Distributed representations of dynamic facial expressions in the superior temporal sulcus

Christopher P. Said Department of Psychology, Princeton University, Princeton, NJ, USA

Christopher D. Moore Department of Psychology, Princeton University, Princeton, NJ, USA

Andrew D. Engell Department of Psychology, Yale University, New Haven, CT, USA

Alexander Todorov Department of Psychology, Princeton University, Princeton, NJ, USA

James V. Haxby Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

Previous research on the superior temporal sulcus (STS) has shown that it responds more to facial expressions than to neutral faces. Here, we extend our understanding of the STS in two ways. First, using targeted high-resolution fMRI measurements of the lateral cortex and multivoxel pattern analysis, we show that the response to seven categories of dynamic facial expressions can be decoded in both the posterior STS (pSTS) and anterior STS (aSTS). We were also able to decode patterns corresponding to these expressions in the frontal operculum (FO), a structure that has also been shown to respond to facial expressions. Second, we measured the similarity structure of these representations and found that the similarity structure in the pSTS significantly correlated with the perceptual similarity structure of the expressions. This was the case regardless of whether we used pattern classification or more traditional correlation techniques to extract the neural similarity structure. These results suggest that distributed representations in the pSTS could underlie the perception of facial expressions.

Keywords: fMRI, faces, facial expressions, superior temporal sulcus, frontal operculum, multivoxel pattern analysis

Citation: Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision, 10*(5):11, 1–12, http://journalofvision.org/content/10/5/11, doi:10.1167/10.5.11.

Introduction

The superior temporal sulcus (STS) is a functionally heterogeneous region of the cortex believed to be responsible for many different cognitive processes (Allison, Puce, & McCarthy, 2000; Binder et al., 1997; Calder et al., 2007; Grossman & Blake, 2002; Hein & Knight, 2008). Among these, there is growing evidence that the STS is involved in the perception of facial expressions (Adolphs, 2002; Calder & Young, 2005; Engell & Haxby, 2007; Furl, van Rijsbergen, Treves, Friston, & Dolan, 2007; Haxby, Hoffman, & Gobbini, 2000; Ishai, 2008; Narumoto, Okada, Sadato, Fukui, & Yonekura, 2001; Pessoa & Padmala, 2007; Tsuchiya, Kawasaki, Oya, Howard, & Adolphs, 2008). Neurons in macaque STS show preferential responses to different facial expressions (Hasselmo, Rolls, & Baylis, 1989). Consistent with this, fMRI adaptation studies in humans have found expression sensitivity in the anterior STS (Winston, Henson, Fine-Goulden, & Dolan, 2004). Furthermore, the STS responds more strongly to facial expressions than to neutral faces (Engell & Haxby, 2007).

In this fMRI study, we extend our understanding of how the STS represents facial expressions in two ways. First, using targeted high-resolution fMRI measurements of the STS and multivoxel pattern analysis (Haxby et al., 2001; Mur, Bandettini, & Kriegeskorte, 2009; Norman, Polyn, Detre, & Haxby, 2006), we show that that the patterns of activation corresponding to individual facial expressions can be decoded in both the posterior STS (pSTS) and the anterior STS (aSTS). This is the first study to demonstrate decoding of the seven canonical expressions—anger, fear, disgust, happiness, sadness, surprise, and neutral—anywhere in the brain. We also show that decoding is possible in the frontal operculum (FO), a structure that has also been shown to respond more to facial expressions than to neutral faces (Engell & Haxby, 2007). While previous

doi: 10.1167/10.5.11 Received September 19, 2009; published May 7, 2010 ISSN 1534-7362 © ARVO

Downloaded from jov.arvojournals.org on 04/25/2024

studies have shown that the STS responds more to facial expressions than to neutral faces, this approach allows us to demonstrate the existence of reliable patterns of activity corresponding to individual expressions. Additionally, while previous studies have also shown that individual neurons encode different expressions (Hasselmo et al., 1989; Winston et al., 2004), our approach demonstrates that this encoding is distributed in spatial patterns at a resolution detectable by fMRI.

Second, we use pattern analysis to measure the similarity structure of the neural representations of facial expressions in the STS. A similarity structure is the collection of similarity relationships between members of a set. Similarity structures can be used to test theories about perception or to relate neural responses to psychological reports and have therefore recently received much interest in shape and object recognition research (Drucker & Aguirre, 2009; Haushofer, Livingstone, & Kanwisher, 2008; Kriegeskorte, Mur, & Bandettini, 2008; Op de Beeck, Torfs, & Wagemans, 2008; Weber, Thompson-Schill, Osherson, Haxby, & Parsons, 2009). In our case, we are interested in the neural activation patterns corresponding to each of the seven canonical facial expressions, a type of stimulus set which is known to have a non-trivial perceptual similarity structure (Calder, Burton, Miller, Young, & Akamatsu, 2001; Dailey, Cottrell, Padgett, & Adolphs, 2002). One possible outcome of our analysis is that all neural patterns are equidistant from each other. Another possibility is that the similarity relations in the patterns of brain activity resemble the perceptual similarity relations reported by observers of the expressions. That is, pairs of facial expressions that are perceived to be similar will show similar neural patterns in the STS, and pairs of facial expressions that are perceived to be dissimilar will show dissimilar patterns in the STS. Yet another possibility is that the STS would show an irregular similarity structure that is completely different from the perceptual similarity structure.

Our analysis will begin by performing seven-way pattern classification in three regions (the aSTS, the pSTS, and the FO) to demonstrate the existence of spatially distributed patterns corresponding to individual expressions. Next, to address the similarity structure, we obtain behavioral similarity ratings for each pair of emotional expressions and compare them to neural similarity measures extracted from pairwise pattern classification. Finally, to confirm that our similarity structure results are not specific to the classifier, we repeat the analysis using pairwise correlations as our measure of neural similarity, instead of using a classifier.

Methods

Behavioral ratings subjects

Twenty-one subjects (15 females) participated in the perceptual similarity study (mean age = 19.7, *SD* = 2.0).

fMRI subjects

Thirty subjects (sixteen females, mean age = 21.3, *SD* = 3.0) participated in the fMRI study. Data from two subjects were excluded from the analysis for containing a mean of more than 1500 outlier voxels per time point. (Using the default options of the AFNI program 3dToutcount, a voxel is considered an outlier at a particular time point if its absolute deviation from its median value exceeds the median absolute deviation by a set proportion.) An additional three were excluded for scoring less than 65% on a memory task performed in the scanner. Consequently, we analyzed the data from 25 subjects.

Stimuli

We used 3-s video clips of seven different facial expressions (anger, disgust, fear, neutral, sadness, happiness, and surprise). The expressions were posed by six volunteers (three females) who were trained to start with a neutral face and then display the expression. All volunteers wore black shirts, and only the head and shoulders were visible in the footage. Stimuli were produced using a Canon XL1s 3CCD digital video camera and were edited using iMovie (Apple Computer, California). Sample frames from one actress are shown in Figure 1, and a complete example of a movie is shown in the Supplementary Movie.

Behavioral ratings to extract perceptual RSMs

To obtain pairwise similarity ratings of the expressions, we conducted a behavioral experiment using Psychtoolbox

Figure 1. Sample frames from videos of one facial identity. From left to right: Anger, disgust, fear, neutral, sadness, happiness, and surprise.

Downloaded from jov.arvojournals.org on 04/25/2024

for Matlab (Brainard, 1997; Pelli, 1997). Participants were asked to rate the perceived similarity in affect between pairs of expressions. Each participant viewed a series of video pairs and was asked to rate their similarity on a 7-point scale. The two videos in each pair were presented sequentially and were separated by a 200-ms fixation point.

The sequences were counterbalanced in the following ways: (1) Each participant viewed all 42 ordered pairs of expression categories exactly twice. An example of an ordered pair of expression categories would be anger followed by disgust. Stimuli presented in the reverse order would constitute a different ordered pair of expression categories. (2) Across all participants, all possible 1722 ordered pairs of videos were shown. We distinguish between ordered pairs of expression categories (which ignores identity) and ordered pairs of videos (which does not). There are more ordered pairs of videos than ordered pairs of expression categories because each expression category was posed by 6 different actors. This counterbalancing was fully achieved halfway through the experiment on the 21st subject. The last half of this subject's data was excluded, as it would disrupt the counterbalancing. In any case, the results (i.e., the mean pair ratings) before the exclusion were highly correlated with the results after the exclusion ($r = 0.999$).

For each pair of expression categories, similarity ratings were averaged across trials and across subjects. These average similarity ratings were then arranged into a Representational Similarity Matrix (RSM; Figure 5A). A representational similarity matrix is a symmetrical n by n matrix containing the similarity measures between pairs of categories, where n is the number of categories (Kriegeskorte et al., 2008).

In addition to their affect, facial expressions can also be rated on their motion and physical appearance. We conducted another behavioral experiment in which subjects rated the perceived motion similarity between members of a pair of expressions. The design was otherwise the same as the affect-based experiment. The RSM obtained from this experiment was highly correlated with the affect-based RSM ($r = 0.93$). All tests relating this RSM to brain data were nearly identical to tests relating the affect-based RSM to brain data. Because of this redundancy, we only report the results from the affect-based RSM in the main body of the paper and show the results from the motion-based RSM in the Auxiliary information.

fMRI task

In the fMRI scanner, expressions were presented sequentially using Psychtoolbox (Brainard, 1997; Pelli, 1997) with a variable intertrial interval (ITI) of 1 s, 3 s, or 5 s. There were thirty presentations of each expression category for a total of 210 movie presentations. Additionally, thirty 3-s rest periods were interleaved among the trials. In order to encourage attention to the stimuli, subjects performed a memory task: Every 4–5 stimulus presentations a question mark appeared, which was then followed by an additional probe expression. Subjects were asked to indicate whether the probe expression appeared in the previous set of 4–5 expressions, regardless of identity. The probe expressions were then followed by a written instruction displayed for 2 s that a new group of expressions would soon appear. In the fMRI data analysis, probes were included in the General Linear Model (GLM) but were not included in pattern classification or neural similarity analyses. Subjects performed the memory task at a mean hit rate of 0.79 ($SD = 0.07$).

To increase subject comfort, the experiment was divided into six "runs" of 6 min and 50 s each. The number of probes, rests, and expression categories was balanced across runs 1–2, runs 3–4, and runs 5–6. Additionally, the order of expression categories was 1-back counterbalanced so that each category was preceded by every other category (including itself) at an equal number of times. Each category also followed probes at an equal number of times. The ITIs were also balanced so that each possible transition between categories was delayed by each possible ITI at an equal number of times. Together these precautions helped ensure that the BOLD responses to particular expression categories were not biased by their position in the experiment or by preceding stimuli, which may carry delayed hemodynamic responses.

fMRI data acquisition

The blood oxygenation level-dependent (BOLD) signal was used as a measure of neural activation (Kwong et al., 1992; Ogawa, Lee, Nayak, & Glynn, 1990). Echo planar images (EPI) were acquired with a Siemens 3.0 Tesla Allegra scanner (Siemens, Erlangen, Germany) and a Nova Medical head transmit coil with receive-only bitemporal array coils (Nova Medical, Wilmington, MA). The receive-only array coils were positioned directly on the lateral surface of the head to achieve high signal-to-noise ratio (SNR) from a local area comprising the STS and FO. Data were obtained at high resolution ($1.5 \times 1.5 \times 1.6$ mm^3) with an interslice gap of 0.36 mm (TR = 2000 ms, TE = 40 ms, flip angle = 76°; matrix size = 192×192). Slices were angled obliquely to be parallel to the Sylvian fissure and centered at the midpoint of the thalamus. Twenty slices were obtained during each TR, allowing coverage of the STS and FO, but not the superior or inferior areas of cortex. Medial parts of the brain were also captured by the slices but achieved only poor SNR due to the use of surface coils. A whole brain high-resolution T1-weighted structural scan was acquired at the beginning of each experiment (TR = 2500 ms, TE = 33 ms, flip angle = 8°, matrix size = 256×256) to permit anatomical localization of regions of interest (ROIs).

Downloaded from jov.arvojournals.org on 04/25/2024

fMRI data preprocessing

All fMRI analysis steps were determined beforehand and informed by our analysis of pilot data. Image analysis was performed with AFNI (Cox, 1996). After discarding the first five functional images from each run to allow the MR signal to reach steady-state equilibrium, the remaining images were slice time-corrected and then motion-corrected to the first image of the first run using a 6-parameter 3-D motion correction algorithm. Transient spikes in the signal were suppressed with the AFNI program 3dDespike. The most superior and most inferior slices were removed to protect against motion artifacts arising from slices shifting into non-excited brain regions. A 3-mm full-width at half-maximum (FWHM) smoothing kernel was then applied to the images before conversion to percent signal change from the mean.

Anatomical masks were created by transforming the T1-weighted images into a standard space using the @auto_tlrc program to match the TT_N27 template (Holmes et al., 1998). The STS was hand drawn on coronal slices (Figure 2) from −64.5 mm to −1.5 mm posterior to the anterior commisure and then separated into posterior and anterior sections at the −33 mm mark. Since the FO is less distinct than the STS, we defined the right FO as a sphere of radius 20 mm centered on a point ($x = 45$, $y = 15$, $z = 22$). This point was the center of a cluster defined by a facial expression versus neutral contrast in a previous experiment (Engell & Haxby, 2007). The left FO was defined as the same sphere reflected across the midline. All ROIs were defined *a priori,* and no other ROIs were considered. This hypothesis-based approach allows for focused tests and avoids the issues of circularity that arise when ROIs are functionally defined by the same data that are subsequently analyzed. All anatomical masks were then transformed back into original space, multiplied by a separate mask of brain versus non-brain areas, and then applied to the functional data. Because of limited functional coverage, most anatomical masks were truncated. Sample masks for one subject are shown in Figure 2, and the average volumes of each mask across subjects are reported in Table 1.

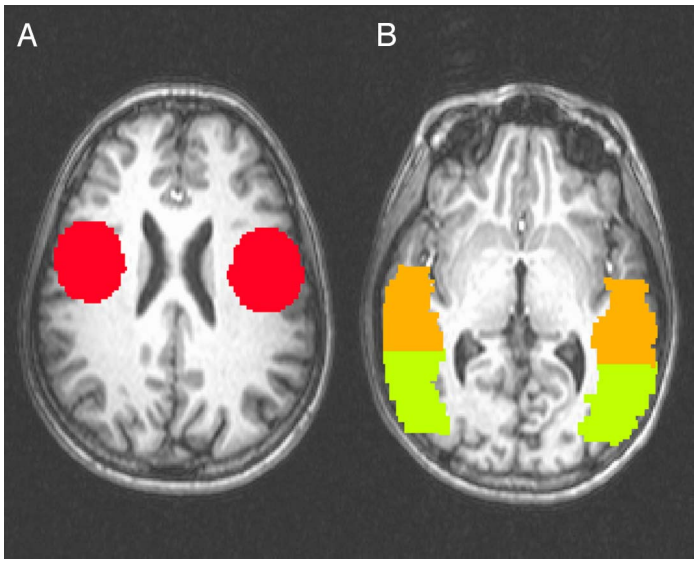

Figure 2. Example of ROIs from one subject. (A) Axial slice at z = 24.5 mm showing the spheres placed over the FO. These spheres were typically truncated at more superior slices because of limited coverage imposed by high-resolution fMRI. (B) Axial slice at z = 1.5 mm showing the aSTS (orange) and pSTS (green) ROIs. The slices shown here are in original space, but the coordinates are provided using the Talairach system.

	Number of voxels	Volume (cm^3)
pSTS	8201 (1594)	36.1 (7.0)
aSTS	6712 (1145)	29.6 (5.0)
FO	5242 (2165)	23.1 (9.5)

Table 1. ROI sizes averaged across subjects. The standard deviations are in parentheses.

Removal of identity effects

Neurons in monkey and human STS are sensitive to both facial expression and facial identity (Calder & Young, 2005; Hasselmo et al., 1989; Winston et al., 2004). Thus, analysis of facial expression is made difficult because of extra variance due to uncontrolled effects of facial identity. We removed the effects of facial identity by modeling the BOLD time series with a General Linear Model (GLM) consisting of nine-parameter tent function expansion regressors for each of the six identities. In addition, the model included two gamma-variate convolved regressors for correct response to probe (yes or no), six regressors for the motion parameters extracted during volume correction, and three regressors for linear, quadratic, and cubic signal drifts during each run. Collectively, these regressors removed variance caused by regressors of no interest. For each ROI, the average time interval between stimulus onset and peak hemodynamic response was noted, and the residual time series for each voxel was saved for further analysis.

Next, we imported the data into Matlab and averaged the residual value at the peak of the response for each trial with the residual values for the immediately prior and immediately adjacent time points. This provided us with one data value per trial per voxel, which could then be associated with an expression label for classification purposes. Pilot data showed that this time point averaging approach gave marginally better results than using the coefficients from a regression that assumed a gamma-variate BOLD response.

From this point onward, the analysis diverges into three separate paths (Figure 3). These paths are briefly summarized in this section, and then described more comprehensively in later sections. In the first path, we use seven-way

Downloaded from jov.arvojournals.org on 04/25/2024

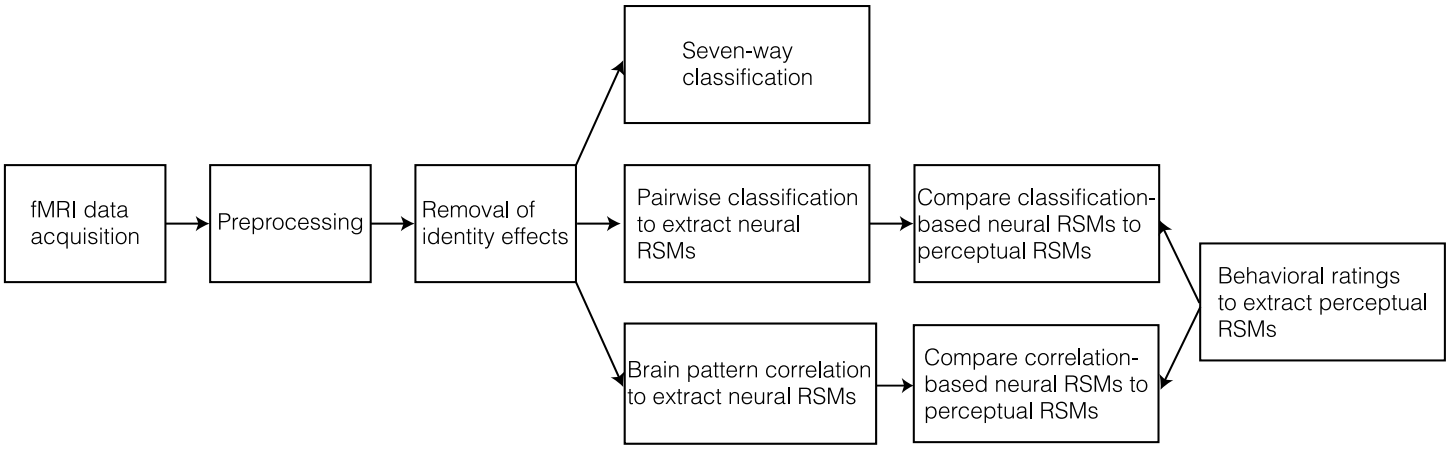

Figure 3. Flowchart of analysis steps. For more detailed information, refer to the Methods section.

classification at the single trial level to test the hypothesis that the aSTS, pSTS, and FO contain information about facial expressions. In the second path, we test whether the perceptual similarity structure of facial expressions is also contained in the neural representations within these ROIs. To do this, we use pairwise classification to extract a neural RSM for each ROI, and then compare these RSMs to perceptual RSM. Pairwise classification, as opposed to multiclass classification, allows us to directly measure the similarity between pairs of expressions in a way that is unbiased by the presence of competing categories during classification (see below). In the third path, we use a more traditional technique (brain pattern correlations) to confirm that the RSM relations we found were not specific to the classification approach. This approach examines the similarity structure of the entire ROI instead of a small subset of voxels. The two methods for measuring neural similarity (classification and correlation) provide complementary pieces of evidence. Classification, with its greater power, can detect information that simpler techniques might miss. The correlation approach weighs all voxels equally and provides assurance that our results are not specific to the classifier.

Seven-way classification

To test for the existence of distinct representations of facial expressions in the pSTS, aSTS, and FO, we first used the Sparse Multinomial Logistic Regression (SMLR; Krishnapuram, Carin, Figueiredo, & Hartemink, 2005) as implemented in the Princeton MVPA Toolbox to perform seven-way classification (Detre et al., 2006). SMLR is based on logistic regression but uses a sparsity-promoting Laplacian prior λ to select only a small set of relevant voxels during training. Classification was cross-validated by run, so that testing for each run was performed using weights trained only on the remaining runs. Based on the results from pilot subjects, we used a λ value of 0.01. Classification was performed on each ROI individually, and the hit rates were then averaged across laterality and across subjects. Of the 50 lateralized pSTS ROIs, the data from one were excluded because the ROI size was less than 1000 voxels (3600 mm^3) due to limited coverage. Seven of the 50 FO ROIs were excluded using the same cutoff. To ensure that our results were not driven entirely by the neutral category, we repeated the procedure with only the six non-neutral expressions.

Pairwise classification to extract neural RSMs

To extract the neural similarity between pairs of facial expressions, we performed 21 binary classification tests. Measures of neural similarity are related to classification because categories that are more similar should be more difficult for a classifier to separate. Binary classification allows us to directly measure the neural similarity between each of the 21 pairs of expressions categories, without bias from other facial expressions that might compete during classification. To illustrate this point, consider two expressions A and B that are neurally distant from each other. These expressions should be easy to classify in a multiclass problem in which all of the remaining expressions are also distant from A and B. In this case, we would correctly infer that A and B are distant from each other. Now consider two expressions C and D that are also neurally very distant from each other, but now assume that a third expression E is very similar to D. A multiclass classifier may have difficulty distinguishing C from D and we may wrongly infer that the two expressions are neurally similar. While this sort of multiexpression confusion is relevant for some types of analyses, it is not relevant to the present analysis, in which we compare neural similarity to a behavioral similarity measure that directly measures the similarity of pairs of expressions.

Like standard logistic regression, binomial SMLR uses maximum likelihood estimation to relate the predictor

Downloaded from jov.arvojournals.org on 04/25/2024

variables (voxels) to a logistic transformation of the binomial category probability known as the logit. For each trial, the logit provides a sensitive measurement of the degree to which the classifier leans toward one category over another. Unlike the probability value, which is a sigmoidal transformation of the logit, the logit is not susceptible to ceiling and floor effects. On trials where the correct category is coded as 1, the classifier should predict a positive logit value, and when the correct category is coded as 0, the classifier should predict a negative logit value. We define the *logit distance* as the negative of the logit on trials when the correct category is coded as 0, and the logit itself on trials when the correct category is coded as 1. For each of the 21 binary classification tests, we took the logit distance on each trial for each ROI, and then averaged across trials, laterality, and subjects. To transform this measure into a *logit similarity,* we multiplied all values by −1. In cases where both the left and right ROIs passed the volume cutoff, we averaged the two ROIs. If only one ROI passed the cutoff, we used only that one. For each expression pair and each ROI, logit similarities were averaged across trials and across subjects. These averaged logit similarities were then arranged as a neural RSM for each ROI. As will be described in a later section, these RSMs were then compared to the perceptual RSM.

Brain pattern correlation to extract neural RSMs

It is possible that the similarity relations determined by SMLR could be specific to the classifier, or only present in the voxels chosen by SMLR's embedded feature selection. To confirm that our results were general across the STS, and not specific to our classifier, we measured the correlations between conditions from the entire set of voxels for each ROI, without using classification. If the correlation approach and the classifier approach yield consistent results, we can be confident that our results are not specific to our classifier. The procedure is given as follows: First, for each subject, we started with the same residual data that we submitted for classification. As described above, this consisted of the near-peak residual brain response to a trial after identity effects had been modeled, averaging across immediately adjacent time points. Next, for each subject, we averaged across trials within expression categories and reshaped the data for each category into vector form. We then computed the correlation between each pair of vectors and averaged across subjects. These average correlations were then arranged as a neural RSM for each ROI.

The two methods for measuring neural similarity (classification and correlation) provide complementary pieces of evidence. Classification, with its greater power, can detect information that simpler techniques might miss. The correlation approach weighs all voxels equally and provides assurance that our results are not specific to the classifier.

Comparing neural RSMs to the perceptual RSM

At this point in the analysis, there is a neural RSM for each ROI, as well as a perceptual RSM. To test whether a brain region represents facial expressions in a way that is consistent with perceptual reports, we put each neural RSM in vector form and correlated them with the perceptual RSM in vector form. We refer to these correlations as the *group r values*.

For significance testing, we compared each fMRI subject's individual neural RSMs to the perceptual RSM. We then tested whether these values, which we refer to as the *individual r values,* were significantly different from zero. To test whether the individual r values were different from zero, we did not assume that the correlations under the null hypothesis were normally distributed, and therefore used the Wilcoxon signed rank test, which makes no assumptions of normality and is generally more conservative than a t-test.

Finally, it is worth noting that the neutral category is qualitatively very different from the remaining categories as it involves only minimal motion. As a result, it is possible that a brain area that can discriminate static stimuli from moving stimuli might present an RSM, which partially correlates with the perceptual RSM, even if the brain area cannot discriminate among the non-neutral categories. Therefore, we report all possible tests on the RSMs both with and without the neutral category included.

Results

Seven-way classification

We conducted a seven-way classification test using SMLR to determine if information about the seven expression categories was present in the bilateral pSTS, aSTS, and FO. Chance performance was approximately 14.2% for all ROIs, as determined by a simulation in which we permuted the labels 50 times for each subject and averaged the hit rate across all simulations. Across subjects, the overall hit rate in the pSTS was significant (mean = 22.6%, $t(24) = 12.0$, $p < 0.001$), as was the hit rate in the aSTS (mean = 19.8%, $t(24) = 9.5$, $p < 0.001$) and the FO (mean = 18.4%, $t(24) = 11.6$, $p < 0.001$). These results are plotted in Figure 4A.[1] The hit rates and false positive rates for each expression are reported in a confusion matrix in the Auxiliary information. In both the pSTS and aSTS, the hit rates were significant for each expression except sadness.

Downloaded from jov.arvojournals.org on 04/25/2024

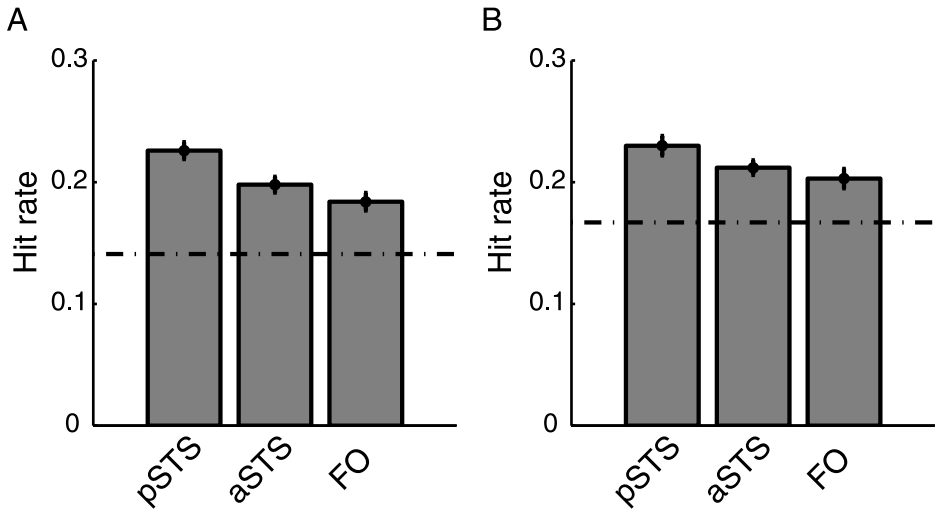

Figure 4. Hit rates in each ROI. The dotted lines represent chance performance, as determined by repeated classification tests on permuted data. (A) Seven-way classification. (B) Six-way classification, where the neutral expression is excluded from the analysis. Error bars represent standard error of the mean.

In the FO, the hit rate was significant for each expression except sadness, disgust, and surprise. In each bilateral ROI, the neutral category had the highest hit rate. To confirm that the overall hit rate was not driven entirely by the neutral category, we reran the classification with only the six non-neutral categories and plot the results in Figure 4B. Here, chance performance was determined by permuted label simulation to be 16.6%. As before, the hit rate in the pSTS was significant (mean = 22.6%, $t(24) = 11.4$, $p < 0.001$), as was the hit rate in the aSTS (mean = 21.2%, $t(24) = 12.5$, $p < 0.001$) and the FO (mean = 20.3%, $t(24) = 7.5$, $p < 0.001$).

The overall hit rate in the pSTS was numerically higher than the hit rates in the FO and in the aSTS. However, since at least part of this difference was driven by a larger number of voxels in the pSTS compared to the other ROIs (Table 1) and by potential regional differences in SNR, we did not test for statistically significant differences between the ROIs.

Representational similarity matrices based on pairwise classification

We ran 21 binary classification tests (one for each expression pair) to directly measure the neural similarity between pairs of categories. Chance performance was 50%. The hit rate in the pSTS was significant (mean = 57.4%, $t(24) = 10.0$, $p < 0.001$), as was the hit rate in the aSTS (mean = 55.7%, $t(24) = 9.8$, $p < 0.001$) and the FO (mean = 53.9%, $t(24) = 5.2$, $p < 0.001$).

Importantly, the average logit similarities (see Methods section) can be arranged into a neural representational similarity matrix (RSM). An RSM is a symmetrical n by n matrix containing the similarity measures between pairs of categories, where n is the number of categories (Kriegeskorte et al., 2008). The neural RSM can then be compared to a perceptual RSM determined by the subjective ratings of a separate group of subjects. There was a significant correlation between the pSTS neural RSM and the perceptual RSM (group $r = 0.49$, average individual $r = 0.23$, $p < 0.01$). When the neutral category was excluded from the analysis, the relationship was still significant (group $r = 0.36$, average individual $r = 0.15$, $p < 0.01$). RSMs are shown in Figure 5.

In the aSTS, the relationship between the neural RSMs and the perceptual RSM was not significant (group $r = 0.33$, average individual $r = 0.09$, $p > 0.05$). Similarly, the relationship was not significant in the FO (group $r = 0.28$, average individual $r = 0.08$, $p > 0.05$). Neither of these tests were significant when the neutral category was excluded.

It is also possible to measure the relationship between neural RSMs from different ROIs. The correlation between the pSTS and the aSTS RSMs was high both when the neutral category was included (group $r = 0.92$, average individual $r = 0.60$, $p < 0.01$) and when it was excluded (group $r = 0.83$, average individual $r = 0.52$, $p < 0.01$), although these measures are somewhat inflated by smoothing across adjacent regions. There were also strong correlations between the RSMs of non-adjacent regions. The correlation between the pSTS and the FO was significant both when the neutral category was included (group $r = 0.73$, average individual $r = 0.41$, $p < 0.01$) and when it was excluded (group $r = 0.44$, average individual $r = 0.33$, $p < 0.01$). Similarly, the correlation between the aSTS and the FO was significant both when the neutral category was included (group $r = 0.84$, average individual $r = 0.40$, $p < 0.01$) and when it was excluded (group $r = 0.63$, average individual $r = 0.36$, $p < 0.01$). In all cases, the correlation between the RSMs from any pair of two

Downloaded from jov.arvojournals.org on 04/25/2024

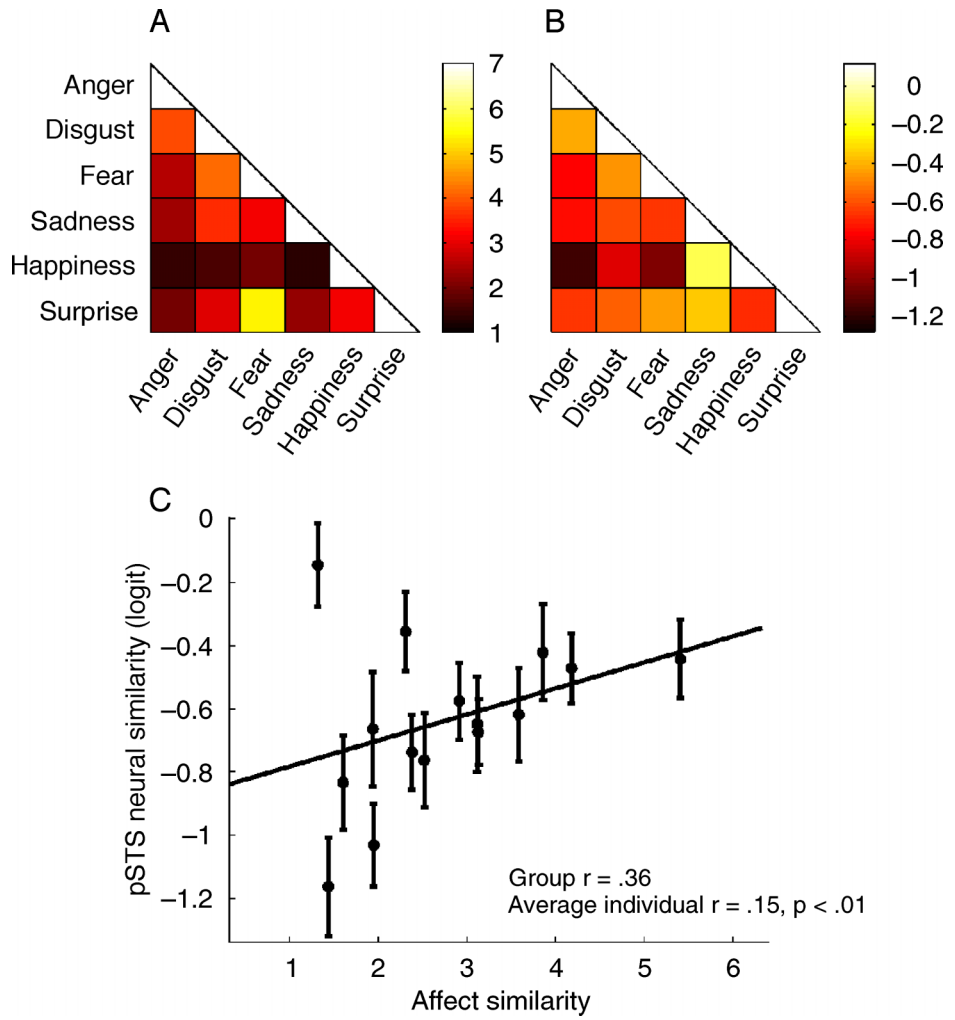

Figure 5. Comparison of the perceptual similarity structure to the neural similarity structure in the pSTS. (A) Perceptual Representational Similarity Matrix (RSM), excluding neutral. The color of each cell indicates the average similarity rating of the pair. Units are on a 7-point Likert scale. (B) Neural RSM in the pSTS, excluding neutral. The color of each cell indicates the average logit similarity of the pair, as described in the Methods section. (C) A scatter plot of pSTS neural similarity against perceptual affect-based similarity, excluding neutral, where each circle represents a pair of expressions. The line is a least-squares fit for all pairs. Error bars represent the standard error of the mean across subjects.

regions was significantly higher than the correlation between the RSM of either of the regions and the perceptual RSM.

To further investigate the role of the pSTS and the variance shared between ROIs, we measured the partial correlation between the pSTS RSM and the perceptual RSM, after the shared variance from other ROIs was removed. This was performed by computing the residuals when the pSTS RSM was regressed on the RSM of a different ROI, and then correlating those residuals with the perceptual RSM. When all emotions were included in the analysis, the partial correlation remained significant when the variance from the FO RSM was removed (group r = 0.40, average individual $r = 0.18$, $p < 0.05$). Similarly, the partial correlation remained significant when the variance from the aSTS RSM was removed (group r = 0.43, average individual $r = 0.21$, $p < 0.05$).

In contrast, the partial correlation between the FO RSM and the perceptual RSM remained non-significant when the variance from the pSTS RSM was removed (group r = 0.07, average individual r = 0, $p > 0.05$). Similarly, the partial correlation between the aSTS RSM and the perceptual RSM remained non-significant when the variance from the pSTS RSM was removed (group r = −0.34, average individual r = −0.07, $p > 0.05$). The significance decisions of all of these results were the same regardless of whether the neutral category was included in the analysis.

RSMs based on correlation

To ensure that our results were not specific to the SMLR classifier, we created new neural RSMs that were based on

Downloaded from jov.arvojournals.org on 04/25/2024

the correlations between patterns of activity for each pair of facial expressions. Using this method, we again found that the pSTS neural RSM significantly correlated with the perceptual RSM (group $r = 0.46$, average individual $r = 0.17$, $p < 0.01$). When the neutral expression was excluded from the analysis, the results were still significant (group $r = 0.40$, average individual $r = 0.14$, $p < 0.01$).

In the aSTS, the relationship between the neural RSM and the perceptual RSM was significant (group $r = 0.26$, average individual $r = 0.06$, $p < 0.05$). The relationship was also significant in the FO (group $r = 0.27$, average individual $r = 0.08$, $p < 0.05$). However, neither of these relationships were significant when the neutral category was excluded.

Discussion

Using pattern classification, we were able decode the seven canonical facial expressions in the pSTS, aSTS, and FO. Whereas previous work has shown that these areas are activated by perception of facial expressions, our results demonstrate that the distinctions among facial expressions are represented in the patterns of response, presumably reflecting distributed population encoding for facial expressions within these three regions. We also showed that in the pSTS, the neural similarity structure significantly correlated with the perceptual similarity structure. We found this to be the case regardless of whether the neutral expression was included, and regardless of whether we extracted the neural similarity structure with pattern classification or with a more traditional correlation technique. It was also the case regardless of whether we used a perceptual RSM based on the affect of the expressions or the physical appearance of the expressions (see Auxiliary information). Furthermore, the partial correlation between the pSTS RSM and the perceptual RSM remained significant after removing the variance of either the FO RSM or the variance of the aSTS RSM. This provides a tighter link between the pSTS and the perception of facial expressions than has been demonstrated before.

When the neutral category was excluded, no significant correlations were found between the perceptual similarity structures and the neural similarity structures in the aSTS and FO. These non-significant correlations do not indicate that the perceptual similarity structure is reflected only in the pSTS. Non-significant correlations in the aSTS and FO may also be due to less statistical power because of a smaller number of voxels (see Auxiliary Table) or a lower SNR as a result of surface coil placement. However, in the seven-way classification analysis, overall hit rates were highest in the pSTS even when we controlled for the number of voxels. It is unknown if differences in SNR were responsible for the remaining difference.

The distributed representation of facial expressions may also extend beyond the STS and FO. In fact, one recent study demonstrated successful decoding of happy and fearful faces using intracranial EEG recorded from the surface of human ventral temporal cortex that was superior to decoding from data collected at lateral temporal sites (Tsuchiya et al., 2008). However, this comparison is tempered by the fact that cortical surface electrodes are less well suited to recording activity from within a sulcus (e.g., the pSTS) than from the surface of a gyrus (e.g., the fusiform gyrus) due to proximity of the neural source to the recording site and to the orientation of the evoked dipole. In our study, we did not measure the fusiform area due to the limited slice coverage necessitated by high-resolution fMRI. Future whole brain fMRI experiments will be necessary to directly compare decoding in these areas.

The correlations between ROI RSMs were high and, in fact, were higher than the correlations between any ROI RSM and the perceptual RSM. The shared variance between ROIs could come from many sources. It is instructive to look at the partial correlation between the pSTS RSM and the perceptual RSM after the variance from either the FO RSM or the aSTS RSM was removed. We found that the correlation remained significant after the removal of variance from other ROIs. This suggests that the variance shared between ROIs is not strongly related to the perceptual RSM. Instead, it could reflect shared noise, or it could perhaps reflect perceptual representations that differ from the one measured here.

The frontal operculum

In addition to the pSTS and aSTS, successful decoding of the facial expressions was achieved using the response from the FO. One possibility for the role of the FO is that it serves as the control region to enforce category distinctions in the STS. There is evidence that the ventrolateral prefrontal cortex, which overlaps with FO, exerts top down control on the temporal lobes by selecting conceptual information during semantic memory tasks (Martin, 2007).

Another possibility is that the response in FO reflects activation of the mirror neuron system. Mirror neurons fire both when an action is observed and when the same action is produced (Ferrari, Gallese, Rizzolatti, & Fogassi, 2003; Montgomery, Seeherman, & Haxby, 2009; Montgomery, Isenberg, & Haxby, 2007; Nakamura et al., 1999; Rizzolatti & Craighero, 2004). While it is important not to overstate the importance or prevalence of mirror neurons (Dinstein, Thomas, Behrmann, & Heeger, 2008), there is reason to believe that they could be involved in our results. Mirror neurons are especially concentrated in monkey area F5, which is believed to be homologous to area FO in humans (Johnson-Frey et al., 2003). Facial expressions are particularly well suited for the mirror neuron hypothesis, since it is known that humans mimic the facial expression of

Downloaded from jov.arvojournals.org on 04/25/2024

people they are interacting with (Buck, 1984) and produce microexpressions when simply looking at expressive face images (Dimberg, 1982; Dimberg, Thunberg, & Elmehed, 2000). It is thus possible that that the activity in FO related to specific facial expressions may be due to mirror neurons, which fire upon the perception of expressions and which might also drive microexpression production in response.

Conclusions

In summary, we showed that information about facial expressions is present in distributed patterns throughout the pSTS, aSTS, and FO, and that these patterns can be decoded at the resolution of fMRI. In the pSTS, the similarity structure of the representations significantly correlated with the perceptual similarity structure, suggesting that the pSTS is an important neural region for the perception of facial expressions.

Acknowledgments

We would like to thank Ken Norman, Per Sederberg, Greg Detre, Francisco Pereira, and Joe McGuire for useful discussions. We would also like to thank three anonymous reviewers for their advice.

Commercial relationships: none.
Corresponding author: Christopher P. Said.
Email: csaid@princeton.edu.
Address: Psychology Department, Princeton University, Princeton, NJ 08540, USA.

Footnote

[1]Collectively, the pSTS, aSTS, and FO cover nearly all of the gray matter in high SNR regions of our acquisition slab. Thus, it is not possible to make a fair comparison to hit rates in other brain regions. Nevertheless, low SNR brain regions can be used to demonstrate that the hit rates in our target ROIs are not classifier artifacts. To test this, we selected 1000 voxels from the frontal region $y > 0$, $-10 < x < 10$ using LPI coordinates in Talairach space. Here, the hit rate was 16.7%, which was significantly below the hit rates of our target ROIs ($p < 0.05$ for all tests).

References

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology, 12,* 169–177. [PubMed]

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences, 4,* 267–278. [PubMed]

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience, 17,* 353–362. [PubMed]

Braddick, O. J., O'Brien, J. M., Wattam-Bell, J., Atkinson, J., & Turner, R. (2000). Form and motion coherence activate independent, but not dorsal/ventral segregated, networks in the human brain. *Current Biology, 10,* 731–734. [PubMed]

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. [PubMed]

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology, 46,* 369–384. [PubMed]

Buck, R. (1984). *The communication of emotion*. New York: Guilford.

Calder, A. J., Beaver, J. D., Winston, J. S., Dolan, R. J., Jenkins, R., Eger, E., et al. (2007). Separate coding of different gaze directions in the superior temporal sulcus and inferior parietal lobule. *Current Biology, 17,* 20–25. [PubMed] [Article]

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research, 41,* 1179–1208. [PubMed]

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews on Neuroscience, 6,* 641–651. [PubMed]

Cox, R. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29,* 162–173. [PubMed]

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience, 14,* 1158–1173. [PubMed]

Detre, G., Polyn, S., Moore, C., Natu, V., Singer, B., Cohen, J., et al. (2006). *The multi-voxel pattern analysis (MVPA) toolbox*. Paper presented at the Organization for Human Brain Mapping Annual Meeting, Florence, Italy.

Dimberg, U. (1982). Facial reactions to facial expressions. *Psychophysiology, 19,* 643–647.

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial

Downloaded from jov.arvojournals.org on 04/25/2024

expressions. *Psychological Science, 11,* 86–89. [PubMed]

Dinstein, I., Thomas, C., Behrmann, M., & Heeger, D. J. (2008). A mirror up to nature. *Current Biology, 18,* R13–R18. [PubMed]

Drucker, D. M., & Aguirre, G. K. (2009). Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cerebral Cortex, 19,* 2269–2280. [PubMed]

Engell, A. D., & Haxby, J. V. (2007). Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia, 45,* 3234–3241. [PubMed]

Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience, 17,* 1703–1714. [PubMed]

Furl, N., van Rijsbergen, N., Treves, A., Friston, K. J., & Dolan, R. J. (2007). Experience-dependent coding of facial expression in superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 13485–13489. [PubMed]

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron, 35,* 1167–1175. [PubMed]

Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual-cortex of the monkey. *Behavioural Brain Research, 32,* 203–218. [PubMed]

Haushofer, J., Livingstone, M. S., & Kanwisher, N. (2008). Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *Plos Biology, 6,* 1459–1467. [PubMed]

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

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4,* 223–233. [PubMed]

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus—It's my area: Or is it? *Journal of Cognitive Neuroscience, 20,* 2125–2136. [PubMed]

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., & Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *Journal of Computer Assisted Tomography, 22,* 324–333. [PubMed]

Ishai, A. (2008). Let's face it: It's a cortical network. *Neuroimage, 40,* 415–419. [PubMed]

Johnson-Frey, S. H., Maloof, F. R., Newman-Norlund, R., Farrer, C., Inati, S., & Grafton, S. T. (2003). Actions or hand–object interactions? Human inferior frontal cortex and action observation. *Neuron, 39,* 1053–1058. [PubMed]

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Front System Neuroscience, 2,* 4. [PubMed]

Krishnapuram, B., Carin, L., Figueiredo, M. A. T., & Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27,* 957–968. [PubMed]

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., et al. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences of the United States of America, 89,* 5675–5679. [PubMed]

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58,* 25–45. [PubMed]

Montgomery, K. J., Seeherman, K. R., & Haxby, J. V. (2009). The well-tempered social brain. *Psychological Science, 20,* 1211–1213. [PubMed]

Montgomery, K. J., Isenberg, N., & Haxby, J. V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Social Cognitive and Affective Neuroscience, 2,* 114–122. [PubMed]

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—An introductory guide. *Social Cognitive and Affective Neuroscience, 4,* 101–109. [PubMed]

Nakamura, K., Kawashima, R., Ito, K., Sugiura, M., Kato, T., Nakamura, A., et al. (1999). Activation of the right inferior frontal cortex during assessment of facial emotion. *Journal of Neurophysiology, 82,* 1610–1614. [PubMed]

Narumoto, J., Okada, T., Sadato, N., Fukui, K., & Yonekura, Y. (2001). Attention to emotion modulates fMRI activity in human right superior temporal sulcus. *Brain Research and Cognitive Brain Research, 12,* 225–231. [PubMed]

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10,* 424–430. [PubMed]

Ogawa, S., Lee, T. M., Nayak, A. S., & Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnitude Reasoning Medical, 14,* 68–78. [PubMed]

Downloaded from jov.arvojournals.org on 04/25/2024

Op de Beeck, H. P., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience, 28,* 10111–10123. [PubMed]

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. [PubMed]

Pessoa, L., & Padmala, S. (2007). Decoding near-threshold perception of fear from distributed single-trial brain activation. *Cerebral Cortex, 17,* 691–701. [PubMed]

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., et al. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society B: Biological Sciences, 265,* 1809–1817. [PubMed]

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192. [PubMed]

Tsuchiya, N., Kawasaki, H., Oya, H., Howard, M. A., 3rd, & Adolphs, R. (2008). Decoding face information in time, frequency and space from direct intracranial recordings of the human brain. *PLoS One, 3,* e3892. [PubMed]

Weber, M., Thompson-Schill, S. L., Osherson, D., Haxby, J., & Parsons, L. (2009). Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia, 47,* 859–868. [PubMed]

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

Downloaded from jov.arvojournals.org on 04/25/2024