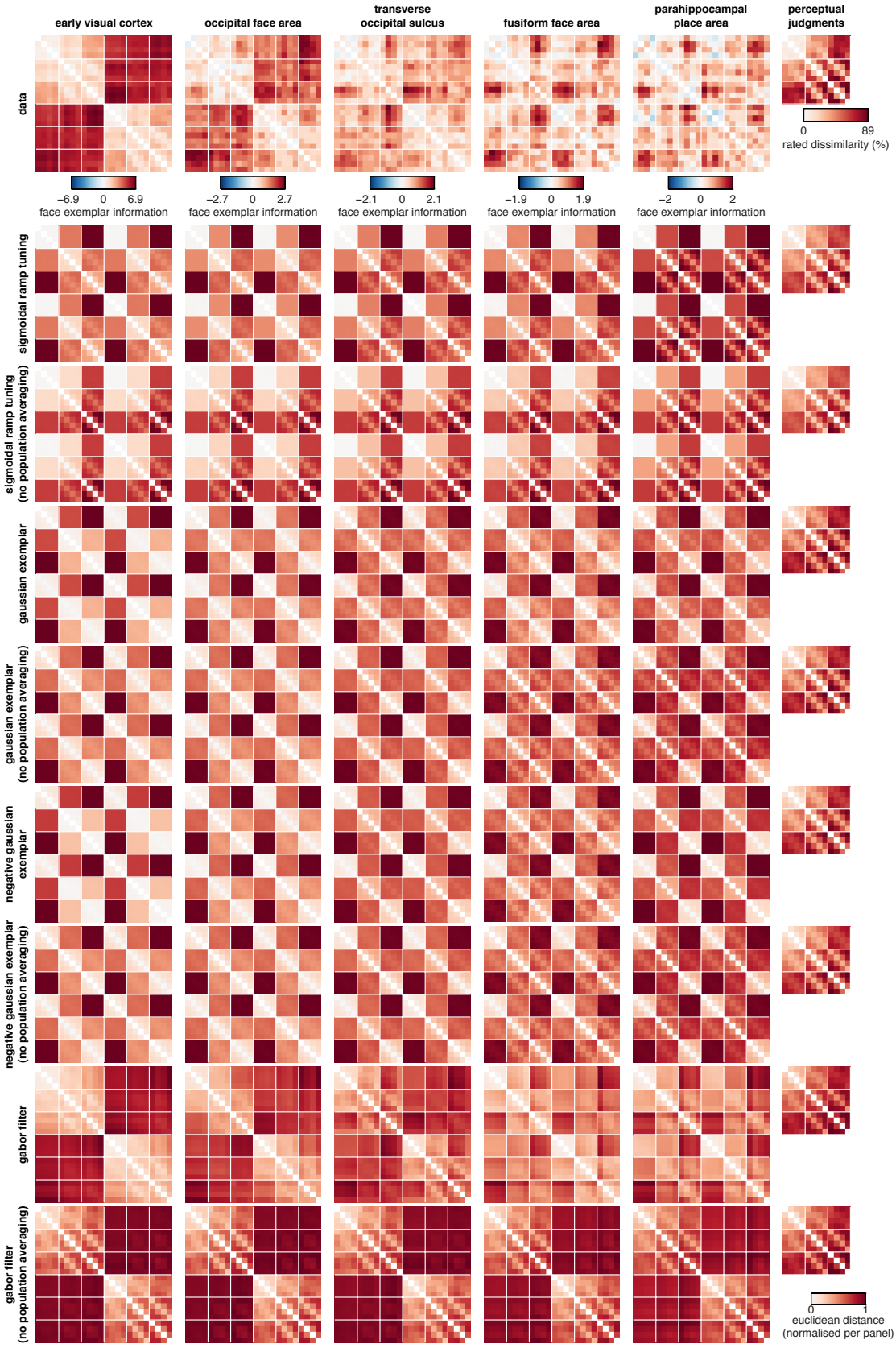
1009

Supporting information

Figure S1

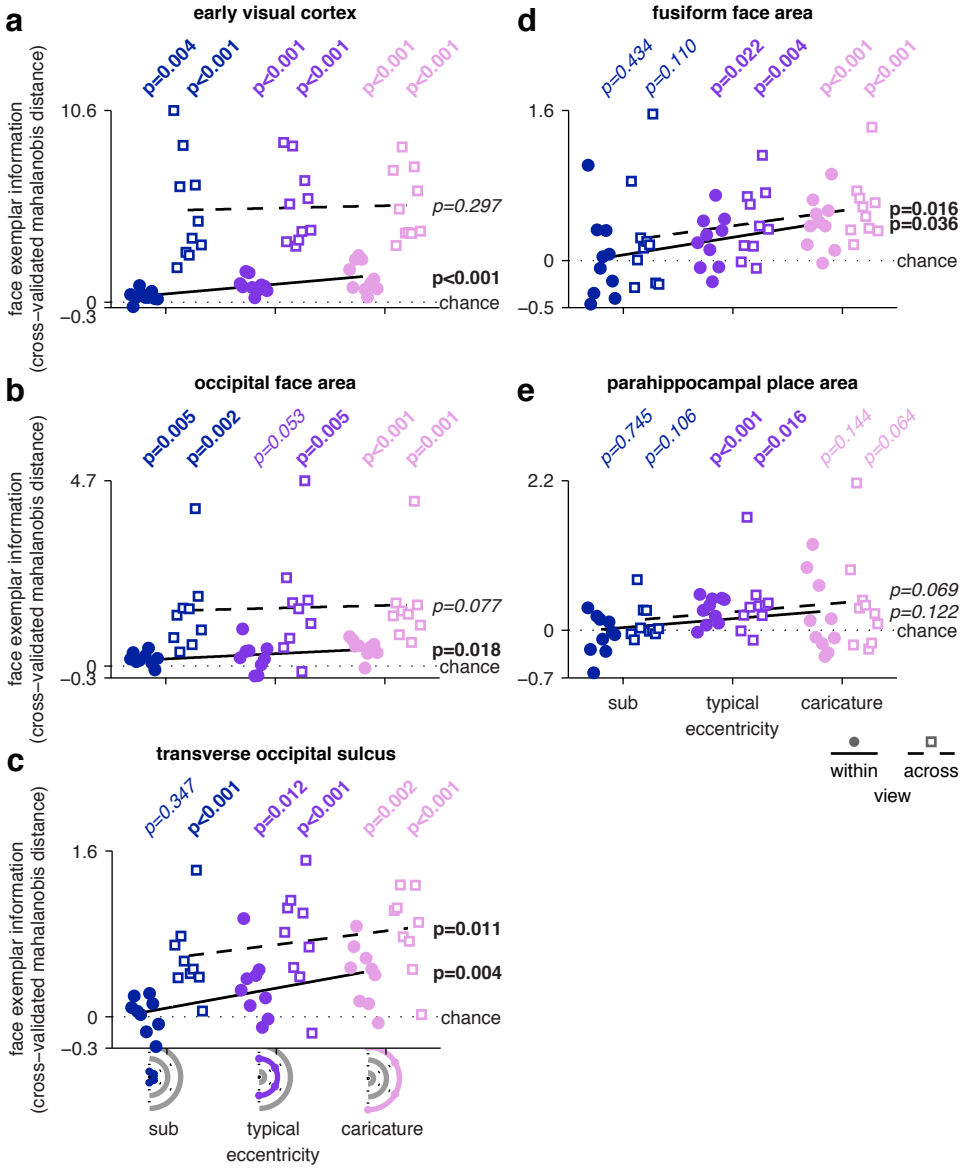


S1 Figure. Example stimulus sets for 4 participants. Each stimulus set shares the same underlying distance matrix in the reference PCA space, while the randomization of the orientation of the plane on which the faces are sampled ensures that each set is visually distinct.



1016

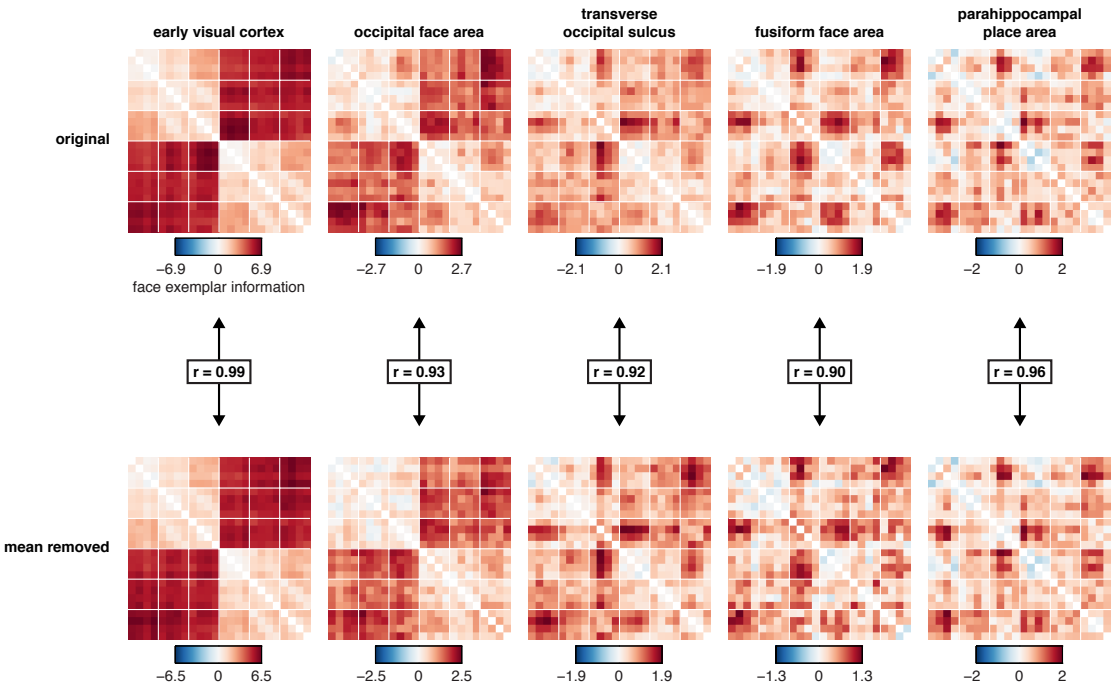
1017 *S2 Figure*. Group-average distance matrices from additional regions of interest and
1018 best-fitting model predictions from each model considered in the main manuscript
1019 (Figure 5).



1020

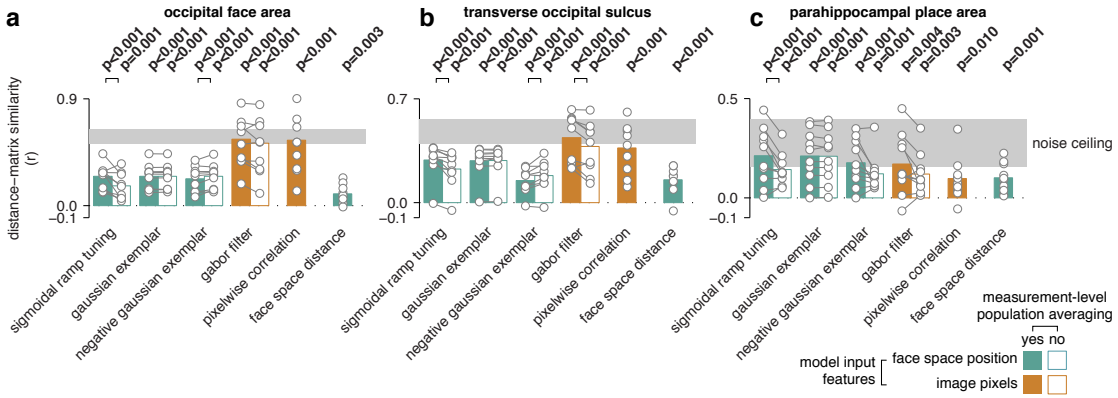
1021 *S3 Figure*. Cortical direction discriminability as a function of eccentricity level. Each
1022 point reflects the mean performance for all directions at a given eccentricity level (4x4
1023 block diagonals in Figure 1) for a single participant. Small random offsets have been
1024 added to each x coordinate for illustrative purposes, and a line shows the least-squares
1025 fit. Performance is plotted separately for distances within viewpoint (round markers,

1026 solid line, left offset) and across viewpoint (square markers, dashed line, right offset). All
1027 plotted p values are obtained through group analysis of single-participant estimates.
1028 Within viewpoint, cortical discrimination performance increases with eccentricity level
1029 in all regions except the parahippocampal place area (e). Across viewpoint, statistically
1030 significant effects are observed in the ventral temporal fusiform face area in the lateral
1031 temporal transverse occipital sulcus, but not in occipital areas (early visual cortex,
1032 occipital face area).



1033

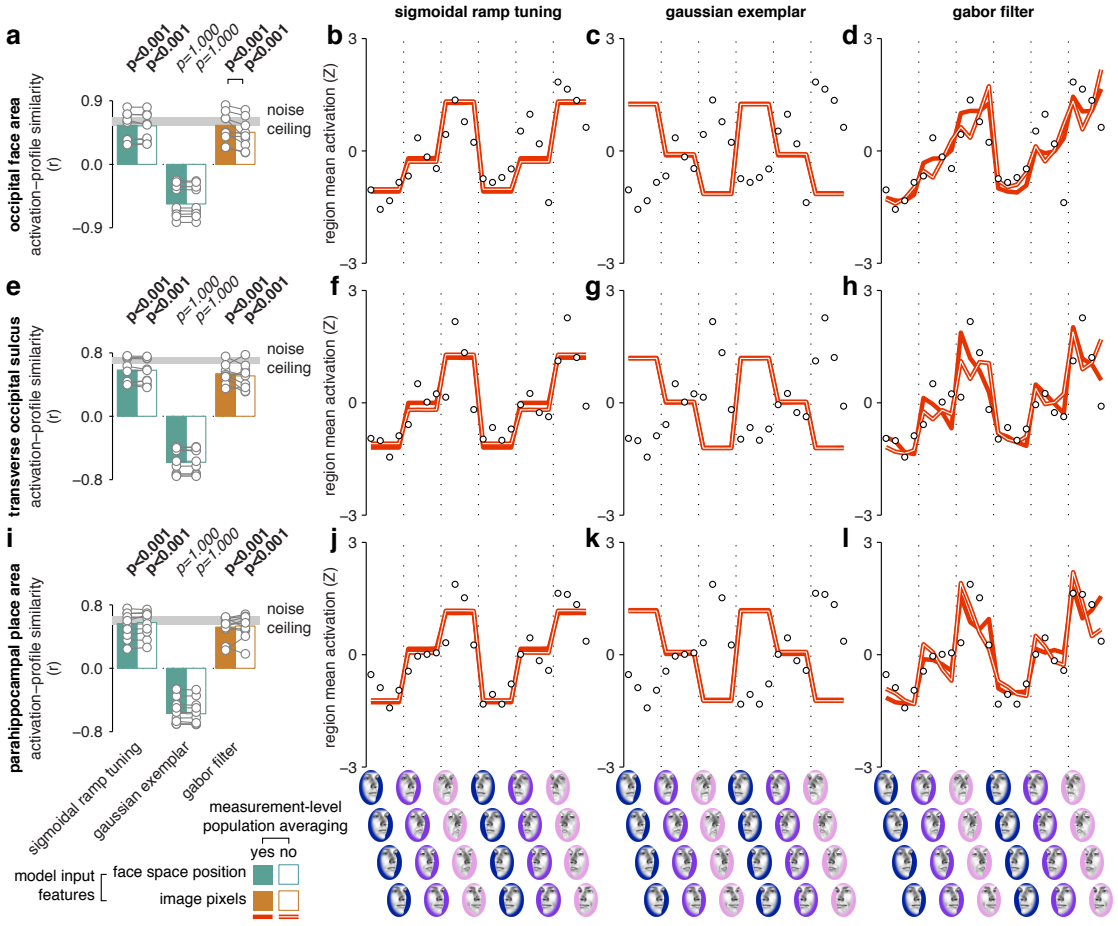
1034 *S4 Figure*. Effect of region-mean removal on cortical face spaces. The top row shows
1035 original distance matrices, while the bottom row shows distance matrices after
1036 removing additive and multiplicative mean pattern effects (Experimental Procedures).
1037 The cited Pearson correlation coefficients are calculated at the group-average level.



1038

1039 *S5 Figure*. Cross-validated distance-matrix generalization performance for

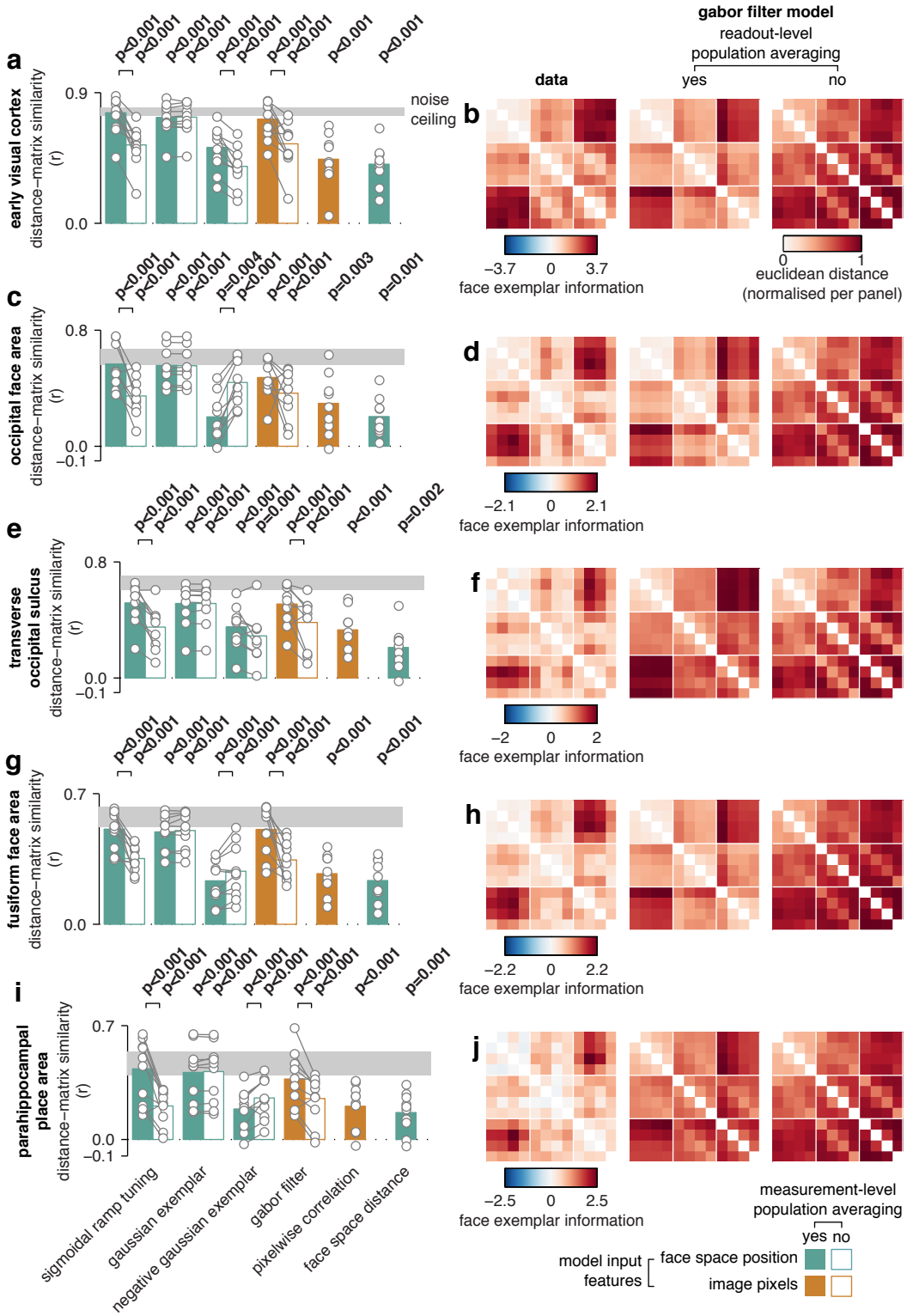
1040 additional cortical regions of interest. Plotted as Figure 5 in main text.



1041

1042 *S6 Figure*. Cross-validated activation-profile generalization performance for

1043 additional cortical regions of interest. Plotted as in Figure 6 in main text.

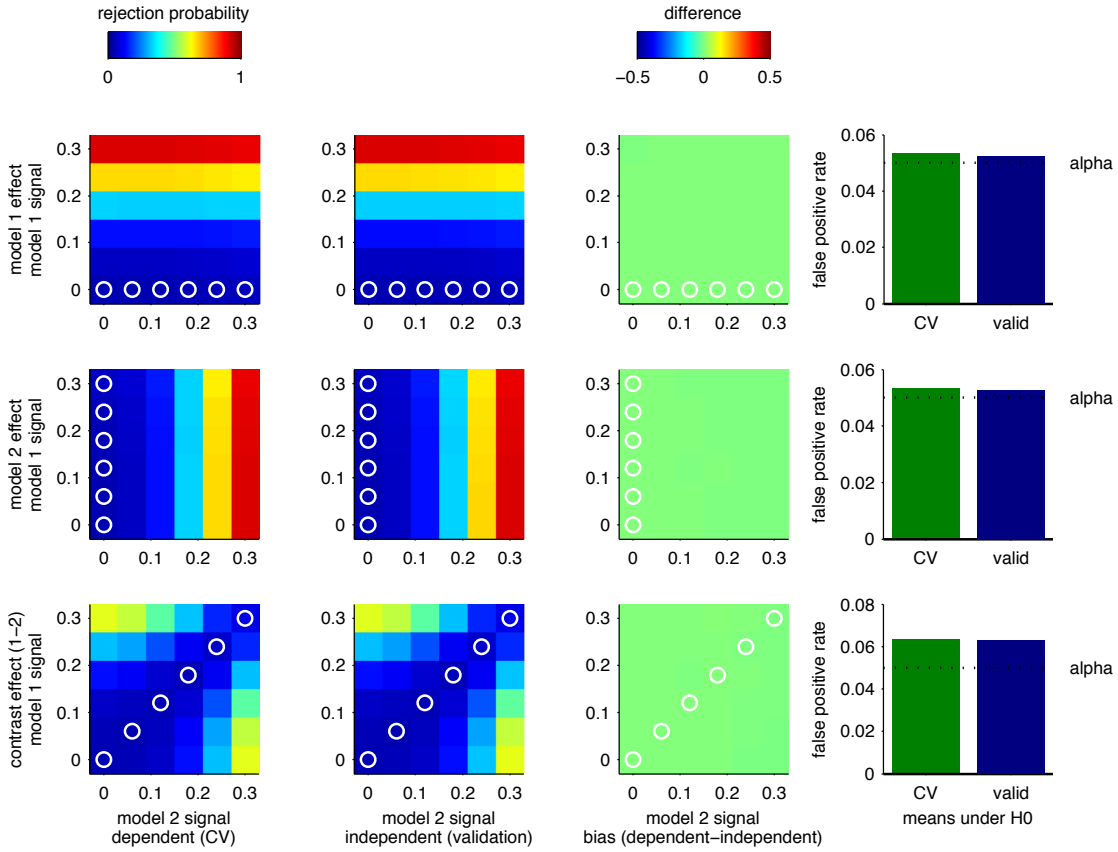


1044

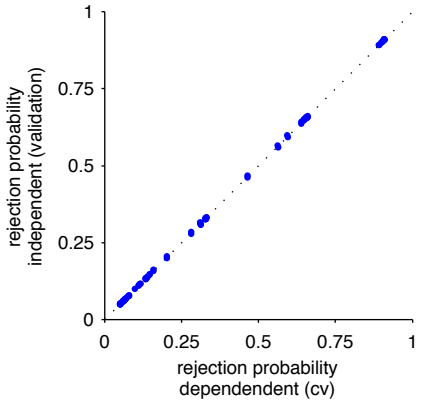
1045 *S7 Figure*. Cross-validated distance-matrix similarity, data distance matrices and

1046 best-fitting predicted matrices for an analysis where the two viewpoints have been

1047 collapsed into a single set of 12 conditions. Plotted as in Figures 1 and 5 in main text.



S8 Figure. Rejection probability simulation. We simulated out-of-sample generalization performance for two arbitrary models, using methods that closely matched the ones described in this manuscript (for details, see S1 Code). We estimated potential bias by comparing cross-validated generalization performance (panels in leftmost column) with generalization to a withheld validation set (left column, subtraction in right column). The probability of rejecting the null hypothesis ($p<0.05$, T test) over 100000 simulations is plotted for tests of either model against zero (one-tailed test, first two rows of panels), and of zero difference between the models' generalization performance (two-tailed test, bottom row). Each color-mapped image shows the rejection probability as a function of signal level for model 1 (vertical axis) and model 2 (horizontal axis). The null hypothesis case is highlighted with white circles. The bars in the rightmost panel summarize the mean rejection probabilities (ie, false positives) for each of these null cases.



1062

1063 *S9 Figure*. Summary plot of the data in Figure S8. Each point represents the mean
1064 rejection probability for a unique set of simulation parameters (values in color-mapped
1065 images in S8 Figure). It can be seen that there is a close to unit relationship between
1066 rejection probability in the dependent, cross-validated case (vertical axis) and the
1067 independent, validation case (horizontal axis).

1068 *S1 Code*. Code to reproduce S8 Figure and S9 Figure. Requires Matlab R2013a.

1069 *S1 Table*. Descriptive and inferential statistics for the distance-matrix correlation
1070 between the face-space models the perceptual and cortical face spaces. We report the
1071 group-average correlation coefficient (mean_r), the group-average Z-transformed
1072 correlation (mean_zr), standard error for the Z-transformed correlation (sterr_zr), one-
1073 tailed p values for the Z-transformed correlation (ppara_zr) and sample sizes (n).
1074 Related to Figure 5.

1075 *S2 Table*. Analysis of variance on parameter estimates from multiple regression RSA
1076 model, with the factors metric (eccentricity, direction), viewpoint (within, across), and a
1077 two-way interaction term. For details, see main text. Related to Figure 2.

1078 *S3 Table*. Descriptive and inferential statistics for the analysis of direction
1079 discriminability as a function of eccentricity level. Mean, standard error (sterr), one-

tailed p values (ppara) and sample sizes (n) are included on separate rows. See also S3 Figure.

S4 Table. Two-tailed parametric p values for all pairwise comparisons between model distance-matrix generalization performances (S1 Table). Related to Figure 5.

S5 Table. Descriptive and inferential statistics for activation-profile similarity analysis. See S1 Table for an account of what the row labels represent. Related to Figure 6.

S6 Table. Two-tailed parametric p values for all pairwise comparisons between activation-profile model fits (S5 Table). Related to Figure 6.

S7 Table. Descriptive and inferential statistics for collapsed-view distance-matrix generalization performances. See S1 Table for an account of what the row labels represent. Related to S7 Figure.

S8 Table. Two-tailed parametric p values for all pairwise comparisons between collapsed-view distance-matrix generalization performances. See S7 Table, S7 Figure.

S1 Movie. Cropped screen capture of the perceptual judgment task as it appeared to participants during data collection.

S2 Movie. Cropped screen capture of the main experiment as it appeared to participants during data collection.