Published in final edited form as:

Neuroimage. 2016 April 15; 130: 77–90. doi:10.1016/j.neuroimage.2016.01.045.

Face-selective regions differ in their ability to classify facial expressions

Hui Zhang, Shruti Japee, Rachel Nolan, Carlton Chu, Ning Liu, and Leslie G Ungerleider

Abstract

Recognition of facial expressions is crucial for effective social interactions. Yet, the extent to which the various face-selective regions in the human brain classify different facial expressions remains unclear. We used functional magnetic resonance imaging (fMRI) and support vector machine pattern classification analysis to determine how well face-selective brain regions are able to decode different categories of facial expression. Subjects participated in a slow event-related fMRI experiment in which they were shown 32 face pictures, portraying four different expressions: neutral, fearful, angry, and happy and belonging to eight different identities. Our results showed that only the amygdala and the posterior superior temporal sulcus (STS) were able to accurately discriminate between these expressions, albeit in different ways: The amygdala discriminated fearful faces from non-fearful faces, whereas STS discriminated neutral from emotional (fearful, angry and happy) faces. In contrast to these findings on the classification of emotional expression, only the fusiform face area (FFA) and anterior inferior temporal cortex (aIT) could discriminate among the various facial identities. Further, the amygdala and STS were better than FFA and aIT at classifying expression, while FFA and aIT were better than the amygdala and STS at classifying identity. Taken together, our findings indicate that the decoding of facial emotion and facial identity occurs in different neural substrates: the amygdala and STS for the former and FFA and aIT for the latter.

Keywords

amygdala; emotional faces; fMRI; STS; SVM

1. Introduction

Facial expressions convey a wealth of social information. The ability to discriminate between expressions is critical for effective social interaction and communication. Although it is apparent that humans discriminate different facial expressions automatically and effortlessly, the underlying neural computations for this ability remain unclear.

Corresponding author: Hui Zhang, Section on Neurocircuitry, Laboratory of Brain and Cognition, NIMH, NIH, Bethesda, Maryland 20892, USA. hui.zhang@nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Researchers have identified seven basic categories of facial expression that can be distinguished and classified; these include neutral, fearful, angry, sad, disgust, surprise and happy (Ekman 1992). Each category of facial expression produces a unique combination of facial musculature, thereby conveying unique social information to the viewer (Ahs et al., 2014). What might be the neural substrates for the classification of emotional facial expressions? One brain structure widely reported to be involved in the representation of emotional expression is the amygdala. Patient S.M., who has bilateral amygdala damage resulting from Urbach-Wiethe Syndrome, is impaired in recognizing fearful, angry and surprised facial expressions; the patient's performance in recognizing fearful faces is especially poor (Adolphs et al., 1994). Another patient, D.R., who sustained partial amygdala damage after undergoing bilateral stereotaxic amygdalotomy for the relief of epilepsy, similarly shows deficits in recognizing several categories of facial expression, including fearful, angry, sad, disgust, surprise and happy (Young et al., 1995). Consistent with these behavioral results, functional magnetic resonance imaging (fMRI) studies in healthy subjects have found that emotional faces, especially fearful expressions, evoke greater activation than neutral faces in the amygdala (Breiter et al., 1996; Whalen et al., 1998; Pessoa et al., 2002 and 2006). Imaging studies have also reported that patients with amygdala lesions show reduced fMRI responses to fearful faces in fusiform and occipital areas as compared with healthy subjects, indicating the critical role played by the amygdala in conveying this information back to visual processing areas (Vuilleumier et al., 2004).

In addition to the amygdala, many studies have suggested that the human posterior superior temporal sulcus (STS) is involved in the discrimination of facial expressions (see Allison et al., 2000 for review). For example, patients with posterior STS damage are reported to have impaired recognition of fearful and angry faces as compared to healthy subjects (Fox et al., 2011). Similarly, TMS to right posterior STS has been shown to impair recognition of facial expressions (Pitcher et al., 2014). Several fMRI studies (Narumoto et al., 2001; Engell et al., 2007) have also found that STS is more strongly activated when subjects viewed faces with emotional expressions than when they viewed neutral faces, and several fMRI adaptation studies (Winston et al., 2004; Andrews et al., 2004) have shown an increased response in STS when the same face was shown with different expressions, indicating the involvement of STS in the processing of facial expressions. In addition, recent multi-voxel pattern analysis (MVPA) of fMRI data has shown that different categories of emotional expression elicit distinct patterns of neural activation in STS (Said et al., 2010a, 2010b).

In addition to the amygdala and posterior STS, a third region, the fusiform face area (FFA) has also been implicated in the processing of facial expressions. Several neuroimaging studies in healthy participants (Dolan et al., 1996; Vuilleumier et al., 2001; Surguladze et al., 2003; Winston et al., 2003; Ganel et al., 2005; Pujol et al., 2009; Pessoa et al., 2002 and 2006; Furl et al., 2013) and electrocorticography recordings in patients (Kawasaki et al., 2012) have found significantly greater responses in FFA to several categories of emotional faces compared to neutral faces.

Taken together, these studies have provided evidence that emotional faces, compared with neutral faces, may be preferentially represented in several face-selective regions, such as the amygdala, STS, and possibly FFA. However, since most of these studies compared

Author Manuscript

activations evoked by emotional faces to activations evoked by neutral faces, they did not directly address whether these brain regions differentiate among the different categories of emotional expression. Additionally, most previous studies have focused only on one or two face-selective regions and very few (Harris et al., 2012; Harris et al., 2014a) have comprehensively examined all face selective regions in the brain.

Author Manuscript

Our current work investigates the contribution of each face-selective region in the human brain to the classification of four categories of facial expression: fearful, angry, happy and neutral. We hypothesized that different categories of facial expression would evoke different patterns of neural activity within the different face-selective regions, and these different patterns could be decoded by combining fMRI with a multivariate machine classification analysis. The advantage of using multivariate classification analysis over the traditional univariate approach is that the former is more efficient at classifying fine-scale spatial differences in neural representations and therefore is expected to yield better classification performance (Kriegeskorte et al, 2006; Misaki et al., 2010).

Author Manuscript

There are several prior studies that have used fMRI and multivariate machine classification methods to investigate facial expression discrimination. Said and colleagues (Said et al., 2010a, 2010b) used Sparse Multinomial Logistic Regression (SMLR) and its seven-way classification method to classify pairs of seven basic categories of facial expressions in STS. They found that facial expressions can be decoded in both posterior STS and anterior STS. Harry and colleagues (Harry et al., 2013) used single class Logistic Regression to classify each of six facial expression categories in FFA and early visual cortex (EVC), and found that facial expressions can be successfully decoded in both regions. Skerry and Saxe (2014) also examined the neural representations of facial expressions with binary SVM, and found that positive and negative expressions can be classified in right middle STS and right FFA. Because these studies used very different stimulus sets (both Said and colleagues, and Skerry and Saxe used dynamic videos, while Harry and colleagues used static images), and focused on different brain regions, it is difficult to compare the facial expression classification performance between these face-selective regions from different studies.

Author Manuscript

Our study, by contrast, examined the ability of each face-selective region to classify four different emotional expressions: fearful, angry, happy and neutral. Subjects participated in a slow event-related fMRI experiment, in which they were repeatedly shown 32 face images belonging to eight different identities. For each face-selective region of interest, we used a one-versus-all Support Vector Machine (SVM) to classify the fMRI activation patterns evoked by each category of facial expression against all other categories, and then calculated the corresponding classification accuracy. The classification accuracy determined how well each category of facial expression was decoded in each face-selective brain region. We extended this one-versus-all classification process in a hierarchical fashion similar to Lee et al. (2011) to investigate how well each region discriminated between the different emotional expressions.

Author Manuscript

In addition to classification of emotional expressions, our experimental design gave us the opportunity to investigate the classification of facial identity using the same dataset, and allowed us to compare the performance of identity classification with that of expression

classification in each face-selective region. One prominent idea in the face literature is that there are two distinct anatomical pathways for the visual analysis of facial expression and identity (Bruce and Young, 1990; Haxby et al., 2000). According to this conceptualization, the changeable aspects of a face, such as emotional expression, and the invariant aspects of a face, such as its identity, are processed in separate neural pathways: STS for expression and FFA for identity. Further, Haxby and his colleagues consider the amygdala as an extension of the system for processing emotional expression. However, the results of imaging studies have not been uniform in their support of this dual system idea (Calder and Young, 2005). We tested the idea in the current study.

Overall, the aim of our study was to uncover the ability of each face-selective region in the human brain to discriminate among four basic categories of facial expression. We also examined whether the discrimination of facial expression and identity are represented in distinct neural structures.

2. Materials and Methods

2.1 Subjects

A total of 25 healthy subjects (12 male) aged 27.0 ± 5.0 (mean $\pm$ SD) years participated in our study. All subjects were right handed, had normal or corrected-to-normal vision, and were in good health with no past neurological or psychiatric history. All participants gave informed consent according to a protocol approved by the Institutional Review Board of the National Institute of Mental Health. Data from two subjects were discarded because of excessive head movement during the fMRI scan, leaving a total of 23 subjects (11 male), aged 26.4 ± 4.8 (mean $\pm$ SD) years, for further analysis.

2.2 Experimental Procedure

2.2.1 Main experiment—Subjects performed a fixation cross color-change task during the presentation of face images with different expressions. The visual stimuli were shown in an event-related design. Each trial began with one of the 32 face stimuli (frontal view) presented simultaneously with a randomly generated colored fixation cross in the center of the image, for 300 ms, this was followed by a white fixation cross centered on the image for the rest of trial (7700 ms; see Fig. 1A). Subjects were instructed to press the left button if the fixation cross was red, and the right button if the fixation cross was green. Subjects were asked to respond as quickly as possible. Each trial lasted for a fixed duration of 8 s and there were 32 trials per run. At the beginning and end of each run, a gray fixation cross was presented at the center of the screen for 8 s. Each run lasted 4 min 32 s.

The 32 face images were selected from the Karolinska Directed Emotional Faces (KDEF) dataset (<http://www.emotionlab.se/resources/kdef>) and belonged to 8 different individuals, each depicting four different facial expressions: neutral, fearful, angry, and happy. Half of the individuals were female and half were male. All face images were cropped beforehand to show only the face on a black background. These images were converted to gray-scale, normalized to have equivalent size, luminance and contrast, and resized to 330×450 pixels. (Fig. 1B). Each face image was presented once in each run. The order of the face images

was randomized across runs, while the order of the runs remained the same across all the subjects.

2.2.2 Identification of face-selective brain regions—To identify face-selective brain regions for each subject, participants also performed a one-back matching task during separate localizer runs. During these runs, subjects viewed blocks of human faces, common objects and scrambled images, and were asked to press the left button if the current image matched the preceding one, and the right button if it did not. Each block lasted 24 seconds, with a 16-sec blank period between blocks. Within a block, each image was presented for 500 ms, with 1-sec blank period between images. There were two blocks for each condition per run, and the order of the blocks was randomized. Each localizer run lasted 5 minutes. The face images used in the localizer runs were all neutral faces, and they differed from those used in the main experiment. The object stimuli included common everyday objects such as eyeglasses, shoes, shirts, scissors, cups, umbrellas, keys, etc. The scrambled images were constructed by randomly permuting the phase of the face images in the Fourier domain. Each of the three categories contained 30 different images. These images were converted to gray-scale, normalized to have equivalent size, luminance and contrast, and resized to 350×350 pixels.

Before subjects were engaged in the experiment, each underwent a brief training session until they were able to adequately perform both tasks. After training, within the MRI scanner, subjects completed 13–16 runs of the main experimental task and two runs of the localizer task.

2.3 Apparatus

The visual stimuli were displayed with Presentation software (www.neurobs.com) and back-projected onto a screen in the dimly lit scanner room using a LCD projector with a refresh rate of 60 Hz and resolution of 1024×768 . Subjects viewed the stimuli via a mirror system installed in the head coil. The images subtended a visual angle of 6.0×8.2 degrees for the main experiment and a visual angle of 6.4×6.4 degrees for the localizer runs.

2.4 fMRI Acquisition

Imaging data were collected using a GE MR 750 3.0 Tesla scanner with a GE 8-channel head coil. Each scan session began with a high resolution T1-weighted Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence (time to echo [TE] = 3.4 ms; time to repetition [TR] = 7 ms; flip angle = 7 degrees; matrix size = 256×256 ; voxel size = $1 \text{ mm} \times 1 \text{ mm} \times 1 \text{ mm}$; 124 slices). The functional images were acquired with a single-shot interleaved gradient-recalled echo planar imaging sequence to cover the whole brain (TE = 35 ms; TR = 2000 ms; flip angle = 77 degrees; matrix size = 72×72 ; voxel size = $3 \text{ mm} \times 3 \text{ mm} \times 3 \text{ mm}$; 41 axial slices).

2.5 fMRI Preprocessing and Analysis

Functional MRI data were analyzed using AFNI (Cox 1996) and in-house MATLAB programs. Data from the first four TRs from each run were discarded. The remaining images were slice-time corrected, and realigned to the first volume of the each session. Signal

intensity was normalized to the mean signal value within each run and multiplied by 100 so that the results after analysis represented percentage signal change from baseline.

For the main experiment scans, no spatial smoothing was performed on the functional data. A gamma function with a peak of one as a model was used for the hemodynamic response function, and a general linear model (GLM) was established for each of the 32 face stimuli across runs. The model included the hemodynamic response predictors for each of the 32 stimuli, and also the regressors of no interest (baseline and head movement parameters). The parameter estimates (beta values) of the hemodynamic response evoked by each face stimulus at each voxel were extracted within the ROIs defined by the localizer scans (see below).

For the localizer scans, the images were spatially smoothed with a 4-mm full-width-half-maximum Gaussian kernel. The conditions of face, object and scrambled images as three regressors of interest were convolved with the gamma function and then input into the GLM for parameter estimation. The regressors of no interest (baseline and head movement parameters) were also included in the GLM model.

2.6 Defining Regions of Interest

The face-selective regions were defined in each individual subject by contrasting the fMRI response to faces with that to objects ($p < 0.001$, uncorrected). Because previous studies have reported that face-selective regions in the right hemisphere are more consistently activated than those in the left hemisphere, here we only considered the face-selective ROIs in the right hemisphere. These regions were selected by drawing a sphere of 6-mm radius around the peak of the face-selective activation. The cortical ROIs defined in this way included the occipital face area (OFA), the fusiform face area (FFA), the anterior inferior temporal cortex (aIT), and the posterior superior temporal sulcus (STS). The ROI for the amygdala was selected by manually drawing an irregularly shaped region around the face-selective activation peak, limited within the amygdala's anatomical boundary. Each ROI covered 35 voxels. As a control ROI, a 6-mm sphere with the same number of voxels, centered at $x = 1$, $y = -79$, $z = 2$ in Talairach space, was selected anatomically for each subject in primary visual cortex (V1) of the right hemisphere. However, the 6-mm ROI centered at this location may have included some voxels in the left hemisphere and V2. The location of this ROI overlapped completely with the face versus baseline activation for each subject.

2.7 fMRI Analysis Using One-versus-all SVM

2.7.1 Classification of facial expression—The response patterns evoked by the 32 face stimuli were organized by expression into 4 groups: neutral, fearful, angry, and happy. Each group contained 8 face samples (one for each of the 8 identities). Since our aim was to differentiate each category of facial expression from all other categories, we re-organized these groups into the following four pairs of partitions: neutral vs. non-neutral (fearful, angry and happy); fearful vs. non-fearful (neutral, angry and happy); angry vs. non-angry (neutral, fearful and happy); and happy vs. non-happy (neutral, fearful and angry). For each of these partition-pairs, we used a one-versus-all Support Vector Machine (SVM) for binary

classification, i.e., of the 32 samples of the face stimuli, the 8 samples in one expression group were classified against the 24 samples in the remaining groups.

To assess the overall performance of each classification, a leave-one-identity-out cross-validation strategy was used, i.e., the 32 samples were split into 28 training samples, and 4 testing samples that were from the same identity. The linear classifiers (with the infinite regulation term c) were trained with the 28 training samples and then tested with the 4 testing samples to evaluate the classification accuracy. Since the face stimulus set covered 8 identities and the 4 testing samples in each round of the cross-validation were from a single identity, a total of 8 rounds of cross-validation were performed. Hence, each face sample was used 7 times as a training sample and once as a testing sample for each classification pair, and all the samples in the stimulus set had an equal chance of being trained and tested.

In a process similar to Lee et al. (2011), the one-versus-all SVM classification was extended in a hierarchical fashion to investigate the ability of the face-selective regions to distinguish between the various expressions, as follows. As outlined above, the first step in the classification was specified as a one-versus-three SVM, which was used to classify the pairs of neutral vs. non-neutral, fearful vs. non-fearful, angry vs. non-angry, and happy vs. non-happy expressions, respectively. Next, for a given ROI, the expression that showed the highest classification accuracy in step 1 was excluded, and the remaining three expressions were entered into a one-versus-two SVM, to classify each of three categories of expression against the remaining two. Finally, as a third step, the expression that showed the highest classification accuracy in step 2 was excluded, and the remaining two expressions were entered into a one-versus-one SVM classification.

2.7.2 Classification of facial identity—The response patterns for the 32 face stimuli were organized by identity into 8 groups, each containing 4 samples (one for each of the 4 expressions). We re-established the GLM model of the 32 face stimuli for even and odd runs separately. This doubled the number of samples for each facial identity from 4 to 8. Since half of the identities were male and half were female, we analyzed the male and female individual faces separately. The four male identity groups were labeled as ID1, ID2, ID3 and ID4, and the four female identity groups were labeled as ID5, ID6, ID7 and ID8.

For the male identities, we re-organized the data into the following four pairs of partition: ID1 vs. (ID2, ID3 and ID4); ID2 vs. (ID1, ID3 and ID4); ID3 vs. (ID1, ID2 and ID4); and ID4 vs. (ID1, ID2 and ID3). For each pair of these partitions, we used a one-versus-all SVM for binary classification, i.e., of the 32 samples of the male face stimuli, 8 samples in one identity group were classified against 24 samples in the remaining 3 identity groups.

In order to assess the overall performance of each classification, we used the same leave-one-out cross-validation strategy that was used for the classification of expression as described above, i.e., the 32 samples were split into 28 training samples and 4 testing samples (that were from the same expression as well as the same even/odd runs). The linear classifiers (with the infinite regulation term c) were trained with the 28 training samples and then tested with the 4 testing samples to evaluate classification accuracy. A total of eight rounds of cross-validation were performed. Hence, each of the 32 male face samples was

used 7 times as a training sample and once as a testing sample for each classification pair, and all the samples in the stimuli set had an equal chance of being trained and tested.

The entire process was repeated for the data for the female identities. For this, the data were re-organized into the following four partition pairs: ID5 vs. (ID6, ID7 and ID8); ID6 vs. (ID5, ID7 and ID8); ID7 vs. (ID5, ID6 and ID8); and ID8 vs. (ID5, ID6, and ID7). All subsequent analyses were the same as those performed for the data for male identities.

2.7.3 Significance test for one-versus-all SVM classification—To determine whether the classification accuracy of any of the classification pairs was significantly greater than chance, we used a permutation test as follows. First we randomly assigned the 4 expression labels to the 32 samples 1000 times and repeated the SVM classification procedure for each such permutation. In this way we created a distribution of classification accuracies that could occur by chance and used it to determine the probability (α) of obtaining a given classification accuracy for a true one-versus-all classification pair. In order to combine data across participants, we transformed these α values into normally distributed z scores in MATLAB ($z_score = -norminv(\alpha, 0, 1)$). Finally, a one sample t-test was performed on these normalized z scores.

3. Results

3.1 Behavioral Data

Analysis of accuracy and reaction time (RT) data showed that subjects performed both tasks well. Average overall accuracy on the main task (fixation cross color-change discrimination task) was $96.7\% \pm 7.0$ (mean $\pm$ SD) and average RT was $557.4\text{ ms} \pm 178.3\text{ ms}$ (mean $\pm$ SD). Average overall accuracy on the localizer task (one-back matching task) was $80.0\% \pm 4$ (mean $\pm$ SD) and average RT was $633.0\text{ ms} \pm 238.3\text{ ms}$ (mean $\pm$ SD). These data indicate that subjects paid attention to the stimuli; hence, all the data were included in subsequent analyses. Further, to confirm that the task did not influence the classification accuracy in face-selective ROIs, we performed a 4 (expression) $\times$ 8 (identity) repeated measures ANOVA separately on the accuracy and RT data, and found no main effects or interactions. This indicates that there was no behavioral difference in viewing the various expressions and identities, and that the task could not explain the difference in classification performance in the brain for expression or identity.

3.2 Localization of Face-Selective Regions of Interest

The contrast of faces vs. objects from the localizer task data was used to identify five face-selective regions in the right hemisphere for each subject ($p < 0.001$, uncorrected). Figure 2 shows the location of all the face-selective regions for a representative subject. Of the 23 subjects, significant activations were found in the inferior lateral occipital gyrus (termed OFA) in 22 subjects, the lateral fusiform gyrus (termed FFA) in 20 subjects, the posterior superior temporal sulcus (termed STS) in 19 subjects, and the dorsal lateral portion of the amygdala in 20 subjects. This is consistent with the results of others who have previously localized face-selective regions (Kanwisher et al., 1997; McCarthy et al., 1997; Harris et al., 2012, 2014a). Additionally, significant activations were found in the anterior inferior

temporal cortex (termed aIT) in 10 of the 23 subjects, which is also consistent with the results of others who have reported the existence of a face-selective region (though less reliably) in this location (Kriegeskorte et al., 2007, Axelrod and Yovel 2013). Our localizer data also revealed some face-selective regions in parietal and frontal cortex, but since the location of these regions was highly variable across subjects, we did not include them in further analyses. As described above, the ROIs were generated for each subject by drawing a sphere of radius 6 mm around the peak activation in each of the face-selective regions. The coordinates of the activation peaks for each of the five ROIs is shown in Table 1, along with the coordinates of a control ROI in V1.

3.3 Classification of Facial Expression

Using one-versus-three SVM, we found two ROIs that were able to accurately classify facial expressions: the amygdala and STS (see Fig. 3).

3.3.1 Amygdala—In the amygdala, the classification accuracy was significant for the classification of fearful relative to non-fearful faces (neutral, angry and happy faces) (mean classification accuracy = 0.72; mean z-normalized accuracy = 1.39; one-sample t-test: $t_{19} = 5.71$; $p = 0.000017$). The amygdala also showed significantly high accuracy for the classification of neutral relative to emotional faces (fearful, angry and happy faces) (mean classification accuracy = 0.68; mean z-normalized accuracy = 1.04; one-sample t-test: $t_{19} = 4.48$; $p = 0.00026$). The accuracy for both these classifications passed the Bonferroni corrected threshold of $p < 0.01$ (correction based on performing a total of 8 one-sample t-tests including 4 one-versus-three, 3 one-versus-two and 1 one-versus-one classification SVMs; see Supplementary Materials Appendix A for details).

To further assess how the amygdala classifies facial expressions, we extended our one-versus-all SVM classification analysis in a hierarchical fashion (see Fig. 4). First, using a paired t-test between classification accuracy for fearful-versus-all relative to neutral-versus-all we determined that the amygdala more accurately classified fearful relative to non-fearful faces than neutral relative to non-neutral faces (mean difference in z-normalized accuracy = 0.35; paired t-test: $t_{19} = 2.67$; one tailed $p = 0.0076$). We then excluded fearful faces in the next stage, to examine how well the amygdala classified the other three expressions. Thus, the remaining three expressions - neutral, angry and happy - were entered into one-versus-two SVMs for this second level analysis. The results from this analysis showed that the amygdala accurately classified neutral relative to non-neutral faces (angry and happy) (mean classification accuracy = 0.62; mean of z-normalized accuracy = 0.88; one-sample t-test: $t_{19} = 3.82$; $p = 0.0012$). The other one-versus-two SVMs did not show significant accuracy in classification.

Similar to the second level analysis, we then excluded the neutral faces and entered the remaining two expressions – angry and happy - into a third level one-versus-one SVM. Here, we found that the amygdala did not accurately classify angry relative to happy faces (mean classification accuracy = 0.44; mean of z-normalized accuracy = 0.16; one-sample t-test: $t_{19} = 0.68$; $p = 0.51$).

3.3.2 STS—In contrast to the amygdala, in the STS, we found that the classification accuracy was only significant for the classification of neutral faces relative to non-neutral faces (fearful, angry and happy faces) (mean classification accuracy = 0.69; mean z-normalized accuracy = 1.11; one-sample t-test: $t_{18} = 4.85$; $p = 0.00013$).

The one-versus-all SVM classification analyses were also extended hierarchically to assess how the STS classifies facial expressions (see Fig. 4). In the STS, since only the SVM of neutral relative to emotional faces (fearful, angry and happy) showed a significantly high classification accuracy, we excluded the neutral faces in the next stage of analysis, to examine how well the STS classified the other three expressions. Thus, the remaining three expressions - fearful, angry and happy, were entered into one-versus-two SVMs for this second level analysis. Here we found that the STS could accurately classify happy relative to fearful and angry faces (mean classification accuracy = 0.60; mean of z-normalized accuracy = 0.72; one-sample t-test: $t_{18} = 3.90$; $p = 0.0011$). The other two SVMs at this second stage of analysis did not yield significant classification accuracy.

Similar to the second level analysis, we then excluded the happy faces and entered the remaining two expressions - fearful and angry - into a third level one-versus-one SVM. Here, we found that the STS did not accurately classify fearful relative to angry faces (mean classification accuracy = 0.43; mean of z-normalized accuracy = 0.03; one-sample t-test: $t_{18} = 0.68$; $p = 0.81$).

The other four regions of interest, namely, OFA, FFA, aIT, and V1, all failed to show significant accuracy in classifying any of the emotional expressions (see Supplementary Table 1).

Besides the one-versus-all classification, we also performed one-versus-one SVM classification on each pair of the four expression categories. The results from one-versus-one SVM analysis were consistent with the one-versus-all SVM results showing high accuracy rates in the amygdala for classification of fearful relative to non-fearful faces, and high accuracy rates in STS for classification of neutral relative to emotional faces (see Supplementary Figures S4 and S5).

3.4 Comparison of Multivariate to Univariate Analysis

To compare the results of our multivariate classification analysis with a traditional univariate approach, we examined the percent signal change of the fMRI response amplitudes evoked by the four expressions in each of the 6 ROIs (see Fig. 5). A random effects ANOVA with expression as a fixed factor and subject as a random factor, was performed separately on each ROI to determine the effect of expression on percent signal change. No significant main effect of expression was found in any of the ROIs.

Because others (Hadj-Bouziane et al., 2008) have previously reported significant enhancement of fMRI signals evoked by fearful relative to neutral faces in regions like the amygdala, we performed three paired t-tests per ROI to examine this specifically. Using this method, we found that the amygdala responded more strongly to fearful relative to neutral

faces (paired t-test: $t_{19} = 3.69$; $p = 0.0015$), but no emotional expression, including fearful faces, evoked enhanced signals in any of the other ROIs.

3.5 Classification of Facial Identity

Similar to the method used for classification of facial expression, we used a one-versus-all SVM to determine if data from any of our ROIs could classify facial identity. Here, data for each of four identities (belonging to the same gender) were classified relative to the remaining three (see Fig. 6). A one-sample t-test of each pair of classifications showed that FFA accurately classified three of the four male identities (ID1: mean classification accuracy = 0.655; mean z-normalized accuracy = 0.98; one-sample t-test: $t_{19} = 4.60$; $p = 0.00020$; ID3: mean classification accuracy = 0.65; mean z-normalized accuracy = 0.93; one-sample t-test: $t_{19} = 4.59$; $p = 0.00020$; and ID4: mean classification accuracy = 0.66; mean z-normalized accuracy = 0.95; one-sample t-test: $t_{19} = 4.58$; $p = 0.00021$), and one of four female identities (ID7: mean classification accuracy = 0.70; mean z-normalized accuracy = 1.20; one-sample t-test: $t_{19} = 5.19$; $p = 0.000052$). A similar analysis in aIT showed that this region accurately classified one of the four male identities (ID1: mean classification accuracy = 0.70; mean z-normalized accuracy = 1.01; one-sample t-test: $t_9 = 4.86$; $p = 0.00090$), while the remaining ROIs -amygdala, OFA, STS, and V1 - did not show any ability to accurately classify identity (see Supplementary Table 2).

Next, for FFA and aIT we performed a hierarchical analysis similar to that used in the classification of expression. Here we did not observe any significant accuracy of classification of one-versus-two SVMs at the second level.

Besides the one-versus-all classification, we also performed one-versus-one SVM classification on each pair of the four identity categories (for each gender category). The one-versus-one SVM revealed additionally that FFA could significantly classify ID5 relative to ID8 (see Supplementary Figure S6 and S7).

As we did for facial expression, to contrast our multivariate analysis to the traditional univariate approach, we examined the percent signal change of the fMRI response amplitudes evoked by the eight individuals in each of the 6 ROIs (see Fig. 7). A random effects ANOVA with identity as a fixed factor and subject as a random factor was performed separately for each gender and each ROI to determine the effect of identity on percent signal change. No significant main effect of identity was found for either gender in any of the ROIs.

3.6 Equivalent analysis of expression classification

Although splitting the data into odd and even runs for identity classification ensured that the same amount of data (8 samples per category) would be used for classifying identity as well as expression, it also introduced a difference in the underlying data that were used to compute the beta estimates. To make sure the exact same beta estimates were used for identity classification and expression classification, we also performed the expression classification analysis by splitting the data into odd and even runs. The results (See Supplementary Figures S8 – S11) were consistent with our original results and support our conclusion that both amygdala and the STS were able to accurately discriminate between

expressions, albeit in different ways: the amygdala can discriminate fearful faces from non-fearful faces, whereas STS can discriminate neutral from emotional (fearful, angry and happy) faces. Classification accuracies from this analysis were used for the subsequent direct comparison of expression to identity classification (see below).

3.7 Comparison of facial expression and identity classification

In order to directly test for a double dissociation of expression and identity classification across face-selective ROIs, we compared the classification performance for expression and identity across different brain regions. To achieve this, we first performed repeated-measures ANOVAs separately for expression and identity classification.

A 4 (expression) $\times$ 6 (ROI) repeated measures ANOVA revealed 1) a main effect of expression ($F_{(3,324)} = 7.19$; $p = 0.00011$), such that overall classification accuracy was significantly higher for neutral-versus-others than angry-versus-others (post-hoc Fisher LSD test; $p = 0.000016$) and happy-versus-others ($p = 0.0052$), and significantly higher for fearful-versus-others than angry-versus-others ($p = 0.0040$). Neutral-versus-others did not differ from Fearful-versus-others classification; 2) a main effect of ROI ($F_{(5,108)} = 2.74$; $p = 0.023$), such that overall classification accuracy for the amygdala was significantly higher than FFA ($p = 0.0047$), aIT ($p = 0.029$), OFA ($p = 0.0024$) and V1 ($p = 0.0067$), but not STS ($p = 0.13$); and 3) an expression by ROI interaction ($F_{(15,324)} = 2.88$; $p = 0.00029$) that was driven by high classification accuracy in the amygdala for neutral and fearful stimuli compared to all other ROIs and expressions except compared to neutral stimuli in STS. Further, classification of neutral stimuli in STS was higher than all other expressions in other ROIs except the amygdala.

A 8 (Identity) $\times$ 6 (ROI) repeated measures ANOVA revealed only a main effect of ROI ($F_{(5,108)} = 6.64$; $p = 0.000020$) such that classification accuracy for identity discrimination was significantly higher in FFA (post-hoc Fisher LSD test; $p = 0.028$) and aIT ($p = 0.0054$) compared to all other ROIs. Further, FFA and aIT showed a similar ability to classify identity ($p = 0.31$).

The results from the above analyses indicate that of all the face-selective ROIs, the amygdala and STS are significantly involved in discrimination of expressions (especially fearful and neutral respectively), and FFA and aIT are significantly involved in discrimination of identity.

Next, in order to directly compare whether the amygdala and STS classified expression as well as FFA and aIT classified identity, we first identified the expression and identity that each region could classify the best (for this we used the numerically highest group averaged z-normalized accuracy rate of one-versus-others SVM based on data split into even and odd runs for both expression and identity, for male and female faces separately). For example, the amygdala could classify fearful-versus-others and ID4 the best for male faces, and fearful-versus-others and ID8 the best for female faces; STS could classify neutral-versus-others and ID1 the best for male faces, and neutral-versus-others and ID8 the best for the female faces; FFA could classify neutral-versus-others and ID1 the best for male faces, and so on. Then using the classification accuracy values for these preferred expressions and

identities, we performed a 3-way 2 (expression and identity) $\times$ 2 (male and female) $\times$ 6 (ROI) repeated measures ANOVA to compare classification ability for each ROI for both expression and identity for each gender. This analysis revealed several effects. First, a main effect of ROI was seen ($F_{(5,108)} = 5.50$; $p = 0.00015$) such that the amygdala, STS, FFA and aIT, while not significantly different from each other, showed higher overall classification accuracies than OFA and V1 (OFA and V1 did not differ from each other in their overall classification accuracy). Second, no overall main effect of classification between expression and identity ($F_{(1,108)} = 1.91$; $p = 0.17$) or gender ($F_{(1,108)} = 0.24$; $p = 0.63$) was seen. Further, there was no 3-way classification $\times$ gender $\times$ ROI interaction ($F_{(5,108)} = 0.57$; $p = 0.72$), nor a 2-way interaction involving gender (gender $\times$ ROI interaction: $F_{(5,108)} = 1.27$; $p = 0.28$; classification $\times$ gender interaction: $F_{(1,108)} = 0.6$; $p = 0.55$). Critically however, we found a 2-way classification by ROI interaction ($F_{(5,108)} = 10.82$; $p < 10^{-7}$) such that the amygdala (post-hoc Fisher LSD test; $p = 0.00042$) and to a certain degree STS ($p = 0.057$) could better classify expression than identity, while FFA ($p = 10^{-6}$) and aIT ($p = 0.0022$) could better classify identity than expression. Further, OFA ($p = 0.87$) and V1 ($p = 0.75$) did not show any difference in their ability to classify expression and identity. In addition, there was no difference between the ability of each of these four regions to classify its preferred attribute, i.e., expression for the amygdala and STS and identity for FFA and aIT.

The results from the above analyses indicate that the amygdala and STS can discriminate facial expression better than facial identity, and FFA and aIT can discriminate facial identity better than facial expression.

Taken together, results from the above three analyses provide strong support for a double dissociation, namely, that the amygdala and STS can better discriminate expressions while FFA and aIT can better discriminate identity.

4. Discussion

In the current study, we used fMRI and SVM multivariate pattern analysis to investigate the ability of face-selective regions in the human brain to classify four categories of emotional expression: neutral, fearful, angry and happy. Our results showed that both the amygdala and the posterior superior temporal sulcus (STS) are able to discriminate between these expressions, albeit in different ways: The amygdala accurately discriminated fearful from non-fearful faces, whereas STS accurately discriminated neutral from emotional (fearful, angry and happy) faces. Further, a hierarchical classification analysis revealed additional differences between the two regions: In the amygdala, fearful relative to non-fearful faces were discriminated at the top level of the hierarchy and neutral relative to emotional faces were discriminated at the second level, while for posterior STS, neutral relative to emotional faces were discriminated at the top level and positive (happy) relative to negative (angry and fearful) faces were discriminated at the second level (see Figure 8). None of the other face-selective regions showed a significant ability to discriminate among the different emotional expressions. In contrast to these findings on the classification of emotional expression, FFA and aIT were able to discriminate among the facial identities: FFA accurately discriminated three of four male individual faces and one of four female individual faces, while aIT accurately discriminated one of four male individual faces. None of the other face-selective

regions showed an ability to discriminate among the different identities. Taken together, our findings indicate that the decoding of facial emotion and facial identity occurs in different neural substrates: the amygdala and STS for the former and FFA and aIT for the latter.

4.1 Face-Selective Regions Differ in Their Ability to Classify Emotional Expressions

4.1.1 The amygdala plays a key role in discriminating between fearful and non-fearful faces—A fearful expression is often thought of as a vigilance reaction to a nearby potential threat, in that it conveys a distinct emotional state intended to avoid bodily harm in the context of social communication. It differs from the negative emotional expression of an angry face, which provides information about the source eliciting the emotional state. A fearful expression, by contrast, reflects the presence of a threat but not its source (Whalen, 2007). Humans automatically and effortlessly distinguish fearful faces from other facial expressions behaviorally, raising the question of which brain region is responsible for fearful face discrimination.

Studies of patient S.M., who has bilateral damage of the amygdala, have found that, of the seven basic facial expressions (neutral, fearful, angry, surprise, disgust, sad and happy), fearful face recognition is the most severely impaired (Adolphs et al., 1994). Previous neuroimaging studies have found that fearful face stimuli, relative to faces expressing other emotions, most consistently and reliably evoke activations in the amygdala (Breiter et al., 1996; Whalen et al., 1998; Pessoa et al., 2002 and 2006). In our current study, the univariate analysis of fMRI response magnitudes showed that only fearful expressions evoked significantly greater activation than neutral faces (see Fig. 5), indicating that the amygdala plays an important role in processing fearful faces. Furthermore, using multi-voxel pattern analysis, we found that fearful faces can be accurately decoded from the evoked patterns of fMRI signals, confirming that the amygdala is critical in discriminating fearful faces from other facial expressions.

To further examine the ability of the amygdala to discriminate between emotions, we extended our classification analysis of the one-versus-all SVM to a hierarchical binary decision tree approach. Our results showed a significant hierarchical structure of classification in the amygdala, such that the partition between fearful and non-fearful faces was at the top of the hierarchy, the partition between neutral and happy/angry faces was at the second level of the hierarchy, and the partition between happy and angry faces was at the third level of the hierarchy (see Fig. 8). The classification tree thus confirms that the dominant categorical classification in the amygdala is that of fearful faces, but the discrimination between neutral and angry/happy faces is also evident in this classification structure. We did not find a significant ability by the amygdala to discriminate angry from happy faces. Our results are consistent with many prior reports in the literature regarding the processing of emotional expressions in the amygdala, though differ in the analysis methods (multivoxel pattern SVM classification vs. univariate magnitude difference test).

Although fearful faces have been shown to evoke the greatest and most consistent activation in the amygdala (Adolphs et al., 1995; Breiter et al., 1996; Morris et al., 1996; Whalen et al., 1998; Pessoa et al., 2002 and 2006), significant activations evoked by happy and/or angry faces relative to neutral faces have also been found in some studies (Breiter et al., 1996;

Whalen et al., 1998; Whalen et al., 2001; Yang et al., 2002, Williams et al., 2004). To our knowledge, no study to date has reported a significant difference in the signals evoked by angry and happy faces in the amygdala.

Overall, our results indicate the dominant role of the amygdala in fearful and non-fearful face discrimination. Moreover, the discrimination of the basic facial expressions is represented in a hierarchical way in amygdala.

4.1.2 STS plays a key role in discriminating between neutral and emotional faces

—A patient with damage affecting the posterior STS was found to have impaired recognition of fearful and angry faces (Fox et al., 2011). Several neuroimaging studies have shown greater fMRI responses evoked by emotional faces (fearful, angry, surprise, disgust and happy) relative to neutral faces in STS (Narumoto et al., 2001; Engell et al., 2007). These findings thus suggest that the posterior STS also plays a role in processing emotional faces. Multi-voxel pattern analysis (MVPA) of fMRI signals has shown that different categories of emotional expression elicit distinct patterns of neural activation in STS (Said et al., 2010a, 2010b), suggesting that different categories of facial expression can be decoded and discriminated in this brain region. In the current study, using SVM, we found that neutral faces relative to emotional faces (fearful, angry and happy) can be accurately decoded from the patterns of fMRI signals, confirming that the posterior STS plays a key role in discriminating between neutral and emotional faces. Furthermore, using a hierarchical binary decision tree approach, we found a hierarchical structure in discriminating the four categories of facial expression in the posterior STS, such that the partition between neutral and emotional (fearful, angry and happy) faces was at the top of the hierarchy, the partition between positive (happy) and negative (fearful/angry) faces was at the second level of the hierarchy, and the partition between fearful and angry faces was at the bottom of the hierarchy (see Fig. 8). These data thus indicate that the posterior STS plays an important role not only in discriminating emotional from neutral faces, but also in discriminating positive from negative emotions, which differs from the decoding ability found in the amygdala. Interestingly, the posterior STS did not show a significant ability to distinguish fearful from angry faces, which the amygdala was able to do accurately.

Our results are consistent with Said and colleagues (Said et al., 2010a and 2010b) who found that various categories of facial expression can be decoded in the STS. However, they reported only the overall decoding performance across all the basic emotional expressions rather than the decoding performance for each category of expression. From our results, we can infer the dominant role of the posterior STS in discriminating between emotional and neutral expressions, but the discrimination between positive (happy) and negative (fearful and angry) faces is also evident in this classification structure. Our findings of the second-level classification are consistent with many previous reports of the significant ability of the STS to discriminate between positive (happy) and negative (angry/fearful) faces (Hooker et al., 2003; Nakato et al., 2011; Rahko et al., 2010; Williams et al., 2008). In contrast, Skerry and Saxe, (2014) recently reported that the STS cannot significantly discriminate positive (happy/smiling) from negative (sad/frowning) emotions. However, the negative emotions in that study (sad/frowning) and ours (fearful/angry) differed, which may account for the discrepant findings.

Using a conventional univariate analysis approach, we found that only the amygdala responded more strongly to fearful faces than neutral faces, matching results from many previous studies. However, using this approach we did not find any other regions that showed an effect of emotion. This is in contrast to some studies that have shown significant effects of emotional expression on fMRI signals in STS (Narumoto et al., 2001; Engell et al., 2007). One possible explanation for this discrepancy is that the slow event-related design used in our study produced much weaker fMRI signals than the block design used in many other studies. Whereas the univariate approach failed to show differential effects of emotion, our multivariate approach brought to light differences in how various face-selective regions process emotional expressions.

One limitation of our study is that neither the amygdala nor STS accurately discriminated every category of facial expression, even though perceptually we can automatically and effortlessly discriminate between the four categories of expression in our stimulus set. This perceptual ability to discriminate facial expressions may reflect not only the specific role played by each brain region in facial expression decoding, but also interactions among these brain regions. Imaging evidence in both humans (Vuilleumier et al., 2004) and monkeys (Hadj-Bouziane et al., 2012) has demonstrated that fMRI responses evoked by emotional faces in visual processing areas are dependent on feedback signals to those areas from the amygdala, i.e., there are strong amygdala-visual cortex interactions in processing emotional expression. Similar to emotional expressions, humans can discriminate individual faces effortlessly at the behavioral level, while our study showed that only some of the individual faces could be correctly decoded by FFA and aIT. Patients with congenital prosopagnosia have been reported to display disrupted functional connectivity between aIT and the ‘core’ face recognition network, including OFA, FFA, and posterior STS, again indicating that interactions among these face-selective regions is important in discriminating facial identity (Avidan et al., 2014).

Very recently, Dubois and colleagues (Dubois et al., 2015) have shown that the use of MVPA of fMRI data failed to decode the identity of human faces in the macaque brain, while the use of MVPA of single-unit data reliably decoded the identity of the same stimulus set in the macaque’s anterior face patch. These findings are relevant to all studies that attempt to use fMRI MVPA decoding techniques to classify neural representations, as they indicate the limitations of fMRI, in terms of both spatial and temporal resolution, and may, in part, explain why perceptual classification of expression is far better than neural classification with fMRI MVPA in our study.

4.1.3 No other face-selective region showed an ability to classify facial expressions—It is widely acknowledged that FFA is involved in face processing, in that it is activated more by faces than by complex non-face objects (e.g. Kanwisher et al., 1997; McCarthy et al., 1997; Rangarajan et al., 2014). Studies using univariate analyses have reported ambiguous results for FFA in discriminating facial expressions. Some studies have reported significantly greater activation evoked by emotional faces than by neutral faces (Dolan et al., 1996; Vuilleumier et al., 2001; Surguladze et al., 2003; Winston et al., 2003; Ganel et al., 2005; Pujol et al., 2009; Pessoa et al., 2002, 2006) and significant fMRI-adaptation effects of facial expressions (Fox et al. 2009; Furl et al., 2013; Ishai et al., 2004;

Cohen Kadosh et al., 2010; Vuilleumier et al., 2001; Xu & Biederman 2010), whereas other studies (Haxby et al., 2000; Winston et al., 2004; Andrews and Ewbank, 2004; Harris et al., 2014b) have not found that facial expressions can be significantly differentiated in FFA. Using multivariate analyses, several recent studies have shown above-chance accuracy of decoding emotional faces in FFA (Nestor et al., 2011, Harry et al., 2013; Skerry and Saxe, 2014). By contrast, neither our SVM classification nor our univariate approach revealed a significant ability of FFA to discriminate facial expressions.

The discrepancy between those reporting positive findings regarding the involvement of FFA in the processing of emotional expressions (Dolan et al., 1996; Vuilleumier et al., 2001; Surguladze et al., 2003; Winston et al., 2003; Pujol et al., 2009; Pessoa et al., 2002 and 2006, Ishai et al., 2004; Fox et al. 2009; Furl et al., 2013; Cohen Kadosh et al., 2010; Vuilleumier et al., 2001; Xu & Biederman 2010) and our negative findings may be due to several, not mutually exclusive, factors. One possible reason for the discrepancy may be due to a difference in tasks the subjects performed during scanning. In all studies reporting positive findings (except Furl et al., 2013), even though the emotional expressions were usually, but not always, irrelevant (e.g., gender discrimination, face working memory, etc.), subjects nonetheless responded to face stimuli, whereas in our study, subjects performed an irrelevant fixation cross color-change task. Therefore, it may be that attention to the face stimuli was reduced in our task compared those involving a facial discrimination, perhaps resulting in weaker fMRI signals and hence our negative findings.

Another possible reason for the discrepancy may be due to a difference in experimental design. While many previous studies reporting significant face expression effects in FFA used a block design (Dolan et al., 1996; Cohen Kadosh et al., 2010; Furl et al., 2013; Harry et al., 2013), we used a slow event-related design. The weaker fMRI signals in event-related designs could also have contributed to our negative findings.

Finally, the discrepancy between those studies reporting significant decoding performance in FFA for facial expression (Nestor et al., 2011, Harry et al., 2013; Skerry and Saxe, 2014) and our negative decoding findings may be due to a difference in statistical tests used to evaluate decoding accuracy. Whereas those reporting positive findings used the theoretical chance level to evaluate statistical significance, we used the permutation test. Recently, Combrisson and Jerbi (2015) compared the various statistical tests used to evaluate decoding performance and showed that the theoretical chance level was less stringent for small numbers of data samples, as found in neuroimaging studies, causing a bias towards statistical significance. The permutation test, by contrast, is not open to this criticism because it is a data-driven method and does not make assumptions about the statistical properties of the dataset. Therefore, the significance test we used may have been more stringent than the one used by those obtaining positive decoding findings in FFA.

In sum, a variety of factors may have contributed to the discrepancy between prior positive findings in FFA regarding significant expression effects and our negative findings, including differences in tasks used, experimental designs, and the evaluation of statistical significance. It is important to note, however, that none of these factors would explain the fact that we did obtain significant decoding of expression in both the amygdala and STS. We also did not

find significant decoding performance for any category of facial expression in aIT, OFA or V1, confirming that these regions are likely not involved in discriminating among emotional expressions.

4.2 Distinct brain pathways for expression and identity processing

One prominent idea in the face processing literature is that there are two distinct neuroanatomical pathways for the visual analysis of facial expression and identity (Bruce and Young, 1990; Young et al., 1993; Hoffman and Haxby, 2000; Haxby et al., 2000): one, located laterally in the brain, is presumed to process the changeable aspects of a face, such as emotional expression, while the other, located ventrally in the brain, is presumed to process the invariant aspects of a face, such as identity (Haxby et al., 2000). According to this conceptualization, STS and its further projections to the amygdala would process emotional expressions, while the OFA, FFA and its further projections to the aIT would process identity (Avidan G et al. 2014; Gilaie-Dotan et al., 2015; Pitcher et al., 2014). Although several studies have provided evidence to support this idea (Adolphs et al., 1995; Haxby et al., 2000; Winston et al., 2004; Andrews and Ewbank, 2004; Harris et al., 2014b; Weiner and Grill-Spector, 2015), others have questioned it (Calder and Young, 2005; also see above discussion regarding FFA). In fact, a recent review of neuroimaging studies of identity and expression processing (with static and dynamic faces), even proposed form and motion as the functional division between the ventral and lateral pathways, rather than identity and expression (Bernstein and Yovel, 2015). However, to date, no one has yet simultaneously classified the neural patterns evoked by different facial expressions and facial identities with MVPA using the same face stimulus set. In our study, we classified each of four categories of facial expression (neutral, fearful, angry and happy) and each of four facial identities of the same gender (a total of eight individual faces) from the neural response patterns evoked by the 32 face stimuli. Our results showed that, of the six regions of interest investigated (five face-selective regions: the amygdala, aIT, STS, FFA, OFA; and one control region in primary visual cortex, V1), only FFA and aIT accurately classified facial identities, which is consistent with results from previous studies investigating classification of facial identity (Kriegeskorte et al., 2007; Nestor et al., 2011; Goesaert and Op de Beeck, 2013; Anzellotti et al., 2014; Avidan and Behrmann, 2014; Axelrod and Yovel 2015). Together with our findings that: 1) only the amygdala and STS accurately classified emotional expressions; 2) the accuracy for classifying expression was greater in the amygdala and STS than in FFA and aIT; and 3) the accuracy for classifying identity was greater in FFA and to some extent aIT than in the amygdala and STS; the current results thus provide strong evidence for the dissociation of neural pathways mediating facial expression and identity discrimination in the human brain.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Intramural Research Program of the National Institute of Mental Health (NCT01087281; 93-M-0170).

References

- Adolphs R, Tranel D, Damasio H, Damasio AR. A Fear and the human amygdala. *J. Neurosci.* 1995; 16:7678–7687. [PubMed: 8922424]
- Adolphs R, Tranel D, Damasio H, Damasio A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature.* 1994; 372:669–672. [PubMed: 7990957]
- Ahs F, Davis CF, Gorka AX, Hariri AR. Feature-based representations of emotional facial expressions in the human amygdala. *Soc Cogn Affect Neurosci.* 2014; 9:1372–1378. [PubMed: 23887817]
- Allison T, Puce A, McCarthy G. Social perception from visual cues: role of the STS region. *Trends CognSci.* 2000; 4:267–278.
- Andrews TJ, Ewbank MP. Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage.* 2004; 23:905–913. [PubMed: 15528090]
- Anzellotti S, Fairhall SL, Caramazza A. Decoding representations of face identity that are tolerant to rotation. *Cereb Cortex.* 2014; 24:1988–1995. [PubMed: 23463339]
- Avidan G, Tanzer M, Hadj-Bouziane F, Liu N, Ungerleider LG, Behrmann M. Selective dissociation between core and extended regions of the face processing network in congenital prosopagnosia. *Cereb Cortex.* 2014; 24:1565–1578. [PubMed: 23377287]
- Avidan G, Behrmann M. Impairment of the face processing network in congenital prosopagnosia. *Front Biosci (Elite Ed).* 2014; 6:236–257. [PubMed: 24896205]
- Axelrod V, Yovel G. The challenge of localizing the anterior temporal face area: A possible solution. *NeuroImage.* 2013; 81:371–380. [PubMed: 23684864]
- Axelrod V, Yovel G. Successful decoding of famous faces in the fusiform face area. *PLoS One.* 2015; 10:e0117126–e0117126. [PubMed: 25714434]
- Bernstein M, Yovel G. Two neural pathways of face processing: A critical evaluation of current models. *Neurosci Biobehav Rev.* 2015; 55:536–546. [PubMed: 26067903]
- Breiter HC, Etcoff NL, Whalen PJ, Kennedy WA, Rauch SL, Buckner RL, Strauss MM, Hyman SE, Rosen BR. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron.* 1996; 17:875–887. [PubMed: 8938120]
- Bruce V, Young A. Understanding face recognition. Comment in *Br J Psychol.* 1990; 81:361–380. [PubMed: 2224396]
- Calder AJ, Young AW. Understanding the recognition of facial identity and facial expression. *Nat Rev Neurosci.* 2005; 6:641–651. [PubMed: 16062171]
- Cohen Kadosh K, Henson RN, Cohen Kadosh R, Johnson MH, Dick F. Task-dependent activation of face-sensitive cortex: an fMRI adaptation study. *J Cogn Neurosci.* 2010; 22(5):903–917. [PubMed: 19320549]
- Combrisson E, Jerbi K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods.* 2015; 250:126–136. [PubMed: 25596422]
- Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res.* 1996; 29:162–173. [PubMed: 8812068]
- Dolan RJ, Fletcher P, Morris J, Kapur N, Deakin JF, Frith CD. Neural activation during covert processing of positive emotional facial expressions. *Neuroimage.* 1996; 4:194–200. [PubMed: 9345509]
- Dubois J, Berker AO, Tsao DY. Single-Unit Recordings in the Macaque Face Patch System Reveal Limitations of fMRI MVPA. *J Neurosci.* 2015; 35:2791–2802. [PubMed: 25673866]
- Ekman P. An argument for basic emotions. *Cognition and Emotion.* 1992; 6:169–200.
- Engell AD, Haxby JV. Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia.* 2007; 45:3234–3241. [PubMed: 17707444]
- Fox CJ, Hanif HM, Iaria G, Duchaine BC, Barton JJ. Perceptual and anatomic patterns of selective deficits in facial identity and expression processing. *Neuropsychologia.* 2011; 49:3188–3200. [PubMed: 21807006]

- Fox CJ, Moon SY, Iaria G, Barton JJ. The correlates of subjective perception of identity and expression in the face network: an fMRI adaptation study. *Neuroimage*. 2009; 44:569–580. [PubMed: 18852053]
- Furl N, Henson RN, Friston KJ, Calder AJ. Top-down control of visual responses to fear by the amygdala. *J Neurosci*. 2013; 33(44):17435–17443. [PubMed: 24174677]
- Ganel T, Valyear KF, Goshen-Gottstein Y, Goodale MA. The involvement of the “fusiform face area” in processing facial expression. *Neuropsychologia*. 2005; 43:1645–1654. [PubMed: 16009246]
- Gilaie-Dotan S, Saygin AP, Lorenzi LJ, Rees G, Behrmann M. Ventral aspect of the visual form pathway is not critical for the perception of biological motion. *Proc Natl Acad Sci U S A*. 2015; 112:E361–E370. [PubMed: 25583504]
- Goesaert E, Op de Beeck HP. Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. *J Neurosci*. 2013; 33(19):8549–8558. [PubMed: 23658192]
- Hadj-Bouziane F, Liu N, Bell AH, Gothard KM, Luh WM, Tootell RB, Murray EA, Ungerleider LG. Amygdala lesions disrupt modulation of functional MRI activity evoked by facial expression in the monkey inferior temporal cortex. *Proc Natl Acad Sci U S A*. 2012; 109:E3640–E3648. [PubMed: 23184972]
- Hadj-Bouziane F, Bell AH, Knusten TA, Ungerleider LG, Tootell RB. Perception of emotional expressions is independent of face selectivity in monkey inferior temporal cortex. *Proc Natl Acad Sci USA*. 2008; 105:5591–5596. [PubMed: 18375769]
- Harris RJ, Young AW, Andrews TJ. Morphing between expressions dissociates continuous from categorical representations of facial expression in the human brain. *Proc Natl Acad Sci USA*. 2012; 109:21164–21169. [PubMed: 23213218]
- Harris RJ, Young AW, Andrews TJ. Dynamic stimuli demonstrate a categorical representation of facial expression in the amygdala. *Neuropsychologia*. 2014a; 56:47–52. [PubMed: 24447769]
- Harris RJ, Young AW, Andrews TJ. Brain regions involved in processing facial identity and expression are differentially selective for surface and edge information. *NeuroImage*. 2014b; 97:217–223. [PubMed: 24747739]
- Harry B, Williams MA, Davis C, Kim J. Emotional expressions evoke a differential response in the fusiform face area. *Front Hum Neurosci*. 2013; 7:692. [PubMed: 24194707]
- Haxby JV, Hoffman EA, Gobbini MI. The distributed human neural system for face perception. *Trends Cogn Sci*. 2000; 4:223–233. [PubMed: 10827445]
- Hooker CI, Paller KA, Gitelman DR, Parrish TB, Mesulam MM, Reber PJ. Brain networks for analyzing eye gaze. *Brain Res Cogn Brain Res*. 2003; 17(2):406–418. [PubMed: 12880911]
- Ishai A, Pessoa L, Bickle PC, Ungerleider LG. Repetition suppression of faces is modulated by emotion. *Proc Natl Acad Sci U S A*. 2004; 101(26):9827–9832. [PubMed: 15210952]
- Kawasaki H, Tsuchiya N, Kovach CK, Nourski KV, Oya H, Howard MA, Adolphs R. Processing of facial emotion in the human fusiform gyrus. *J CognNeurosci*. 2012; 24:1358–1370.
- Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*. 1997; 17:4302–4311. [PubMed: 9151747]
- Kriegeskorte N, Formisano E, Sorger B, Goebel R. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci USA*. 2007; 104:20600–20605. [PubMed: 18077383]
- Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. *Proc Natl Acad Sci USA*. 2006; 103:3863–3868. [PubMed: 16537458]
- Lee C, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*. 2011; 53:1162–1171.
- McCarthy G, Puce A, Gore JC, Allison T. Face-specific processing in the human fusiform gyrus. *J. Cogn. Neurosci*. 1997; 9:605–610. [PubMed: 23965119]
- Misaki M, Kim Y, Bandettini PA. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*. 2010; 53:103–118. [PubMed: 20580933]
- Morris JS, Frith CD, Perrett DI, Rowland D, Young AW, Calder AJ, Dolan RJ. A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature*. 1996; 383:812–815. [PubMed: 8893004]

- Nakato E, Otsuka Y, Kanazawa S, Yamaguchi MK, Kakigi R. Distinct differences in the pattern of hemodynamic response to happy and angry facial expressions in infants—a near-infrared spectroscopic study. *Neuroimage*. 2011; 54(2):1600–1606. [PubMed: 20850548]
- Narumoto J, Okada T, Sadato N, Fukui K, Yonekura Y. Attention to emotion modulates fMRI activity in human right superior temporal sulcus. *Brain Res Cogn Brain Res*. 2001; 12:225–231. [PubMed: 11587892]
- Nestor A, Plaut DC, Behrmann M. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proc Natl Acad Sci U S A*. 2011; 108:9998–10003. [PubMed: 21628569]
- Pessoa L, Japee S, Sturman D, Ungerleider LG. Target visibility and visual awareness modulate amygdala responses to fearful faces. *Cereb Cortex*. 2006; 16:366–375. [PubMed: 15930371]
- Pessoa L, McKenna M, Gutierrez E, Ungerleider LG. Neural processing of emotional faces requires attention. *Proc Natl Acad Sci USA*. 2002; 99:11458–11463. [PubMed: 12177449]
- Pitcher D. Facial expression recognition takes longer in the posterior superior temporal sulcus than in the occipital face area. *J Neurosci*. 2014; 34:9173–9177. [PubMed: 24990937]
- Pujol J, Harrison BJ, Ortiz H, Deus J, Soriano-Mas C, López-Solà M, Yücel M, Perich X, Cardoner N. Influence of the fusiform gyrus on amygdala response to emotional faces in the non-clinical range of social anxiety. *Psychol Med*. 2009; 39:1177–1187. [PubMed: 19154647]
- Rangarajan V, Hermes D, Foster BL, Weiner KS, Jacques C, Grill-Spector K, Parvizi J. Electrical stimulation of the left and right human fusiform gyrus causes different effects in conscious face perception. *J Neurosci*. 2014; 34:12828–12836. [PubMed: 25232118]
- Rahko J, Paakki JJ, Starck T, Nikkinen J, Remes J, Hurtig T, Kuusikko-Gauffin S, Mattila ML, Jussila K, Jansson-Verkasalo E, Kätsyri J, Sams M, Pauls D, Ebeling H, Moilanen I, Tervonen O, Kiviniemi V. Functional mapping of dynamic happy and fearful facial expression processing in adolescents. *Brain Imaging Behav*. 2010; 4(2):164–176. [PubMed: 20502991]
- Said CP, Moore CD, Engell AD, Todorov A, Haxby JV. Distributed representations of dynamic facial expressions in the superior temporal sulcus. *J Vis*. 2010; 10:11. [PubMed: 20616141]
- Said CP, Moore CD, Norman KA, Haxby JV, Todorov A. Graded representations of emotional expressions in the left superior temporal sulcus. *Front Syst Neurosci*. 2010; 4:6. [PubMed: 20305753]
- Salton, G.; McGill, M. *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc; 1983.
- Skerry AE, Saxe R. A Common Neural Code for Perceived and Inferred Emotion. *J Neurosci*. 2014; 34:15997–16008. [PubMed: 25429141]
- Surguladze SA, Brammer MJ, Young AW, Andrew C, Travis MJ, Williams SC, Phillips ML. A preferential increase in the extrastriate response to signals of danger. *Neuroimage*. 2003; 19(4):1317–1328. [PubMed: 12948690]
- Vuilleumier P, Armony JL, Driver J, Dolan RJ. Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron*. 2001; 30(3):829–841. [PubMed: 11430815]
- Vuilleumier P, Richardson MP, Armony JL, Driver J, Dolan RJ. Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nat Neurosci*. 2004; 7:1271–1278. [PubMed: 15494727]
- Weiner KS, Grill-Spector K. The evolution of face processing networks. *Trends Cogn Sci*. 2015; 19:240–241. [PubMed: 25840651]
- Whalen PJ, Rauch SL, Etcoff NL, McInerney SC, Lee M, Jenike MA. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *J Neurosci*. 1998; 18:411–418. [PubMed: 9412517]
- Whalen PJ. The uncertainty of it all. *Trends Cognit Sci*. 2007; 11:499–500. [PubMed: 18024182]
- Whalen PJ, Shin LM, McInerney SC, Fischer H, Wright CI, Rauch SL. A functional MRI study of human amygdala responses to facial expressions of fear versus anger. *Emotion*. 2001; 1:70–83. [PubMed: 12894812]

- Winston JS, Henson RN, Fine-Goulden MR, Dolan RJ. fMRI-Adaptation Reveals Dissociable Neural Representations of Identity and Expression in Face Perception. *J Neurophysiol.* 2004; 92:1830–1839. [PubMed: 15115795]
- Winston JS, O'Doherty J, Dolan RJ. Common and distinct neural responses during direct and incidental processing of multiple facial emotions. *Neuroimage.* 2003; 20(1):84–97. [PubMed: 14527572]
- Williams MA, Morris AP, McGlone F, Abbott DF, Mattingley JB. Amygdala responses to fearful and happy facial expressions under conditions of binocular suppression. *J Neurosci.* 2004; 24:2898–2904. [PubMed: 15044528]
- Williams MA, McGlone F, Abbott DF, Mattingley JB. Stimulus-driven and strategic neural responses to fearful and happy facial expressions in humans. *Eur J Neurosci.* 2008; 27(11):3074–3082. [PubMed: 18540880]
- Xu X, Biederman I. Loci of the release from fMRI adaptation for changes in facial expressions, identity, and viewpoint. *J. Vis.* 2010
- Yang TT, Menon V, Eliez S, Blasey C, White CD, Reid AJ, Gotlib IH, Reiss AL. Amygdalar activation associated with positive and negative facial expressions. *Neuroreport.* 2002; 13:1737–1741. [PubMed: 12395114]
- Young AW, Aggleton JP, Hellawell DJ, Johnson M, Brooks P, Hanley JR. Face processing impairments after amygdalotomy. *Brain.* 1995; 118:15–24. [PubMed: 7895001]
- Young AW, Newcombe F, de Haan EH, Small M, Hay DC. Face perception after brain injury. Selective impairments affecting identity and expression. *Brain.* 1993; 116:941–959. [PubMed: 8353717]

Highlights

Classification performance evaluated in human face-selective regions

Amygdala and pSTS accurately classified facial expressions

FFA and aIT accurately classified facial identities

Classification of expression and identity occurs in different neural pathways

Figure 1.

Schematic showing the main task (A) and the stimulus set used for the main task. A) The trial began with the presentation of a face image for 300 ms with a colored fixation cross centered on the image, followed by a white fixation cross centered on the screen for the rest of the trial (7700 ms). The colored fixation-cross appeared when a face image appeared and changed to a white fixation cross when the face image disappeared. Each trial lasted a fixed time duration of 8 seconds. There were 32 face trials in one run and each run lasted 4 min 32 seconds. In the MRI scanner, subjects viewed the visual stimuli projected onto a screen and pressed the left button when the fixation cross changed from white to red, and the right button when the fixation cross changed from white to green. B) The face stimulus set contained 32 faces, which belonged to four categories of expression: neutral, fearful, angry, and happy, each consisting of eight different facial identities. Half of the stimuli were female and half were male. Each face image was presented once in each run. The order of the face images was randomized across runs.

Figure 2.

Example of localization of face-selective regions by contrasting the fMRI response evoked by face stimuli compared to the response evoked by object stimuli (face > object) for a single subject. This contrast identified five face-selective regions in the right hemisphere: the dorsal-lateral part of the amygdala; anterior inferior temporal cortex - aIT; fusiform face area - FFA; posterior superior temporal sulcus - posterior STS and occipital face area - OFA. Primary visual cortex - V1 was selected as a control region of interest.

Figure 3.

The decoding performance (average z-normalized accuracy rate across subjects) of each face-selective region of interest (ROI; panels A - E) as well as V1 (panel F) for classification of facial expression. For each ROI, the decoding was performed by using one- versus-three SVM to classify each category of facial expression (neutral, fearful, angry and happy) relative to all other facial expression categories.

Figure 4.

Hierarchical classification analysis of facial expression in the amygdala and posterior STS. The hierarchical structure was calculated by applying one-versus-all SVM classification at each level of the binary tree architecture. In the amygdala (panels A–C), the classification of fearful and non-fearful faces was at the top of the hierarchy and showed significant accuracy of discrimination between the two (A); the classification of neutral and emotional faces (angry and happy) was at the second level and showed significant accuracy of discrimination (B); the classification of happy and angry faces was at the bottom of the hierarchy but did

not show significant accuracy of discrimination (C). In contrast, in STS (panels D–F), the classification of neutral and emotional faces (fearful, angry and happy) was at the top of the hierarchy and showed significant accuracy of discrimination (D); the classification of positive (happy) and negative (fearful and angry) expressions was at the second level and showed significant accuracy of discrimination (E); the classification of angry and fearful faces was at the bottom of the hierarchy and did not show significant accuracy of discrimination (F).

Figure 5. Percent signal change of the fMRI response amplitude for each category of facial expression (neutral, fearful, angry and happy) in face-selective ROIs (panels A–E) and V1 (panel F) calculated using the conventional univariate analysis approach.

Figure 6.

The decoding performance (average of z-normalized accuracy rate across subjects) of each face-selective ROIs (panels A–E) as well as V1 (panel F) for facial identity classification. For each ROI, the decoding was performed by using one-versus-three SVM to classify each individual face relative to the other three individuals of the same gender. Dark gray bars indicate male faces; light gray bars indicate female faces.

Figure 7.

Average percent signal change of the fMRI response amplitude for each face identity in face-selective ROIs (panels A–E) and V1 (panel F) calculated using the conventional univariate analysis approach.

Figure 8.

Summary of the hierarchical classification structure in amygdala and posterior STS. Solid black lines outline the facial expressions that can successfully be classified; dashed black lines outline the facial expressions that cannot be classified.

Table 1

Mean ROI Coordinates

ROI	Talairach Peak Coordinates			Number of Subjects
	x	y	z	
Amygdala	20	−4	−9	20
aIT	29	−4	−30	10
FFA	40	−44	−21	20
STS	44	−46	14	19
OFA	38	−74	−12	22
V1	1	−79	2	23