Social Cognitive and Affective Neuroscience, 2020, 803–813

doi:10.1093/scan/nsaa110
Advance Access Publication Date: 2 October 2020
Original Manuscript

Statistical pattern recognition reveals shared neural signatures for displaying and recognizing specific facial expressions

Sofia Volynets[1], Dmitry Smirnov[1], Heini Saarimäki[1,2] and Lauri Nummenmaa[3,4]

[1]Department of Neuroscience and Biomedical Engineering, School of Science, Aalto University, FI-0076 Aalto, Finland, [2]Faculty of Social Sciences, Tampere University, FI-33014 Tampere, Finland, [3]Turku PET Centre and Department of Psychology, University of Turku, FI-20520 Turku, Finland and [4]Turku University Hospital, University of Turku, FI-20520 Turku, Finland

Correspondence should be addressed to Lauri Nummenmaa, Turku PET Centre c/o Turku University Hospital, Kiinamyllynkatu 4-6, 20520 Turku, Finland. E-mail: latanu@utu.fi

Abstract

Human neuroimaging and behavioural studies suggest that somatomotor 'mirroring' of seen facial expressions may support their recognition. Here we show that viewing specific facial expressions triggers the representation corresponding to that expression in the observer's brain. Twelve healthy female volunteers underwent two separate fMRI sessions: one where they observed and another where they displayed three types of facial expressions (joy, anger and disgust). Pattern classifier based on Bayesian logistic regression was trained to classify facial expressions (i) within modality (trained and tested with data recorded while observing or displaying expressions) and (ii) between modalities (trained with data recorded while displaying expressions and tested with data recorded while observing the expressions). Cross-modal classification was performed in two ways: with and without functional realignment of the data across observing/displaying conditions. All expressions could be accurately classified within and also across modalities. Brain regions contributing most to cross-modal classification accuracy included primary motor and somatosensory cortices. Functional realignment led to only minor increases in cross-modal classification accuracy for most of the examined ROIs. Substantial improvement was observed in the occipito-ventral components of the core system for facial expression recognition. Altogether these results support the embodied emotion recognition model and show that expression-specific somatomotor neural signatures could support facial expression recognition.

Key words: emotion; facial expression; pattern recognition; fMRI

Introduction

Humans convey their internal states, motives and needs with facial expressions. So-called basic emotion theories propose that the evolution has carved out discrete neural circuits and physiological systems that support distinct survival functions (Panksepp, 1982; Ekman, 1992), and that activation of each discrete system would be associated with specific pattern of facial

Received: 3 August 2020; **Revised**: 25 May 2020; **Accepted**: 3 August 2020

muscle activity resulting in emotion-specific facial expressions (Ekman, 1999 for a more recent formulation, see Cowen *et al.*, 2019). These expressions are used for communication in social interaction, as they convey information about the expressers' internal states that can be 'read out' from the face.

Expression recognition is however not carried only by the brain's visual systems (Haxby *et al.*, 2000). Models of embodied emotion recognition (see reviews in Niedenthal, 2007; Wood *et al.*, 2016) have proposed that sensorimotor simulation of seen facial expressions could support their recognition, especially when the recognition cannot be achieved via the straightforward automated visual pattern-recognition strategies (Winkielman *et al.*, 2015; Calvo and Nummenmaa, 2016; Wood *et al.*, 2016). In other words, recognition of emotional expressions is supported by reinstating a partial somatosensory and motor representation of the implied emotion in the observer, serving as a recall cue for assigning the category label to the expression. In line with the embodiment hypothesis, seeing others' facial expressions triggers automatic facial mimicry (see, e.g. Dimberg and Thunberg, 1998). Neuroimaging studies have also established that displaying and observing facial expressions activates overlapping brain regions including premotor, somatosensory, and gustatory cortices (Carr *et al.*, 2003; Hennenlotter *et al.*, 2005; Van der Gaag *et al.*, 2007; Kircher *et al.*, 2012; Wicker *et al.*, 2003). Furthermore, both damage to somatosensory cortex (Adolphs *et al.*, 2000) and its inactivation by transcranial magnetic stimulation (TMS; Pourtois *et al.*, 2004) impairs recognizing of emotions from facial expressions, suggesting that somatomotor embodiment of seen emotions supports their recognition.

The embodied simulation model is further supported by multivariate pattern recognition studies, showing that emotional facial expressions and internal emotional states can be successfully decoded from motor brain regions (Saarimäki *et al.*, 2016; Liang *et al.*, 2017). Yet strong support for the embodied recognition view would require showing that (i) both displaying and seeing different facial expressions would trigger expression-specific, discrete neural signatures in the somatomotor system and that (ii) these expression-specific neural signatures would be similar to those displaying and observing the expressions. Previous studies have found that different seen facial expressions elicit discernible neural activation patterns in visual areas and multisensory temporal cortical areas (e.g. Said *et al.*, 2010; Peelen *et al.*, 2010; Harry *et al.*, 2013; Wegrzyn *et al.*, 2015), but it remains unresolved whether these and somatomotor activation patterns converge with those elicited while displaying specific facial expressions.

Here we tested the embodied emotion recognition model directly by using functional magnetic resonance imaging (fMRI) and statistical pattern recognition techniques. Participants observed and displayed three types of facial expressions (joy, anger and disgust) while their brain activity was measured with fMRI. These expressions were chosen because they are accurately recognized (Calvo *et al.*, 2014; Calvo and Nummenmaa, 2016) and based on distinct (muscular) Facial Action Coding System (FACS) activation patterns (Ekman and Friesen, 1978), and are thus easy to pose by naïve volunteers. Subsequently, a pattern classifier based on Bayesian logistic regression was trained to classify the observed or displayed emotions from the fMRI signals using regions-of-interest (ROIs) in the face perception system (Haxby *et al.*, 2000), emotion circuit (Saarimäki *et al.*, 2016) and visual and somatomotor areas. The critical test involved training the classifier with data recorded while displaying expressions and testing with data recorded while observing the expressions. Successful classifier performance in this condition, and particularly when using data from the somatosensory and motor cortices, would provide support for expression-specific embodiment during perception of facial expressions. We show that different facial expressions have distinguishable somatomotor neural signatures that are activated similarly both when viewing and displaying the expressions.

Materials and methods

Participants

Twelve healthy right-handed female volunteers (mean age 21 years, range 20–26 years) with normal or corrected to normal vision and normal hearing (self-reported) volunteered for the study. None had a history of neurological or psychiatric diseases, or current medication affecting the central nervous system. All subjects signed informed consent forms, approved by the Aalto University Institutional Review Board, and were compensated for their time. All subjects were pre-tested for their ability to recognize emotional facial expressions: Subjects viewed photographs of unfamiliar models displaying facial expressions of emotions (anger, fear, disgust, happiness, sadness and surprise) as well as morphed versions (30% and 60% morphs with neutral expressions) of the same photos. The test stimuli were derived from the Karolinska Directed Emotional Faces database (KDEF; Lundqvist *et al.*, 1998). Subjects were asked to identify emotion on the given photo in a six-alternative forced choice test. Mean accuracy was 74% and exceeded chance level (16.6%) for all expressions.

Experimental design and statistical analysis

Training phase

At least one day prior to the fMRI experiment, all subjects participated in an individual training session, where the experimental setup and tasks were explained. Main facial features of the expressions of joy, anger and disgust were described according to the FACS (Ekman and Friesen, 1978). Participants were then instructed to (i) select one triggering memory/image for eliciting each emotion (joy, anger and disgust) to make the expressions more genuine (e.g. favourite joke to elicit smile) and (ii) rehearse displaying the corresponding facial expressions when prompted with minimal head motion.

Displaying expressions task

Experimental design is summarized in Figure 1. During the Displaying Expressions Task (Figure 1A) participants displayed facial expressions of joy, anger and disgust while being scanned with fMRI. Each trial started with an auditory instruction, specifying the facial expression to be displayed (spoken words 'joy', 'anger' or 'disgust'). Next, a beep indicated that the subject should display the facial expression and keep the full-blown pose until a second beep (after 5 s) indicated the end of the trial. During the 10 to 15 s of intertrial interval, the subject was instructed to relax their face. A fixation cross was shown at the centre of the screen throughout the experiment. The Displaying part consisted of 4 runs with 24 trials in each, and the subjects displayed each expression 8 times per each run.

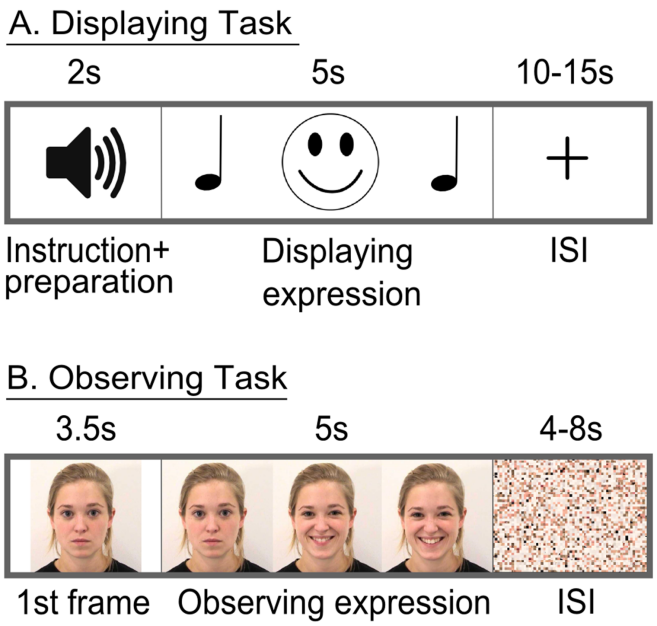

Fig. 1. Experimental design. (A) In the Displaying task, participants heard a word describing the facial expression to be performed next. After 2 s, they heard a 500 ms of beep sound, which marked the start of expression execution. During the next 5 s, participants went from neutral face to the fully blown target facial expression and kept it until the second auditory cue, which prompted the participants to relax their facial muscles for the subsequent 10 to 15 s of interstimulus interval (ISI). (B) In the Observing task, participants were first shown a neutral facial expression of the model from upcoming video for 3.5 s. After that, they viewed a 5 s of dynamic facial expression video, where the model went from neutral face to fully blown emotional expression and kept it until the end of the video. During the 4 to 8 s of ISI, a scrambled facial expression was shown.

Observing expressions task

In the Observing Expressions Task (Figure 1B) participants viewed short video clips (5 s) of six models (three females) displaying facial expressions of joy, anger and disgust. The stimuli were selected from ADFES database (Van der Schalk *et al.*, 2011), and the models in the clips were unfamiliar to the participants. All clips begun with a neutral face, followed by a dynamic display of the facial expression. Prior to each clip, subjects were shown the first frame of the video (i.e. neutral face) for 3.5 s to avoid peaks in low-level visual activation due to simultaneous visual stimulus and motion onset. This was followed by the dynamic expression, which was held in its full-blown phase until the end of the clip. Each stimulus was followed by a random 4 to 8 s of rest period. Again, to avoid peaks in low-level visual cortical activations, a scrambled picture of the upcoming model was shown during the rest period.

To keep participants focused on the task, three trials per run contained a still picture of the neutral face instead of the video clip. Participants were asked to press the response button as soon as they detected a trial without any facial motion. These trials were excluded from the analysis. The Observing part consisted of 4 runs with $24+3$ trials in each, and the subjects observed the videos of each facial expression 8 times per each run.

Stimulus presentation

Stimulus delivery was controlled using Presentation software (Neurobehavioral Systems Inc., Albany, CA, USA). Visual stimuli were back-projected on a semi-transparent screen using a three-micromirror data projector (Christie X3, Christie Digital Systems Ltd, Mönchengladbach, Germany) and reflected via a mirror to the subject. Auditory cues were delivered with Sensimetrics S14 insert earphones (Sensimetrics Corporation, Malden, MA, United States). Sound intensity was adjusted for each subject to be loud enough to be heard over the scanner noise. Three subjects were also recorded with an fMRI-compatible face camera to ensure that the facial expressions were displayed successfully during the experiment.

fMRI acquisition and preprocessing

MRI scanning was performed on two sessions with observing and displaying tasks done on separated days. The order of the tasks was counterbalanced across participants. The data were acquired with 3T Siemens Magnetom Skyra scanner at the Advanced Magnetic Imaging Centre, Aalto NeuroImaging, Aalto University, using a Siemens head coil of 20 channels. Functional images were collected using a whole brain T2*-weighted echo-planar imaging (EPI) sequence, sensitive to blood oxygenation level-dependent (BOLD) signal contrast, with the following parameters: 33 axial interleaved slices, TR $= 1.7$ s, TE $= 24$ ms, flip angle $= 70°$, voxel size $= 3 \times 3 \times 4.0$ mm3, matrix size $= 64 \times 64 \times 33$. A total of 245 volumes were acquired in each run for the Observing task, and 275 volumes were acquired in each run for the Making task. The first 5 volumes of each run were discarded. High-resolution anatomical images with isotropic 1 mm^3 of voxel size were collected using a T1-weighted MP-RAGE sequence.

The fMRI data were preprocessed FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl) with additional in-house signal clean-up tools implemented in MATLAB with SPM8 (www.fil.ion.ucl.ac.uk/spm/). After slice timing correction, the functional images were realigned to the middle scan by rigid-body transformations with MCFLIRT (Jenkinson *et al.*, 2002) to correct for subject motion. Non-brain matter was next removed

using BET (Smith, 2002). Functional images were then registered to the MNI152 standard-space template (Montreal Neurological Institute) with 2 mm of resolution. The transformation parameters were acquired by first calculating transformations from structural to standard space and from functional to structural space, and then concatenating these parameters. Next, these transformation parameters were used to co-register functional datasets to the standard space. Both registration steps were performed using FLIRT (Jenkinson *et al.*, 2002). Motion artefacts were cleaned from the functional data using 24 motion regressors (Power *et al.*, 2014). None of the subjects were excluded from analysis due to excessive head motion during the Displaying task, as frame-wise displacement values (Power *et al.*, 2012) exceeded 0.5 mm in less than 1.81% of all timepoints per subject. Signal from white matter, ventricles and cerebrospinal fluid were cleaned from the data as implemented in our in-house BraMiLa pipeline (https://git.becs.aalto.fi/bml/bramila). For the general linear model (GLM) analysis, additional spatial smoothing step with a Gaussian kernel of FWHM 8 mm was also applied; the subsequent classification analysis was run on unsmoothed data.

Task-evoked BOLD responses

Task-evoked responses to execution and observation of facial expressions were analysed using the two-stage random effects analysis with GLM implemented in SPM8 (www.fil.ion.ucl.ac.uk/spm). Boxcar regressors (displaying and observing facial expressions) were used to model fMRI voxel time series. The regressors included the time points when the facial expressions were observed or displayed, respectively. Regressors were convolved with the canonical hemodynamic response function (HRF) to account for hemodynamic lag. The input data were high-pass filtered with 128 s of cut-off. After generating subject-wise contrast images, a second level (random effects) analysis was applied to these contrast images in a new GLM to allow population-level inference. Statistical threshold was set at $P < 0.05$, false discovery rate (FDR) corrected at cluster level.

Region-of-interest selection

For the classification analysis, three sets of ROIs were employed. First, we used the whole grey matter (derived from MNI standard brain template; Grabner *et al.*, 2006) as a single ROI. Second, we used functional ROIs based on the experimental tasks. The ROIs were derived from the group level activations for (i) observing and (ii) displaying facial expressions, (iii) their intersection and (iv) their union. The contrast images from second-level GLM analysis were thresholded at $T > 2$ for Displaying task and at $T > 3$ for Observing task. Liberal threshold was chosen as the maps were not used for statistical inference, but rather as a priori feature-selection filter that would capture the expression display- and observation-dependent neural activation. Third, we used anatomically defined ROIs in the emotion, somatosensory, and face perception circuits defined using the Harvard-Oxford cortical and subcortical structural atlases (Desikan *et al.*, 2006). The emotion circuit ROIs included insula, thalamus and anterior cingulate cortex (ACC; Saarimäki *et al.*, 2016). The somatomotor ROIs included primary somatosensory cortex (SI), secondary somatosensory cortex (SII), primary motor cortex and premotor cortex. The face perception circuit ROIs included the key nodes of the core system for face perception (Haxby *et al.*, 2000): inferior occipital cortex, fusiform cortex, superior temporal sulcus (STS) and MT/V5 region.

Pattern classification

Pattern classification was performed in three ways. First, we wanted to test whether each of the displayed and observed expressions is associated with distinct neural signatures. To that end, we performed a conventional within-modality classification, where the classifier was trained and tested with data from the same modality (displaying or observing). Second, to test whether displaying and observing facial expressions would be associated with similar, expression-specific neural signatures in the brain, we initially performed cross-modal classification without functional realignment. In this approach, the data from the Displaying condition were used to train the pattern classifier to distinguish between the three different expressions, and then, the classifier was validated using corresponding data from the Observing condition. Third, to test whether the neural codes for observing and displaying facial expressions would be similar, yet anatomically misaligned, we performed cross-modal classification analysis where an additional realignment step was employed to allow functional coregistration of Observing and Displaying data. Statistical significance of mean classification accuracies was tested by comparing them against chance level. Since classification accuracies across participants were not normally distributed, we used one-tailed Wilcoxon signed rank test. Multiple comparisons were corrected for with Benjamini-Hohchberg false discovery rate (BH-FDR) correction (Benjamini and Hochberg, 1995).

For all tested classifiers, the data comprised all trials, with 3 TRs per trial, recorded during displaying and observing phases of the experiment, and shifted by 6 s to account for the hemodynamic lag, with each TR independently used as a training or testing example. For all tested classifiers, we evaluated the performance of the classification model in leave-one-run-out cross-validation framework, where three runs were used to train the classifier and the left-out run was used in testing, and the process was repeated iteratively for each run. In cross-modal classification analyses, the training runs were taken from Displaying data, and the testing run was taken from the Observing data. This training-testing protocol was implemented, because we specifically wanted to test whether of expression display-related activity would be predictive of the seen facial expressions.

Classification was performed with Bayesian logistic regression with a sparsity promoting Laplace prior to classify brain activity patterns measured during displaying and observing facial expressions (Van Gerven *et al.*, 2010). Each individual voxel weight was given a univariate Laplace prior distribution with a common scale hyperparameter to promote sparsity in the posterior distribution (Williams, 1995). The multivariate posterior distribution was approximated using the expectation propagation algorithm (Van Gerven *et al.*, 2010) implemented in the FieldTrip toolbox (Oostenveld *et al.*, 2011). Four binary classifiers were trained to discriminate between each expression category *vs* the others. The classification performance was tested by collecting the class probabilities for each pattern in the testing set using the binary classifiers and assigning the class with the maximum probability to each pattern.

Functional realignment and cross-modal classification

To test for possible differences in regional organization of displaying and observing facial expressions, an additional

functional realignment step was introduced in the analysis pipeline (see Smirnov et al., 2017, for details and validation of the approach). Briefly, we used Bayesian canonical correlation analysis (BCCA; Klami et al., 2013) to perform the realignment step prior to cross-modal classification. BCCA was implemented using R CCAGFA package (Virtanen et al., 2012; Klami et al., 2013). The BCCA model separates the correlation patterns in the simultaneous brain-activity spaces of display and observation of the same facial expression into three types of components: display-specific, observation-specific and shared. The shared components provide a low-rank linear mapping for the realignment of the brain-activity spaces. Classifier was trained on the displaying data from the three runs and tested on functionally realigned observation data run.

Results

Task-related BOLD responses

Figure 2 shows brain regions activated during Displaying and Observing tasks in the main experiment (all expressions combined). Both Displaying and Observing tasks engaged precentral gyrus (motor strip), supramarginal gyrus, supplementary motor area (SMA), superior and inferior frontal gyri, temporal pole, as well as parietal operculum (S2), caudate, insula and right thalamus. In general, observing *vs* displaying facial expressions yielded more widespread activity, yet 50% of the voxels activated during the Displaying task were also activated during the Observing task. Significant observation-specific activations were found in frontal pole, putamen, amygdala, lateral occipital

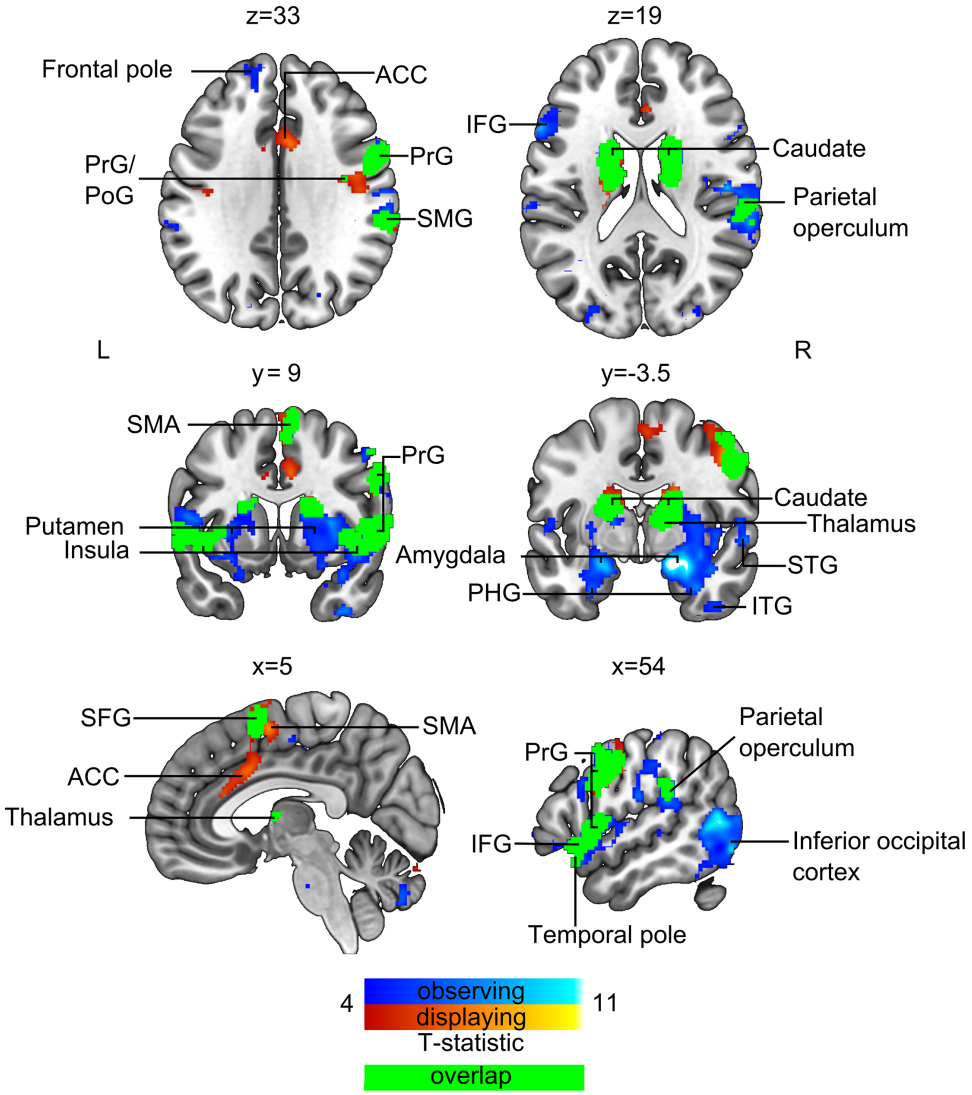

Fig. 2. Brain regions responding to displaying (hot colours) and observing (cool colours) facial expressions. Green colouring shows the overlap between displaying and observing dependent activations at the used statistical threshold ($P < 0.05$, FDR corrected). ACC, anterior cingulate cortex; PrG, precentral gyrus; PoG, postcentral gyrus; SMG, supramarginal gyrus; SMA, supplementary motor area; SFG, superior frontal gyrus; IFG, inferior frontal gyrus; STG, superior temporal gyrus; ITG, inferior temporal gyrus; PHG, parahippocampal gyrus.

cortex, parahippocampal, superior temporal and inferior temporal gyri. The only area uniquely activated during the Displaying task was ACC. Additionally, somatomotor responses (precentral and postcentral gyri, SMA) were more widespread in the Displaying *vs* Observing task.

Within-modality classification for displaying and observing facial expressions

First, we ran conventional within-modality pattern classification analyses to test whether the three displayed or observed facial expressions have distinct neural signatures. For the Displaying condition, all functional and anatomical ROIs yielded statistically significantly above chance level accuracy (Figures 1–4 and Figure 5). Mean accuracies varied from 39% to 64% against 33% chance level ($Ws = 78$, $ps < 0.001$). Accuracy was highest for whole grey matter and Displaying–Observing union ROIs (64%), followed by Displaying (62%) and Observing (61%) ROIs. Importantly, all facial expressions could be classified from each other significantly above chance level (Figure 3A). Within-modality classifier for Observing condition also yielded above chance level accuracy in all ROIs with accuracies ranging from 36% to 54% against 33% chance level ($Ws = 61–78$, $ps < 0.05$; Figures 1–4). Accuracy was highest for inferior occipital cortex (54%), followed by whole grey matter (52%) and Displaying–Observing union (51%) ROIs. Again, all facial expressions could be classified from each other significantly above chance level (Figure 3B). Within-modality classification was significantly more accurate for displaying *vs* observing emotions in all ROIs except for inferior occipital cortex, V5 and ACC ($Ws = 59–78$, $ps < 0.05$).

Cross-modal classification with and without functional realignment

Accuracy of cross-modal classification exceeded the chance level already before functional realignment in most of the ROIs ($Ws = 62–76$, $ps < 0.05$; Figures 4 and 5). Accuracy was highest in the whole grey matter ROI (42%), followed by Displaying–Observing union (40%) and Displaying ROI (40%). Chance level was not exceeded significantly for thalamus (35%), fusiform cortex (35%), amygdala (35%) and inferior occipital cortex (36%). After functional realignment, classification accuracy improved statistically significantly in the inferior occipital cortex, fusiform cortex, whole grey matter and all functional ROIs derived from experimental data ($Ws = 54–78$, $ps < 0.05$, mean increase 2%–14%; Figure 4). Accuracy was highest in inferior occipital cortex (50%), whole grey matter (49%), Displaying–Observing union (49%) and fusiform cortex (46%) ROIs, followed by the rest of functional ROIs with accuracies ranging from 43% to 45% ($Ws = 77–78$, $ps < 0.001$) and anatomically defined ROIs with accuracies ranging from 37% to 40% ($Ws = 73–78$, $ps < 0.01$). After functional realignment, chance level was not exceeded significantly only for thalamus (33%) and insula (34%).

Finally, we validated that neural responses to displaying and observing of each individual expression could be classified across modalities with approximately the same precision. After functional realignment, the classifier was able to distinguish expressions from each other with accuracies comparable across the ROIs: disgust 44% to 50%, anger 49% to 54% and joy 46% to 55% (Figure 3C). The accuracy ranges for cross-modal classification before functional realignment were considerably lower: disgust 31% to 40%, anger 33% to 41% and joy 29% to 46% (Figure 3D). However, the best discrimination was achieved in within-modality classification on Displaying data, specifically, in the functional ROIs and motor-related ROIs: mean accuracies were for disgust 57% to 62%, for anger 62% to 65% and for joy 62% to 69% (Figure 3A).

Discussion

Our main finding was that expression-specific neural codes are shared between displaying and observing specific emotional facial expressions. Classifier trained on haemodynamic data acquired while subjects displayed different facial expressions successfully predicted neural activation patterns triggered while viewing each of those expressions displayed by unfamiliar models. Highest cross-modal classification accuracies were observed in functional ROIs derived from the experimental data, and in primary motor and somatosensory cortices. Hyperalignment increased cross-modal classification accuracy only modestly outside the visual regions. This suggests that neural codes for displaying and recognizing facial expressions are sufficiently similar so that cross-modal classification is possible without such additional functional warping step. These results support the embodied emotion recognition model and show that automatically activated, expression-specific neural signatures in sensorimotor and face perception regions of the brain, as well as in emotion circuit ROIs, could support facial expression recognition.

Discrete neural basis of viewing and displaying facial expressions

We first established that displaying different facial expressions was associated with discrete neural activation patterns, similarly as has been previously demonstrated for hand actions (Dinstein *et al.*, 2008; Smirnov *et al.*, 2017). As expected, highest regional classification accuracies were found in the motor and somatosensory cortices. However, accurate classification of the displayed facial expressions was also possible in the regions involved in visual (inferior occipital and fusiform cortices; STS) and affective (amygdala, insula, thalamus, ACC) analysis of facial expression (Haxby *et al.*, 2000). These data thus suggest that displaying emotional facial expressions involved engagement of their affective and visual representations in the brain, likely resulting in visual-motor and affective-motor integration during volitional generation of facial expressions.

In line with previous work (e.g. Said *et al.*, 2010; Peelen *et al.*, 2010; Harry *et al.*, 2013; Wegrzyn *et al.*, 2015), we also confirmed that viewing the facial expressions was associated with distinct expression-specific activation patterns, particularly in the fusiform and inferior occipital cortices, V5 and STS. However, seen expressions could also be successfully decoded from regional activation patterns in the somatosensory (see also Kragel and LaBar, 2016) and motor cortices (see also Liang *et al.*, 2017), and components of the emotion circuit (amygdala, ACC), suggesting that expression-specific affective and somatomotor codes are also activated during facial expression perception.

Observing *vs* displaying facial expressions triggered more widespread brain activation mainly extending to inferior occipital cortex and medial temporal lobe. Despite this, within-modality classification was, in general (with the exception of inferior occipital cortex, V5 and ACC), significantly more accurate for displaying *vs* observing emotions in all ROIs. Moreover, observed expressions could not be successfully classified from insula and thalamus, suggesting stronger limbic

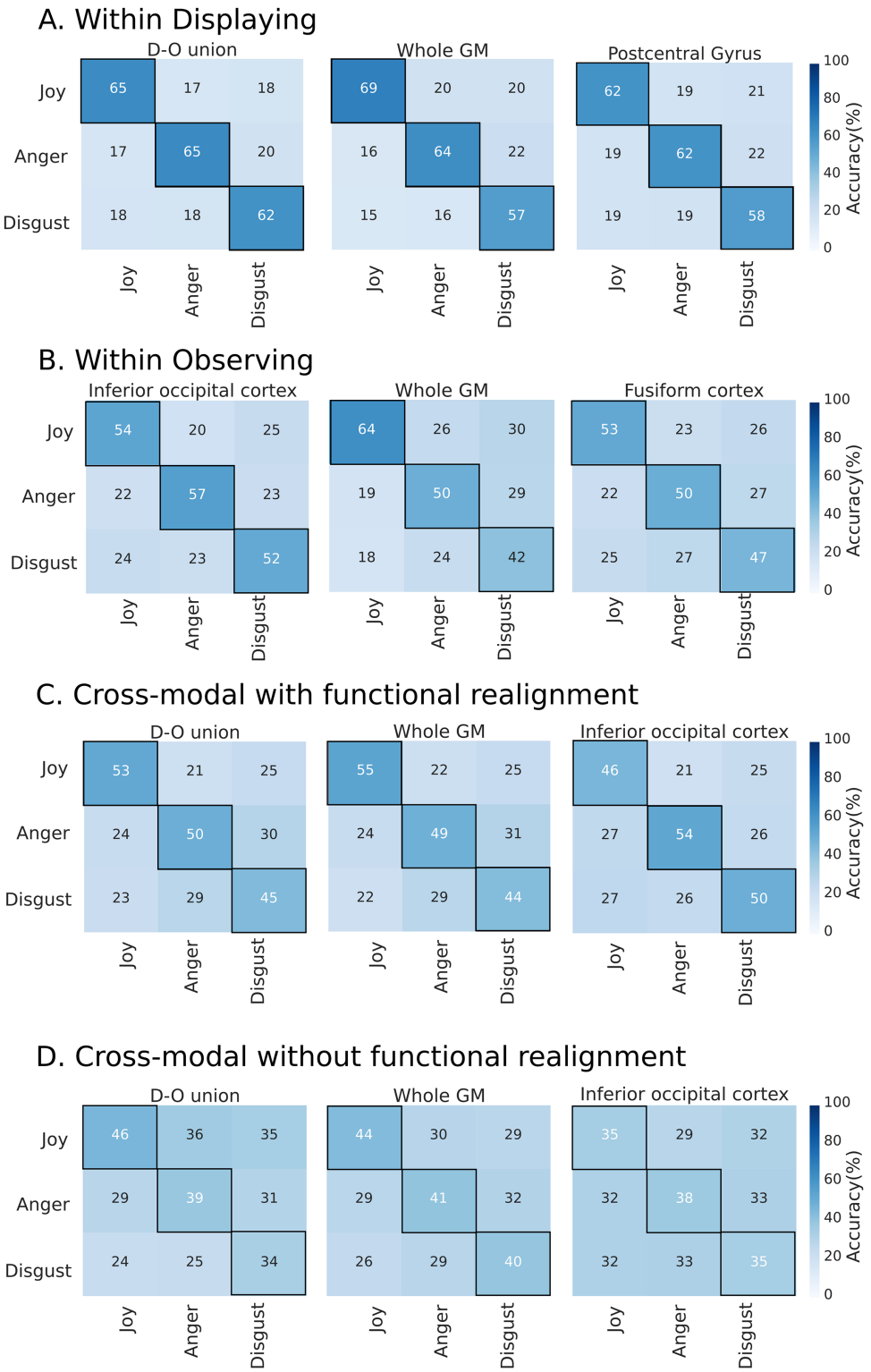

Fig. 3. Confusion matrices from representative regions of interest for within-modality (A-B) and cross-modal classification with (C) and without (D) functional realignment in representative regions of interest.

involvement while displaying *vs* seeing expressions. Altogether these data suggest that the expression-specific neural codes are significantly more discrete when actually generating the expressions (and possibly experiencing the emotion), than when the observers decode the seen facial expressions. Importantly, none of the ROIs alone surpassed accuracy of the whole-brain classification of seen or displayed facial expressions. This suggests that distributed cerebral activation patterns contain the

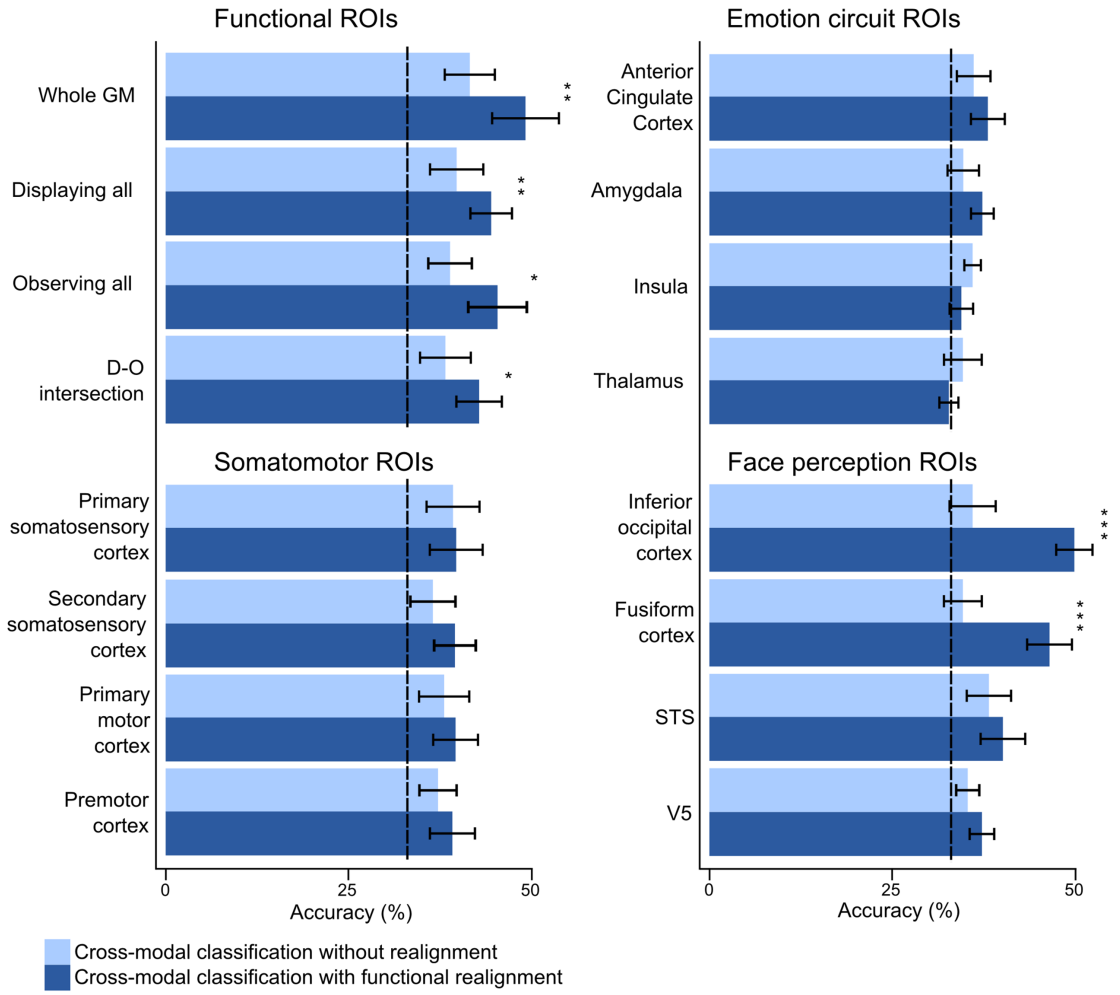

Fig. 4. Mean classification accuracies (in %) and 95% confidence intervals for cross-modal classification with and without functional realignment in functional ROIs as well as in key components in the emotion, somatomotor, and face perception related regions. Dashed line represents the 33% chance level. Asterisks reflect statistical significance of the difference between classification accuracies before and after functional realignment of the data accordingly to the results of Wilcoxon tests (***$P < 0.001$, ** $P < 0.01$, * $P < 0.5$). D-O intersection, intersection of Displaying expressions and Observing expressions ROIs derived from experimental data; STS, superior temporal gyrus; V5, visual cortex area V5.

most accurate neural representation of the seen/displayed facial expression. In general sense, these data support the notion that different emotions have discrete neural bases (Panksepp, 1982; Ekman, 1992; Kragel and LaBar, 2015; Wager *et al.*, 2015; Saarimäki *et al.*, 2016) also in visual and somatomotor domains.

Shared somatomotor signatures for displayed and observed facial expressions support embodied models of emotion recognition

Both displaying and viewing facial expressions triggered overlapping activity in motor (motor strip and SMA) and somatosensory (S2) cortices. Activation patterns within these regions could be used to predict which facial expression the participants had displayed or viewed, and critically, classifier trained with activation patterns elicited by displaying facial expressions could successfully predict which facial expressions the participants saw in the expression observation condition. Such high cross-modal classification accuracy in primary motor and somatosensory cortices suggests strong embodied component in facial expression recognition: These data highlight that the somatomotor codes elicited by viewing facial emotions are expression specific and thus detailed enough to support embodied emotion recognition.

These data agree with previous neuroimaging studies, which have found common activation patterns in motor and somatosensory cortices during observing and producing facial expressions (Carr *et al.*, 2003; Hennenlotter *et al.*, 2005; Kircher *et al.*, 2012; Van der Gaag *et al.*, 2007; Wicker *et al.*, 2003). These regions also synchronize across a group of individuals seeing or hearing emotional episodes (Nummenmaa *et al.*, 2012, 2014b), possibly providing means for shared somatomotor representations of emotions. Furthermore, damage to somatosensory cortex (Adolphs *et al.*, 2000) and its deactivation with TMS (Pourtois *et al.*, 2004) impairs facial expression recognition, while insular cortex supports interoceptive awareness (Critchley *et al.*, 2004). Altogether these data support the models of embodied emotion recognition, which propose that perceiving emotion involves perceptual, somatovisceral and motoric re-experiencing of the relevant emotion in oneself, and one possible mechanism for that is reactivating modality-specific brain areas without actual

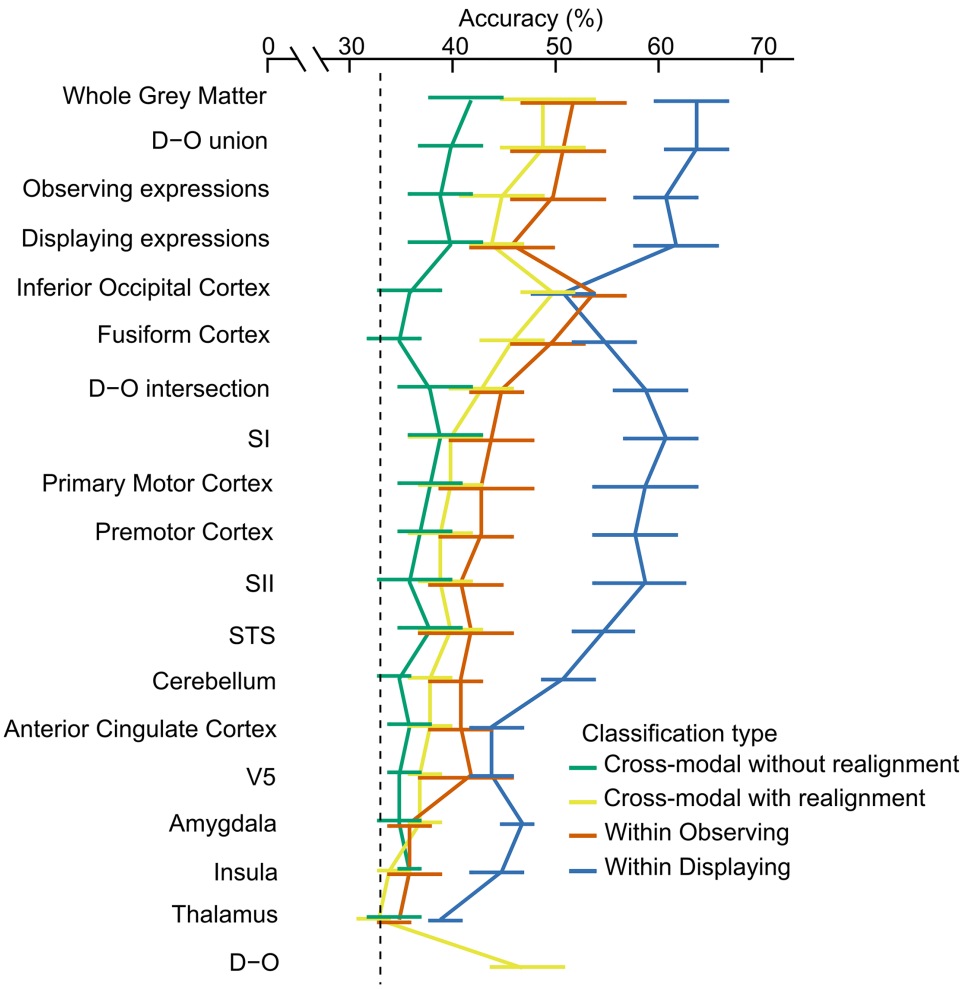

Fig. 5. Mean classification accuracies (in %) and 95% confidence intervals for all types of classification analysis across all the ROIs. Dashed line represents the 33% chance level. D-O union and intersection, union and intersection of Displaying expressions and Observing expressions ROIs derived from experimental data; SI, primary somatosensory cortex; SII, secondary somatosensory cortex; STS, superior temporal gyrus; V5, visual cortex area V5; D-O, specific ROI for cross-modal classification with functional realignment, where observing data are masked with Observing expressions ROI; displaying data, with Displaying expressions ROI (see Methods section for details).

behavioural output (Niedenthal, 2007). Indeed, behavioural studies have shown that a wide variety of emotions are represented in embodied somatosensory format (Nummenmaa *et al.*, 2014, 2018; Volynets *et al.*, in press), and the present study shows how such embodied signatures of emotions can also contribute to recognizing others' expressions.

In addition to the somatosensory and motor regions, both seen and displayed facial expressions could be successfully decoded from key regions of the emotion circuit including ACC and amygdala. Critically, cross-modal classification (after functional realignment) was also accurate in these regions. Previously it has been established that these components of the emotion circuit contain discrete neural signatures of experienced emotions (Saarimäki *et al.*, 2016; Saarimaki *et al.*, 2018). Accordingly, these nodes of the emotion system (particularly ACC and amygdala) are involved in both displaying and viewing facial expressions, likely due to their central role in generating the specific emotional states on the basis of internal and external signals. Taken together these data support the position that facial expression recognition is, in addition to visual and somatomotor mechanisms, also supported by affective analysis of the facial signals (Calvo and Nummenmaa, 2016).

Neural signatures for displayed and observed facial expressions are anatomically aligned

For cross-modal classification, we employed pattern recognition after functional realignment of the data from the observing and making facial expressions condition. This was used to test whether the expression-specific neural codes during displaying and viewing facial expressions would be similar yet misaligned across modalities. However, we found that cross-modal classification was already almost as accurate before the functional realignment for most of the ROIS. Realignment only led to a modest increase in classification accuracy in select brain regions (inferior occipital cortex, fusiform cortex, whole grey matter, functional ROIs derived from Observing and Displaying data, their union and intersection). This suggests a direct linkage between the neural systems engaged while displaying and observing facial expressions.

In our experiment, brain regions where functional realignment increased classification accuracy most were occipito-ventral components of the core system for facial expression recognition—inferior occipital and fusiform cortices (Haxby *et al.*, 2000; Said *et al.*, 2010). In turn, little improvement was observed in the emotion and somatomotor systems Within-modalities classification was highly successful in these regions too. We propose that in these regions neural signatures of displaying and observing certain facial expressions are more anatomically misaligned than in motor and somatosensory cortices, but still distinct and expression specific.

Limitations

Because the experiment involved a long and complicated multi-session setup (with two long fMRI sessions, separate behavioural testing and a practice session one day before the scans), we only scanned 12 individuals. However, the results were consistent across the tested subjects, with above chance level cross-modal classification in all subjects. Although it is still possible that this would reflect some special features (such as exceptional face recognition ability) in the current sample, it is very unlikely that such a high accuracy could be a randomly occurring property of the sample. Nevertheless, future studies need to assess the generalizability of the cross-modal representations of facial expressions in more divergent and larger samples. We only included three facial expressions into the study; thus, it is not certain if the results generalize to other facial expressions such as those proposed by Cowen and Keltner (in press). Yet, pattern classification work on facial expression recognition (e.g. Said *et al.*, 2010) and 'hyperclassification' of seen and executed hand actions (Smirnov *et al.*, 2017) suggests that our results should be generalizable. Finally, we could not successfully record facial movements from all subjects and thus verify their task performance. The MRI data (both GLM and classifiers) however suggests that they performed the task as instructed.

Conclusions

We conclude that displaying and observing emotional facial expressions are supported by shared expression-specific neural codes. A shared set of regions activated during observing and displaying emotional facial expressions includes somatosensory and motor cortices, parts of orbitofrontal cortex, temporal pole and parietal operculum, as well as bilateral caudate, right insula and right thalamus. Classification analysis successfully distinguished between all tested emotions both within and across modalities. Accurate classification in primary motor and somatosensory cortices and STS suggests strong embodied component in facial expression recognition, while limbic regions are also strongly involved in both displaying and observing emotional facial expressions. Taken together, our results support the embodied emotion recognition model and suggest that expression-specific neural signatures could underlie facial expression recognition.

Acknowledgements

We also thank Marita Kattelus and Tuomas Tolvanen for their help with the data acquisition.

Funding

This work was supported by Centre for International Mobility (grant TM-14-9213 to S.V.); Academy of Finland (grants #294897 and #265917 to L.N.) and European Research Council Starting Grant (#313000 to L.N.).

Conflict of interest

The authors declare no competing financial interests.

References

Adolphs, R., Damasio, H., Tranel, D., Cooper, G., Damasio, A.R. (2000). A role for somatosensory cortices in the visual recognition of emotion as revealed by three dimensional lesion mapping. *Journal Neuroscience*, *20*, 2683–90.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Calvo, M.G., Gutiérrez-García, A., Fernández-Martín, A., Nummenmaa, L. (2014). Recognition of facial expressions of emotion is related to their frequency in everyday life. *Journal of Nonverbal Behavior*, *38*, 549–67.

Calvo, M.G., Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: an integrative review. *Cognition and Emotion*, *30*, 1081–106.

Carr, L., Iacoboni, M., Dubeau, M.C., Mazziotta, J.C., Lenzi, G.L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(9), 5497–502.

Cowen, A., Sauter, D., Tracy, J.L., Keltner, D. (2019). Mapping the passions: toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, *20*, 69–90.

Cowen, A.S., Keltner, D. (2020). What the face displays: mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*. *75*, 349–364

Critchley, H.D., Wiens, S., Rotshtein, P., Öhman, A., Dolan, R.J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*, 189–95.

Desikan, R.S., Ségonne, F., Fischl, B., *et al.* (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*, 968–80.

Dimberg, U., Thunberg, M. (1998). Rapid facial reactions to emotional facial expressions. *Scandinavian Journal of Psychologist*, *39*, 39–45.

Dinstein, I., Gardner, J.L., Jazayeri, M., Heeger, D.J. (2008). Executed and observed movements have different distributed representations in human aIPS. *Journal Neuroscience*, *28*, 11231–9.

Ekman, P., Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto (CA): Consulting Psychologists Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200.

Ekman, P. (1999). Facial expressions In: Dalgleish, T., Power, M., editors. *Handbook of Cognition and Emotion*. New York (NY): John Wiley & Sons Ltd, 301–20.

Grabner, G., Janke, A.L., Budge, M.M., Smith, D., Pruessner, J., Collins, D.L. (2006). Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Medical Image Computing and Computer-Assisted Intervention*, *9*(Pt 2), 58–66.

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

Haxby, J.V., Hoffman, E.A., Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*, 223–33.

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

Jenkinson, M., Bannister, P., Brady, M., Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–41.

Kircher, T., Pohl, A., Krach, S., *et al.* (2012). Affect-specific activation of shared networks for perception and execution of facial expressions. *Social Cognitive and Affective Neuroscience*, *8*, 370–7.

Klami, A., Virtanen, S., Kaski, S. (2013). Bayesian canonical correlation analysis. *Journal of Machine Learning and Research*, *14*, 965–1003.

Kragel, P.A., LaBar, K.S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. *Social Cognitive and Affective Neuroscience*, *10*(11), 1437–48.

Kragel, P.A., LaBar, K.S. (2016). Somatosensory representations link the perception of emotional expressions and sensory experience. *Eneuro*, *3*, 2.

Liang, Y., Liu, B., Xu, J., *et al.* (2017). Decoding facial expressions based on face-selective and motion-sensitive areas. *Human Brain Mapping*, *38*(6), 3113–25.

Lundqvist, D., Flykt, A., Öhman, A. (1998). The Karolinska Directed Emotional Faces–KDEF [CD-ROM]. Stockholm: Karolinska Institutet, Department of Clinical Neuroscience, Psychology section.

Niedenthal, P.M. (2007). Embodying emotion. *Science*, *316*, 1002–5.

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I.P., Hari, R., Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 9599–604.

Nummenmaa, L., Glerean, E., Hari, R., Hietanen, J.K. (2014a). Bodily maps of emotions. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(2), 646–51.

Nummenmaa, L., Hari, R., Hietanen, J.K., Glerean, E. (2018). Maps of subjective feelings. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 9198–203.

Nummenmaa, L., Saarimäki, H., Glerean, E., Gotsopoulos, A., Hari, R., Sams, M. (2014b). Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks. *Neuroimage*, *102*(2), 498–509.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, ID156869.

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioural Brain Science*, *5*, 407–22.

Peelen, M., Atkinson, A., Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal Neuroscience*, *30*(30), 10127–34.

Pourtois, G., Sander, D., Andres, M., *et al.* (2004). Dissociable roles of the human somatosensory and superior temporal cortices for processing social face signals. *European Journal of Neuroscience*, *20*, 3507–15.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*, *59*(3), 2142–54.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, *84*, 320–41.

Saarimaki, H., Ejtehadian, L.F., Glerean, E., *et al.* (2018). Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience*, *13*, 471–82.

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I.P., *et al.* (2016). Discrete neural signatures of basic emotions. *Cerebral Cortex*, *26*(6), 2563–73.

Said, C.P., Moore, C.D., Engell, A.D., Todorov, A., Haxby, J.V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Visualization*, *10*(5), 11.

Smirnov, D., Lachat, F., Peltola, T., *et al.* (2017). Brain-to-brain hyperclassification reveals action-specific motor mapping of observed actions in humans. *PLoS One*, *12*(12), e0189508.

Smith, S.M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–55.

Van der Gaag, C., Minderaa, R.B., Keysers, C. (2007). Facial expressions: what the mirror neuron system can and cannot tell us. *Society Neuroscience*, *2*, 179–222.

Van der Schalk, J., Hawk, S.T., Fischer, A.H., Doosje, B.J. (2011). Moving faces, looking places: the Amsterdam Dynamic Facial Expressions Set (ADFES). *Emotion*, *11*, 907–20.

Van Gerven, M.A., Cseke, B., de Lange, F.P., Heskes, T. (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *Neuroimage*, *50*, 150–61.

Virtanen, S., Klami, A., Khan, S.A., Kaski, S. (2012). Bayesian Group Factor Analysis. *AISTATS, JMLR W&CP.*, *22*, 1269–77.

Wager, T.D., Kang, J., Johnson, T.D., Nichols, T.E., Satpute, A.B., Barrett, L.F. (2015). A Bayesian model of category-specific emotional brain responses. *PLoS Computational Biology*, *11*(4), e1004066.

Wegrzyn, M., Riehle, M., Labudda, K., *et al.* (2015). Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. *Cortex*, *69*, 131–40.

Wicker, B., Keysers, C., Plailly, J., Royet, J.P., Gallese, V., Rizzolatti, G. (2003). Both of us disgusted in my insula. *Neuron*, *40*(3), 655–64.

Williams, P.M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, *7*(1), 117–43.

Winkielman, P., Niedenthal, P., Wielgosz, J., Eelen, J., Kavanagh, L.C. (2015). Embodiment of cognition and emotion. In: Mikulincer, M., Shaver,P.R., Borgida, E., Bargh, J.A., editors. *APA Handbook of Personality and Social Psychology, Vol. 1. Attitudes and Social Cognition*. Washington, D.C. (US): American Psychological Association, 151–75.

Volynets, S., Glerean, E., Hietanen, J.K., Hari, R., Nummenmaa, L. (in press). Bodily maps of emotions are culturally universal. *Emotion*.

Wood, A., Rychlowska, M., Korb, S., Niedenthal, P. (2016). Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends Cognitive Science*, *20*(3), 227–40.