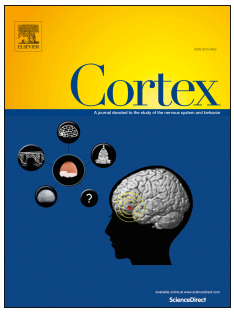
Accepted Manuscript



Spatially generalizable representations of facial expressions: Decoding across partial face samples

Steven G. Greening, Derek G.V. Mitchell, Fraser W. Smith

PII: S0010-9452(17)30400-8

DOI: 10.1016/j.cortex.2017.11.016

Reference: CORTEX 2190

To appear in: *Cortex*

Received Date: 10 August 2017

Revised Date: 2 November 2017

Accepted Date: 28 November 2017

Please cite this article as: Greening SG, Mitchell DGV, Smith FW, Spatially generalizable representations of facial expressions: Decoding across partial face samples, *CORTEX* (2018), doi: 10.1016/j.cortex.2017.11.016.

Spatially generalizable representations of facial expressions: Decoding across partial face samples

Steven G. Greening[1,2], Derek G.V. Mitchell[3,4,5,6*] & Fraser W. Smith[7*]

[1]Department of Psychology, Louisiana State University, Baton Rouge, USA

[2]Pennington Biomedical Research Institute, Louisiana State University, Baton Rouge, USA

[3]Department of Psychiatry, [4]Department of Anatomy & Cell Biology, [5]Neuroscience Program, Schulich School of Medicine & Dentistry, University of Western Ontario, London, CA

[6]Brain and Mind Institute, Natural Sciences Centre, University of Western Ontario, London, CA

[7]School of Psychology, University of East Anglia, Norwich, UK

*Correspondence should be addressed to FWS OR DGVM

Fraser W Smith (Fraser.Smith@uea.ac.uk)

School of Psychology

LSB Building

University of East Anglia

Norwich Research Park

Norwich, UK, NR4 7TJ

Derek GV Mitchell (dmitch8@uwo.ca)

Brain and Mind Institute

University of Western Ontario

London, Ontario, Canada

N6A 5B7

Number of Pages: 39

Number of figures and tables: Figures = 5; Tables = 0

Number of words: Abstract = 201, Body = 6,500

Acknowledgments: The authors would like to thank Gesine Alders, Thida Han, and the staff at the Centre for Functional and Metabolic Mapping (CFMM) for their help in collecting the data. Funding for data collection was provided by a SSHRC grant to D.G.V.M.

Highlights

- Facial expressions can be classified from patterns of brain activity
- Classification generalized across spatially independent samples of face information
- Significant cross-classification in early and higher visual areas and dorsal PFC
- Cross-classification in STS and dorsal PFC correlated with behavioural accuracy
- Cortical feedback may facilitate reactivation of occluded parts of face stimuli

Abstract

A network of cortical and sub-cortical regions is known to be important in the processing of facial expression. However, to date no study has investigated whether representations of facial expressions present in this network permit generalization across independent samples of face information (e.g. eye region Vs mouth region). We presented participants with partial face samples of five expression categories in a rapid event-related fMRI experiment. We reveal a network of face-sensitive regions that contain information about facial expression categories regardless of which part of the face is presented. We further reveal that the neural information present in a subset of these regions: dorsal prefrontal cortex (dPFC), superior temporal sulcus (STS), lateral occipital and ventral temporal cortex, and even early visual cortex, enables reliable generalization across independent visual inputs (faces depicting the 'eyes only' versus 'eyes removed'). Furthermore, classification performance was correlated to behavioral performance in STS and dPFC. Our results demonstrate that both higher (e.g. STS, dPFC) and lower level cortical regions contain information useful for facial expression decoding that go beyond the visual information presented, and implicate a key role for contextual mechanisms such as cortical feedback in facial expression perception under challenging conditions of visual occlusion.

Introduction

Facial affect recognition (FAR) is a critical component of healthy social cognition in humans (Adolphs, 2003). Impairments in recognition and related neural dysfunction are found in several disorders associated with socio-affective dysfunction, including psychopathy and conduct disorder (Dadds et al., 2006; Contreras-Rodriguez et al., 2014), mood disorders (Surguladze et al., 2004), and autism (Swartz et al., 2013). Prominent neurocognitive models of FAR focus both on core face regions such as fusiform face area (FFA), superior temporal sulcus (STS), and occipital face area (OFA; Ishai, 2008; Park et al., 2012), and on the extended face network, which includes the amygdala, and areas of prefrontal cortex such as inferior frontal gyrus (IFG) and dorsal prefrontal cortex (dPFC; Stein et al., 2007; Dal Monte et al., 2013; Ferrari et al., 2016). While previous studies have found increased activity in ventral visual cortex & STS for emotional versus neutral faces (Vuilleumier et al., 2001; Engell and Haxby, 2007; Fusar-Poli et al., 2009), recent studies using MVPA have addressed whether the information present in particular regions can discriminate **particular** facial expression categories (Wegrzyn et al., 2015; Zhang et al., 2016). Wegrzyn et al (2015), for instance, revealed that facial expression category can be read out from multiple regions (e.g. amygdala, STS, fusiform gyrus and inferior occipital gyrus). However, while some of these studies have investigated the invariance of expression representations to differing identities (Zhang et al., 2016), none have investigated whether the evoked representation of facial expressions generalizes across the specific parts of the face used to signal it (e.g. eyes Vs mouth; cf. Anzellotti and Caramazza, 2016). Such invariance across visual features would be useful because the ability to recognize a facial expression is important even when specific facial features are occluded (e.g. in the presence of sun glasses or scarf; cf. Tang et al., 2014).

Moreover, while regions in the extended face network may be contributing unique information to FAR, it is also likely that they are providing feedback to earlier areas, particularly under conditions of partial

occlusion (O'Reilly et al., 2013; Tang et al., 2014). Such information may feed-back even to the earliest visual areas (Clark, 2013; Muckli et al., 2013). For example, recent work has revealed that early visual regions contain information about occluded parts of visual scenes (Smith and Muckli, 2010; Muckli et al., 2015). Similarly, we showed that retinotopically mapped sub-regions (eye & mouth) of V1 changed activity as a function of the specific face categorization task at hand (Petro et al., 2013).

Behavioral and patient work has demonstrated that different features of the face are diagnostic for particular expression categorizations (e.g. the eyes in fear; Adolphs et al., 2005; Smith et al., 2005; Smith and Schyns, 2009), which might lead one to predict that generalization across face features is not possible. On the other hand, recurrent or top-down connections within the visual system may allow for an effective filling in of missing information from occluded visual stimuli (Smith and Muckli, 2010; Clark, 2013; O'Reilly et al., 2013; Tang et al., 2014) and hence lead to such generalizations. The present study tests these competing predictions by measuring whether representations within key nodes of the core and extended face networks contain measurably similar activity to independent samples of face information: expressions revealing the eyes only versus eyes removed (Han et al., 2012).

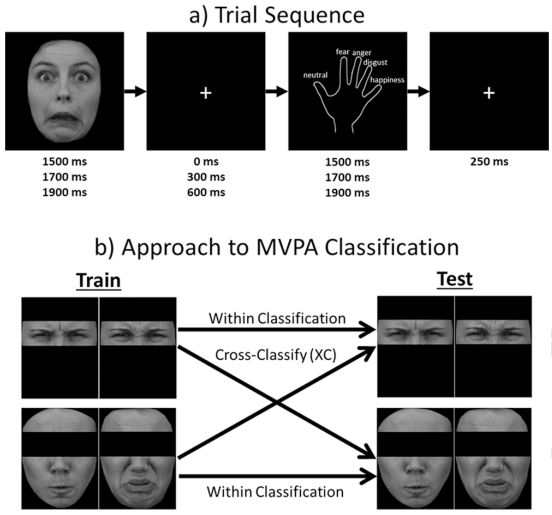


Figure 1: Exemplars of Stimuli, Timing of Experiment, and Analytic Approach

Top (a): Trial sequence used in the fMRI experiment. Bottom (b): Logic of the MVPA Classification analyses used in the present analyses, depicting both within-condition classification of expression and cross-classification across different parts of the face (i.e., train on 'eyes only' data, test on 'eyes removed').

Methods

Participants and experimental task

We analyzed data from 19 randomly selected participants who completed the 'partial face encoding' task, which has been described in detail previously (Han et al., 2012). The data presented in the current work were collected in the context of a larger study examining the relationship between individual differences in distinct facets of empathy and social cognition. With the exception of data from 7 subjects (part of the sample presented in Han et al., 2012), none of the data presented here have been published

elsewhere. Briefly, while in an fMRI scanner, participants were asked to label the facial expression shown (i.e., angry, disgusted, fearful, happy, neutral) of either a 'whole face', a face with 'eyes only', or a face with 'eyes removed'. Facial stimuli were taken from 36 actors (18 female, 18 male) in the Karolinska Directed Emotional Faces set (Lundqvist et al., 1998), grey scaled, and had their hair and ears cropped to remove potentially distracting idiosyncratic stylistic features. Participants viewed each facial expression the same number of times, and each actor included displayed each of the five facial expressions. Each trial consisted of an initial fixation cross (1200ms), a face, or partial face, presentation (1500, 1700, or 1900 ms), followed by a fixation inter-stimulus interval (0, 300, or 600 ms), a response screen (1500, 1700, or 1900 ms), and a fixation inter-trial interval (250 ms). Additionally, twenty fixation-only trials (10 of 2500ms and 10 of 3000ms) were included in each run **(Figure 1)**. This trial structure and design was used to maximize the accuracy of the regression analysis given the event related nature of the task. Participants completed six runs, each run consisting of 90 images, balanced for emotion, face sample (i.e., whole faces, eyes only, or eyes removed), and actor's gender. **Specifically, there are 5 emotional labels one can assigned to a given trial type and there are 15 trial types (i.e., combinations of emotion and face parts; see Supplementary Figure 1).** Run order and emotion-response finger designations were randomized across participants. Prior to beginning the experimental runs, all participants completed two practice runs, which trained participants on which button corresponded to which emotion and acclimatized them to the pace of the task.

Behavioural Analysis

To assess participants' FAR accuracy we first conducted a 5 (Emotion: angry, disgusted, fearful, happy, neutral) by 3 (Face parts: 'whole face', 'eyes only', 'eyes removed') repeated measures ANOVA on accuracy for identifying the viewed emotion, followed by Bonferroni-adjusted *t*-tests. Second, we quantified the percentage of responses participants made that were both 'hits' (e.g., labeling an angry

face as angry) and 'false alarms' when labeling the faces (e.g., labeling an angry face as disgust) relative to chance responding (i.e., 1/5, or 20%). **Specifically, there are 5 emotional labels one can assigned to a given trial type and there are 15 trial types (i.e., combinations of emotion and face parts; see Supplementary Figure 1)**. This analysis resulted in 75 one-sample t-tests and was corrected for multiple comparisons using the Holm-Bonferroni correction (Holm, 1979; Gaetano, 2013). This analytic approach is consistent with the approach taken in Han et al. (2011), which focused only on recognition accuracy and not reaction times, as accuracy of recognition was the focus of the study and emphasized to participants.

MRI acquisition

The experimental task was completed at the Centre for Metabolic Mapping, in the Robarts Research Institute's 3T Siemens scanner equipped with a 32-channel head coil. The session began with a high-resolution anatomical scan that covered the whole brain (time to repetition=2,300 ms, time to echo=4.25 ms; Field of View=25.6 cm; 192 axial slices; voxel dimensions=1mm isovoxels; 256x256mm matrix). Participants completed six functional MRI runs during which blood-oxygenation-level-dependent (BOLD) changes were measured using a T2*-gradient echo-planar sequence (EPI; time to repetition=3000 ms, time to echo=30 ms; 120x120mm matrix; field of view=24 cm; flip angle = 90°). One-hundred-forty-seven volumes were collected per run, resulting in run durations of 7.35 minutes. Complete brain coverage was obtained with 45 interleaved slices of 2mm by 2mm in plane and a slice thickness of 2.5mm (forming voxels of 2x2x2.5mm).

MRI data processing

Standard individual preprocessing was carried out using Analysis of Functional NeuroImages software (Cox, 1996). After removing the first four volumes of each run, the functional data was slice time

corrected, and motion corrected by registration to the first volume of the first run, and transformed into the standard space of Talairach and Tournoux. No spatial smoothing was applied. The time series of each voxel was scaled such that the coefficient produced by the regression analysis represented the percent signal change from the mean voxel activity. A first-level general linear model regression was performed including a single regressor for each trial of each condition of interest. Regressors of no interest were also included to avoid potential confounds, these included modelling the response epoch for each condition as a single regressor and modelling volumes with excessive motion (where the Euclidean Norm of the motion derivative exceeded 1.0).

All regressors were produced by convolving the trial onsets with the gamma-variate hemodynamic response function. To account for voxel-wise correlated drifting, baseline plus linear drift and quadratic trend were also modeled. This produced trial-by-trial beta coefficients for each of our **conditions of interest**, which were then used in the MVPA analyses described below. Additionally, a first-level contrast of all whole-face conditions (i.e, irrespective of facial expression) versus baseline was carried out in each individual. The resulting t-values were used in the feature selection step described below.

Region-of-interest Definition

A probabilistic V1 mask was defined from the cytoarchitectonic map in the Juelich Anatomy toolbox in FSL (Eickhoff et al., 2005), and was thresholded at 50%. Face-related ROIs were generated using the automated meta-analysis toolbox Neurosynth (www.neurosynth.org, Yarkoni et al., 2011) from a meta-analysis of previous fMRI studies that frequently use the word 'face' (i.e., all studies in which the word "face" was mentioned more than once per 1000 words were included). We generated our ROI mask by identifying brain regions that were significantly active given the use of the word "face" (i.e., the 'forward

inference' analysis). In order to isolate the ROIs and avoid larger clusters that incorporated multiple regions of interest, we thresholded the image at a z-value of 5 and removed voxels corresponding to early visual cortices. In this manner in addition to the V1 mask we produced four bilateral ROIs corresponding to the lateral occipital/ventral temporal cortex (LO/VT), Inferior frontal gyrus (IFG), dorsal prefrontal cortex (dPFC: including dlPFC and dmPFC), and amygdala. The LO/VT mask also included right STS, thus we isolated right STS from LO/VT; however, the resulting right STS mask was <350 voxels. Furthermore, there was a left STS cluster that was <15 voxels. In order to derive an independent bilateral ROI of the STS we dilated the edges of the right STS cluster using a 2mm kernel which produced a resulting mask of 1,304 voxels. Finally, we mirrored the right STS mask in the left hemisphere and subtracted the LO/VT mask to ensure independence. Note that the final left STS mask included all voxels that were in the original small left STS cluster. The frontal regions above contribute to the extended face network (Ishai, 2008; Park et al., 2012), and have each been identified as contributing to certain aspects of facial affect recognition (Stein et al., 2007; Dal Monte et al., 2013; Ferrari et al., 2016). Notably, the brain regions identified using the 'reverse inference' in Neurosynth produced similar ROIs in the LO/VT, STS and amygdala regions, but did not include the prefrontal cortex regions.

Feature selection

We selected the top 1,000 voxels per ROI per participant for inclusion in the MVPA analysis based on their t-value from the first-level within-subject analyses of all whole-faces versus baseline, with only those voxels with positive t-values being included. This decision was based on a preliminary within-condition classification analysis of whole-face facial expressions on data from a broad early visual cortex mask, which demonstrated that classification accuracy reached an asymptote at ~1,000 features (i.e.,

voxels) in the model. Thus, for all analyses we used 1,000 voxels per bilateral ROI per participant. Lastly, we produced 'heat maps' for each bilateral ROI to display the number of participants for which a given voxel was included in each classification analysis. In order to confirm that our effects were not dependent on voxel number, we repeated the analyses with 400 voxels per ROI (see Wegryzn et al 2016), and note that these analyses produced near identical classification results.

Data analysis

Within-subject pattern classification analysis (Figure 1). Linear support vector machine classification with a leave-one-run-out cross-validation scheme was carried out within each participant (Smith and Muckli, 2010; Smith and Goodale, 2015). This involved iteratively training the SVM classifier on data from **five** runs and testing on the held out, independent, **sixth** run. This produced **six** folds of training and testing, in which each run served as the testing set exactly one time. In this manner multiclass classification on the five facial expressions was carried out four times for each ROI. The first three classification analyses involved within-condition classification such that it was carried out independently on each of the three visual categories (whole faces, eyes only, and eyes removed). Although results in early visual regions produced by this analysis could be driven by basic visual features, it was important for the non-visual ROIs and provided an upper bound to our classification accuracy. To test our main questions, we performed the fourth classification analysis, which involved cross-classification such that an SVM model was trained on the pattern of brain activity elicited by facial expression of eyes only or eyes removed, and tested on data elicited by eyes removed or eyes only, respectively (Smith and Muckli, 2010; Vetter et al., 2014; Kaplan et al., 2015). Thus, **while the low-level features in the training versus testing data set are spatially independent, the presented images still share the same high order expression category.** The accuracy of these two cross-classifications were averaged together to produce one cross-classification (XC) accuracy score. It should be noted that for the cross-classification

analysis we also used the leave-one-run-out cross-validation scheme for training and testing the model to ensure independence between the training and testing set. For each multiclass classification analysis we always used all five facial expressions, thus chance performance was expected to be ~20%. We indeed verified that chance performance was ~20% using permutation testing by repeating the within subject leave-one-run-out cross validation with permuted data labels in each run. This also confirmed there was no systematic bias in our classification.

The linear SVM algorithm was implemented using the LIBSVM toolbox (Chang and Lin 2011), with default parameters (notably C=1). Note that the activity of each voxel in the training data was normalized within a range of -1 to 1 prior to input to the SVM. The test data were normalized using the same parameters (min, max) as obtained from the training set normalization in order to optimize the classification performance (Smith and Muckli, 2010; Chang and Lin, 2011; Vetter et al., 2014; Smith and Goodale, 2015). The five-alternative forced choice (5AFC) classification task is implemented by LIBSVM as a series of 1 Vs 1 binary comparisons (10 here) that allows for the generation of a full confusion matrix, similar to that obtained from experiments on human facial expression categorization. LIBSVM by default uses a non-optimal solution for dealing with ties in multiclass classification when the confusion matrix (and not just accuracy) is important (simply choosing the first class defined during training). To overcome this limitation, when ties were present we randomly chose one of the five possible categories, and repeated the whole cross-validation procedure fifty times in order to better reflect the nature of the errors the classifier typically made.

Group-wise analysis. Independently for each of the classification analyses, classification performance was averaged across each fold for inclusion in the group analysis. Significance was determined for each

classification analysis in each ROI by comparing the classifier accuracy to the chance level using a non-parametric Wilcoxon signed rank test (two tailed) followed by FDR correction ($q<0.05$) to protect against type one errors due to repeated testing across multiple ROIs. Moreover, given the nature of multiclass classification, we also produced 5x5 confusion matrices for each of the four main classification analyses (i.e., within whole faces, within eyes only, within eyes removed, and cross-classification). Confusion matrices allow for the quantification of the classifiers accuracy for identifying a target facial expression and the probability that each of the other four facial expressions is mistakenly labelled as the target emotion. **As we had no a priori predictions regarding the contribution of specific expressions to classifier performance, we performed *post hoc* analyses comparing the hit rate of each specific expression to chance (i.e., the top-left to bottom-right diagonal elements of the confusion matrices) using non-parametric Wilcoxon signed rank test (two tailed). This identifies which specific facial expressions are discriminated above chance.**

Relationship between brain activity and behaviour. We sought to determine whether there was a relationship between decoding accuracy in the critical cross-classification case (eyes only to minus eyes or vice versa) with the behavioural categorization performance of our participants. We were specifically interested in the cross-classification case as, if significant, this would presumably reflect a high-level effect as there is no overlap in the low level visual information present in the stimuli. Decoding effects found within each face condition, on the other hand, may rely on a strong contribution from low level features (see our Computational Model Analysis section below and in results). Thus, we first averaged participants' behavioural accuracy across the two relevant conditions (eyes only and eyes removed) that underlie the key cross-decoding analysis. **We then Spearman correlated this with mean cross-classification decoding performance, to protect against the influence of outliers on the standard Pearson correlation coefficient (Pernet et al., 2012). As a further test of the strength of these**

correlations, we also re-ran our correlations with the robust 'skipped correlation' Pearson coefficient measure (see Pernet et al., 2012). We used the Robust Correlation toolbox (Pernet et al., 2012) to implement this test and note here that as it only computes whether a correlation is significant or not at alpha = .05 using bootstrapped confidence intervals, multiple comparisons correction could not be applied.

Computational Model Analysis

To explore the extent to which low level image features could account for our cross-classification results of interest, we applied the same linear classifier and trained it to decode expression category based on the image pixel values. We trained the classifier to decode expression both within each face condition (eyes only, eyes removed and whole faces) and also on the key cross-classification analysis: i.e. where the classifier is trained on one partial face condition (i.e. eyes only) and tested on the complementary case (i.e. rest of face minus the eyes) or vice versa. We used a leave one identity out procedure as is standard in computational models of facial expression perception (see e.g. Dailey et al., 2002 and Susskind et al., 2007; for related approaches) with analyses based on down-sampled images (1/4 of original size). We also conducted a more elaborate analysis using a V1 Gabor model (see Mutch and Lowe, 2006; Serre et al., 2007) which produced comparable results to the simpler analysis described above and as with the simpler analysis produced no significant decoding in the key cross-classification analysis. In the results section, we present the simpler image pixel analyses which do not require the multiple specification of parameters that are required for the V1 Gabor model.

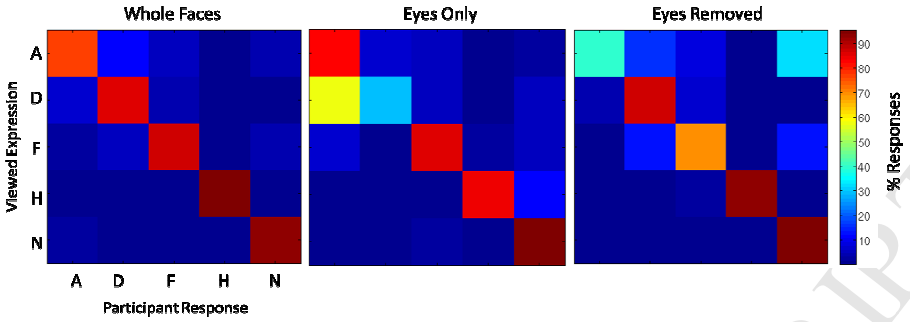


Figure 2: Behavioural results for facial affect recognition accuracy.

Participants' behavioural response rates **are displayed as** confusion matrices for each category of face parts: (Left) Whole Faces, (Middle) Eyes Only, and (Right) Eyes Removed. For each individual confusion matrix, the facial expression that was presented is represented by each row and the participant response by each column [A = anger, D = disgust, F = fear, H = happy, N = neutral]. The scale represents the percentage of a particular response by the participant during the presentation of a specific facial expression. The diagonal, moving from top-left to bottom-right, of each matrix represents the correct response rate ('hits') and all other cells in the matrix represent 'false alarms'.

Results

Behaviour

We first quantified participants' pattern of FAR using a 5 (Emotion) x 3 (Face Part) repeated measures ANOVA (see figure 2). This **analysis revealed a** main effect of emotion, $F(4,72) = 62.46$, $p < .001$, $\eta 2 = .78$. Follow-up Bonferroni-corrected pairwise comparisons demonstrated that irrespective of face presentation condition, participants were most accurate at identifying neutral faces compared to every

other emotion, angry ($p_{corrected}$<.001), disgust ($p_{corrected}$<.001), fear ($p_{corrected}$<.001), and happy ($p_{corrected}$<.05). Accuracy for happy was greater than angry ($p_{corrected}$<.001), disgust ($p_{corrected}$<.001), and fear ($p_{corrected}$<.001); Accuracy for fear was great than for angry ($p_{corrected}$<.001) and disgust ($p_{corrected}$<.005); and there was no difference in accuracy for angry versus disgust ($p_{corrected}$>.05). The ANOVA also revealed a main effect of face presentation condition, F(2,36) = 83.28, p < .001, η2= .82. Bonferroni-corrected pairwise comparisons demonstrated that irrespective of emotion participants were more accurate identifying emotions from the whole face compared to both the eyes only ($p_{corrected}$<.001) and eyes removed ($p_{corrected}$<.001) conditions. There was no difference in FAR accuracy between the eyes only and eyes removed conditions ($p_{corrected}$>.05). Lastly, the **ANOVA revealed** a significant emotion by face presentation condition interaction, F(8,144) = 91.45, p < 0.001, η2= .84.

To unpack this interaction, we focused on all pairwise comparisons between facial expressions within a given face presentation condition (e.g., accuracy for angry versus disgust within the whole face category) and all pairwise comparisons between face presentation conditions within a given emotional expression (e.g., accuracy for angry faces recognition for 'eyes only' versus 'eyes removed'). This produced 45 pairwise comparisons, which we evaluated against a Bonferroni correct alpha value of 0.05/45 (i.e., ~0.0011). For 'whole face' trials we found that compared to angry faces recognition of happy, $t(18) = 5.61$ ($p < .001$), and neutral faces, $t(18) = 4.90$ ($p < .001$), was significantly greater . For 'eyes only' trials accuracy for recognizing disgust was quite poor, as recognition of anger, $t(18) = 9.58$, fear, $t(18) = 12.34$, happy, $t(18) = 13.41$, and neutral, $t(18) = 14.20$, were all significantly greater than disgust recognition ($p < .001$). Also for 'eyes only' trials we found that recognition of neutral faces was significantly greater than both anger, $t(18) = 5.18$, and fear, $t(18) = 4.87$, recognition ($p < .001$). For 'eyes removed' trials recognition accuracy of anger was very poor, as recognition of disgust, $t(18) = 10.28$, fear, $t(18) = 7.14$, happy, $t(18) = 13.48$, and neutral, $t(18) = 13.72$, were all significantly greater than

anger recognition ($p < .001$). Additionally for 'eyes removed' trials, recognition of disgust, $t(18) = 3.98$, happy, $t(18) = 7.12$, and neutral, $t(18) = 8.06$, were all significantly greater than recognition of fear ($p < .001$), and recognition of neutral was significantly greater than disgust, $t(18) = 5.12$ ($p < .001$). On the other hand, within each emotion we found that accuracy for anger 'whole face', $t(18) = 14.61$, and anger 'eyes only', $t(18) = 10.22$, was significantly higher than for anger 'eyes removed' ($p < .001$); accuracy for disgust 'whole face', $t(18) = 15.00$, and disgust 'eyes removed', $t(18) = 11.77$, was significantly higher than for disgust 'eyes only' ($p < .001$); accuracy for fear 'whole face', $t(18) = 7.77$, and fear 'eyes only', $t(18) = 3.96$, was significantly higher than for fear 'eyes removed' ($p < .001$); accuracy for happy 'whole face', $t(18) = 5.12$, and happy 'eyes removed', $t(18) = 4.10$, was significantly higher than for happy 'eyes only' ($p < .001$); and, there were no significant differences in recognition accuracy for neutral expressions across the face part trial types.

As a follow-up, exploratory analysis, we performed *post hoc* exploratory analysis using one sample t-tests on the accuracy for each individual trial type (i.e., 75 one-tailed tests, including 'hits' and 'false alarms') versus chance (i.e., 1/5 or 0.2% accuracy). These analyses were corrected for multiple comparisons (Holm, 1979; Gaetano, 2013). Of note, on 'eyes only' trials participants labeled disgust eyes as angry eyes significantly above chance, but did not identified disgust eyes as disgust significantly better than chance. Additionally, on eyes removed trials, participants labelled angry faces as angry, as was expected, however, they also labelled angry faces as neutral on a proportion of trials significantly greater than chance. Otherwise, all the other significant effects were participants correctly labeling the target emotion.

Brain Imaging

We defined regions of interest for six areas we expected to be involved in neural processing and representation of facial expressions (Early Visual Cortex, VT/LO, STS, dPFC, IFG, and Amygdala) in the partial faces task (see Methods). Figure 3 (A-D, 1st column) shows the sampling density of the group of subjects for the EVC, VT/LO, STS, and dPFC ROIs, respectively, which is indicative of how likely a given voxel was selected for use in the MVPA analysis across participants. Figure 3 (A-D, 2nd column) shows the decoding results for each ROI. Additionally, results for the IFG and amygdala are not included in the figure but are reported in text below. To assess whether decoding was greater than chance (1/5 or 20%) in each region we performed two-tailed non-parametric Wilcoxon signed rank tests. Correction for multiple testing was made across all regions and all comparisons, using the false discovery rate ($q < .05$).

MVPA Within condition classification: As expected, decoding of facial expressions was highly significant for each face presentation condition (whole faces, eyes only, eyes removed) in multiple ROIs. This was true in early visual cortex (Med = 33.9%, 33.2% and 29.3% respectively; all *p*'s < .0002; chance = 20%, figure 3a, 2nd column) , LO/VT (Med = 24.6%, 23.4%, and 23.5% respectively for whole faces, eyes only and eyes removed; all *p's* < .0029, figure 3b, 2nd column) and in STS (Med = 23.8%, 23.0%, and 21.8% respectively for whole faces, eyes only and eyes removed; all *p's* < .0025. figure 3c, 2nd column). Furthermore, both dPFC (Med = 28.5%, 24.1%, and 26.5% respectively for whole faces, eyes only and eyes removed; all *p's* < .0013, figure 3d, 2nd column) and IFG (Med = 24.1%, 23.8%, and 23% respectively for whole faces, eyes only and eyes removed; all *p's* < .0033) also showed reliable decoding of facial expression within each face condition. The amygdala, however, did not show reliable decoding in any of the abovementioned conditions (Med = 19.9%, 19.9%, and 20.1% respectively for whole faces, eyes only and eyes removed; all *p's* > .52), and therefore the amygdala was excluded from the cross-classification analyses reported below. Figure 4, columns 1-3, shows the full confusion matrices underlying the multiclass classification results for each of EVC, LO/VT, STS, and dPFC. ***Post hoc* examination of the**

confusion matrices reveals variation of classifier performance across the five facial expressions. We report both the uncorrected and FDR corrected results of these contrasts in full in Figure 4, columns 1-3.

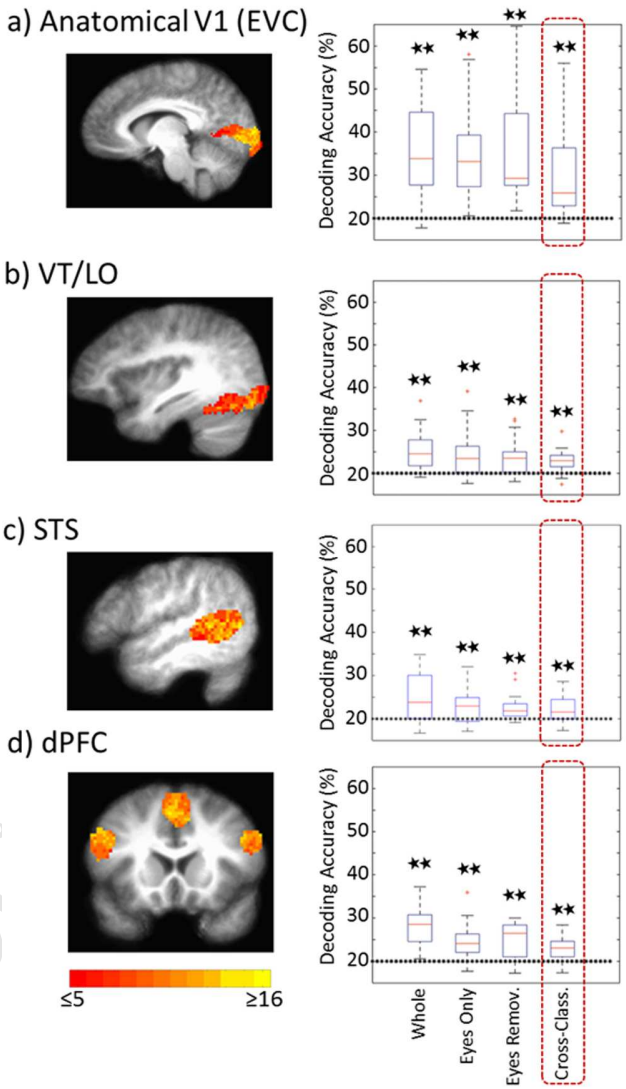


Figure 3: Within- and Cross-Condition Decoding Results

A) Anatomical V1 (EVC): Density map (First Column) showing the likelihood of each selected voxel being present across participants. Second column: Decoding results showing median accuracy for each classification analysis performed (chance = 20%, stars indicate significant after FDR correction) are displayed using box plots. The critical cross-classification results are highlighted by the red-dashed box, and represent the median classifier accuracy (across participants) for decoding expression averaged across the two possible train-test permutations: training on 'eyes only' and testing on 'eyes removed', and training on 'eyes removed' and testing on 'eyes only'. **Rows B-D, as in A, but for VT/LO, STS and dPFC regions respectively.**

MVPA Cross-classification: It is important to note that the within condition decoding of expression may rely on low level features to a significant extent (see Low Level Feature Analysis below). Hence, we were particularly interested in testing whether significant cross-classification would be present generalizing across completely non-overlapping sets of facial features (i.e. trained on eyes only then tested on eyes removed, and trained on eyes removed then tested on eyes only). We found such an effect in STS (Med = 22%, p = .008, figure 3c, 2nd column), VT/LO (Med = 23%, p = .0008, figure 3b, 2nd column), dPFC (Med = 23%, p = .001, figure 3d, 2nd column) and even early visual cortex (Med = 26%, p = .0002, figure 3a, 2nd column). Thus, the activity patterns generated in these regions to independent parts of a face stimulus go beyond merely encoding the statistics of the visual input, suggesting a key role for higher-level context in shaping the representations present. This effect, however, did not reach significance in IFG (Med = 21%, p = 0.11). Figure 4, column 4, shows the full confusion matrices underlying the multiclass cross-classification results for each of EVC, LO/VT, STS, and dPFC. **We also performed *Post hoc* analyses comparing the hit rate of the classifier performance to chance so as to ascertain which expressions,**

considered independently, achieve significant decoding. Both uncorrected and FDR corrected results of these contrasts are reported in Figure 4, column 4.

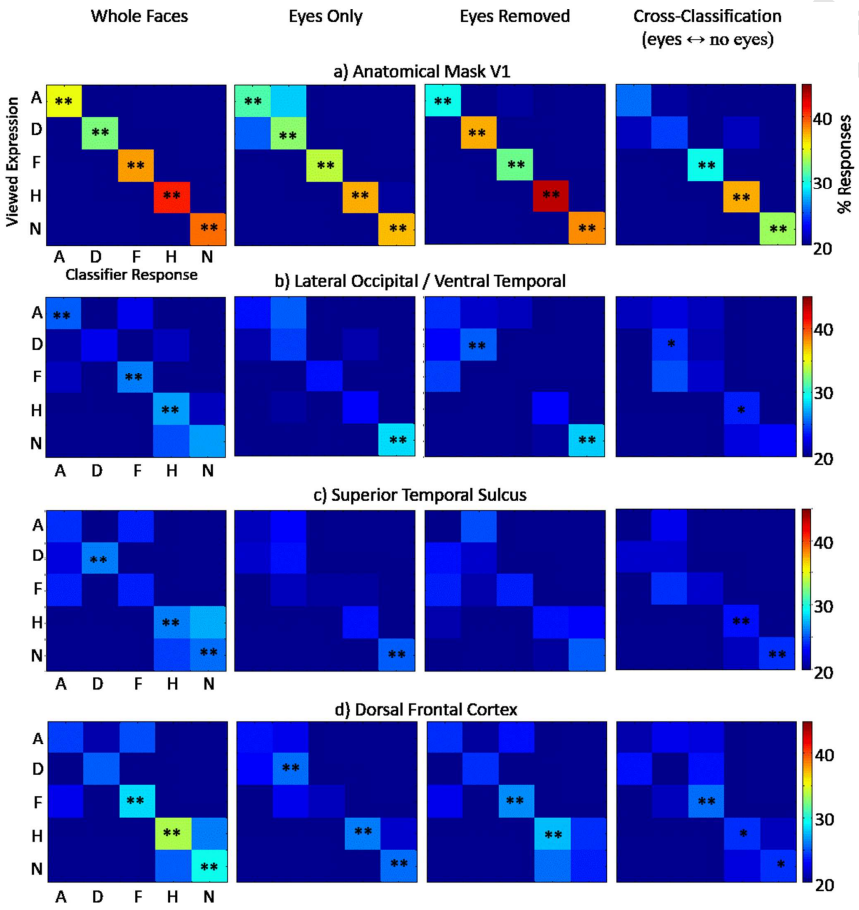


Figure 4: Confusion Matrices Underlying Classification

Full Confusion matrices underlying the main classification analyses. Rows (A-D) of the figure depict region of interest, EVC, LO/VT, STS, and DPFC, respectively. From left to right, the columns depict a particular classification analysis: 'whole faces', 'eyes only', 'eyes removed', and cross-classification. For each individual confusion matrix, the facial expression that was

presented is represented by each row and the response the classifier chose by each column [A = anger, D = disgust, F = fear, H = happy, N = neutral]. The scale represents the frequency of a particular response by the classifier during the presentation of a specific facial expression. The diagonal, moving from top-left to bottom-right, of each matrix represents the correct response rate. ***Post hoc*** **analysis of the classifier hit rate for specific emotions (i.e., the top-left to bottom-right diagonal elements of the matrices) to chance using non-parametric Wilcoxon signed rank test (two tailed) reveals the expressions contributing the most to the overall classification accuracy: ** *p < .05 FDR corrected (q < .05); * p < .05 uncorrected. The FDR correction was applied independently to the within condition classification and the cross-classification.***

Low level Feature Analysis

To what extent might the cross-decoding observed above, depend on low level visual features? We expect any contribution of this to be minimal given the absence of visual overlap between the key conditions for each face image (i.e. eye region Vs minus eye region stimulus). To assess this empirically, we trained a classifier to decode expression based on the image pixel values (see Methods). Crucially while a decoder built on image statistics could robustly decode expression within each partial face condition (Med = 60% whole faces; Med = 40% eyes only and Med = 60% eyes removed; all *p's* <= .0015; signed rank two tailed test, chance 20%), it provided no evidence whatsoever of being able to decode expression in the key cross-classification analyses (Med = 19%, *p* = 0.5; signed rank two-tailed test, chance = 20%). Hence low-level statistics cannot explain the successful cross-decoding of expression across complementary sets of visual information demonstrated above (i.e. eyes to minus eyes). In

addition, we note that taking expression decoding of whole faces as a baseline, then the reduction in performance to the key cross-decoding analysis is *largest* in early visual cortex (approximately 8%) and much smaller in higher order visual regions such as STS and LOC/VT (2%, and 1.6% respectively). This suggests that these higher visual regions may have more generalizable representations and again speaks against a low-level feature explanation of our cross-decoding effect in early visual cortex.

Brain-behavior relationship: Finally, we wanted to test whether the decoding accuracy observed in the critical cross-classification analysis was related to participants' behavioural categorization accuracy. If contextual mechanisms generate information going beyond the bottom-up stimulus information present, and that information is useful for behavior, then we would expect to observe such a relationship, perhaps even extending back to areas early in the visual processing stream via feedback. We computed Spearman correlations (FDR corrected for number of regions) between the decoding performance for the critical cross-classification analysis and the mean behavioral performance averaged across the two key partial face conditions, eyes only and eyes removed (see Methods). Crucially, significant correlations (FDR corrected for multiple comparisons) were found in STS (rS = .66, p = .002, figure 5a) and in dPFC (rS = .54, p = .017, figure 5b). These correlations are depicted in Figure 5a-b. The same pattern was found in IFG (rS = .60, p = .007, Figure 5c) but after FDR multiple-comparisons correction no significant effects were found in either ventral temporal/lateral occipital cortex (rS=.47, p = .044) or in early visual cortex (rS = .475, p = .04), see **Supplementary Figure 2**. **For corroboration, we re-ran our correlation analyses using the robust and powerful 'skipped correlation' measure (see Pernet et al., 2012 & methods). Note that correction for multiple comparison is not available at present for this measure (see Methods). These analyses, however, revealed a broadly similar pattern of results: i.e. high and significant (uncorrected) correlations in STS (r=.63, 95% CI [.23, .84], dPFC**

(r=.62, 95% CI [.11, .87]), IFG (r=.64, 95% CI [.17, .86]) and V1 (r=.5, 95% CI [.07, .76]) but not LO/VT (r=.34, 95% CI [-.11, .73]). The results of these analyses suggest that in several higher tier brain areas, high-level contextual information is consistently related to, and hence may facilitate, behavioural categorization performance.

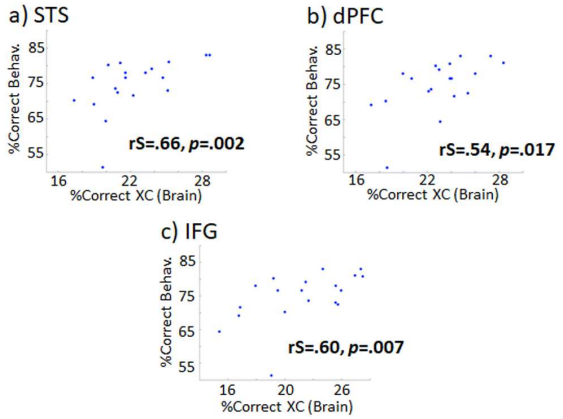


Figure 5: Brain-behavior Spearman correlations of cross-classification performance against behaviour.

The spearman correlation between cross-classification performance on the x-axis and participants behavioural accuracy for identifying facial expressions from partial faces on the y-axis is displayed for (a) STS, (b) dPFC, **and (c) IFG**.

Discussion

The present findings demonstrate that a network of areas important in face categorization contain information about facial expression categories, namely EVC, LO/VT, STS, dPFC and IFG, with all areas except IFG permitting decoding of facial expression category across non-overlapping samples of visual information. This suggests that these regions contain information relevant to FAR, in a manner

that generalizes, to a degree, beyond the specific visual information provided. We further show that the magnitude of this effect is related to behavioral performance in an expression categorization task, in STS and dPFC.

Spatial generalization of expression decoding

We note that while the magnitude of some within condition and cross-classification effects were relatively small in percentage terms, they were statistically reliable. In fact, using traditional effect size measures (Cohen's d) all of our significant effects (including the cross-classification case) produced moderate to large effect sizes: i.e. our smallest effect size was d = 0.69 (whereas d > 0.80 is conventionally taken as evidence for a large effect, d=0.5 is a moderate effect size). Our classification results were also robust to trial-level noise which can undermine trial-level analyses. Similar small but significant effects have been observed in similar trial-level classification studies involving within-category classification of faces and emotions (Harry et al., 2013; Skerry and Saxe, 2014; Anzellotti and Caramazza, 2016). Hence, we interpret the current results as providing evidence that STS, LO/VT, and dPFC contain high-level representations of facial expression that generalize *to a degree* across changes in the visual input.

Representation of Facial Expression in Face Sensitive Regions

Reliable decoding of facial expression category was found in a network of brain areas important in face processing tasks: LO/VT, STS, dPFC, and IFG all showed decoding of facial expressions within each face presentation condition (whole faces, eyes only or rest of face minus eyes). STS has often been proposed as the key region in representing *dynamic* facial expression of emotion (Haxby et al., 2000; Furl et al., 2013), although some previous studies have also found evidence of facial expression sensitivity in ventral temporal regions (Tsuchiya et al., 2008; Kawasaki et al., 2012; Harry et al., 2013;

Wegrzyn et al., 2015). In the present study, we did find the highest correlation with behavior in the STS (see also Said & Haxby, 2010), but we also found expression information present within VT/LO regions. Wegrzyn et al. (2015) also found reliable decoding of expression in both fusiform and STS regions, as well as the amygdala, while Zhang et al. (2016) in contrast, found reliable decoding only within the amygdala (Fear Vs other categories) and pSTS (neutral Vs other categories) but not within fusiform regions. Importantly Zhang et al.'s analyses involved generalizing across identity, which may provide a reason for the lack of fusiform decoding observed in their study. While an earlier study, Said et al. (2010), revealed reliable decoding of each basic category in both anterior and posterior STS (anatomically defined) with dynamic stimuli and a correlation with perceptual ratings (pSTS), a more recent study that attempted to decode expressions from the combination of constituent facial action units, failed to find reliable expression category decoding in STS (Srinivasan et al., 2016). Hence, while several studies, including our own, have found decoding within the STS for static or dynamic faces, which specific expressions underlie successful decoding is likely to be dependent on several factors: Whether dynamic or static faces were used; which particular expressions were classified (as certain expressions are more readily confused with other expressions, see Smith & Schyns, 2009); and, whether the analyses attempted to generalize either across identity of the faces, across different subsets of visual information, or across different sensory modalities.

Crucially, in the present work we found that face responsive STS & LO/VT cortex contain representations of facial expression that generalize to a degree across independent samples of visual information from a face (see Anzelotti & Caramazza, 2015, for a similar case with respect to representation of identity in the anterior temporal lobe). We further reveal that the representations present are related to human *categorization* performance under these challenging conditions in STS. As mentioned earlier, recurrent or top-down connections within the face network may allow for

partial reactivation of missing information from occluded visual stimuli (Smith and Muckli, 2010; Clark, 2013; O'Reilly et al., 2013; Tang et al., 2014) and hence permit such a generalization to be possible.

Previous research, furthermore, has provided evidence for supramodal representations of emotion using cross-classification across different sensory or stimulus modalities. Peelen et al. (2010) found evidence for the cross-classification of emotional expressions between facial, bodily, and vocal stimuli in both STS and parts of the anterior-dorsomedial PFC. In the present study, it could certainly be the case that our partial face stimuli are tapping into a similar high-level representation in these regions. However, our study shows evidence of such representation also in VT/LO region, which may not necessarily be expected when generalizing across sensory modalities. Recently, Skerry and Saxe (2014) found that while the STS does contain representations that generalize across variations in facial expression stimuli (e.g. variable positions and variable lighting conditions) that allow for the decoding of pleasant versus unpleasant expressions, it does not generalize across more conceptually disparate categories (such as positive and negative social situations of non-human agents).

Many previous MVPA studies, in addition, did not consider the role of frontal regions in FAR, primarily as the focus was on perceptual representations (cf., Said et al., 2010). In the present work, we show reliable decoding both within IFG and dPFC regions under explicit processing conditions (with the latter also showing the key cross-decoding effect). This is consistent with the meta-analysis by Fusar-Poli et al. (2009), which found that explicit FAR consistently recruits dorsal and lateral aspects of PFC. Given the potential top-down nature of the PFE task, especially on trials in which facial expressions are judged based on little diagnostic information (e.g., fear on ‘eyes removed' trials), we interpret the dPFC decoding consistent with its role in selective attention and cognitive control. Indeed, a recent study using the PFE paradigm found greater activity in the dPFC in low-callous trait individuals when judging

fearful faces in the 'eyes removed' relative to the 'eyes only' condition (Han et al., 2012). Moreover, our explicit FAR task required active maintenance of participants' behavioural choice from the choice encoding to the response phase, which is expected to require dorsal and lateral PFC regions involved in the representation of such higher-order information (D'Esposito and Postle, 2015).

While the current study did not find significant decoding of facial expressions in the amygdala, unlike previous studies (Wegrzyn et al., 2015; Zhang et al., 2016) we used an explicit emotion recognition task. While the lack of decoding within the amygdala indicates a lack of a perceptual or abstract representations that distinguish whole or partial facial expressions, it does not rule out the possibility that the amygdala makes a general contribution to FAR (Adolphs, 2010). Evidence indicates that the amygdala contributes to the allocation of attention towards biologically relevant stimuli (Gamer et al., 2013; Peck et al., 2013), rather than discriminating between emotions per se (cf. Wegrzyn et al., 2015; Zhang et al., 2016). The amygdala also has efferent connections throughout the ventral visual cortex (Amaral, 2002) that contribute to emotion-induced visual enhancement (Pessoa et al., 2002; Vuilleumier et al., 2004; Amting et al., 2010). A complimentary alternative is that the amygdala facilitates attention selectivity via indirect connections to the lateral PFC that do not involve the ventral visual pathway (Mohanty et al., 2009; Zikopoulos and Barbas, 2012; Pessoa, 2013; Greening and Mitchell, 2015). Thus, the amygdala could contribute to neural filling-in within perceptual regions or attentional reorienting in the absence of significant neural decoding of facial expressions.

Early Visual Cortex contains reliably similar information across independent visual inputs

The present study agrees with many others that show how high-level contextual influences can shape processing within the earlier regions of sensory cortex, even in a content-specific manner (e.g. Meyer et

al., 2010; Smith and Muckli, 2010; Man et al., 2012; Muckli et al., 2015). Our results demonstrate that reliably similar information about facial expression category is present in EVC across spatially independent samples of visual information. What type of mechanism might explain the present results? As there is no visual information in common across the two critical conditions (eyes only and eyes removed), low level stimulus based processing does not provide a straightforward explanation of the present findings (we note this is in contrast to decoding observed within a specific face presentation condition). For example, a typical V1 model (e.g. HMAX layer C1, Serre et al., 2007) would not contain any differential expressive information in common between the critical conditions. Indeed our analysis of low level features explicitly demonstrated that while a decoder trained on image pixels could robustly decode expression within each face condition, it was exactly at chance in the key cross-decoding analysis. Hence our results support the interpretation that cortical feedback or lateral interactions transmit contextual information relating to the expression category of the occluded parts of the face stimulus. Cortical feedback, for instance, could arise from higher face processing areas (e.g. FFA; see Strother et al., 2011) or areas within the PFC, as observed during mental imagery of faces (Kim et al., 2007; Diekhof et al., 2011). PFC regions have been proposed to generate high-level expectations (e.g., task instructions, or goals) about perceptual stimuli (e.g., Bar, 2003; Summerfield et al., 2006; Trapp and Bar, 2015). Summerfield et al. (2006) also found such effects for faces within the FFA and again, in the present study, our decoding analyses revealed generalization across visual information in face specific regions in VT/LO. Thus, feedback from sectors of the PFC to higher-level face areas in the ventral stream, could feedback expectation of missing features to early visual cortex. Thus, we tentatively suggest that feedback from PFC likely via STS/FFA to EVC may aid perception of occluded facial expressions.

Wider Implications

FAR deficits are found in several disorders; however, the neurocognitive mechanisms of this deficit appear to vary across disorders. In a number of clinical cases, including neurological patients with amygdala lesions (Adolphs et al., 2005; Gamer et al., 2013) and youth with conduct disorder (Dadds et al., 2006) directing attention to the most diagnostically meaningful regions of a face reduces the FAR deficit. This highlights the importance of selective attention in FAR (Gamer and Buchel, 2009; Han et al., 2012). These findings are consistent with our observation that dPFC was a significant predictor of behavioral accuracy in FAR for partial faces. Conversely, FAR deficits in frontotemporal dementia cannot be improved by manipulating selective attention (Oliver et al., 2015), suggesting the importance of a different mechanism. Selective attention may also provide top-down influence on representations in VT & STS, which is not possible if neural dysfunction is present in aspects of the temporal-occipital regions (Virani et al., 2013). Nevertheless, our observation of cross-classification from differing face parts could also suggest that partial faces generate simulations of facial affect, particularly within parts of the IFG and lateral PFC (Carr et al., 2003; Hennenlotter et al., 2005; Jabbi and Keysers, 2008). Such processes may be sufficient for driving our cross-classification results, though they are deficient is some clinical populations. Future research employing cross-classification in clinical populations could help clarify the mechanisms contributing to FAR deficits, and to socio-emotional deficits more generally.

References

Adolphs R (2003) Cognitive neuroscience of human social behaviour. Nat Rev Neurosci 4:165-178.

Adolphs R (2010) What does the amygdala contribute to social cognition? Ann N Y Acad Sci 1191:42-61.

Adolphs R, Gosselin F, Buchanan TW, Tranel D, Schyns P, Damasio AR (2005) A mechanism for impaired fear recognition after amygdala damage. Nature 433:68-72.

Amaral DG (2002) The primate amygdala and the neurobiology of social behavior: implications for understanding social anxiety. Biol Psychiatry 51:11-17.

Amting JM, Greening SG, Mitchell DG (2010) Multiple mechanisms of consciousness: the neural correlates of emotional awareness. J Neurosci 30:10039-10047.

Anzellotti S, Caramazza A (2016) From Parts to Identity: Invariance and Sensitivity of Face Representations to Different Face Halves. Cereb Cortex 26:1900-1909.

Bar M (2003) A cortical mechanism for triggering top-down facilitation in visual object recognition. J Cogn Neurosci 15:600-609.

Carr L, Iacoboni M, Dubeau MC, Mazziotta JC, Lenzi GL (2003) Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. Proc Natl Acad Sci U S A 100:5497-5502.

Chang CC, Lin CJ (2011) LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:27:21-27:27.

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181-204.

Contreras-Rodriguez O, Pujol J, Batalla I, Harrison BJ, Bosque J, Ibern-Regas I, Hernandez-Ribas R, Soriano-Mas C, Deus J, Lopez-Sola M, Pifarre J, Menchon JM, Cardoner N (2014) Disrupted neural processing of emotional faces in psychopathy. Soc Cogn Affect Neurosci 9:505-512.

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162-173.

D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. Annu Rev Psychol 66:115-142.

Dadds MR, Perry Y, Hawes DJ, Merz S, Riddell AC, Haines DJ, Solak E, Abeygunawardane AI (2006) Attention to the eyes and fear-recognition deficits in child psychopathy. Br J Psychiatry 189:280-281.

Dal Monte O, Krueger F, Solomon JM, Schintu S, Knutson KM, Strenziok M, Pardini M, Leopold A, Raymont V, Grafman J (2013) A voxel-based lesion study on facial emotion recognition after penetrating brain injury. Soc Cogn Affect Neurosci 8:632-639.

Diekhof EK, Kipshagen HE, Falkai P, Dechent P, Baudewig J, Gruber O (2011) The power of imagination--how anticipatory mental imagery alters perceptual processing of fearful facial expressions. Neuroimage 54:1703-1714.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25:1325-1335.

Engell AD, Haxby JV (2007) Facial expression and gaze-direction in human superior temporal sulcus. Neuropsychologia 45:3234-3241.

Ferrari C, Lega C, Vernice M, Tamietto M, Mende-Siedlecki P, Vecchi T, Todorov A, Cattaneo Z (2016) The Dorsomedial Prefrontal Cortex Plays a Causal Role in Integrating Social Impressions from Faces and Verbal Descriptions. Cereb Cortex 26:156-165.

Furl N, Henson RN, Friston KJ, Calder AJ (2013) Top-down control of visual responses to fear by the amygdala. J Neurosci 33:17435-17443.

Fusar-Poli P, Placentino A, Carletti F, Landi P, Allen P, Surguladze S, Benedetti F, Abbamonte M, Gasparotti R, Barale F, Perez J, McGuire P, Politi P (2009) Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. J Psychiatry Neurosci 34:418-432.

Gaetano J (2013) Holm-Bonferroni Sequential Correction: An EXCEL Calculator. In.

Gamer M, Buchel C (2009) Amygdala activation predicts gaze toward fearful eyes. J Neurosci 29:9123-9126.

Gamer M, Schmitz AK, Tittgemeyer M, Schilbach L (2013) The human amygdala drives reflexive orienting towards facial features. Curr Biol 23:R917-918.

Greening SG, Mitchell DG (2015) A network of amygdala connections predict individual differences in trait anxiety. Hum Brain Mapp 36:4819-4830.

Han T, Alders GL, Greening SG, Neufeld RW, Mitchell DG (2012) Do fearful eyes activate empathy-related brain regions in individuals with callous traits? Soc Cogn Affect Neurosci 7:958-968.

Harry B, Williams MA, Davis C, Kim J (2013) Emotional expressions evoke a differential response in the fusiform face area. Front Hum Neurosci 7:692.

Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4:223-233.

Hennenlotter A, Schroeder U, Erhard P, Castrop F, Haslinger B, Stoecker D, Lange KW, Ceballos-Baumann AO (2005) A common neural basis for receptive and expressive communication of pleasant facial affect. Neuroimage 26:581-591.

Holm S (1979) A simple sequential rejective method procedure. Scandinavian Journal of Statistics 6:65-70.

Ishai A (2008) Let's face it: it's a cortical network. Neuroimage 40:415-419.

Jabbi M, Keysers C (2008) Inferior frontal gyrus activity triggers anterior insula response to emotional facial expressions. Emotion 8:775-780.

Kaplan JT, Man K, Greening SG (2015) Multivariate cross-classification: applying machine learning techniques to characterize abstraction in neural representations. Front Hum Neurosci 9:151.

Kawasaki H, Tsuchiya N, Kovach CK, Nourski KV, Oya H, Howard MA, Adolphs R (2012) Processing of facial emotion in the human fusiform gyrus. J Cogn Neurosci 24:1358-1370.

Kim SE, Kim JW, Kim JJ, Jeong BS, Choi EA, Jeong YG, Kim JH, Ku J, Ki SW (2007) The neural mechanism of imagining facial affective expression. Brain Res 1145:128-137.

Lundqvist D, Flykt A, Ohman A (1998) The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Man K, Kaplan JT, Damasio A, Meyer K (2012) Sight and sound converge to form modality-invariant representations in temporoparietal cortex. J Neurosci 32:16629-16636.

Meyer K, Kaplan JT, Essex R, Webber C, Damasio H, Damasio A (2010) Predicting visual stimuli on the basis of activity in auditory cortices. Nat Neurosci 13:667-668.

Mohanty A, Egner T, Monti JM, Mesulam MM (2009) Search for a threatening target triggers limbic guidance of spatial attention. J Neurosci 29:10563-10572.

Muckli L, Petro LS, Smith FW (2013) Backwards is the way forward: feedback in the cortical hierarchy predicts the expected future. Behav Brain Sci 36:221.

Muckli L, De Martino F, Vizioli L, Petro LS, Smith FW, Ugurbil K, Goebel R, Yacoub E (2015) Contextual Feedback to Superficial Layers of V1. Curr Biol 25:2690-2695.

Mutch J, Lowe DG (2006) Multiclass Object Recognition with Sparse, Localized Features. IEEE Conference on Computer Vision and Pattern Recognition (CVPR):11-18.

O'Reilly RC, Wyatte D, Herd S, Mingus B, Jilk DJ (2013) Recurrent Processing during Object Recognition. Front Psychol 4:124.

Oliver LD, Mitchell DG, Dziobek I, MacKinley J, Coleman K, Rankin KP, Finger EC (2015) Parsing cognitive and emotional empathy deficits for negative and positive stimuli in frontotemporal dementia. Neuropsychologia 67:14-26.

Park J, Carp J, Kennedy KM, Rodrigue KM, Bischof GN, Huang CM, Rieck JR, Polk TA, Park DC (2012) Neural broadening or neural attenuation? Investigating age-related dedifferentiation in the face network in a large lifespan sample. J Neurosci 32:2154-2158.

Peck CJ, Lau B, Salzman CD (2013) The primate amygdala combines information about space and value. Nat Neurosci 16:340-348.

Peelen MV, Atkinson AP, Vuilleumier P (2010) Supramodal representations of perceived emotions in the human brain. Journal of Neuroscience 30:10127-10134.

Pernet CR, Wilcox R, Rousselet GA (2012) Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox. Frontiers in psychology 3.

Pessoa L (2013) Chapter 7: Dual Competition Model. In: The cognitive-emotional brain: From interactions to integration. Cambridge, MA: Massachusetts Institute of Technology Press.

Pessoa L, McKenna M, Gutierrez E, Ungerleider LG (2002) Neural processing of emotional faces requires attention. Proc Natl Acad Sci U S A 99:11458-11463.

Petro LS, Smith FW, Schyns PG, Muckli L (2013) Decoding face categories in diagnostic subregions of primary visual cortex. Eur J Neurosci 37:1130-1139.

Said CP, Moore CD, Engell AD, Todorov A, Haxby JV (2010) Distributed representations of dynamic facial expressions in the superior temporal sulcus. J Vis 10:11.

Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci U S A 104:6424-6429.

Skerry AE, Saxe R (2014) A common neural code for perceived and inferred emotion. Journal of Neuroscience 34:15997-16008.

Smith FW, Schyns PG (2009) Smile through your fear and sadness: transmitting and identifying facial expression signals over a range of viewing distances. Psychol Sci 20:1202-1208.

Smith FW, Muckli L (2010) Nonstimulated early visual areas carry information about surrounding context. Proc Natl Acad Sci U S A 107:20099-20103.

Smith FW, Goodale MA (2015) Decoding visual object categories in early somatosensory cortex. Cereb Cortex 25:1020-1031.

Smith ML, Cottrell GW, Gosselin F, Schyns PG (2005) Transmitting and decoding facial expressions. Psychol Sci 16:184-189.

Srinivasan R, Golomb JD, Martinez AM (2016) A neural basis of facial action recognition in humans. Journal of Neuroscience 36:4434-4442.

Stein JL, Wiedholz LM, Bassett DS, Weinberger DR, Zink CF, Mattay VS, Meyer-Lindenberg A (2007) A validated network of effective amygdala connectivity. Neuroimage 36:736-745.

Strother L, Mathuranath PS, Aldcroft A, Lavell C, Goodale MA, Vilis T (2011) Face inversion reduces the persistence of global form and its neural correlates. PLoS One 6:e18705.

Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J (2006) Predictive codes for forthcoming perception in the frontal cortex. Science 314:1311-1314.

Surguladze SA, Young AW, Senior C, Brebion G, Travis MJ, Phillips ML (2004) Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. Neuropsychology 18:212-218.

Swartz JR, Wiggins JL, Carrasco M, Lord C, Monk CS (2013) Amygdala habituation and prefrontal functional connectivity in youth with autism spectrum disorders. J Am Acad Child Adolesc Psychiatry 52:84-93.

Tang H, Buia C, Madhavan R, Crone NE, Madsen JR, Anderson WS, Kreiman G (2014) Spatiotemporal dynamics underlying object completion in human ventral visual cortex. Neuron 83:736-748.

Trapp S, Bar M (2015) Prediction, context, and competition in visual recognition. Ann N Y Acad Sci 1339:190-198.

Tsuchiya N, Kawasaki H, Oya H, Howard MA, 3rd, Adolphs R (2008) Decoding face information in time, frequency and space from direct intracranial recordings of the human brain. PLoS One 3:e3892.

Vetter P, Smith FW, Muckli L (2014) Decoding sound and imagery content in early visual cortex. Curr Biol 24:1256-1262.

Virani K, Jesso S, Kertesz A, Mitchell D, Finger E (2013) Functional neural correlates of emotional expression processing deficits in behavioural variant frontotemporal dementia. J Psychiatry Neurosci 38:174-182.

Vuilleumier P, Richardson MP, Armony JL, Driver J, Dolan RJ (2004) Distant influences of amygdala lesion on visual cortical activation during emotional face processing. Nat Neurosci 7:1271-1278.

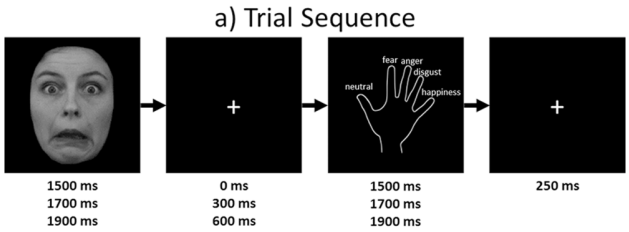
Vuilleumier P, Sagiv N, Hazeltine E, Poldrack RA, Swick D, Rafal RD, Gabrieli JDE (2001) Neural fate of seen and unseen faces in visuospatial neglect: A combined event-related functional MRI and event-related potential study. P Natl Acad Sci USA 98:3495-3500.

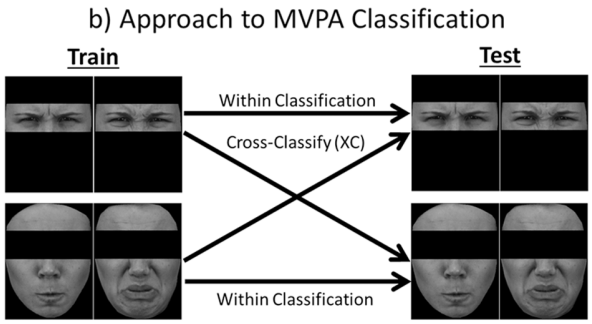
Wegrzyn M, Riehle M, Labudda K, Woermann F, Baumgartner F, Pollmann S, Bien CG, Kissler J (2015) Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. Cortex 69:131-140.

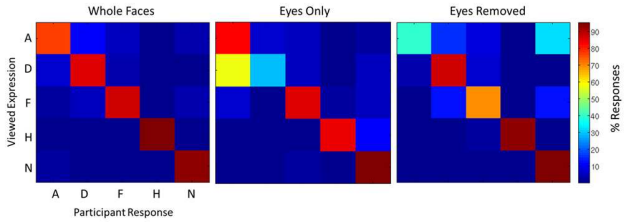
Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8:665-670.

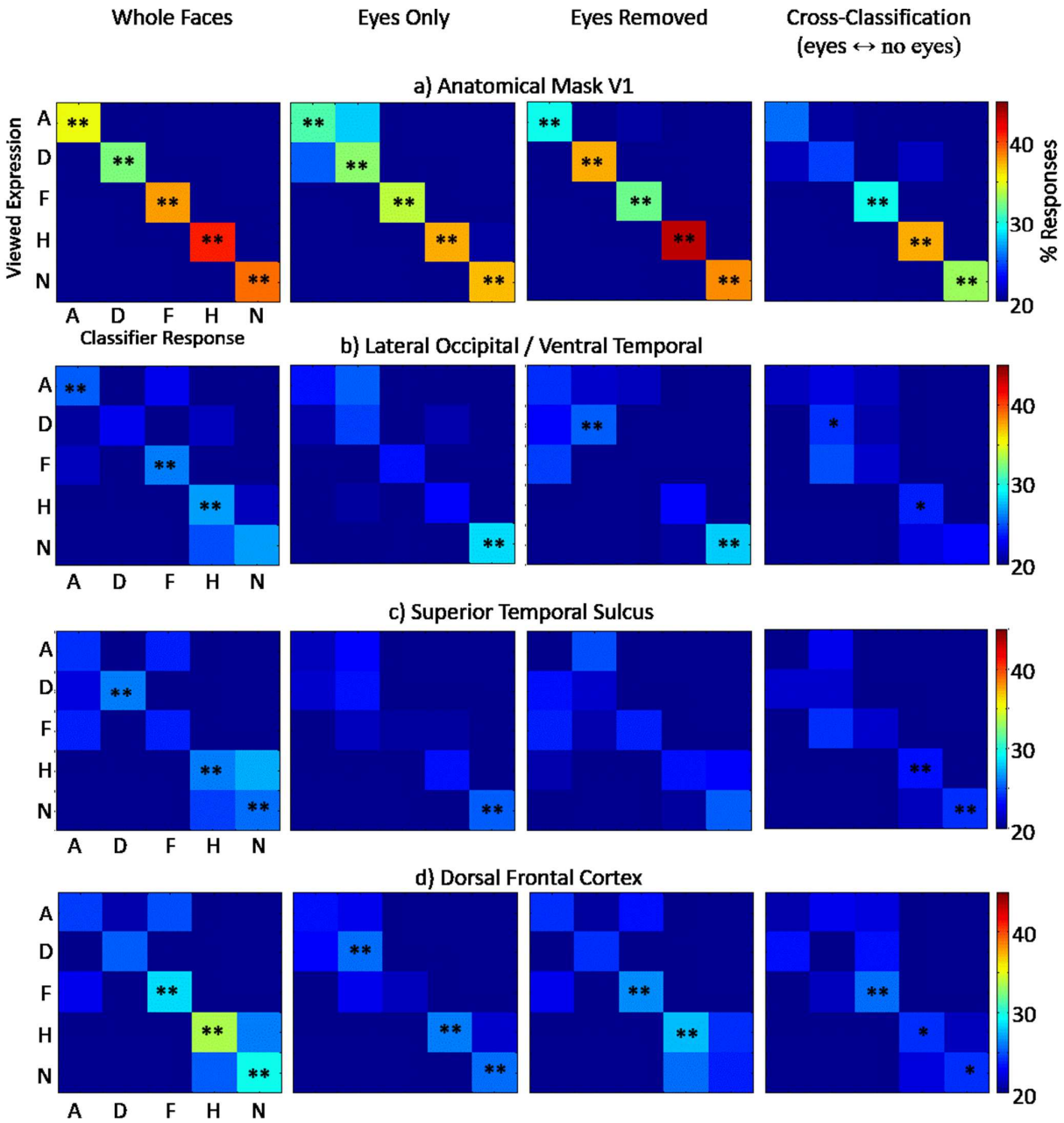
Zhang H, Japee S, Nolan R, Chu C, Liu N, Ungerleider LG (2016) Face-selective regions differ in their ability to classify facial expressions. Neuroimage 130:77-90.

Zikopoulos B, Barbas H (2012) Pathways for emotions and attention converge on the thalamic reticular nucleus in primates. J Neurosci 32:5338-5350.








Whole Faces
Eyes Only
Eyes Removed
Cross-Classification
(eyes ↔ no eyes)
a) Anatomical Mask V1
b) Lateral Occipital / Ventral Temporal
c) Superior Temporal Sulcus
d) Dorsal Frontal Cortex
Viewed Expression
Classifier Response
A D F H N
% Responses
40
30
20