MODELLING THE PERCEPTUAL SIMILARITY OF FACIAL EXPRESSIONS FROM IMAGE STATISTICS AND NEURAL RESPONSES

Mladen Sormaz, David M. Watson, William A.P. Smith, Andrew W. Young and Timothy J. Andrews*

Department of Psychology and York Neuroimaging Centre,

University of York, York YO10 5DD United Kingdom

*Corresponding author: timothy.andrews@york.ac.uk

Abstract: 170

Introduction: 754

Discussion: 965

Figures: 4

Table: 1

Keywords: MVPA, face, expression, STS, FFA

Acknowledgments: This work was supported by a grant from the Wellcome Trust (WT087720MA).

ABSTRACT

The ability to perceive facial expressions of emotion is essential for effective social communication. We investigated how the perception of facial expression emerges from the image properties that convey this important social signal, and how neural responses in face-selective brain regions might track these properties. To do this, we measured the perceptual similarity between expressions of basic emotions, and investigated how this is reflected in image measures and in the neural response of different face-selective regions. We show that the perceptual similarity of different facial expressions (fear, anger, disgust, sadness, happiness) can be predicted by both surface and feature shape information in the image. Using block design fMRI, we found that the perceptual similarity of expressions could also be predicted from the patterns of neural response in the face-selective posterior superior temporal sulcus (STS), but not in the fusiform face area (FFA). These results show that the perception of facial expression is dependent on the shape and surface properties of the image and on the activity of specific face-selective regions.

INTRODUCTION

The ability to visually encode changes in facial musculature that reflect emotional state is essential for effective social communication (Ekman, 1972; Bruce & Young, 2012). A full understanding of the mechanisms that underpin the perception of facial expression requires understanding both the way in which these processes are driven by visual properties of the image and the way in which different brain regions are involved (Haxby, Hoffman, & Gobbini, 2000; Bruce & Young, 2012).

Any facial image consists of a set of edges created by abrupt changes in reflectance that define the shapes and positions of facial features and a broader pattern of reflectance based on the surface properties of the face, also known as the albedo or texture (Bruce & Young, 1998, 2012). Shape can be defined by the spatial location of fiducial points that correspond to key features of the face. In contrast, surface properties reflect the reflectance of light that is caused by pigmentation and shape from shading cues. Shape and surface properties have both been proposed to contribute to the perception of identity and expression (Bruce & Young, 1998; Calder, Young, Perrett, Etcoff, & Rowland, 1996), but with the perception of familiar identity being relatively dominated by surface cues (Burton et al., 2005; Russell & Sinha, 2007) and feature shapes being relatively dominant in perceiving facial expressions (McKelvie, 1973; Etcoff & Magee, 1992; Butler, Oruc, Fox, & Barton, 2008). This differential use of image properties in the perception of identity and expression is consistent with models of face perception which propose that they are processed independently (Bruce & Young, 1998, 2012; Haxby, Hoffman, & Gobbini, 2000).

Support for the critical role of shape information in the perception of facial expression is found in studies that show manipulations of the image that degrade surface information, but leave shape information intact, have little impact on perceptual and neural

responses to facial expression (Bruce & Young, 1998; Magnussen, Sunde, & Dyrnes, 1994; White, 2001; Pallett & Meng, 2013; Harris, Young, & Andrews, 2014). Similarly, image manipulations that completely remove surface information, such as line drawings of faces, also show relatively preserved expression perception (McKelvie, 1973; Etcoff & Magee, 1992).

Although previous studies have suggested that feature shape is the dominant cue for the perception of facial expressions, there is some evidence to suggest that surface information may also play a role. Calder, Burton, Miller, Young & Amakatsu (2001) found that Principal Components (PCs) that convey variation in surface information could be used to categorize different facial expressions, albeit to a lesser extent than PCs that convey variation in shape. More recently, Benton (2009) found a decrease in the emotional expression aftereffect to facial expressions when images were negated, suggesting that the perception of facial expression can be affected by changes in surface information. So, it remains uncertain how different image properties contribute to the perception of facial expression.

The first aim of this study was therefore to explore the relative importance of shape and surface properties to the perception of facial expression. Specifically, we asked whether the perceptual similarity of different facial expressions could be predicted by corresponding similarities in the shape or surface properties of the image. The perceptual similarity task involved rating the degree of similarity in expression between pairs of pictures of facial expressions. This task was used to generate a matrix of perceived (rated) similarities between exemplars of facial expressions of five basic emotions. This is equivalent to the procedure used to establish widely-adopted perceptual models such as Russell's circumplex (Russell, 1980), where expressions of emotion lie proximally or distally on a two-dimensional

surface based on their perceived similarity, with the distance between expressions reflecting their similarity or confusability to human observers.

Our second aim was to determine if the perceptual similarity of facial expressions is reflected in the patterns of neural responses in face-selective regions of the brain. Neural models of face perception suggest that a network of face-selective brain regions underpins the perception of faces (Allison, Puce, & McCarthy, 2000; Haxby, Hoffman, & Gobbini, 2000; Ishai, 2008), with the posterior superior temporal sulcus (STS) playing a key role in processing facial expression (Winston, Henson, Fine-Goulden, & Dolan, 2004; Engell & Haxby, 2007; Harris, Young, & Andrews, 2012; Baseler, Harris, Young, & Andrews, 2014; Psalta, Young, Thompson, & Andrews, 2014). Recent evidence has shown that it is possible to successfully decode some properties of facial expressions from face responsive brain regions (Wegrzyn et al., 2015, Said, Moore, Engell, & Haxby, 2010). Nevertheless, the extent to which the neural response can predict the fine-grained perception of facial expression remains unclear. Using multi-voxel pattern analysis (MVPA) techniques, we asked whether the perceptual similarity of expressions could be explained by the neural response in different face-selective regions. Our prediction was that patterns of response in regions associated with processing of facial expression should predict the perception of facial expression.

METHODS

Participants

Twenty-four healthy volunteers took part in the fMRI experiment and the behavioural similarity ratings experiment (12 female, mean age = 25.2 years). All participants were right-handed and had normal or corrected to normal vision with no history of neurological illness. The fMRI work was approved and conducted following the guidelines of the York Neuroimaging Centre Research Ethics Committee, University of York, and the behavioural study by the Department of Psychology Ethics Committee. All participants gave written consent prior to their participation.

Stimuli

Figure 1 shows all the stimuli from the five expression conditions. Static images of expressions were presented as these are well-recognised as long as they represent the apex of the pattern of muscle movements involved in producing the expression (see Bruce & Young, 2012). By using well-validated images from the Radboud Face database (Langner et al., 2010) we ensured that this criterion was met. Images were selected on the basis of high recognisability of their facial expressions and the similarity of the action units (muscle groups) used to pose each of the expressions. Only male faces were used to avoid any confounds from characteristics introduced by gender differences in the images themselves. For each of five models, images of expressions of fear, anger, disgust, sadness and happiness were used.

Perceptual Similarity Experiment

First, we determined the perceptual similarity of different facial expressions. Participants carried out a perceptual similarity rating task. Pairs of images were presented either side of

a fixation cross and participants were asked to rate the images on the similarity of expression on a scale of 1-7 (1: not very similar expressions, 7: very similar expressions). Each possible combination of pairs of different images from the set of expressions was displayed once in the perceptual similarity rating experiment, excluding pairs of images from the same identity. This resulted in 200 trials in total. From these we were able to derive the average rated similarity between examples of expressions of same or different basic emotions. These similarity ratings were z-scored and then incorporated into a similarity matrix for each participant.

Image Properties

To determine whether the patterns of perceptual similarity found in our behavioural task could be explained by shape information in the face images, we defined the locations of 140 fiducial points corresponding to expressive features in each of the face images using PsychoMorph software (Tiddeman, Burt, & Perrett, 2001). This produced a 2 x 140 matrix for facial feature positions in 2D image space, with x and y co-ordinates for each fiducial point (Figure 2). These fiducial locations were then used to provide a measure of facial feature shape by entering the fiducial location matrices into a procrustean comparison (Schönemann, 1966) to measure the similarity in feature locations between every possible pair of images. The procrustean analysis rigidly aligns fiducial points allowing shape translation, rotation or scaling to correct for image position or size without morphing or non-linear image distortion. After alignment of a pair of images in this way, the procrustean metric computes the averaged squared distance between each pair of aligned points giving a value between 0-1. To create a similarity matrix, each value was subtracted from 1 and then z-scored.

We also calculated a surface measure of image differences that controlled for the position of the facial features in the image. To do this each of the 25 original images was reshaped (using a wavelet-based Markov random field sampling method) to the average shape across all 25 images (Tiddeman, Stirrat, & Perrett (2005). This removed any underlying shape cues to expression (as all images now shared exactly the same set of fiducial points), but left the surface information relatively unchanged. We then correlated the pixel values from the face for the same image pair combinations as for our procrustean analysis. These pixel correlations were transformed using Fisher's Z-transform. The values were z-scored to create an average surface similarity measure between each expression pairing.

fMRI experiment

To determine whether the patterns of perceptual similarity response in our behavioural task could be explained by patterns of response in face-selective regions, we measured the response in face-selective regions to different facial expressions. A block design was used with each block comprising a series of face images depicting one of the five expressions (fear, anger, disgust, sadness and happiness). Within each block, 5 images were each presented for 1 second followed by a 200 ms fixation cross, giving a block duration of 6s (Peirce, 2008). Stimulus blocks were separated by a fixation cross on a grey screen for 9s. Each condition was repeated eight times in a counterbalanced order, giving a total of 40 blocks. To minimise any influence of task effects on the patterns of neural response to expression, participants were not required to respond to the facial expressions during the fMRI scan. Instead, an irrelevant task of pressing a button when a red spot appeared was used to ensure that they paid attention to the stimuli without responding to their

expressions per se. A small red spot appeared on 1 or 2 images in each block and participants were instructed to press a response button whenever they saw the red spot. Participants correctly detected the red spot on over 90% of trials (mean accuracy = 95.3 $\pm$ 2 %, SD = 2).

Scanning was performed at the York Neuroimaging Centre at the University of York with a 3 Tesla HD MRI system with an eight channel phased array head coil (GE Signa Excite 3.0 T, High resolution brain array, MRI Devices Corp., Gainesville, FL). Axial images were acquired for functional and structural MRI scans. For fMRI scanning, echo-planar images were acquired using a T2*-weighted gradient echo sequence with blood oxygen level-dependent (BOLD) contrast (TR = 3 s, TE = 32.7 ms, flip-angle = 90°, acquisition matrix 128 x 128, field of view = 288 mm x 288 mm). Whole head volumes were acquired with 38 contiguous axial slices, each with an in-plane resolution of 2.25 mm x 2.25 mm and a slice thickness of 3 mm. T1-weighted images were acquired for each participant to provide high-resolution structural images using an Inversion Recovery (IR = 450 ms) prepared 3D-FSPGR (Fast Spoiled Gradient Echo) pulse sequence (TR = 7.8 s, TE = 3 ms, flip-angle = 20°, acquisition matrix = 256 x 256, field of view = 290 mm x 290 mm, in-plane resolution = 1.1 mm x 1.1 mm, slice thickness = 1 mm). To improve co-registration between fMRI and the 3D-FSPGR structural image a high resolution T1 FLAIR was acquired in the same orientation planes as the fMRI protocol (TR = 2850 ms, TE = 10 ms, acquisition matrix 256 x 224 interpolated to 512 giving effective in-plane resolution of 0.56 mm). First-level analysis of the facial expression scan was performed with FEAT v 5.98. The initial 9s of data were removed to reduce the effects of magnetic stimulation saturation. Motion correction (MCFLIRT, FSL) was applied followed by temporal high-pass filtering (Gaussian-weighted least-squares straight line fitting, sigma = 120s). Spatial smoothing (Gaussian) was applied

at 6 mm (FWHM). Individual participant data were entered into a higher-level group analysis using a mixed-effects design (FLAME, http://www.fmrib.ox.ac.uk/fsl). Parameter estimate maps were generated for each experimental condition; fear, anger, disgust, sadness and happiness. These maps were then registered to a high-resolution T1-anatomical image and then onto the standard MNI brain (ICBM152). Regions defined by the localiser scan were used to constrict MVPA analyses to face-responsive regions only.

To identify face-selective regions, data from a series of localizer scans with a different set of participants (n = 83) was used (Flack et al., 2014). The localizer scan included blocks of faces and scrambled faces. Images from each condition were presented in a blocked design with five images in each block. Each image was presented for 1 s followed by a 200-ms fixation cross. Individual participant data were entered into a higher-level group analysis using a mixed-effects design (FLAME, http://www.fmrib.ox.ac.uk/fsl). Face-responsive regions of interest were defined by the contrast of faces>scrambled faces at the group level and spatially normalised to an MNI152 standard brain template. The peak voxels for the OFA, FFA and STS in each hemisphere were determined from the resulting group statistical maps. Then the 500 voxels with the highest z-scores within each region were used to generate a mask. Masks were combined across hemispheres to generate 3 masks for the OFA, FFA and posterior STS, which form the core face-selective regions in Haxby et al's (2000) neural model (Supplementary Figure 1).

Parameter estimates in the main experimental scan to each expression were normalised independently in each voxel by subtracting the mean parameter estimate across all expressions and then registered onto the standard MNI152 brain. Pattern analyses were then performed using the correlation-based MVPA method devised by Haxby and colleagues (Haxby, Gobbini, Furey, Ishai, Schouten & Pietrini, 2001). After separating the data across

odd and even blocks for each participant (as was done by Haxby, et al., 2001), we determined the reliability of the patterns within participants by correlating patterns across odd and even runs for each condition. This procedure was performed 24 times (i.e. once for each participant) for each of the 15 possible combinations of basic emotions. The final correlation matrix provides a measure of the similarity in the pattern of response across different combinations of facial expressions. These neural correlations were transformed using Fisher's Z-transform and then converted into z scores.

Regression analyses

To then determine whether the pattern of perceptual similarity responses was best predicted by variance in facial shape or surface information, a linear regression analysis was performed using the similarity matrix for shape and surface analyses as independent regressors and the perceptual similarity rating correlation matrices from each individual as outcomes. Our linear regression method is similar to a Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, 2009) which can characterise the information carried by a given representation in behavioural response patterns, neural activity patterns or a representational model. By analyzing the correspondence between participant responses and neural response we can test and compare different models. For example if either the shape or surface regressors are able to explain a significant amount of the variance in the corresponding perceptual similarity rating matrices, the model regression coefficient can be expected to be significantly greater than zero. All regressor and outcome variables were Z-scored prior to the regression analysis. However, it is important to note that the similarity responses are not fully independent. The same method was used to measure similarity between predictor models based on neural response

patterns in OFA, FFS and STS regions and perceptual ratings of expression similarity as outcomes.

RESULTS

Perception of facial expression is predicted by shape and surface properties of the image

Figure 3 shows the average perceptual similarity scores for each of 15 possible combinations of facial expression across all participants. We then determined the extent to which perceptual similarity of facial expressions could be predicted by the normalized shape and surface properties of the image, by generating a corresponding similarity matrix for these image properties. The group averaged matrix for perception was significantly correlated with both shape ($r(15) = .61$, $p = .016$) and surface ($r(15) = .77$, $p < .001$) properties. In these analyses, images were normalized through rigid realignment of fiducial positions in the shape (procrustes) analysis and through a non-rigid transform to create fixed-shape images for the measure of surface similarity.

An important question concerns whether these transforms were necessary, or superfluous because the same characteristics were present in low-level properties of the untransformed images. A similar analysis with the raw images failed to show a significant relationship between perception and either shape ($r(15) = .27$, $p = .31$) and surface ($r(15) = .37$, $p = .16$) properties. This suggests that the mechanism underlying the perception of facial expression involves some form of equivalent normalization process.

To measure the reliability across participants, a regression analysis was performed in which the models derived from the shape or surface analyses were independently used as predictor variables and the perceptual similarity ratings matrices from each individual as outcomes (Kriegeskorte et al. 2008). First, we checked that our image property models (shape model and surface model) were not colinear. The variance inflation factor (VIF) value for the shape and surface models was 3.17, which does not exceed the recommended threshold of 5 (Montgomery, Peck, & Vining, 2012). The output of the regression analysis

shows that the perceptual similarity of the facial expression could be explained by both the shape ($F(1,358) = 178$, $\beta=.58$, $p<.001$) and the surface ($F(1,358) = 399.5$, $\beta=.73$, $p<.001$) properties in the images.

In Figure 3 it is clear that the perceptual similarity between expressions can in part be driven by high similarity ratings along the diagonal (where one fear expression is seen as very similar to another fear expression, and so on). We will refer to these as within-category comparisons. To determine the extent to which these within-category comparisons were responsible for the result of the regression analysis, we repeated the analysis with just the between-category (off-diagonal) comparisons, looking to see whether the pattern of perceptual similarities between different expressions might still be tracked by the image properties. Again, we found that the perceptual similarity of the expressions was significantly predicted by both the shape ($F(1,238) = 51.81$, $\beta=.42$, $p<.001$) and the surface ($F(1,238) = 61.14$, $\beta=.46$, $p<.001$) properties of the image, offering strong evidence of their importance.

Perception of facial expression is predicted by neural responses in face-selective regions

Figure 4 shows the average correlation matrix for expressions involving each of the 15 possible combinations of basic emotions in each of the core face-selective regions. To measure the reliability of the neural response to each facial expression, the data were analysed in each face responsive region with a 5 x 2 repeated measures ANOVA with Comparison (within-category, between-category) and Expression (Fear, Anger, Disgust, Sadness and Happiness) as factors. There was a significant main effect of Comparison in the STS ($F(1, 23) = 5.27$, $p=.03$) and OFA ($F(1, 23) = 6.45$, $p=.018$), but not in the FFA ($F(1, 23) = 0.067$, $p=.8$). This suggests that there are reliable patterns of response to facial expression

in STS and OFA. We did not find any effect of Expression (STS: $F(4, 92) = .59$, $p = .67$, OFA: $F(4, 92) = 1.2$, $p = .31$, FFA: $F(4, 92) = 1.38$, $p = .248$) or any interaction between Comparison and Expression (STS: $F(4, 92) = .77$, $p = .55$, OFA: $F(4, 92) = .94$, $p = .45$; FFA: $F(4, 92) = 1.32$, $p = .27$) in any of the core face-selective regions. This suggests that the ability to discriminate expressions was not driven by any specific expressions, but rather by a generalised ability to discriminate all patterns of neural response to expressions.

Next, we determined how the pattern of perceptual similarity might be linked to the patterns of response in different face-selective regions. We compared the similarity of patterns of response to different facial expressions in each face-selective region (see Fig. 4) with perceived similarity of the expressions (see Fig. 3). There was a significant correlation between perception and patterns of response in the STS ($r(15) = 0.62$, $p = .014$) and OFA ($r(15) = 0.67$, $p < .001$). However, there was no significant correlation between perception and patterns of neural response in the FFA ($r(15) = -0.08$, $p = .77$).

To measure the reliability across participants, a linear regression analysis was used with the neural responses in the different face responsive regions (OFA, FFA and posterior STS) responses as individual regressors and the perceptual similarity ratings matrices from each individual as the outcome. The perceptual similarity of the facial expressions could be predicted by neural response to facial expressions in STS ($F(1,358) = 181.2$, $\beta = .58$, $p < .001$) and OFA ($F(1,358) = 235.7$, $\beta = .63$, $p < .001$) regions but not in the FFA region ($F(1,358) = 2.11$, $\beta = -.08$, $p = .15$).

Again, one possible interpretation of these results is that they might be driven primarily by the higher within-condition compared to between-condition correlations. To determine if this was the case, we repeated the analysis only using the off-diagonal elements of the correlation matrices. As before, results showed that the perceptual

similarity of the facial expressions could be predicted by neural response to facial expressions in the STS ($F(1,238) = 7.18$, $\beta=.17$, $p<.008$) and OFA ($F(1,238) = 9.96$, $\beta=.25$, $p<.002$), but not in the FFA region ($F(1,238) = 1.5$, $\beta=.08$, $p = .22$).

To determine whether the patterns of response in the face-selective regions could be explained by the magnitude of response to different expressions, we performed a univariate analysis on each region of interest. Table 1 shows the % MR signal to each expression. In contrast to the MVPA, Table 1 shows that similar levels of activation were evident to all expressions within each region. A repeated measures ANOVA showed that there was an effect of Region ($F=63.0$, $p<0.0001$), which was due to lower responses in the STS. However, there was only a marginal effect of Expression ($F=2.38$, $p=0.073$) and a marginal interaction between Region and Expression ($F=2.11$, $p=0.066$). This marginal interaction likely reflects a relatively larger response to happiness compared to other expressions in the OFA and STS, but a relatively larger response to fear compared to other expressions in the FFA. It may also reflect the low response to sadness in the FFA but the high response to sadness in the OFA and STS.

Finally, we determined how the pattern of perceptual similarity might be linked to the patterns of response in regions outside the core face-selective regions. The localiser scan was able to define other face-selective regions in the inferior frontal gyrus (IFG), amygdala and precuneus, which are part of the extended face processing network. We compared the similarity of patterns of response to different facial expressions in each face-selective region with perceived similarity of the expressions. There was a significant correlation between perception and patterns of response in the IFG ($r (15) = 0.63$, $p =0.01$), but not in the amygdala ($r (15) = 0.28$, $p=0.31$) or precuneus ($r (15) = -0.25$, $p=0.37$). However, when only the between-category comparisons were measured we did not see any

significant correlations in any of these face regions (IFG: r (10) = -0.01, p = 0.97; amygdala: r (10) = -0.33, p = 0.23; precuneus: r (10) = 0.02, p = 0.94).

To determine whether regions outside the face-selective ROIs could also predict patterns of response to facial expression, we repeated the analysis using the Harvard Oxford anatomical masks (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases). First we asked whether there were distinct patterns of response to different facial expressions. From the 48 anatomical regions, only the Inferior Temporal Gyrus posterior (ITGp, F = 17.9, p<.001) and the Middle Temporal Gyrus posterior division (MTGp, F = 4.3, p=.048) showed distinct patterns (Suppl. Table 1). Next, we compared the similarity of patterns of response to different facial expressions in each region with perceived similarity of the expressions. In contrast to the face-selective ROIs, neither the ITGp (r (15) = 0.15, p = .59), the MTGp (r (15) = 0.48, p = .077) nor any other anatomical region showed a significant correlation between patterns of response and perceptual similarity.

DISCUSSION

Facial expressions are signalled by complex patterns of muscle movements that create changes in the appearance of the face. The aims of the present study were to determine how our perception of expression is linked to (1) the image properties of the face and (2) the neural responses in face-selective regions. Together, our findings show that the mechanisms that underpin the perception of facial expression are tightly linked to both shape and surface properties of the image and to the pattern of neural response in specific face-selective regions.

Our use of a measure of the perceptual similarity between expressions allows a more fine-grained analysis than the more standard method of categorizing each expression as one of the basic emotions (e.g. Mattavelli et al., 2013). Instead, we were able to track the magnitude of perceived differences between emotions, and to demonstrate that this pattern of between-category differences could still be modelled both from normalized image properties and from neural responses in STS. The fact that the link between image properties and perception was still evident when the within-category correlations (fear with fear, etc.) were removed from the analysis shows that the findings are not driven solely by the relatively high within-category relationships. Rather, it suggests a more continuous representation of facial expression involving a distinct between-category structure.

Different facial expressions can be defined by edge-based shape cues that result from changes in the shape of the internal features (Ekman, 1972; Bruce & Young, 1998, 2012). Previous studies have suggested that these shape cues are important for the perception of facial expressions (Bruce & Young, 1998; Magnussen et al., 1994; White, 2001; Harris et al., 2014). Although changes in facial expression also affect the surface properties of the face (Calder et al., 2001), this information has not been thought to be particularly

diagnostic for discriminating facial expression (Bruce & Young, 1998). In this study, we found that both the shape and surface properties correlated highly with perceptual judgements. So, while the present findings provide further support for the long held assertion that shape cues are important for the perception of expression, the novel finding from this study is that surface properties are as important. This usefulness of both types of cue may reflect the natural intercorrelation between shape and surface cues within many expressions. For example, fear expressions involve opening the mouth and widening the eyes (shape cues) and this creates salient contrast changes in the eye and mouth regions (surface cues).

Neuroimaging studies have previously revealed a number of regions that respond selectively to facial expression (Haxby et al., 2000; Allison et al., 2000). We found that the perceptual similarity of different facial expressions could be predicted by the similarity in the pattern of neural response in the OFA and STS. That is, facial expressions that were perceived as being similar had more similar neural patterns of response in these regions, which is of course consistent with Haxby et al.'s (2000) idea that they are important to the analysis of changeable aspects of faces such as expression. Our findings are also consistent with a recent study showing that patterns of neural response correlated with the perceptual similarity of dynamic facial expressions in the posterior superior temporal sulcus (Said, et al 2010). Indeed, the correspondence between perception and neural response in the superior temporal region is consistent with the role of this region in the perception of facial expression (Haxby et al., 2000; Winston et al., 2004; Engell & Haxby, 2007; Harris, Young & Andrews, 2012; Harris et al., 2014; Baseler, Harris, Young & Andrews , 2014; Pitcher, 2014; Psalta et al., 2014; Wegrzyn et al., 2015).

The OFA is thought to be the primary input area in the face processing network and has projections to both the STS and FFA (Haxby et al., 2000). However, more recently there

is evidence that face processing can occur in the absence of input through the OFA (Rossion et al., 2003). Our finding that the OFA can decode expression and contains representations of perceived similarity of these images suggests that it is involved in representing facial expression. This fits with other studies showing that the OFA adapts to facial expression (Fox et al., 2009) and that applying TMS to the OFA disrupts the perception of facial expression (Pitcher, 2014).

In contrast to the STS and OFA, patterns of response in the FFA did not predict the perception of facial expression. Although our findings are consistent with neural models that suggest that this region is important for the representation of relatively invariant facial characteristics associated with recognition of identity (Allison et al., 2000; Haxby et al., 2000), they contrast with more recent studies that have shown responses in the FFA can be linked to the perception of facial expression (Harry, Williams, Davis & Kim, 2013; Wegrzyn et al., 2015). One potentially crucial difference between our study and these previous studies is that they asked only whether patterns of response to different facial expressions were distinct. In our study, we addressed the more fine-grained question of whether the perceptual similarity of different facial expressions can be explained by the similarity in the patterns of neural response.

In conclusion, we show that perceptual patterns of response to facial expression are correlated with statistical properties of face images and with neural responses. We found that changes in both the shape and surface properties of the face predict perceptual responses to facial expression and that difference in the neural patterns of response in the STS, but not the FFA can also predict perceptual responses to facial expressions. Together, these results show the importance of image properties in understanding higher level

perceptual judgements and suggest that these factors may be an important organizing principle for the neural representations underlying the perception of facial expression.

REFERENCES

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, *4*(7), 267–278.

Baseler, H. A., Harris, R. J., Young, A. W., & Andrews, T. J. (2014). Neural responses to expression and gaze in the posterior superior temporal sulcus interact with facial identity. *Cerebral Cortex*, *24*(3), 737–744.

Benton, C. P. (2009). Effect of photographic negation on face expression aftereffects. *Perception*, *38*(9), 1267–1274.

Bruce, V., & Young, A. (1998). *In the eye of the beholder : the science of face perception*. Oxford University Press.

Bruce, V., & Young, A. (2012). *Face perception*. Psychology Press.

Burton, A.M., Jenkins, R., Hancock, P.J.B., & White, D. (2005). Robust representations for face recognition: The power of averages. Cognitive Psychology, 51, 256-284.

Butler, A., Oruc, I., Fox, C. J., & Barton, J. J. S. (2008). Factors contributing to the adaptation aftereffects of facial expression. *Brain Research*, *1191*, 116–126.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208.

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Catgorical perceptions of morphed facial expressions. *Visual Cognition*, *3*(2), 81–117.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, *19*, 207–283.

Engell, A. D., & Haxby, J. V. (2007). Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia*, *45*(14), 3234–3241.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227–240.

Flack, T. R., Watson, D. M., Harris, R. J., Hymers, M., Gouws, A., Young, A. W., & Andrews, T. J. (2014). Distinct Representations for Rigid and Non-Rigid Facial Movements in Face-Selective Regions of the Human Brain. *Journal of Vision*, *14*(10), 1383–1383.

Fox, C. J., Moon, S. Y., Iaria, G., & Barton, J. J. S. (2009). The correlates of subjective perception of identity and expression in the face network: An fMRI adaptation study. *NeuroImage*, *44*(2), 569–580.

Harris, R. J., Young, A. W., & Andrews, T. J. (2012). Morphing between expressions dissociates continuous from categorical representations of facial expression in the

human brain. *Proceedings of the National Academy of Sciences*, *109*(51), 21164–9.

Harris, R. J., Young, A. W., & Andrews, T. J. (2014). Brain regions involved in processing facial identity and expression are differentially selective for surface and edge information. *NeuroImage*, *97*, 217–223.

Harry, B., Williams, M. A., Davis, C., & Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in Human Neuroscience*, *7*, 692.

Haxby, J., Hoffman, E., & Gobbini, M. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233.

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

Ishai, A. (2008). Let's face it: It's a cortical network. *NeuroImage*, *40*(2), 415–419.

Kriegeskorte, N. (2009). Relating Population-Code Representations between Man, Monkey, and Computational Models. *Frontiers in Neuroscience*, *3*(3), 363–73.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008b). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), 1–28.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, *24*(8), 1377–1388.

Magnussen, S., Sunde, B., & Dyrnes, S. (1994). Patterns of perceptual asymmetry in processing facial expression. *Cortex*, *30*(2), 215–229.

Mattavelli, G., Sormaz, M., Flack, T., Asghar, A. U. R., Fan, S., Frey, J., … Andrews, T. J. (2013). Neural responses to facial expressions support the role of the amygdala in processing threat. *Social Cognitive and Affective Neuroscience*, 1–6.

McKelvie, S. J. (1973). The meaningfulness and meaning of schematic faces. *Perception & Psychophysics*, *14*(2), 343–348.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis, 5th Edition*. New York: Wiley.

Pallett, P. M., & Meng, M. (2013). Contrast negation differentiates visual pathways

underlying dynamic and invariant facial processing. *Journal of Vision*, *13*(14), 1–18.

Pitcher, D. (2014). Facial expression recognition takes longer in the posterior superior temporal sulcus than in the occipital face area. *The Journal of Neuroscience : 34*(27), 9173–7.

Psalta, L., Young, A. W., Thompson, P., & Andrews, T. J. (2014). The thatcher illusion reveals orientation dependence in brain regions involved in processing facial expressions. *Psychological Science*, *25*(1), 128–136.

Rossion, B., Caldara, R., Seghier, M., Schuller, A. M., Lazeyras, F., & Mayer, E. (2003). A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain*, *126*(11), 2381–2395.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Russell, R., & Sinha, P. (2007). Real-world face recognition: The importance of surface reflectance properties. *Perception*, *36*(9), 1368–1374.

Said, C. P., Moore, C. D., Engell, A. D., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, *10*, 1–12.

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*.

Tiddeman, B. P., Burt, D. M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, *21*, 42–50.

Tiddeman, B. P., Stirrat, M. R., & Perrett, D. I. (2005). Towards Realism in Facial Image Transformation: Results of a Wavelet MRF Method. *Computer Graphics Forum*, *24*(3), 449–456.

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

White, M. (2001). Effect of photographic negation on matching the expressions and identities of faces. *Perception*, *30*(8), 969–981.

Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, *92*(3), 1830–1839.

	Fear	Anger	Disgust	Sad	Happy
OFA	0.87 + 0.08	0.79 + 0.07	0.79 + 0.08	0.79 + 0.09	0.88 + 0.09
STS	0.34 + 0.08	0.33 + 0.07	0.28 + 0.08	0.31 + 0.09	0.34 + 0.09
FFA	0.82 + 0.08	0.73 + 0.07	0.72 + 0.07	0.70 + 0.08	0.81 + 0.08

Table 1 % MR signal in face-selective regions to different facial expressions.

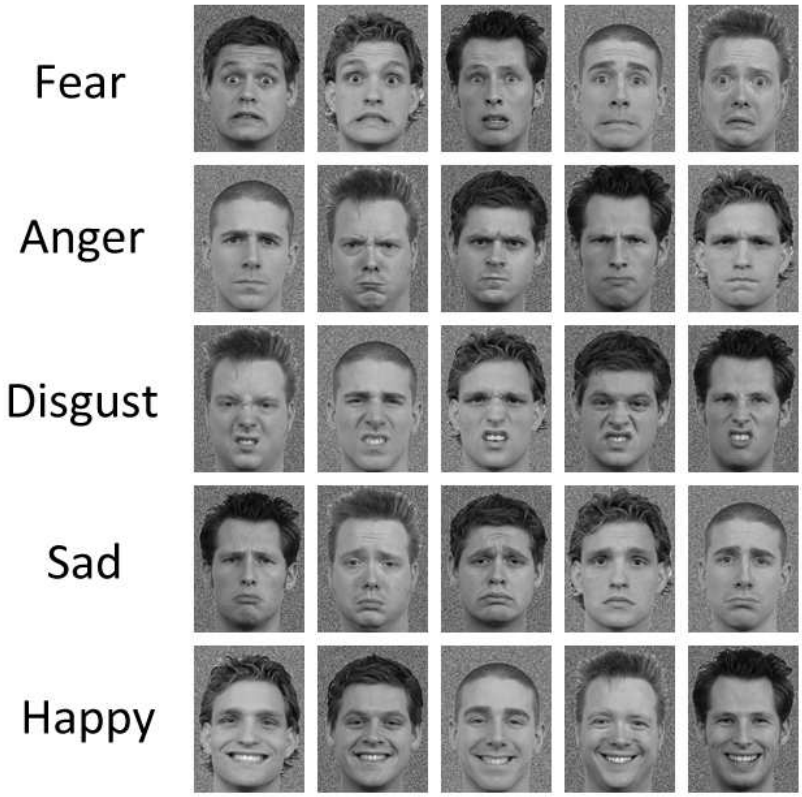


Figure 1 Images used in behavioural and fMRI experiments. Images were taken from 5 identities posing expressions of 5 basic emotions. Each row shows a typical sequence that might form a block in the fMRI experiment (presenting the images from left to right one at a time).

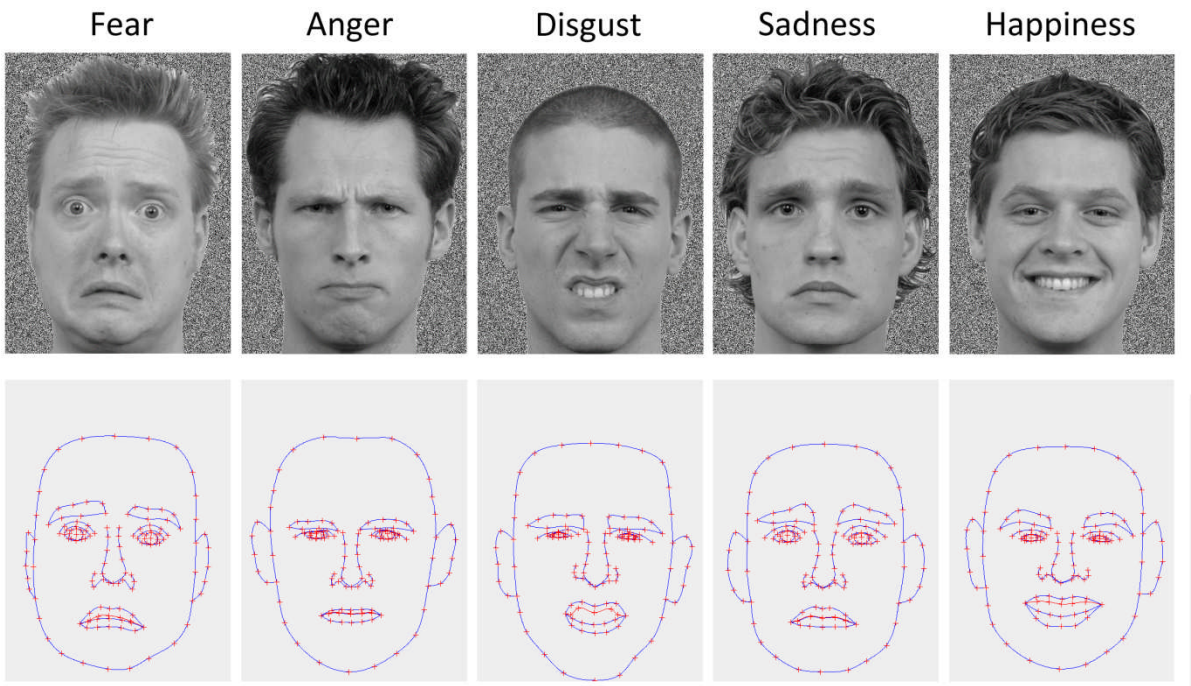


Figure 2 Exemplars of faces posing different expressions (top) and the location of the key fiducial points in each face (bottom).

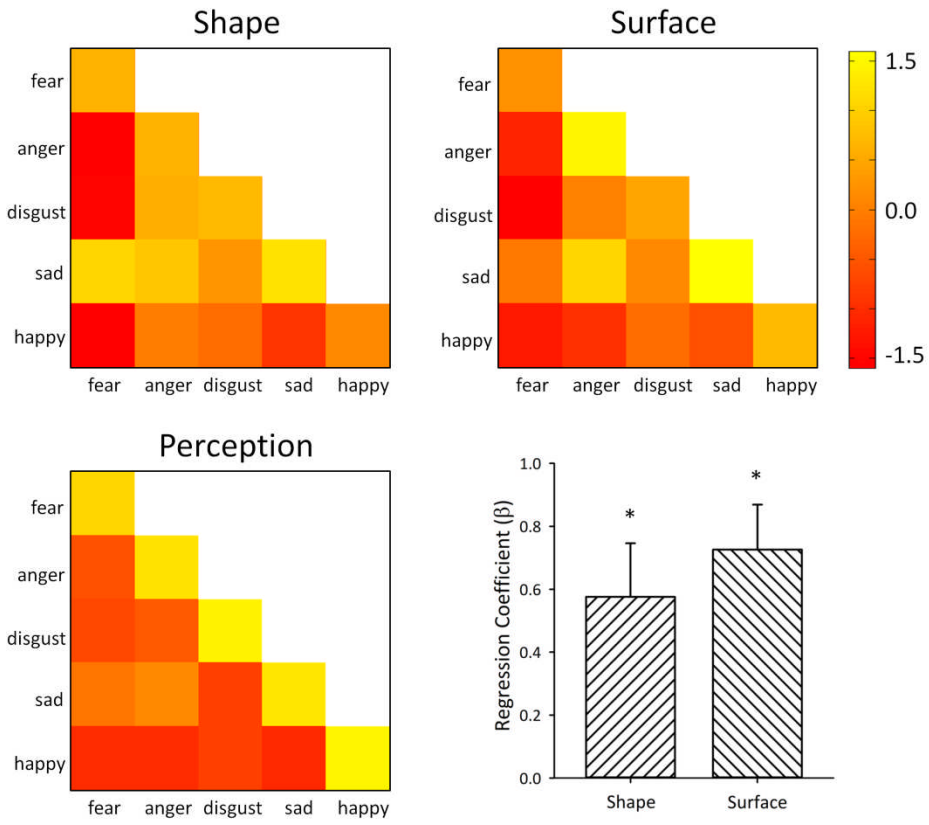


Figure 3 Regression analyses of the perceptual similarity data with shape and surface properties of the image. The analysis shows that the perceptual similarity of facial expressions can be predicted by both the shape and surface properties of the face. Error bars represent 95% confidence intervals. * denotes p<.001. Colour bars for each grid represent z score scale.

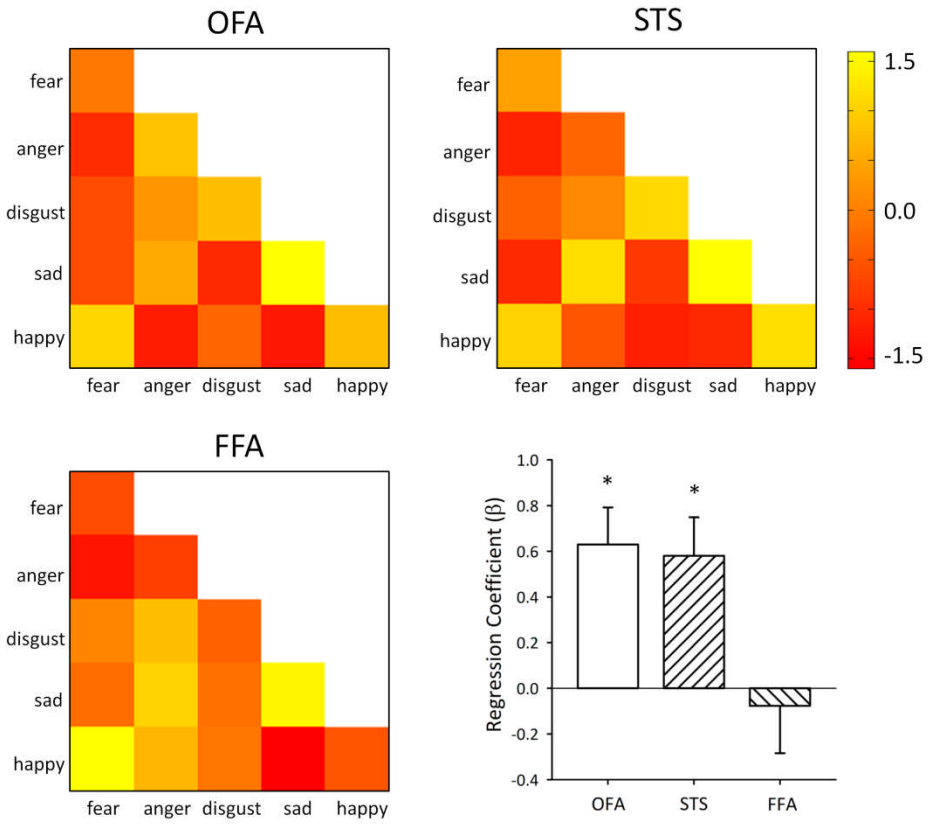
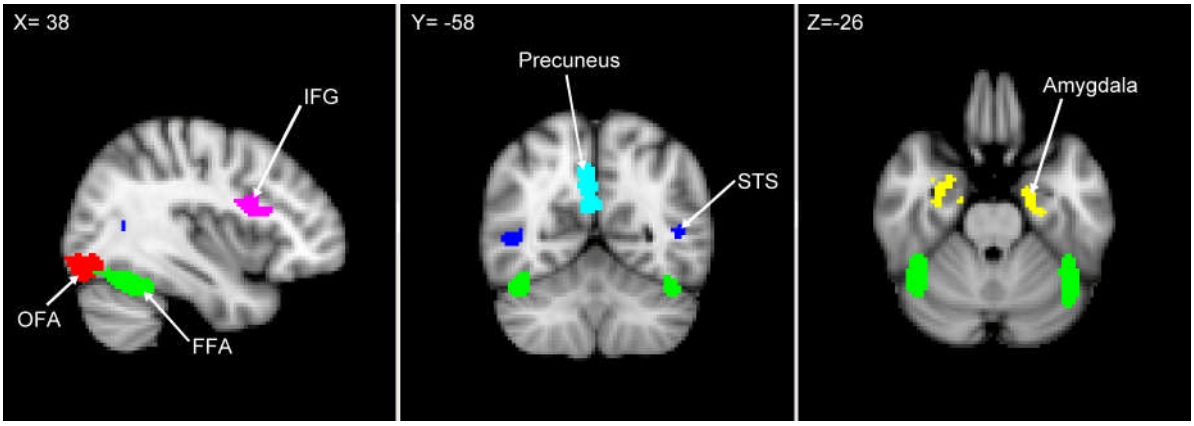


Figure 4 Regression analyses of the perceptual similarity data (shown in Figure 2) with the fMRI data from different face-selective regions. The analysis shows that the perceptual similarity of facial expressions can be predicted by the pattern of response in the OFA and STS, but not in the FFA. Error bars represent 95% confidence intervals. * denotes p<.001. Colour bars for each grid represent z score scale.



Supplementary Figure 1 Location of the core (superior temporal sulcus: STS, occipital face area: OFA, fusiform face area: FFA) and extended (inferior frontal gyrus: IFG, amygdala: AMG, PC: precuneus) face-selective regions

Harvard Oxford Brain Region	Within vs between level discrimination (p value)
Angular Gyrus	.29
Central Opercular Cortex	.78
Cingulate Gyrus anterior division	.07
Cingulate Gyrus posterior division	.18
Cuneal Cortex	.09
Frontal Medial Cortex	.75
Frontal Operculum Cortex	.81
Frontal Orbital Cortex	.60
Frontal Pole	.49
Heschls Gyrus includes H1 and H2	.4
Inferior Frontal Gyrus pars opercularis	.33
Inferior Frontal Gyrus pars triangularis	.22
Inferior Temporal Gyrus anterior division	.35
Inferior Temporal Gyrus posterior division	.001
Inferior Temporal Gyrus temporo occipital part	.61
Insular Cortex	.37
Intracalcarine Cortex	.45
Juxtapositional Lobule Cortex formerly SMA	.77
Lateral Occipital Cortex inferior division	.12
Lateral Occipital Cortex superior division	.07
Lingual Gyrus	.61
Middle Frontal Gyrus	.63
Middle Temporal Gyrus anterior division	.51
Middle Temporal Gyrus posterior division	.05
Middle Temporal Gyrus temporooccipital part	.16
Occipital Fusiform Gyrus	.92
Occipital Pole	.59
Paracingulate Gyrus	.44
Parahippocampal Gyrus anterior division	.95
Parahippocampal Gyrus posterior division	.29
Parietal Operculum Cortex	.29
Planum Polare	.65
Planum Temporale	.99
Postcentral Gyrus	.53
Precentral Gyrus	.24
Precuneus Cortex	.06
Subcallosal Cortex	.95
Superior Frontal Gyrus	.97
Superior Parietal Lobule	.75
Superior Temporal Gyrus anterior division	.47
Superior Temporal Gyrus posterior division	.80
Supracalcarine Cortex	.86
Supramarginal Gyrus anterior division	.07
Supramarginal Gyrus posterior division	.24
Temporal Fusiform Cortex anterior division	.87
Temporal Fusiform Cortex posterior division	.60
Temporal Occipital Fusiform Cortex	.82
Temporal Pole	.98

Suppl. Table 1 Significance values for within>between category expression discrimination in 48 cortical brain regions as defined by the Harvard Oxford brain atlas