The Visual Features of Emotional Faces That Predict Forced Choice Selection of Faces Download Text Copy to Clipboard

Sjoerd Stuit (s.m.stuit@uu.nl)
Utrecht University
Timo Kootstra
Utrecht University
David Terburg
Utrecht University
Carlijn van den Boomen
Utrecht University
Maarten van der Smagt
Utrecht University
Leon Kenemans
Utrecht University
Stefan Van der Stigchel
Utrecht University

Research Article

Keywords: emotional facial expressions, anger superiority, happiness superiority, visual features

Posted Date: November 30th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-106995/v1

Version of Record: A version of this preprint was published at Scientific Reports on April 15th, 2021. See the published version at https://doi.org/10.1038/s41598-021-87881-w.

Abstract

Emotional facial expressions are important visual communication signals that indicate a sender's intent and emotional state to an observer. As such, it is not surprising that reactions to different expressions are thought to be automatic and independent of awareness. What is surprising, is that studies show inconsistent results concerning such automatic reactions, particularly when using different face stimuli. We argue that automatic reactions to facial expressions can be better explained, and better understood, in terms of quantitative descriptions of their visual features rather than in terms of the semantic labels (e.g. angry) of the expressions. Here, we focused on overall spatial frequency (SF) and localized Histograms of Oriented Gradients (HOG) features. We used machine learning classification to reveal the SF and HOG features that are sufficient for classification of the first selected face out of two simultaneously presented faces. In other words, we show which visual features predict selection between two faces. Interestingly, the identified features serve as better predictors than the semantic label of the expressions. We therefore propose that our modelling approach can further specify which visual features drive the behavioural effects related to emotional expressions, which can help solve the inconsistencies found in this line of research.

Introduction

In any social species, the ability to convey an internal state to nearby members of the social group provides adaptational value. Displaying a particular facial expression is an important way of doing just that. The necessary underlying facial musculature that evolved in non-human primate species is thought to be important for the development of complex social structures (Parr, Winslow, Hopkins & de Waal, 2000; Burrows, Waller, Parr & Bonar, 2006). The adaptive nature of this social ability is thought to result in increasingly pronounced forms of facial expressions. Consequently, facial expressions signalling internal states became more distinctive and, importantly, more prototypical (Darwin, 1896; Ekman, 2006; Eibl-Eibesfeldt, 1989). While humans demonstrate the ability to express a multitude of emotional expressions, the general consensus among research into emotional expressions is that humans invariably display six discrete affects: anger, fear, disgust, happiness, surprise and sadness (Batty & Taylor, 2003; Sprengelmeyer, Rausch, Eysel & Przuntek, 1998). These expressions deviate from the standard facial musculature configuration; the neutral expression. The effects of facial muscular deviation go beyond their effects on the sender of the expression; for instance, faces with emotional expressions attract and hold more visual attention compared to neutral expressions (Vuilleumier & Schwartz, 2001; Palermo & Rhodes, 2006).

Not all expressions affect observers equally. For instance, Hansen and Hansen (1998) observed search asymmetries between particular combinations of emotional expressions. Participants detected angry expressions faster among happy distractors than vice-versa. Additionally, angry expressions were detected faster among neutral distractors than happy expressions among neutral ones. This behavioural finding, an angry superiority effect which was dubbed 'the face in the crowd effect', has been widely replicated since then (e.g. Lundqvist & Öhman, 2005; LoBue, 2009; Ceccarini & Caudek, 2013). However, a

fair amount of research has also found a seemingly opposite effect: an emotional superiority effects for happy facial expressions (e.g. Juth, Lundqvist, Karlsson & Öhman, 2005; Calvo & Nummenmaa, 2008; Hodsoll, Viding & Lavie, 2011). Still, others have argued that there are no emotional superiority effects at all and that reports of such effects reflect differences in the mouth area (Horstmann, Lipp & Becker, 2012). Specifically, visual search was found to be more efficient for emotional expressions with open- versus closed-mouth expressions. The authors conclude that the state of the mouth alone might be sufficient for explaining differences in search efficiency across happy and angry expression. Consequently, they propose the display of teeth as the primary candidate mechanism for this difference (Horstmann, Lipp & Becker, 2012). Note that the influence of teeth is not necessarily about the presence of teeth as objects, but the increase in contrast in the image associated with the presence of teeth (Frischen, Eastwood & Smilek, 2008).

Both the difference in emotional superiority effects and the effect of displayed teeth may be explained by the basic image properties, such as contrast differences, between the expressions (Purcell, Stewart & Skov, 1996; Purcell & Stewart, 2010). In an attempt to explain the variance found across emotional superiority effects, Savage and colleagues (2013) ran a series of experiments using face images from both the NimStim and Ekman & Friesen databases (Tottenham et al, 2009; Ekman & Friesen, 1971). Savage and colleagues found a range of both happy- and angry emotional superiority effects as a function of the stimulus set used. They conclude that emotional superiority effects are not related to the emotional expressions per se, but must be associated with stimulus properties present in the face images. This suggesting was further strengthened by a second set of experiments showing that happy and angry superiority effects depend on the stimulus set used, which remained when faces were presented upside-down (Savage & Lipp 2015). Furthermore, Frischen, Eastwood & Smilek (2008) suggest that displaying teeth in emotional expressions produces a detection advantage because of increased contrast in the mouth area relative to a closed mouth. These findings suggest that attention effects towards emotional expressions may be better explained in terms of basic perceptual features of the images, instead of holistic representations of faces. This raises questions about which stimulus properties are responsible for these attentional effects regarding facial expressions.

A suitable candidate of such a stimulus property appears to be the spatial frequency content of an image, defined by luminance variations cycling over different amounts of space. Note that visual sensitivity depends on both the spatial frequencies and the orientations within an image (Appelle, 1974). Holistic face perception may not be equally reliant on all frequencies in each emotion (Goffaux, & Rossion, 2006; See Jeantet et al, 2018 for a review). Previous research has shown that identification and recognition for different expressions can depend on different spatial frequencies. For example, while identification of happy facial expressions relies on lower spatial frequency content, identification of sad facial expressions relies on higher spatial frequencies (Kumar & Srinivasan, 2011). Moreover, the exact range within lower or higher spatial frequencies that drive emotion identification and recognition varies between emotional expressions (Jeantet et al, 2018). However, as noted by Jeantet and colleagues, the common approach to understanding the relevance of the specific spatial frequency content of faces may have several limitations. In this approach, faces are filtered to contain a specific range of spatial frequencies,

which in turn affects the ecological validity of the results. Moreover, since studies vary widely in their ranges of spatial frequencies, their conclusions depend on what is defined as ‘higher’ versus ‘lower’ spatial frequencies (Jeantet et al, 2018).

Another candidate stimulus property is its local edge orientations: not all oriented edges within a face image are thought to be equally relevant for emotion recognition. For example, horizontal edges are thought to be among the most relevant for recognition (Huynh & Balas, 2014). However, this information was based on the Fourier content of the images and as such does not specify to what structure in the face the horizontal edges belong to. A possible solution to this ambiguity is to extract orientation information locally from the images, which is possible with Histograms of Oriented Gradients (HOGs). In recent years, HOGs have been found suitable descriptors of objects in general, and faces in particular. While, like spatial frequencies features, HOGs are also based on contrast energy, HOGs represent locally formed orientation descriptors of an image (Déniz, Bueno, Salido & De la Torre, 2011). Note that, relative to spatial frequency content, HOG is highly spatially specific. As such, the HOG features can be used to analyse images with great(er) spatial specificity. Although both HOG and Fourier content reflect contrast energy for along different orientations within the images, they are fundamentally different. Specifically, while HOG features are based on Sobel filters and therefore best capture edges, Fourier features are based on waveforms and therefore best capture repeating patterns such as textures. Also, while Fourier content can be transformed back into the original image perfectly using an inverse Fourier transform, this is not possible using HOG features extracted at a single resolution. This shows that the amount of information captured in the HOG features, although highly spatially specific, also comes with the cost of losing much of the information in the original image. Taken together, HOG and Fourier content may capture different aspects of the image and we will therefore use both as complementary descriptions.

In the current study, we aimed to better understand emotional superiority effects by examining the image features associated with forced-choice selection between two expressions. We instructed participants to make an eye-movement to the first face they perceive when two faces are presented simultaneously. Our assumption here is that the face that attracted more attention would be selected first. The unconventional nature of the task is directly related to our analyses, since this simple design allows us to apply a custom feature selection algorithm that aims the find the features of the faces that best predict the participants’ selection. Our main interest is in happy and angry expressions due to previously reported conflicting superiority effects in visual-search. However, we also used sad and neutral faces to increase the variance in the features of the images. In contrast to other feature selection and decoding algorithms, decoding will only serve as the tool for finding behaviourally relevant features. In fact, above chance decoding is mainly relevant because it means we can interpret the selected features used for this decoding as relevant to behaviour. An advantage of using basic, perceptual information is that both spatial frequencies- and HOG-features result in a more detailed representation of a facial expression, one that goes far beyond its semantic category which we will refer to as the holistic expression (e.g. angry, happy, sad and neutral). Consequently, these methods enable more sophisticated, data-driven prediction methods for explaining selection effects for emotional expressions.

Methods

Participants

A total of 102 participants (17 males), 11 of which were left-handed, were included in this study in return for course credit in the Bachelor Psychology program. The mean age of the participants was 21.09 years (*SD = 2.01*). The main reason for the large sample size is that for the current study, both in its paradigm and analyses, no comparable material was available. Therefore we collected data from all participant who applied within a 3-month time frame. All participants indicated normal- or corrected to normal vision, no history of visually triggered epilepsy and no colour blindness. All participants signed an informed consent form before commencing the experiment. This form emphasized that data of the participants would be analysed anonymously and that they were free to leave at any time, without giving any form of formal explanation and without losing their course credits (although these would be scaled to the time spent in the experiment). The study was approved by the local ethical committee of the faculty of social and behavioural sciences at Utrecht University. Furthermore, this research was conducted according to the principles expressed in the Declaration of Helsinki.

Apparatus

The experiment was run on a computer running Windows 7. The monitor used was a linearized 22-inch PHILIPS 202P70/00 with a refresh rate of 85Hz and dimensions of 2048x1536 pixels. To allow for a fast and easy means to perform our task (selecting one out of two faces), eye movements were recorded using an EyeTribe eye-tracker with a 60hz sampling rate and an average spatial accuracy of 0.5 degrees. For integration between the EyeTribe and MATLAB, we used a custom solution developed by Edwin Dalmaijer and colleagues (Dalmaijer, Mathôt & Van der Stigchel, 2014). To provide for optimal measurements during eye-tracking, a metal headrest was used to stabilize the participant's heads. Both this headrest and the seating for this experiment were adjusted in height for to allow participants to look straight ahead while sitting comfortably.

Stimuli

All stimuli were presented via MATLAB 2016a and Psychtoolbox 3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). The stimuli used consisted of greyscale photographs of happy, angry, neutral and sad frontal-facing facial expressions with a frontal gaze, with 39 different identities, from the Radboud Faces database (Langner et al., 2010). All stimuli were of adult Caucasians. Each trial contained a central circular fixation point (0.6 degrees of visual angle diameter). Two face images were presented 14.6 degrees of visual angle into the periphery in a circular apertures of 14.6 x 14.6 degrees of visual angle (632 by 632 px) at a viewing distance of 57 centimetres. The size of the faces was matched against the perception of faces in real life situations based on research by Miller (2014).

Procedure

Prior to the experiment, participants were given both written and spoken instructions concerning the procedure of, and instructions for, the experiment. Next, the EyeTribe was calibrated for the participant. This procedure was repeated after each break in which the participant left the experimental environment. Breaks were built in once every 125 trials (4-7 minutes). It was up to the participant's discretion to either take a break, or continue with the experiment. At the start of each trial (Figure 1), a grey background was presented with a pseudo-random duration between 500-1500ms. Next, a fixation point was added to the centre of the screen. This fixation point was presented until the participant's gaze was registered at this location by the EyeTribe. Subsequently, two images of faces with emotional expressions (angry, happy, sad or neutral) were presented on both the left and right side of the screen. Participants were instructed to make an eye movement to the first face they perceived. As soon as gaze-location overlapped with one of the two presented faces, the trial was ended and the faces were removed from the screen. Note that the motivation for using an eye-movements as a means for selection instead of manual responses is to allow participants to make a rapid and natural response. The faces were always from different identities and could also be different in their displayed expression, resulting in a total of 16 conditions (4 possible expressions on the left side of the screen times four possible expressions on the right side of the screen). Conditions were counterbalanced for stimulus location and emotional content and each condition was presented 56 times. To aid compliance with the task, the experiment also contained additional trials which were identical to the trials described above with one exception, there was a temporal offset (between 17 and 134ms) between the presentations of the two faces. Each of the conditions was presented 19 times, resulting in a total of 1120 trials for the full experiment.

A schematic representation of one trial. The upper right arrow represents the maximum duration of the trial. The bottom left arrows represent the trial events. Every trial started with a grey background. After small delay, the fixation point was presented. When the participant's gaze was registered on this fixation point, two faces were presented. The faces always had different identities and could have different or the same expression. Participants were instructed to make an eye movement to the first face they perceived after which the trial ended. Note that only 20% of the trials contained an actual temporal offset between the presentation of the two faces.

Feature Extraction & Labelling

Spatial frequency information was estimated using the Fourier Magnitude Spectrum (Spatial Frequency, SF). Note that magnitude spectrum is not informative of spatial position of a particular contrast. The Fourier Magnitude Spectrum was subdivided into 24 spatial frequencies and 16 orientations, with the magnitudes summed for each oriented SF, resulting in 384 features describing the contrast energy in the images for different SFs and orientations (Figure 2). Note that magnitudes for frequencies higher than the minimal Nyquist frequency (the Nyquist frequency for the cardinal axes) are excluded.

For HOG, we subdivided the images into 10 x 10 px non-overlapping sections (Figure 3). For each section we extracted the power in 9 orientations, resulting in a total of 3600 features describing the spatial orientation structure of the image. Of these, 1116 were excluded from the analysis since they represented

sections of the images outside of the aperture used when presenting the face images, resulting a total of 2484 HOG features used in our analyses. Since the spatial resolution of the HOG features (their cell size) affects what structural components of images are represented, which in turn may affect performance, we extracted two additional sets of HOG features. One with a 20 x 20 px cell size which, after excluding those that corresponded to locations outside of the aperture, resulted in 540 HOG features. The third set used a 40 x40 px cell size which, after excluding those that corresponded to locations outside of the aperture, resulted in 81 HOG features.

Since we used two faces next to each other in each trial, we subtracted the feature values of the left images from the feature values of the right image (Figure 2 & Figure 3). If the right image received the first eye-movement, this trial was labelled as a 1, if not it was labelled as a 0.

Visualisation of the Fourier feature extraction showing an example of two images used in a trial, their respective, down sampled Fourier features and the Fourier feature differences map. Note that in the Fourier Maps each location corresponds to a particular combination of a spatial frequency and an orientation. The Fourier maps all have horizontal contrasts along the horizontal axes, and vertical contrast along the vertical axes. The radial axes are for cycles per image (abbreviated to cpi in the figure; ranging from low in the centre to high near the edges). Luminance intensity, from black to white, indicates the relative strength of the contrast for the corresponding section of the map. The Fourier feature differences map was calculated by subtracting the down sampled Fourier features from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

Visualisation of the HOG feature extraction showing an example of two images used in a trial, their respective HOG features and the HOG feature differences map using the highest resolution (10x10 cell size). All HOG maps use the same x and y axes as the original images, meaning position in the HOG map is directly coupled with position in an image. The HOG maps show 20x20 grids where each position in the grids represents an area of 10x10 pixels. For each 10x10 area of pixel in an image, the weights for 9 differently oriented gradients is calculated. The 9 weights are visualized by white bars where the length reflects the weights. The 9 bars are then superimposed on the 10x10 pixel area there are based on. The HOG feature differences map was calculated by subtracting these HOG features weights from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

Data Splitting & Cross-validation

After feature extracting and labelling the data for a participant, the full data set was divided into 8 partitions, with each partition containing approximately the same balance between the two classes (class 1: left image received eye-movement, class 0: right image received eye-movement) of trials (those labelled as 0 and those labelled as 1), for 8 fold cross-validation. To avoid any possible temporal order effect, each partition contained trials from throughout the experiment. For each fold, one partition was set aside for cross-validation (referred to as the hold out set), the other 7 were used for feature selection. During the

feature selection procedure, these seven partitions will be further split up into train sets and feature selection validations sets.

Feature Selection

The feature selection algorithm used is a custom built algorithm combining a filter method and a wrapper method (Kohavi & John, 1997). Note that, at each step of the feature selection procedure, that is every time features were tested for their additive value to a model, we pseudo randomly selected approximately 50% of the data as a train set with equal representations of each class (selection of the left versus the right face) . The residual data, roughly 50% of the data from the 7 partitions, was used as a validation set to obtain indicators of the relevance of each tested feature. Data from the hold out set was not part of this subdivision and only used for cross-validation after feature selection was complete. Specifically, for each fold, we first rank each feature based on its chi-squared statistic using Kruskal-Wallis analysis of variance (Figure 4, Step 2). Next, the algorithm uses an iterative wrapper approach to select features to be included into a model based on classification performance, using a linear support vector machine (SVM), in a stepwise additive manner. (Figure 4, Step 3). The features that are tested are a subsample of all available features, based on the current ranking. Using the validation set, performance associated with the tested features is ranked based on their F1 score (a measure of accuracy taking into account true positive, false positive and false negatives) since, unlike any training set, the validation set does not necessarily contain equal representation of trials with each of our labels. The best performing features are selected for inclusion (Figure 4, Step 4). Note that the ranking for selection for testing is continuously updated based on the validation performance (Figure 4, Step 5). This process repeats until the models consisted of a maximum number of features, set individually for each fold of each participant (Figure 4, Step 6). There is, however, one additional step: When 25% of the maximum number of features has been included, the currently included features are all tested separately to see if their inclusion adds anything to the performance of the current model. Features that do not aid the performance of the model are excluded from it. The maximum number of features is based on the initial chi-squared ranking; this is the minimum number of features required to collectively contain 10% of the total sum of chi-squared scores of all features in the analysis. In other words, very high chi-squared values found in the initial ranking result in a lower number of features used for a model. Our assumption here is that, even though the initial ranking does not take into account interactions between features, higher degrees of separability of the data reflected in the chi-square statistics means that decoding should be possible with fewer features. Note that analyses using HOG and spatial frequency features are done separately.

Schematic representation of the feature selection algorithm used in the current project. Step 1) a random collection of features (indicated by the red bars) is selected for a random model. Step 2) The features are ranked based on Chi-Square scores. The top of the ranking is used to determine the filter model. Step 3) A search space is defined from the top-ranking features and the features in this search space are tested for inclusion into the wrapper model (Step 4). Step 5) The Chi-Square based ranking of search space features is updated based on their performance. Step 6) Steps 3-5 are repeated until enough features have been selected for the wrapper model. Step 7) From the residual features, unused by the wrapper model, a

random selection is made for a pseudo random model. Step 8) Each of the four combinations of features are used to train classification models and cross-validation performance is subsequently estimated using the hold-out data. The final model is selected based on highest cross-validation performance (P). See the above section Feature Selection, and the below section Final Model Selection for additional details.

Final Model Selection

The feature selection algorithm results in four sets of features on which classification models are trained using the current train and validation data and cross-validated on the holdout data. The first model, referred to as the filter model (Figure 4, step 2), uses the features required to collectively contain 10% of the sum of the chi-squared scores of all features. The second model, referred to as the wrapper model, uses the same number of features but with the more extensive selection procedure described above (Figure 4, step 6). The last two models also contain the same number of features, but here the features are selected either randomly or pseudo randomly. The random model simply selects a random collection of features for training (Figure 4, step 1). The pseudo random model selects a random collection of the residual features not used in the current filter or wrapper model (Figure 4, step 7). These last two models were included as back up options for when the wrapper or filter model failed to predict the hold out data well. Moreover, if feature selection is irrelevant, all models should perform equally well. Our final model (Figure 4, step 8), referred to as the minimal model, was the best performing, based on the accuracy score, of these models (excluding the full model) since it reflected, empirically, the features that best predicted our participants behaviour. For comparison, we trained one additional model, referred to as the full model, which includes all available features. If all features are relevant for prediction, the full model should perform best.

Estimating Chance Performance

Based on the procedure for model selection described above, comparisons between minimal model performances and chance levels are problematic. Specifically, since the final model performance is the highest performance associated with any of the four models, the final model performance can, theoretically, reach an above chance performance even when all models perform around chance level, simply because the algorithm always selects the highest performing model. To overcome this issue, we estimated the empirical chance level under the same conditions as the selection of the minimal model. For this estimation, all feature values are shuffled into a random arrangement such that there should be no residual relationship between the class of an example and the associated features. The four models are then trained and crossed validated using the shuffled data. From the four resulting cross validation performances, the highest accuracy is as used as empirical chance performance. In this way, the ability to exceed chance level performances in the absence of a relationship between the classes and the features is the same for the minimal model and the final control model and their performances can be compared directly. Analyses files available at https://osf.io/ms8df/.

Results

Expression Selection Behaviour

To test for biases in selection based on the holistic expressions, we first analysed selection behaviour for trials where the two faces were presented at the same time and expressed different emotions. We found a bias in the probability of selecting happy faces (Friedman ANOVA, Chi-Squared (3, n=408) = 14.52, $p < 0.01$; Figure 5A). These results suggest a happy superiority effect as previously reported by (Juth, Lundqvist, Karlsson, & Öhman, 2005; Savage, Lipp, Craig, Becker, & Horstmann, 2013; Savage & Lipp, 2015). Average median reaction times for trials where dissimilar expressions were presented (Mdn = 0.23 s) did not differ from reaction times where similar expressions were presented (Mdn = 0.24 s; two-sided Wilcoxon Signed-Ranks test, Z = -1.65, p = 0.10; Figure 5B). Finally, consistent with previous reports (Ossandón, Onat & König, 2014), we found that 61% of participants were biased to make leftward eye-movements (Figure 5C). Note that our decoding procedure cannot learn from these biases as it always balances the training to contain equally amounts of data from leftward and rightward eye-movements.

In figure 5A we show the average fraction of selecting, across participants, (y-axis) a particular expression (x-axis; HA: Happy, AN: Angry, NE: Neutral & SA: Sad) for trials in which two different expressions were displayed. Errorbars represent the Standard Error of the Mean. Results show a significant positive bias for happy facial expressions. Figure 5B shows the distributions of the average median reaction times of all participants for the trials with different, and the trials with the same expressions. Figure 5C shows the distributions of left- and right-wards biases for all participants. Overall, 61% of the participants had a biases towards making leftward eye-movements.

Expression Selection Decoding – Different Expressions

To uncover which stimulus features best predict the first face selected, we first looked at the feature selection and decoding results when using only trials in which two different expressions were presented. Specifically, we trained the linear SVM using all HOG feature differences (referred to as HOGfull), using a subset of relative HOG features picked by our selection algorithm (referred to as HOGmin), using all SF feature differences (referred to as SFfull) and finally, using a subset of relative SF features as selected by our algorithms (referred to as SFmin). Since HOG spatial resolution (10x10, 20x20, 40x40) did not affect decoding performance for either the full model (Friedman ANOVA, Chi-Squared (2,n=305) = 2.44, p = 0.29), nor the minimal model (Friedman ANOVA, Chi-Squared (2,n=305) = 0.31, p = 0.73), results for the three resolutions were averaged for each participant. The HOGmin models used 30.23% filter models, 25.29% wrapper models, 23.28% random models and 21.20% pseudo random models. An average of 3.5% of the available features were used in these models. For the SFmin model these percentages were 30.64, 25.12, 22.55 and 21.69%, respectively. The SFmin model used an average of 2.1% of the available features. All full and final models were able to decode selection above chance, although negligibly for the full model, and performance was significantly higher when using our minimal model procedure compared to using all image-features (Figure 6 & Table 1). The relevance of different locations was estimated based on the average performances associated with the HOG features and are shown as a heatmap in Figure 7A. The relevance of different spatial frequencies and orientations, expressed as their contribution to over

decoding performances are shown in green in Figure 7C&D. Results suggest mainly horizontal, low spatial frequencies, and the mouth and cheek areas are used for predicting selection.

Table 1: Statistical Test Outcome – Accuracies Different Expressions Trials

Comparison	*Median(s)*	*test*	*outcome*
HOGfull against control	0.51017/0.50025	two-sided Wilcoxon Signed-Ranks	Z = 4.145, p < 0.001
SFfull against control	0.50893/0.50	two-sided Wilcoxon Signed-Ranks	Z = 3.4981, p < 0.001
HOGmin against control	0.5558/0.50025	two-sided Wilcoxon Signed-Ranks	Z = 8.768, p < 0.001
SFmin against control	0.56399/0.50	two-sided Wilcoxon Signed-Ranks	Z = 8.7101, p < 0.001
HOGmin against HOGfull	0.5558/0.51017	two-sided Wilcoxon Signed-Ranks	Z = 8.733, p < 0.001
SFmin against SFfull	0.56399/0.50893	two-sided Wilcoxon Signed-Ranks	Z = 8.7117, p < 0.001

In figure 6A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on the trials where different expressions were presented to the participants. The dotted line represents the overall empirical chance level performance. Errorbars represent the Standard Error of the Mean. 6B-C) Confusion Matrices for the minimal models. For all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that, performance is nearly equal for all combinations of expressions.

Visual representations of the most relevant features for decoding. 7A-B) Heatmaps reflecting the relevance of spatial locations of the HOG features to decoding either trials with different expressions (A) or the same expression (B) overlaid on the averages of all neutral expressions. As relative importance of a location increases, colour changes from blue through green to yellow. 7C) Here we show the weight, reflecting the percentage of contribution to overall performance, for each band of spatial frequencies used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. 7D) Here we show the weight for each band of orientations used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. Note that, for both spatial frequency and orientation, the only clear difference is a larger weight for horizontal orientations in trials where the expressions differ.

Expression Selection Decoding – Same Expressions

If selection is based on basic stimulus properties, a difference between the holistic expressions should not be required. To test this, we next looked at the face selection behaviour when choosing between two faces displaying the same expression but with different identities. We again trained four linear SVMs (HOGfull, HOGmin. SFfull, SFmin) Because HOG spatial resolution did not affect decoding performance for either the full model (Friedman ANOVA, Chi-Squared (2,n=305) = 0.66, p = 0.51), nor the minimal model (Friedman ANOVA, Chi-Squared (2, n=305) = 0.33, p = 0.72), results for the three resolutions were averaged for each participant. The HOGmin models were based on the filter method in 30.69% of the folds. The wrapper model was used in 25.08%, the random model in 22.59% and the pseudo random model in 21.36%. An average of 3.5% of the available features were used in these models. For the SFmin model these percentages were 31.37, 25.12, 23.28 and 20.22% respectively. The SFmin model used an average of 2.1% of available features. We found that only the minimal models, not the full models, resulted in significant decoding performance and again found that performance was significantly higher when using our minimal model procedure compared to using all image-features (Figure 8 & Table 2). The relevance of different locations are shown as a heatmap in Figure 7B. The relevance of different spatial frequency orientations are shown in green in Figure 7C&D. Results suggest mainly horizontal, low spatial frequencies, and the edges around the nose, cheeks and forehead areas are used for predicting behaviour.

Table 2: Statistical Test Outcome – Same Expressions Trials

Comparison	*Median(s)*	*test*	*outcome*
HOGfull against control	0.50595/0.49702	two-sided Wilcoxon Signed-Ranks	Z = 1.8142, p = 0.06964
SFfull against control	0.50893/0.50223	two-sided Wilcoxon Signed-Ranks	Z = 2.2627, p = 0.05
HOGmin against control	0.59524/0.49702	two-sided Wilcoxon Signed-Ranks	Z = 8.7683, p < 0.001
SFmin against control	0.59821/0.50223	two-sided Wilcoxon Signed-Ranks	Z = 8.7529, p < 0.001
HOGmin against HOGfull	0.59524/0.50595	two-sided Wilcoxon Signed-Ranks	Z = 8.7684, p < 0.001
SFmin against SFfull	0.59821/0.50893	two-sided Wilcoxon Signed-Ranks	Z = 8.6867, p < 0.001

In figure 8A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on the trials where the same expression was present to both the left and the right side of the screen, leaving only a difference in identity. The dotted line represents the overall empirical chance level performance. Errorbars represent the Standard Error of the Mean. 8B-C) Confusion Matrices for the minimal models. For

all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that the differences in performance are very small.

Decoding the expressions in the images using the features selected to decode behaviour

To better understand the image content captured by the selected features, we tested how well these features decoded the semantic label of the four used expressions (39 images per expression). As a reference point for performance, we first used the same feature selection method as used for decoding behavioural selection to estimate performance when the feature selection algorithm had the same degrees of freedom as the feature selection algorithm used to decode selection behaviour in the current experiment. When decoding expressions, the minimal model based on HOG features used 4.2% filter models, 83.3% wrapper models and 12.5% random models, using an average of 3.9% of all available features. For expression decoding based on SF features, the minimal model used 62.5% filter models and 37.5% wrapper models with an average of 3.3% of the features. Next, using the same exact same folding as used for this expression decoding, meaning the separations into train and test data were identical, we trained expression decoding models using only the top n features based on decoding face selection in the current experiment. Here, n is based on, and the same as, the number of features used in each fold of the reference's expression decoding procedure. While performances for expression decoding were well above chance level (25%) when the algorithm was free to choose the features used for decoding, expression decoding using the features found for our behavioural data was below our reference analysis (where the feature selection algorithm decided the features to use for decoding). Moreover, only the high-resolution HOG features based on decoding eye-movement behaviour are relevant for decoding expressions which is striking since decoding selection behaviour in our task was unaffected by the spatial resolution of the HOG features..

Average decoding performances across all folds (y-axis) based on different sets of HOG and Fourier features. The dotted black lines represent chance level performance. Errorbars represent the Standard Error of the Mean. Figure 9A shows average performance for decoding expressions based on HOG features using three different feature sets at three different spatial resolutions (x-axis). EDf (Expression Decoding features) uses a feature set based on feature selection for expression decoding (i.e. decoding the semantic label of the expression, not selection behaviour), DETf (Different Expressions Trials features) uses the features based on decoding eye-movements towards faces with different expressions and SETf (Same Expressions Trials features) uses the features based on decoding eye-movements towards faces with the same expression. Figure 9B shows average performance for decoding expressions based on Fourier features, again using three different feature sets (x-axis). Note that, for both HOG and Fourier features, the features based on decoding selection behaviour are suboptimal for decoding expressions. Moreover, only the high-resolution HOG features based on decoding eye-movement behaviour are relevant for decoding expressions.

Comparison between predictions based on visual features and expression preferences

Finally, we evaluated what can best predict selection: the holistic expressions or the differences in visual features. Therefore, we aimed to compare the predictive value from our modelling approach to the predictive value based on behavioural biases between expressions as seen in the face selection data (Figure 5). If our participants simply have a set bias to select a particular expression over another in each combination of two expressions, and those behavioural biases are the reason we can decode selection, we would expect prediction accuracy based on these biases towards particular expressions to be as large as or larger than the prediction accuracy of our models. In other words, the maximum decoding performance would be determined by the degree of preference for one expression over another. However, if the feature differences, independent of the expressions, provide better insight into the face selection of our participants than the holistic expression, our models should be better at predicting which expression receives the first eye-movement compared to the biases seen in behaviour based on the holistic expression. To be able to directly compare cross-validation performance of the models (seen in Figure 6) to the biases seen in behaviour (Figure 5), we created an additional decoding procedure. For each individual participant, using the same folding as was used for that individual, we determined the preference of one expression over another (for each expression pair separately) based on the training data and used that preference to predict face selection in the cross-validation data. The eight resulting performances of each participant for each expression pair were averaged, resulting in one average performance for each pair of expressions. Likewise, we extracted the associated feature selecting decoding performance for each participant and each expression pair. This procedure was performed separately for HOGmin performances and folding and the SFmin performances and folding. The results for the two analyses were averaged and plotted against each other in Figure 10. The diagonal line indicates the points where percentage correct for the predictions based on features and those based on biases towards particular expressions are equal. Note that most decoding performances based on visual features are above the diagonal. This suggests that our models were better at predicting the first eye-movement from the differences in the image features than preferences for particular expressions seen in the behaviour of our participants. In fact, results show that performance based on image features is significantly higher than when based on biases within expression pairs (Mdn = 55.95 against Mdn = 52.09 respectively, two-sided Wilcoxon Signed-Ranks test, $Z = 15.70$, $p < 0.001$). Furthermore, we also compared performances based on image features and biases for each expression pair separately and found that only for pairs of happy and neutral faces we can predict selection based on previous biases (Mdn= 53.09 against chance (50), two-sided Wilcoxon Signed-Ranks test, $Z = 2.20$, $p < 0.001$).

Figure 10A shows the relation between the average percentage correct prediction for each participant and each combination of expressions based on the biases towards expressions (x-axis) and the average corresponding percentage correct prediction based the visual features of the images (y-axis). The dotted vertical line represents chance level performance for expression-based decoding. The dotted horizontal line represents chance level performance for visual feature-based decoding. The solid diagonal line shows where performance would be equal. Figure 10B shows, for each combination of expressions separately, the percentage of the predictions where visual feature-based prediction out-performed

prediction based on biases towards expressions. Note that for all expressions this is the case (all values well above 50%).

Discussion

Here, we aimed was to find the specific visual features in emotional expressions that predict rapid selection between two emotional expressions. We first show that initial selection between two faces is biased towards happy facial expressions (Figure 5) compared to angry, sad and neutral faces. More importantly, we found that selection behaviour can be predicted using the differences in either the spatial-structure information (represented with HOG features) or the spatial-frequency contrast information (represented by the Fourier Magnitude Spectrum) in the face images (Figures 6 & 7). Note that selection could also be predicted for trials where the faces did not differ between expressions, but still had different identities (Figures 7 & 8). The spatial frequency features relevant for decoding behaviour are also sufficient for classifying the emotional expressions in the images used in the current experiment (Figure 9). For the HOG features however, only HOG features sampled at the highest resolution were relevant for classifying the emotional expressions used in our task, even though HOG spatial resolution did not affect the ability to decode *selection* during the behavioural task. This suggests that decoding behavioural selection based on the lower resolution HOG features is not based on features that capture the emotional expression. We go on to show that we can predict behaviour better when based on image features than when based on the emotion represented by the expressions (Figure 10). Taken together, these results suggest that basic visual features can serve as better predictors for visual selection than the holistic emotional expression itself. Crucially, we also show *what* aspects of the images have predictive value in a data driven manner (Figure 7). As such, our results give insight into what aspects of faces affect selection in our current task. However, we deliberately kept our task rather minimalistic to allow for a (relatively straight-forward), feature selection analysis. Although this means that generalisation of our specific findings to other tasks may be limited, we argue this approach should be applied to other visual attention paradigm as well (e.g. visual-search tasks, emotion-interference task, etc.), as it provides the possibility to identify those visual features that underlie behavioural effects related to emotional expressions in other tasks. Since we show significant decoding of behaviour, the selected features can be interpreted as relevant to that behaviour. Before we discuss our main interpretations of the results, we will briefly discuss the unconventional nature of our task and approach such that the strengths and limitation can be taken into account.

The task designed for the current experiment was made specifically for the current feature selection procedure. Since it resulted in a single selection based on only two images per trial, the data could be analysed using the current feature selection procedure, something that would not have been possible using a more standard visual search task involving many different faces. The simplistic nature of the task may result in more reflexive behaviour from the participants, in turn favouring predicting via visual features over prediction via holistic expressions. It is possible that when behavioural biases become more obvious, the predictive value of our models will no longer surpass the biases based on expressions seen in behaviour and will start to align with them instead. Furthermore, the idea that selection in our task is

related to attention remains an assumption. Our main aim was to validate a data driven method to find visual features of emotional expressions that predicts human behaviour. As such, the current results provide a means to investigate behaviours related to emotional faces in a highly specific, data driven manner.

Still, we argue there are several advantages to the current approach. First and foremost is the level of detail attained. We do not only show which holistic expressions are preferred for selection, or what coarse structures (such as the mouth) influence selection, we show which specific features of the images are relevant for predicting selection. For example, the mouth area was previously suggested by Savage and colleagues (2013) to be relevant in order to explain inconsistencies between studies reporting on happy and/or angry superiority effects. However, here we show the effect of the mouth may mostly be indirect. Specifically, we show that the nasolabial folds in the cheeks appeared to be more important than the mouth itself. Note that, when similar expressions are presented to our participants, the results show that the relevant features are much less focussed on any specific area of the faces. This suggests that when the differences between the two faces are smaller, participants base their selection on differences throughout the face. Other approaches, such as filtering images (e.g. Deruelle & Fagot, 2005; Kumar & Srinivasan, 2011; Goffaux, Hault, Michel, Vuong & Rossion, 2005), eye tracking (Frischen, Eastwood & Smilek, 2008) and using composite images (Lundqvist & Ohman, 2005; Stein & Sterzer, 2012) have not yet attained this level of detail, nor do they give any indication on predictive value.

Our main consideration comes back to emotional superiority effects. Why does the literature consistently show emotional superiority effects but with inconsistent directions (See Savage et al, 2013 and Savage & Lipp, 2015)? Although we cannot generalize our results in a way that would answer this question directly, we now know that both spatial frequency and HOG differences are relevant for selection between expressions and that there is a dissociation between prediction via features and via holistic expressions. Emotional expressions are inherently visual, and expressions are inherently prototypical (Ekman, & Friesen, 1971). Therefore, even though our results show the features of an expression are better predictors than its holistic emotional expression, a clear dissociation between the expression and the features that form that expression is impossible. Here we show that, even though spatial frequency content is down-sampled and phase information is ignored, we can use it to decode the expressions used in the current experiment. This suggests that spatial frequency and orientation patterns are prototypical for different expressions. Since contrast sensitivity varies with spatial frequency and orientation, and spatial frequency and orientation content are predictive for both selection between two expressions and decoding the expressions themselves, some faces may simply result in stronger visual signals than others based on their respective spatial frequency content. Note that the influence of spatial frequency is often accounted for by adding control conditions containing inverted faces with the assumption that the spatial frequency is unaffected by face inversion. However, while this holds for images with perfect vertical symmetry, without perfect symmetry, this assumption is only correct for contrasts in the cardinal orientations. Even though contrast sensitivity for diagonal orientations is relatively low (Appelle, 1972), as is their relevance for predicting selection in the current experiment (Figure 7C), neither sensitivity nor the relevance for prediction should be ignored. Note though, that face-inversion did not influence the ability to

decode the expressions used in our experiment (data not shown). That being said, face-inversion does also not reliably remove effects related to expressions (Savage & Lipp 2014). As such, the influence of spatial frequency and face inversion on expression selection behaviour needs to be tested empirically to fully disentangle the influence of spatial frequencies from those directly related to holistic expressions.

At high and medium spatial resolutions, HOG features decode expressions better than spatial frequency content. These features reflect structural aspects of the images and are therefore more likely to be directly related to human expression recognition. However, with exception of the HOG features selected for decoding selection using the highest spatial resolution, the HOG features found relevant to our task are not able to decode the expressions of the images used in the current experiment. This is in stark contrast to the effect HOG spatial resolution had on the ability to decode selection behaviour. For behaviour, all spatial scales worked equally well for decoding behaviour. This suggest that we can decode selection behaviour also with HOG features that cannot be used to decode expressions. Moreover, we show we can also decode face selection between two faces with the same expression. Taken together, this suggests that structural parts of the face, specifically configurations of oriented edges, are sufficient to predict selection (even more so than the holistic expression) and these feature are not likely related to specific expressions. Taking this into account, as well as our results concerning spatial frequency features, the inconsistency problem with emotional superiority effects may lie in heterogeneity both within and between the holistic expressions. These inconsistencies may be resolved when taking the visual features of the images into account. Our current approach is of course only a first step in that direction.

In conclusion, here we show what subsets of the spatial frequency and HOG feature content of emotional expressions predicts selection between two faces. Our results suggest that such features of emotional expressions serve as a better predictor compared to the holistic expressions. We suggest that the current approach allows for a better understanding of how different expressions affect behaviour by focussing on data driven stimulus features rather course expression category labels that obscure important variations within the faces.

Declarations

Acknowledgements

This work was supported by a research grant of the focus area Applied Data Science (Utrecht University, www.uu.nl/ads)

Author Contributions

S.M.S., Wrote the initial manuscript, designed the analysis.

T.M.K., Build the behavioural experiment, wrote the initial experimental methods sections.

D.T. Revised the manuscript.

C.B., Acquired the data & revised the manuscript.

M.J.S. Revised the manuscript.

L.J.K. Revised the manuscript.

S.S. Revised the manuscript.

Competing interests
The author(s) declare no competing interests.

References

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin, 78*(4), 266–278. https://doi.org/10.1037/h0033117

Batty, M., & Taylor, M. J. (2003). Early processing of the six basic facial emotional expressions. *Cognitive Brain Research*. *17*(3), 613 – 620.

Burrows, A. M., Waller, B. M., Parr, L. A., & Bonar, C. J. (2006). Muscles of facial expression in the chimpanzee (Pan troglodytes): descriptive, comparative and phylogenetic contexts. *Journal of anatomy*, *208*(2), 153-167.

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

Calvo, M. G., & Nummenmaa, L. (2008). Detection of emotional faces: salient physical features guide effective visual search. *Journal of Experimental Psychology: General*, *137*(3), 471.

Ceccarini, F., & Caudek, C. (2013). Anger superiority effect: The importance of dynamic emotional facial expressions. *Visual Cognition*, *21*(4), 498-540.

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014). *Behavioural Research Methods*, *46*, 913. https://doi.org/10.3758/s13428-013-0422-2

Darwin, C. R. (1896). The expression of emotions in man and animals. New York: Philosophical Library.

Déniz, O., Bueno, G., Salido, J., & De la Torre, F. (2011). Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, *32*(12), 1598-1603.

Deruelle, C., & Fagot, J. (2005). Categorizing facial identities, emotions, and genders: Attention to high- and low-spatial frequencies by children and adults. *Journal of experimental child psychology*, *90*(2), 172-184.

Eibl-Eibesfeldt, I. (1989). Foundations of human behavior. *Human ethology. Hawthorne, NY, US: Aldine de Gruyter*.

Ekman, P. (Ed.). (2006). *Darwin and facial expression: A century of research in review*. Ishk.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124–129. https://doi.org/10.1037/h0030377

Frischen, A., Eastwood, J. D., & Smilek, D. (2008). Visual search for faces with emotional expressions. *Psychological bulletin*, *134*(5), 662.

Goffaux, V., Hault, B., Michel, C., Vuong, Q. C., & Rossion, B. (2005). The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception*, *34*(1), 77-86.

Goffaux, V., & Rossion, B. (2006). Faces are "spatial"--holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology: Human Perception and Performance, 32*(4), 1023–1039. https://doi.org/10.1037/0096-1523.32.4.1023

Hansen, C. H., & Hansen, R. D. (1988). Finding the face in the crowd: an anger superiority effect. *Journal of personality and social psychology*, *54*(6), 917.

Hodsoll, S., Viding, E., & Lavie, N. (2011). Attentional capture by irrelevant emotional distractor faces. *Emotion*, *11*(2), 346.

Horstmann, G., Lipp, O.V., & Becker, S. (2012). Of toothy grins and angry snarls -Open mouth displays contribute to efficiency gains in search for emotional faces. *Journal of Vision*, *12*(5), 1-15. doi: 10.1167/12.5.7

Huynh, C. M. & Balas, B. (2014). Emotion recognition (sometimes) depends on horizontal orientations. *Attention, Perception, & Phychophysics, 76,* 1381-1392.

Jeantet, C., Caharel, S., Schwan, R., Lighezzolo-Alnot, J., and Laprevote, V. (2018). Factors influencing spatial frequencies extraction in faces: a review. *Neuroscience and Biobehavioral Reviews*, *93*, 123-138.

Juth, P., Lundqvist, D., Karlsson, A., & Öhman, A. (2005). Looking for foes and friends: perceptual and emotional factors when finding a face in the crowd. *Emotion*, *5*(4), 379

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox 3? *Perception*, *36*, ECVP Abstract Supplement.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273-324

Kumar, D., & Srinivasan, N. (2011). Emotion perception is mediated by spatial frequency content. *Emotion*, *11*(5), 1144.

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 and*

Emotion, *24*(8), 1377-1388, DOI: 10.1080/02699930903485076

LoBue, V. (2009). More than just another face in the crowd: Superior detection of threatening facial expressions in children and adults. *Developmental Science*, *12*(2), 305-313.

Lundqvist, D., & Ohman, A. (2005). Emotion regulates attention: The relation between facial configurations, facial emotion, and visual attention. *Visual Cognition*, *12*(1), 51-84.

Miller, R. S. (2014). *Intimate Relationships* (7th ed.). McGraw-Hill.

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, *45*(1), 75-92.

Parr, L. A., Winslow, J. T., Hopkins, W. D., & de Waal, F. (2000). Recognizing facial cues: individual discrimination by chimpanzees (Pan troglodytes) and rhesus monkeys (Macaca mulatta). *Journal of Comparative Psychology*, *114*(1), 47.

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

Purcell, D. G., Stewart, A.L. & Skov, R.S. (1996). It Takes a Confounded Face to Pop Out of a Crowd. *Perception*, *25*,1091-1108.

Purcell, D.G., & Stewart, A.L. (2010). Still another confounded face in the crowd. *Attention, Perception & Psychophysics*, *72*(8), 2115-2127.

Ossandón, J. P., Onat, S., & König, P. (2014) Spatial biases in viewing behavior. Journal of Vision, 14(2):20, 1–26, http:// www.journalofvision.org/content/14/2/20, doi:10.1167/14.2.20.

Savage, R. A., Lipp, O. V., Craig, B. M., Becker, S. I., & Horstmann, G. (2013). In search of the emotional face: Anger versus happiness superiority in visual search. *Emotion*, *13*(4), 758.

Savage, R., & Lipp, O. V. (2015). The effect of face inversion on the detection of emotional faces in visual search. *Cognition and Emotion*, 29:6, 972-991, DOI: 10.1080/02699931.2014.958981

Sprengelmeyer, R., Rausch, M., Eysel, U. T., & Przuntek, H. (1998). Neural structures associated with recognition of facial expressions of basic emotions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1409), 1927-1931.

Stein, T., & Sterzer, P. (2012). Not just another face in the crowd: detecting emotional schematic faces during continuous flash suppression. *Emotion*, *12*(5), 988.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, BJ, & Nelson, C. (2009). The NimStim set of facial expressions: Judgements from untrained research participants. *Psychiatry Research*, *168*(13), 242-249.

Vuilleumier, P., & Schwartz, S. (2001). Emotional facial expressions capture attention. *Neurology*, *56*(2), 153-158.

Figures

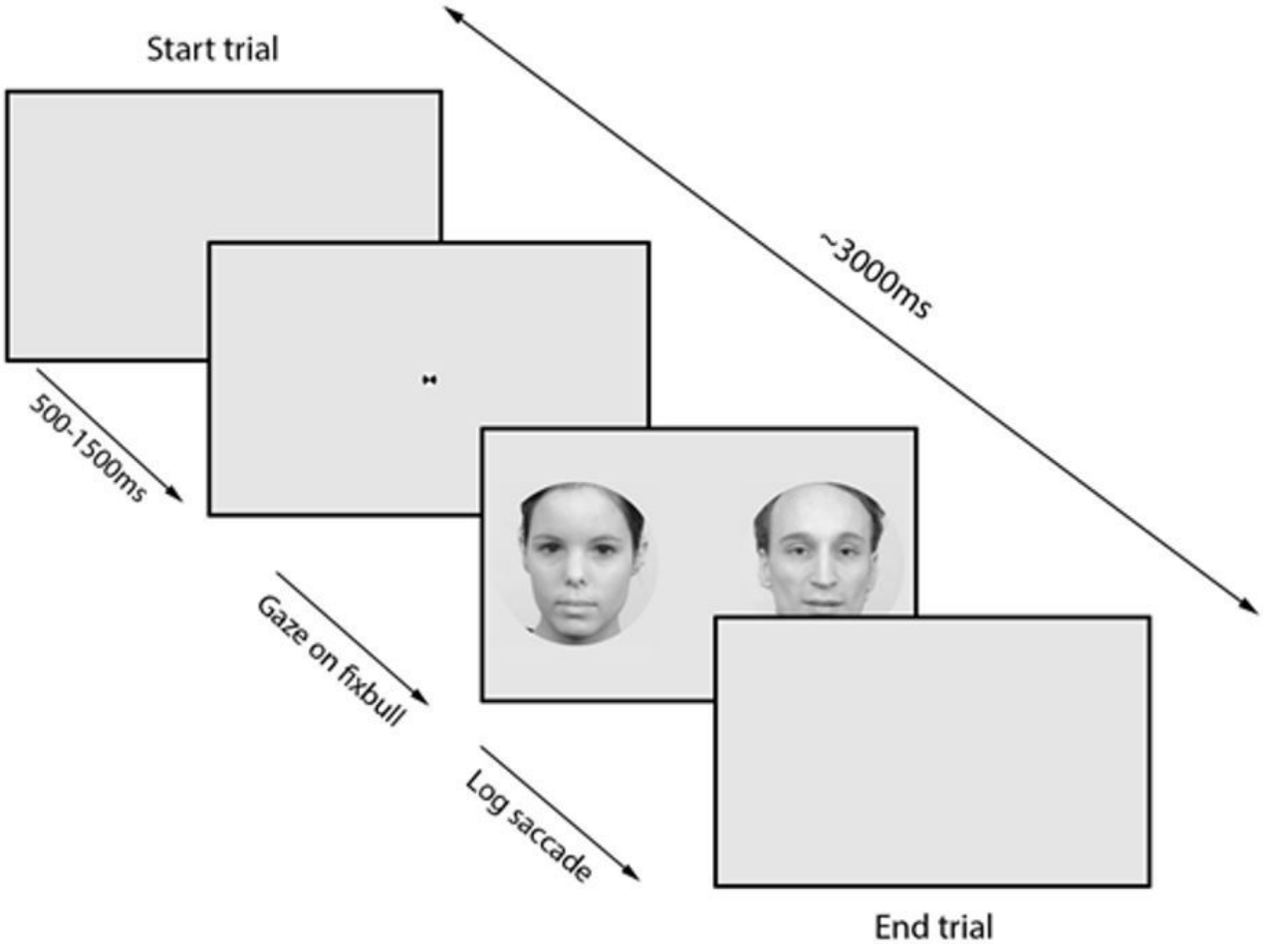

Figure 1

A schematic representation of one trial. The upper right arrow represents the maximum duration of the trial. The bottom left arrows represent the trial events. Every trial started with a grey background. After small delay, the fixation point was presented. When the participant's gaze was registered on this fixation point, two faces were presented. The faces always had different identities and could have different or the same expression. Participants were instructed to make an eye movement to the first face they perceived after which the trial ended. Note that only 20% of the trials contained an actual temporal offset between the presentation of the two faces.

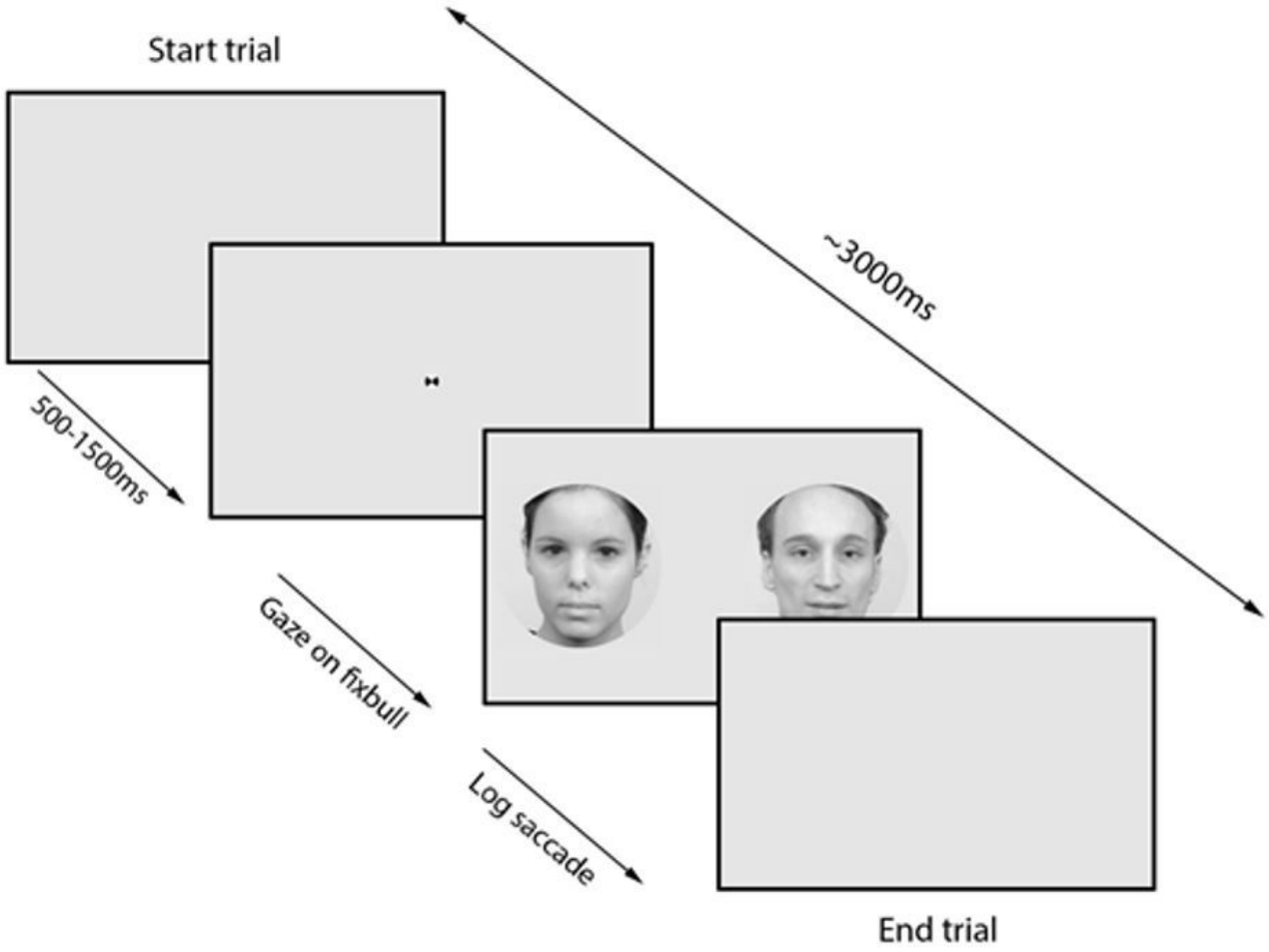

Figure 1

A schematic representation of one trial. The upper right arrow represents the maximum duration of the trial. The bottom left arrows represent the trial events. Every trial started with a grey background. After small delay, the fixation point was presented. When the participant's gaze was registered on this fixation point, two faces were presented. The faces always had different identities and could have different or the same expression. Participants were instructed to make an eye movement to the first face they perceived after which the trial ended. Note that only 20% of the trials contained an actual temporal offset between the presentation of the two faces.

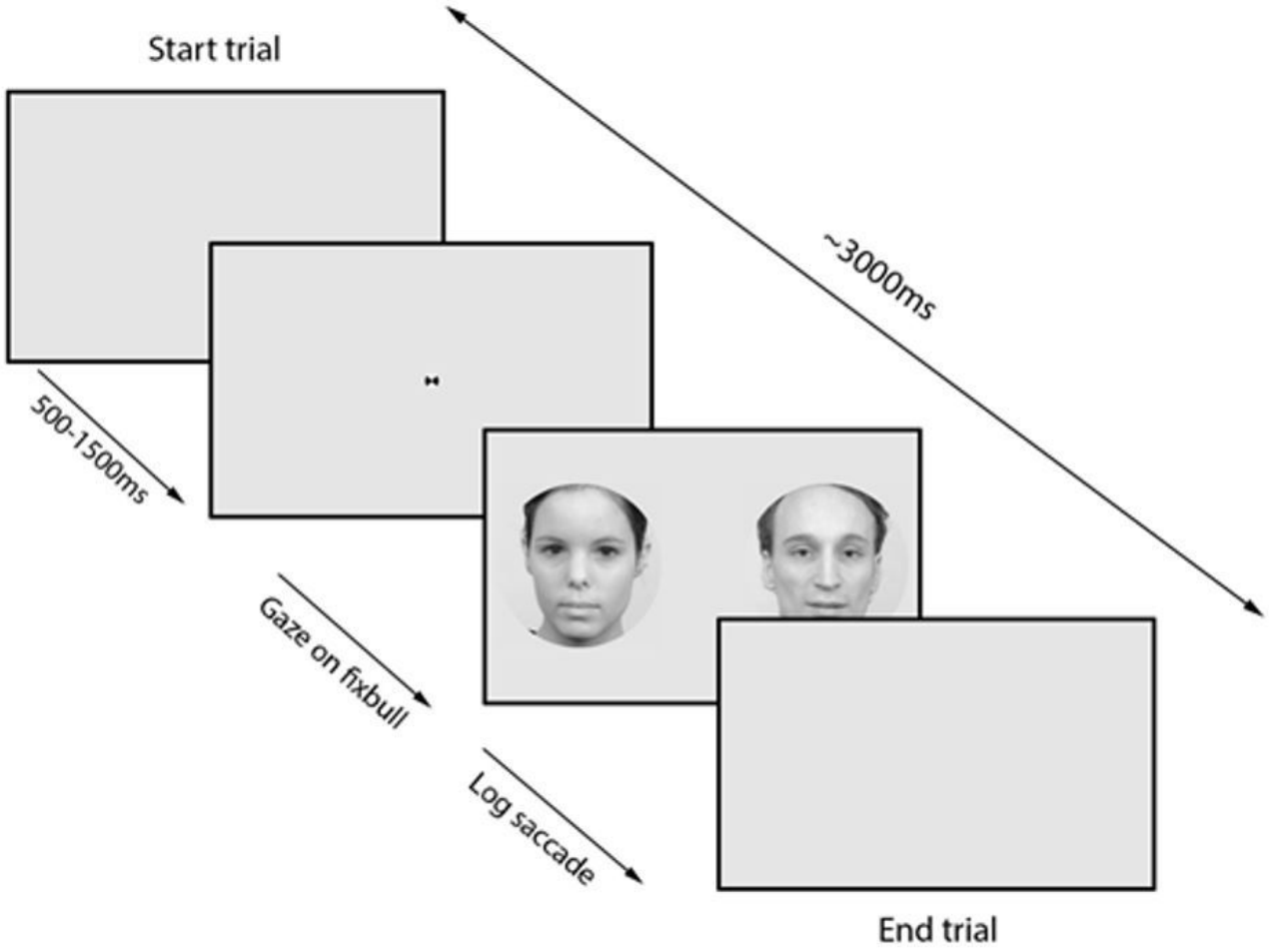

Figure 1

A schematic representation of one trial. The upper right arrow represents the maximum duration of the trial. The bottom left arrows represent the trial events. Every trial started with a grey background. After small delay, the fixation point was presented. When the participant's gaze was registered on this fixation point, two faces were presented. The faces always had different identities and could have different or the same expression. Participants were instructed to make an eye movement to the first face they perceived after which the trial ended. Note that only 20% of the trials contained an actual temporal offset between the presentation of the two faces.

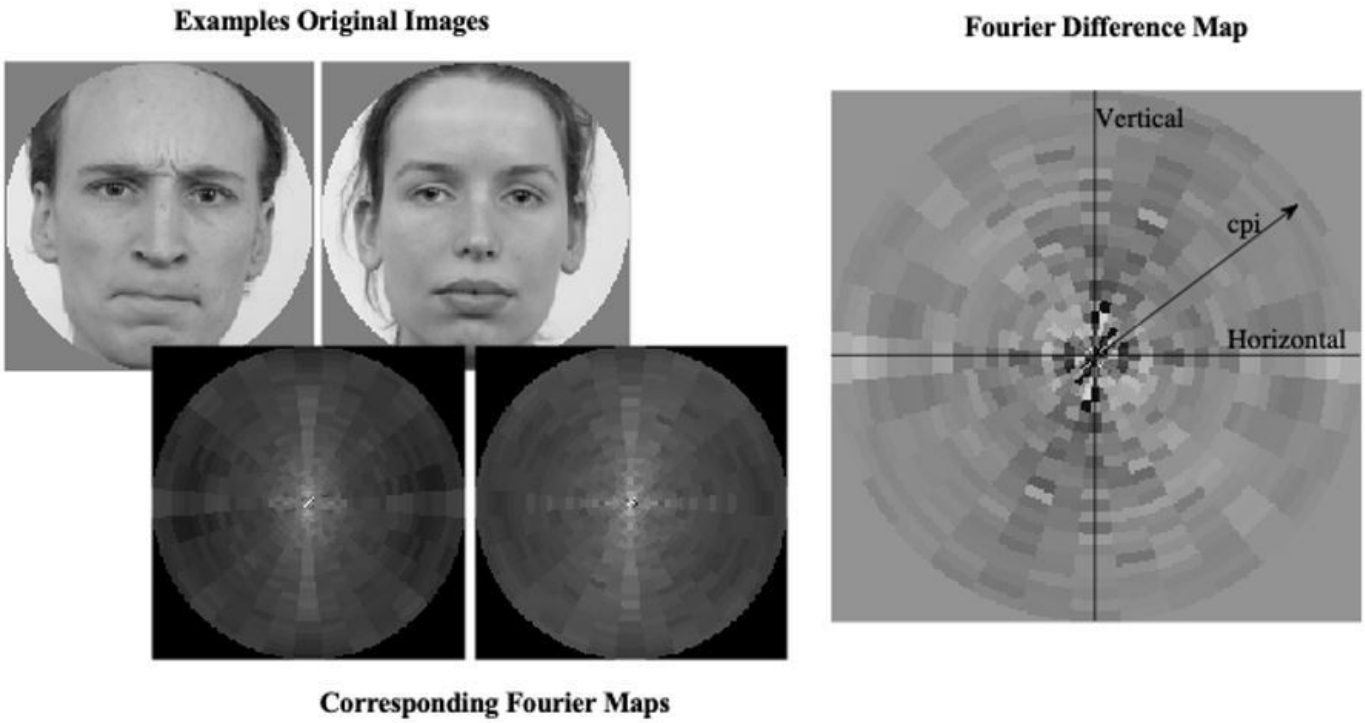

Figure 2

Visualisation of the Fourier feature extraction showing an example of two images used in a trial, their respective, down sampled Fourier features and the Fourier feature differences map. Note that in the Fourier Maps each location corresponds to a particular combination of a spatial frequency and an orientation. The Fourier maps all have horizontal contrasts along the horizontal axes, and vertical contrast along the vertical axes. The radial axes are for cycles per image (abbreviated to cpi in the figure; ranging from low in the centre to high near the edges). Luminance intensity, from black to white, indicates the relative strength of the contrast for the corresponding section of the map. The Fourier feature differences map was calculated by subtracting the down sampled Fourier features from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

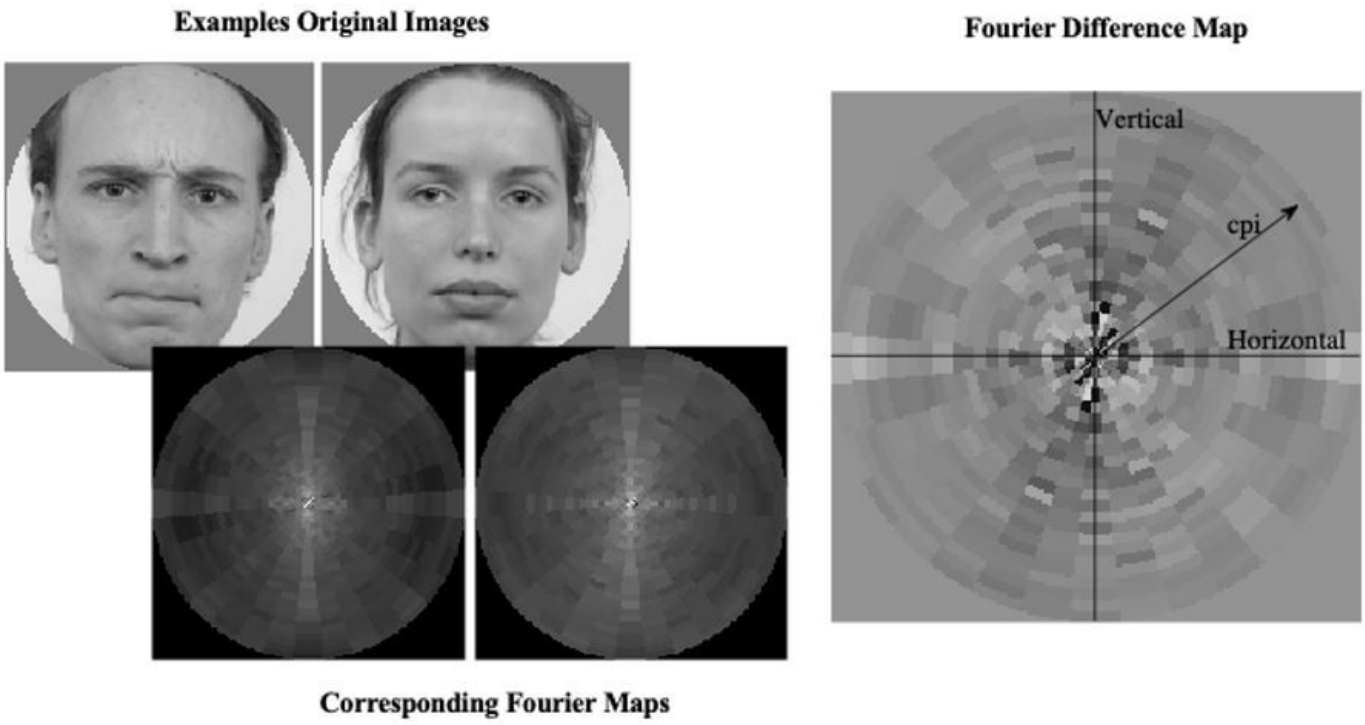

Figure 2

Visualisation of the Fourier feature extraction showing an example of two images used in a trial, their respective, down sampled Fourier features and the Fourier feature differences map. Note that in the Fourier Maps each location corresponds to a particular combination of a spatial frequency and an orientation. The Fourier maps all have horizontal contrasts along the horizontal axes, and vertical contrast along the vertical axes. The radial axes are for cycles per image (abbreviated to cpi in the figure; ranging from low in the centre to high near the edges). Luminance intensity, from black to white, indicates the relative strength of the contrast for the corresponding section of the map. The Fourier feature differences map was calculated by subtracting the down sampled Fourier features from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

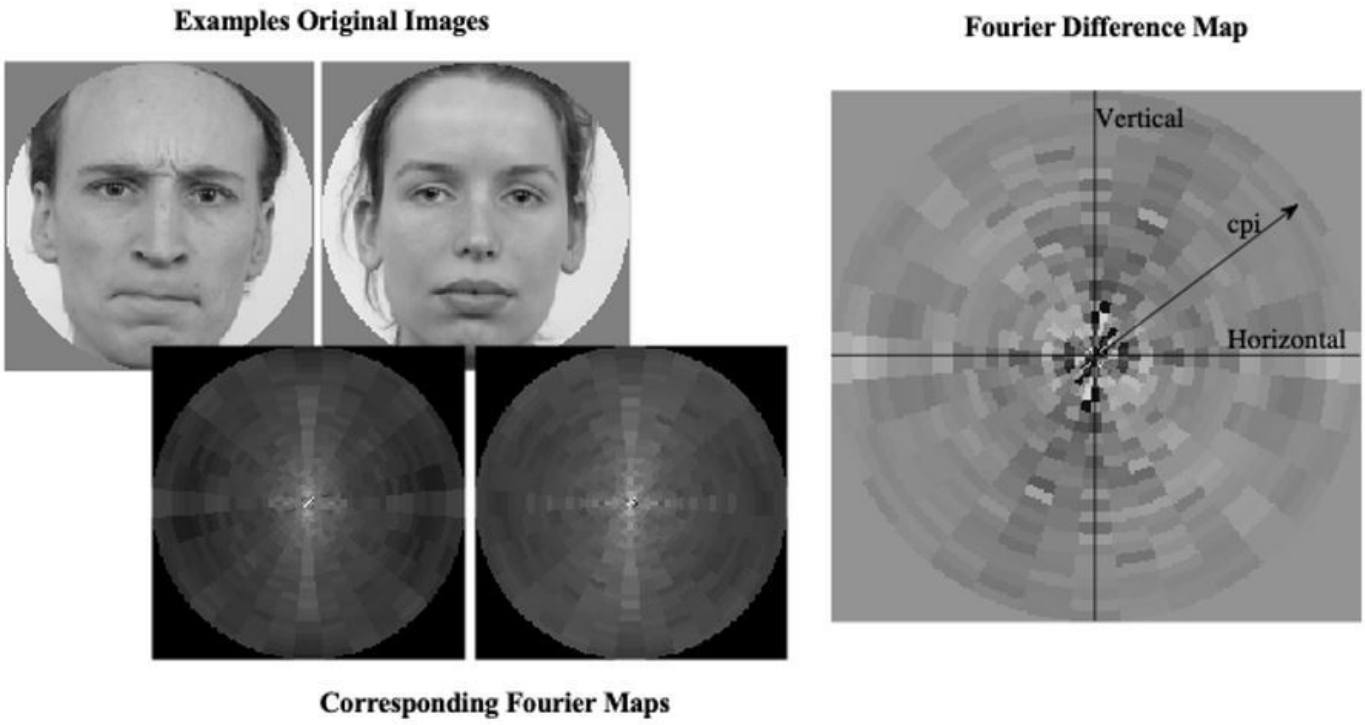

Figure 2

Visualisation of the Fourier feature extraction showing an example of two images used in a trial, their respective, down sampled Fourier features and the Fourier feature differences map. Note that in the Fourier Maps each location corresponds to a particular combination of a spatial frequency and an orientation. The Fourier maps all have horizontal contrasts along the horizontal axes, and vertical contrast along the vertical axes. The radial axes are for cycles per image (abbreviated to cpi in the figure; ranging from low in the centre to high near the edges). Luminance intensity, from black to white, indicates the relative strength of the contrast for the corresponding section of the map. The Fourier feature differences map was calculated by subtracting the down sampled Fourier features from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

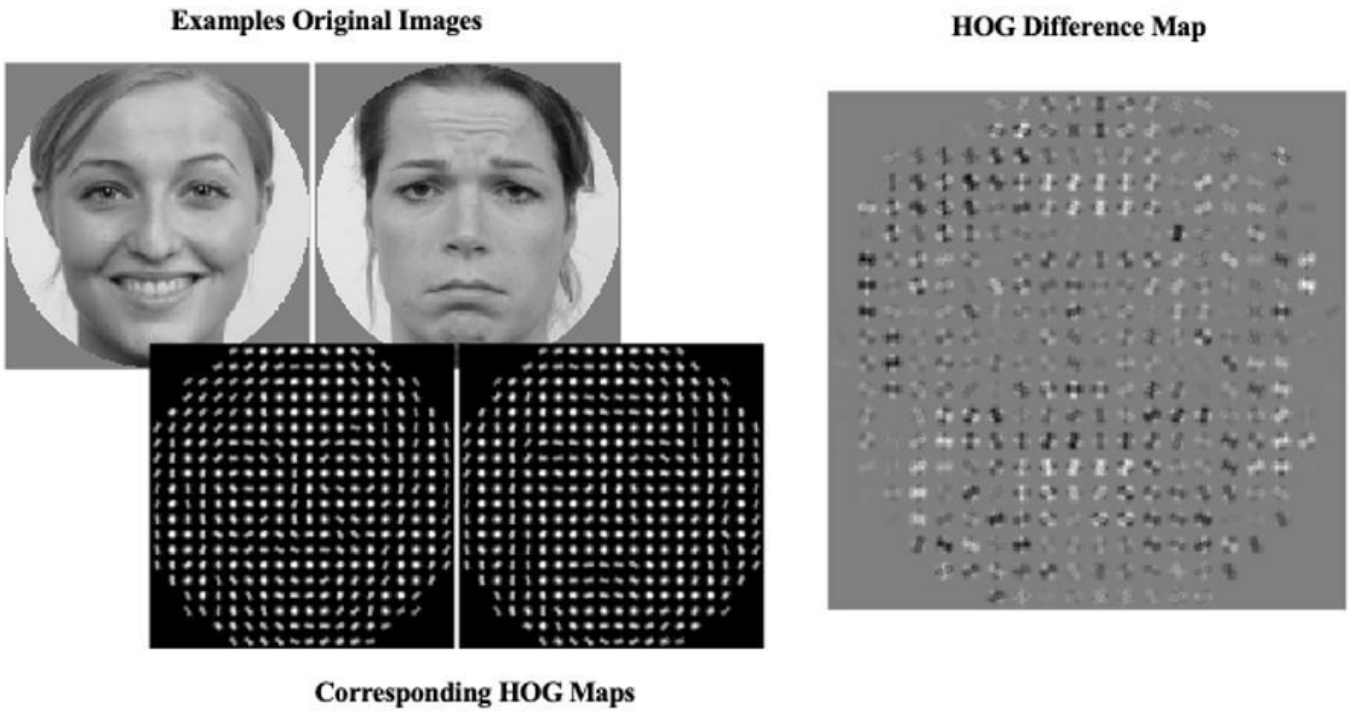

Figure 3

Visualisation of the HOG feature extraction showing an example of two images used in a trial, their respective HOG features and the HOG feature differences map using the highest resolution (10x10 cell size). All HOG maps use the same x and y axes as the original images, meaning position in the HOG map is directly coupled with position in an image. The HOG maps show 20x20 grids where each position in the grids represents an area of 10x10 pixels. For each 10x10 area of pixel in an image, the weights for 9 differently oriented gradients is calculated. The 9 weights are visualized by white bars where the length reflects the weights. The 9 bars are then superimposed on the 10x10 pixel area there are based on. The HOG feature differences map was calculated by subtracting these HOG features weights from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

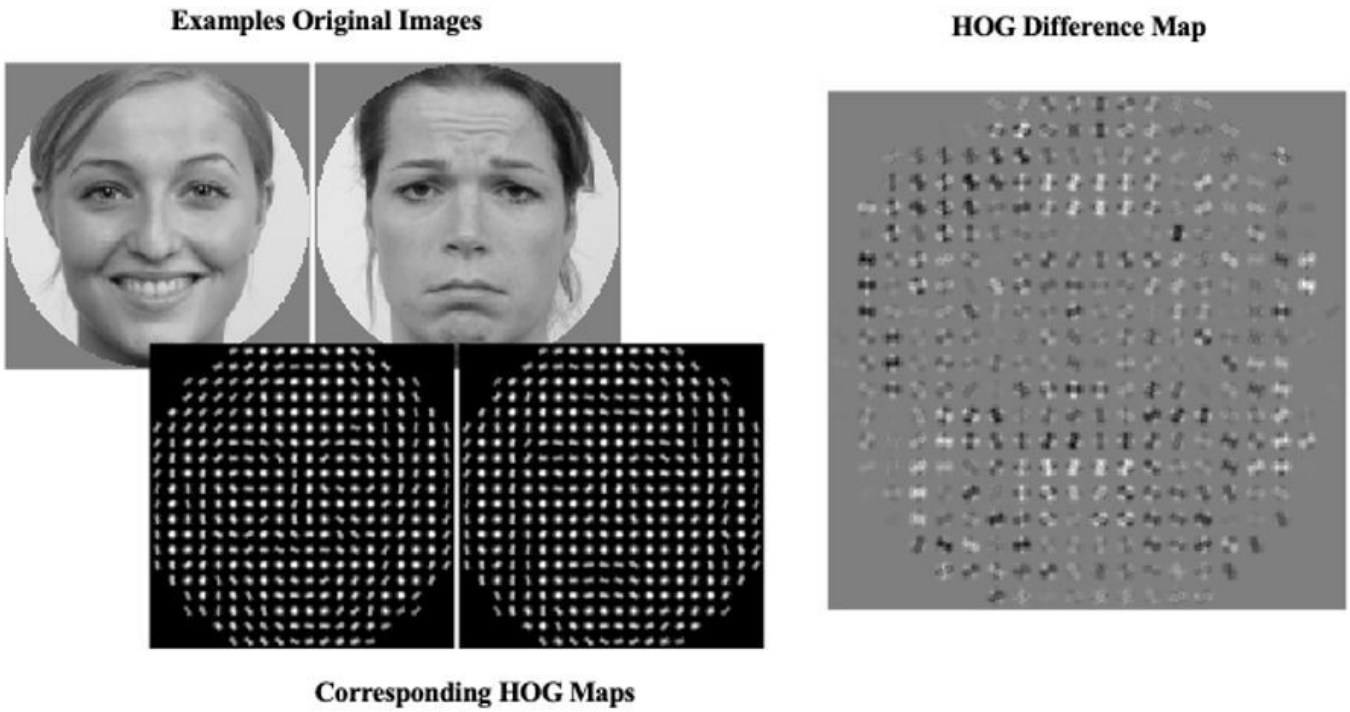

Figure 3

Visualisation of the HOG feature extraction showing an example of two images used in a trial, their respective HOG features and the HOG feature differences map using the highest resolution (10x10 cell size). All HOG maps use the same x and y axes as the original images, meaning position in the HOG map is directly coupled with position in an image. The HOG maps show 20x20 grids where each position in the grids represents an area of 10x10 pixels. For each 10x10 area of pixel in an image, the weights for 9 differently oriented gradients is calculated. The 9 weights are visualized by white bars where the length reflects the weights. The 9 bars are then superimposed on the 10x10 pixel area there are based on. The HOG feature differences map was calculated by subtracting these HOG features weights from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

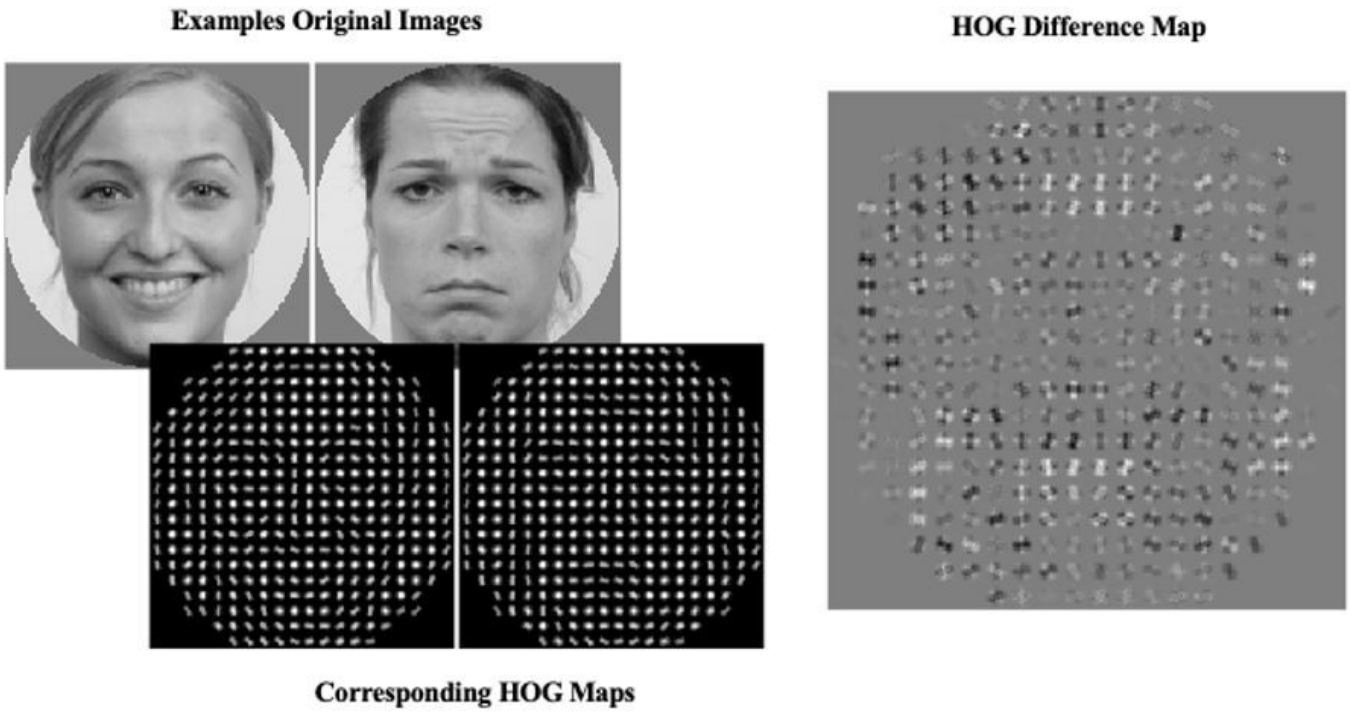

Figure 3

Visualisation of the HOG feature extraction showing an example of two images used in a trial, their respective HOG features and the HOG feature differences map using the highest resolution (10x10 cell size). All HOG maps use the same x and y axes as the original images, meaning position in the HOG map is directly coupled with position in an image. The HOG maps show 20x20 grids where each position in the grids represents an area of 10x10 pixels. For each 10x10 area of pixel in an image, the weights for 9 differently oriented gradients is calculated. The 9 weights are visualized by white bars where the length reflects the weights. The 9 bars are then superimposed on the 10x10 pixel area there are based on. The HOG feature differences map was calculated by subtracting these HOG features weights from the image presented on the right, from those of the image of the left. Note that the feature difference map is scaled such that dark regions indicate negative values and light regions indicate positive values.

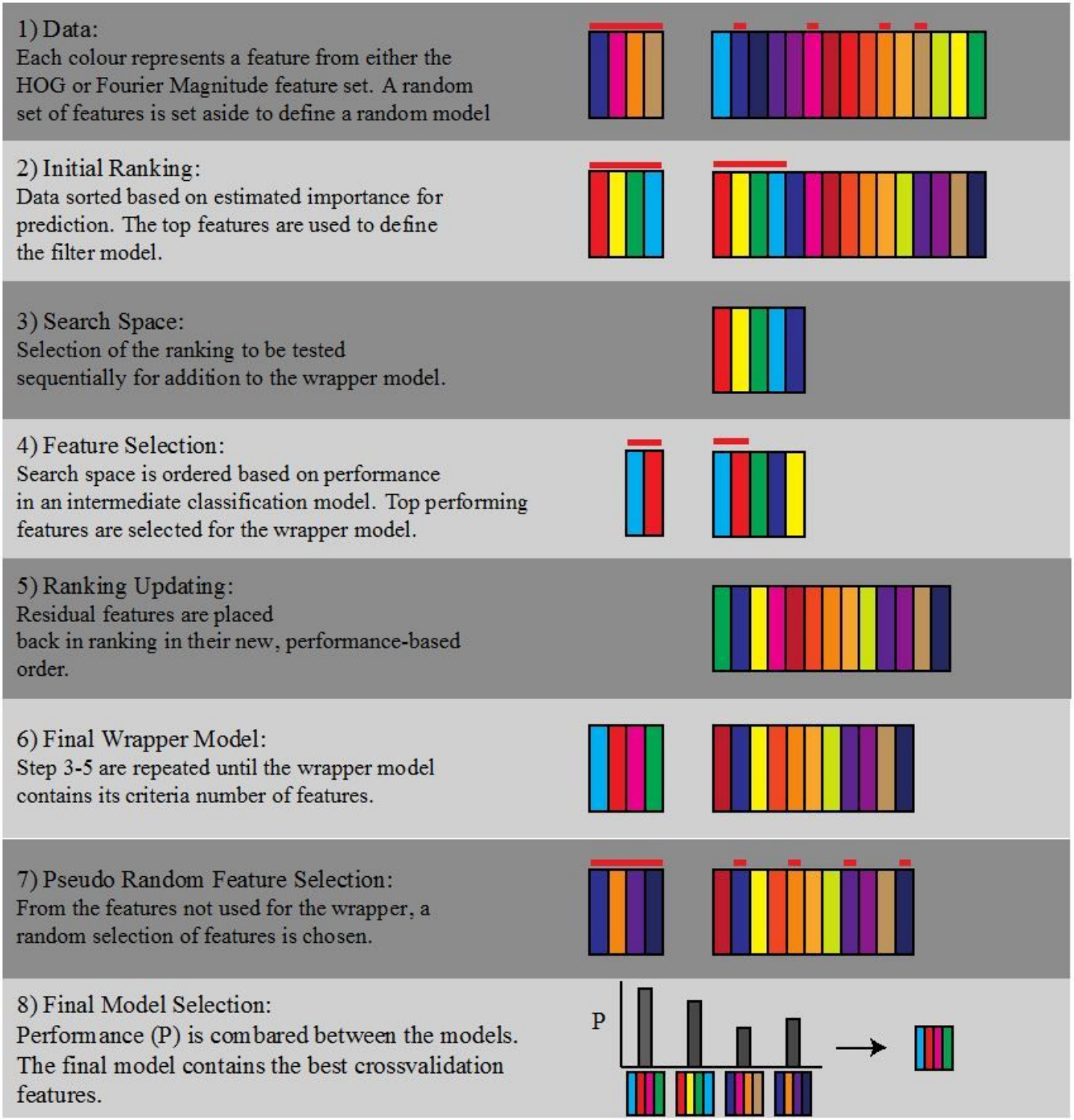

Figure 4

Schematic representation of the feature selection algorithm used in the current project. Step 1) a random collection of features (indicated by the red bars) is selected for a random model. Step 2) The features are ranked based on Chi-Square scores. The top of the ranking is used to determine the filter model. Step 3) A search space is defined from the top-ranking features and the features in this search space are tested for inclusion into the wrapper model (Step 4). Step 5) The Chi-Square based ranking of search space

features is updated based on their performance. Step 6) Steps 3-5 are repeated until enough features have been selected for the wrapper model. Step 7) From the residual features, unused by the wrapper model, a random selection is made for a pseudo random model. Step 8) Each of the four combinations of features are used to train classification models and cross-validation performance is subsequently estimated using the hold-out data. The final model is selected based on highest cross-validation performance (P). See the above section Feature Selection, and the below section Final Model Selection for additional details.

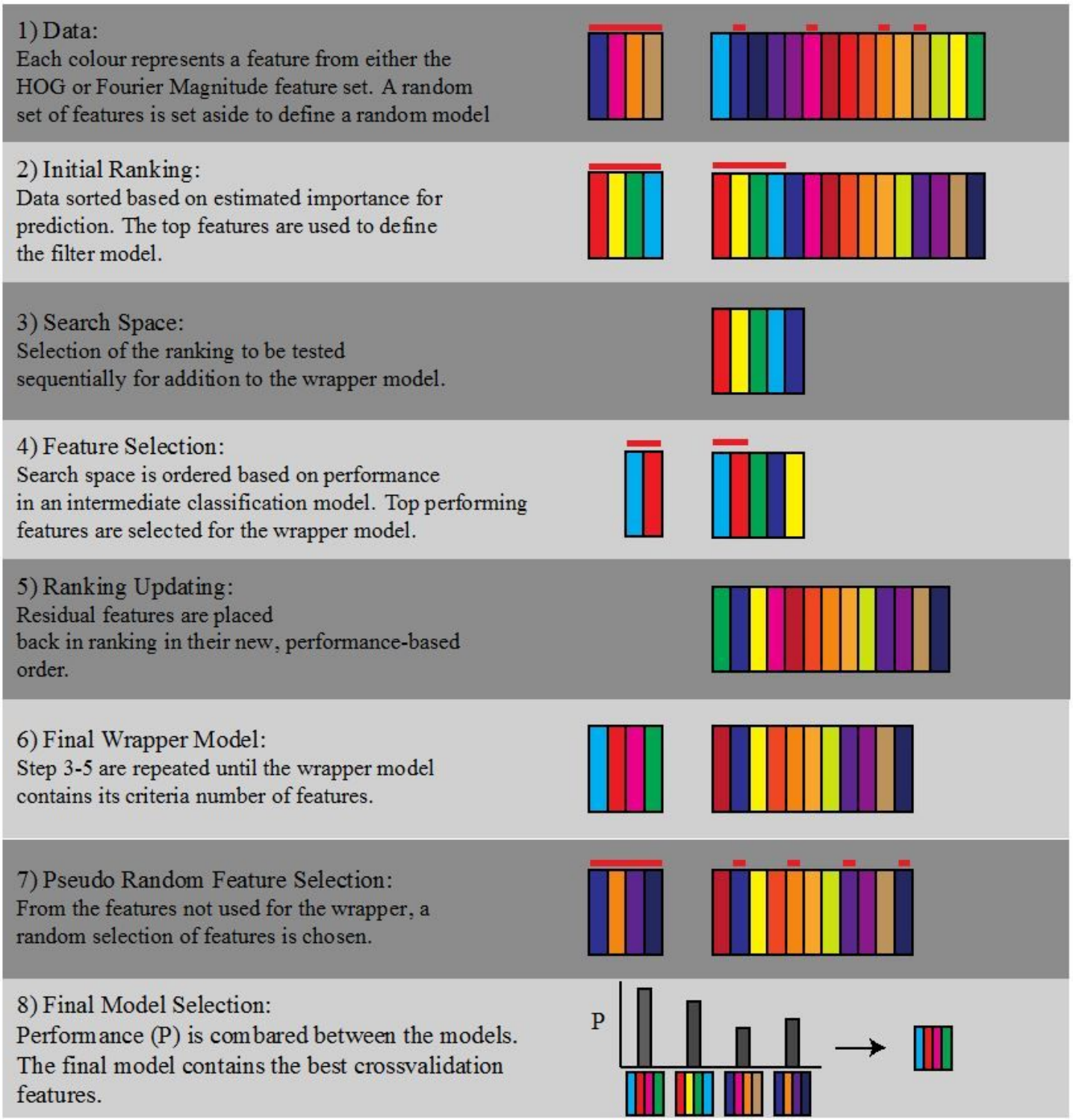

Figure 4

Schematic representation of the feature selection algorithm used in the current project. Step 1) a random collection of features (indicated by the red bars) is selected for a random model. Step 2) The features are ranked based on Chi-Square scores. The top of the ranking is used to determine the filter model. Step 3) A search space is defined from the top-ranking features and the features in this search space are tested for inclusion into the wrapper model (Step 4). Step 5) The Chi-Square based ranking of search space features is updated based on their performance. Step 6) Steps 3-5 are repeated until enough features have been selected for the wrapper model. Step 7) From the residual features, unused by the wrapper model, a random selection is made for a pseudo random model. Step 8) Each of the four combinations of features are used to train classification models and cross-validation performance is subsequently estimated using the hold-out data. The final model is selected based on highest cross-validation performance (P). See the above section Feature Selection, and the below section Final Model Selection for additional details.

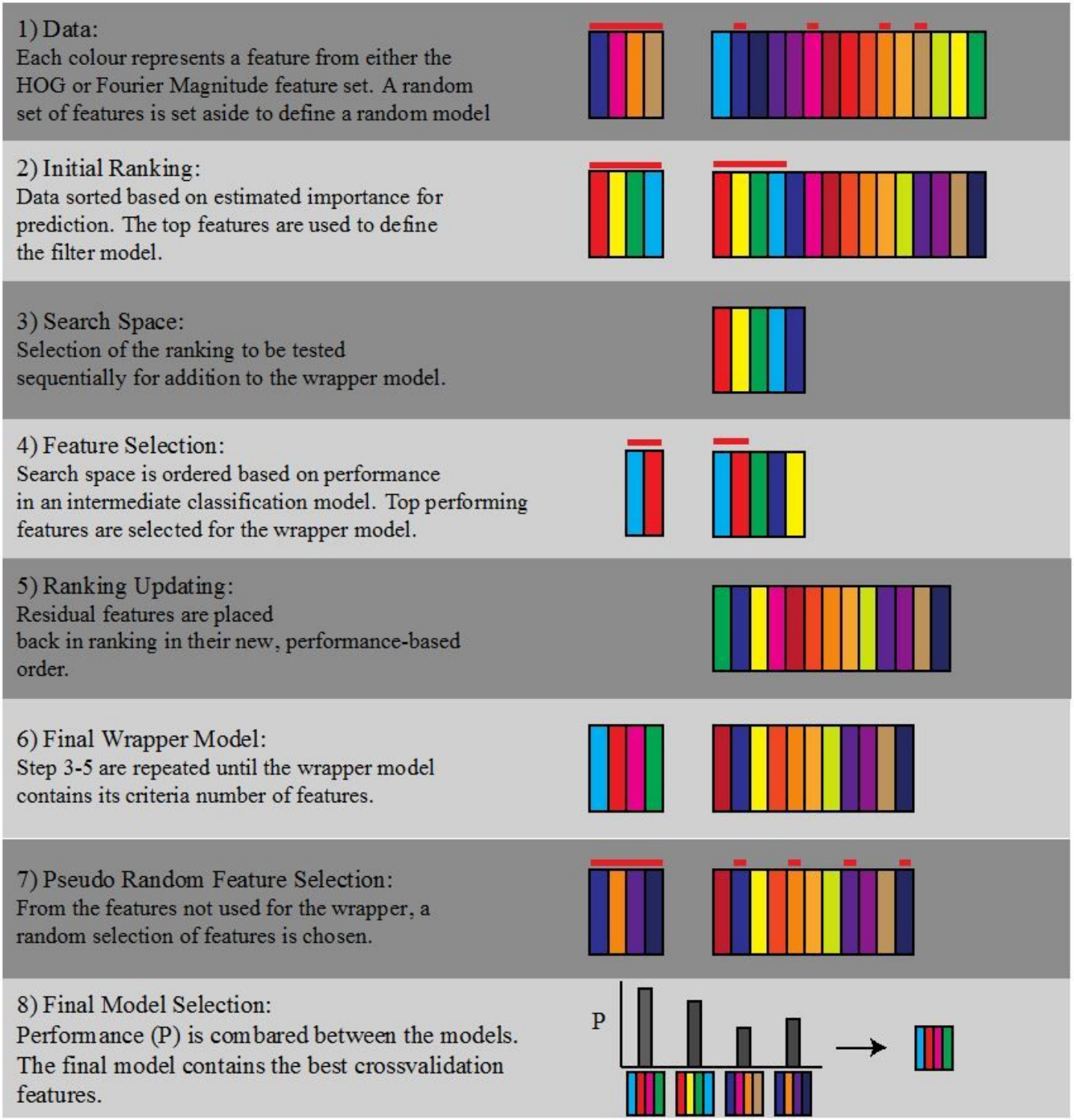

Figure 4

Schematic representation of the feature selection algorithm used in the current project. Step 1) a random collection of features (indicated by the red bars) is selected for a random model. Step 2) The features are ranked based on Chi-Square scores. The top of the ranking is used to determine the filter model. Step 3) A search space is defined from the top-ranking features and the features in this search space are tested for inclusion into the wrapper model (Step 4). Step 5) The Chi-Square based ranking of search space

features is updated based on their performance. Step 6) Steps 3-5 are repeated until enough features have been selected for the wrapper model. Step 7) From the residual features, unused by the wrapper model, a random selection is made for a pseudo random model. Step 8) Each of the four combinations of features are used to train classification models and cross-validation performance is subsequently estimated using the hold-out data. The final model is selected based on highest cross-validation performance (P). See the above section Feature Selection, and the below section Final Model Selection for additional details.

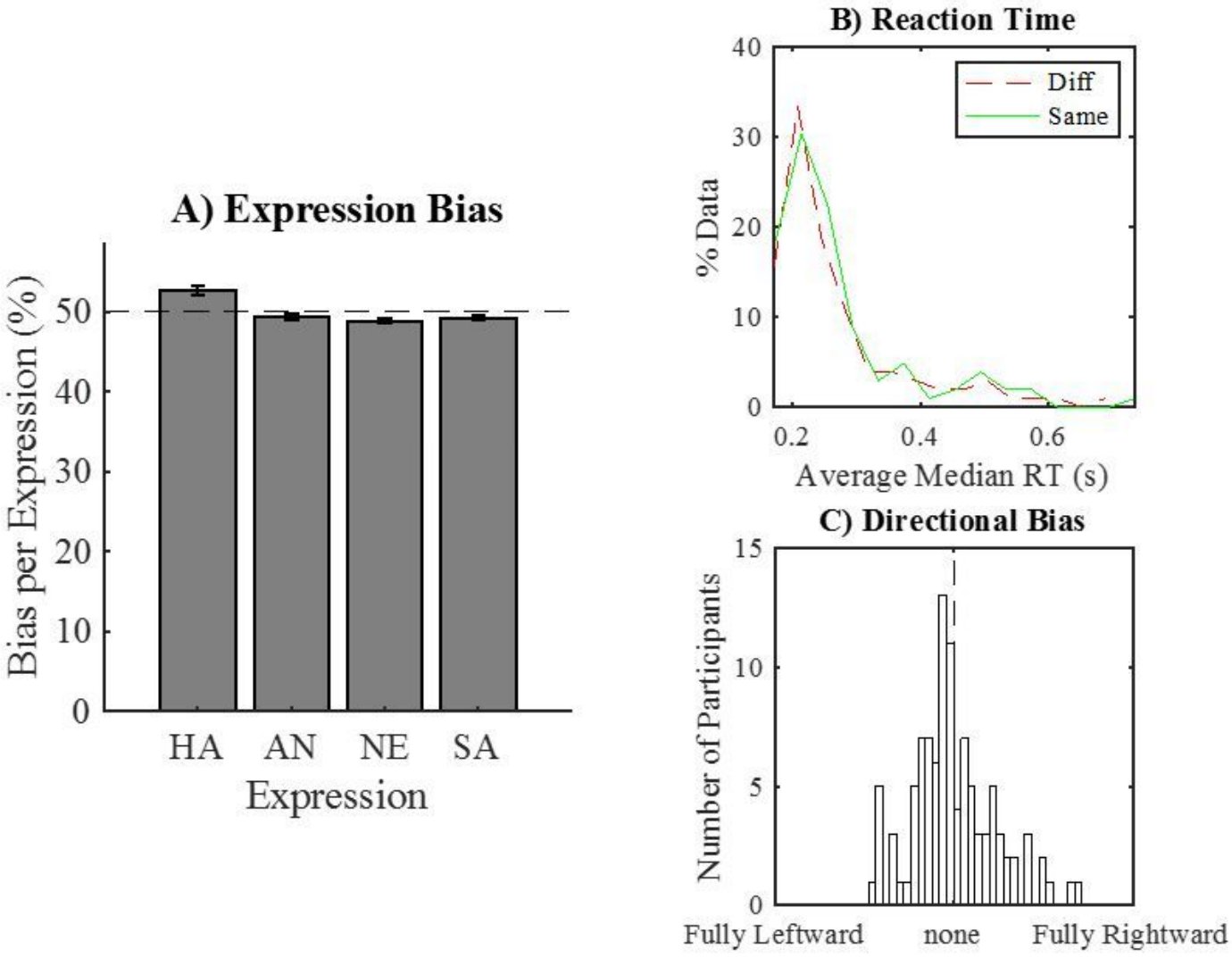

Figure 5

In figure 5A we show the average fraction of selecting, across participants, (y-axis) a particular expression (x-axis; HA: Happy, AN: Angry, NE: Neutral & SA: Sad) for trials in which two different expressions were displayed. Errorbars represent the Standard Error of the Mean. Results show a significant positive bias for happy facial expressions. Figure 5B shows the distributions of the average median reaction times of all participants for the trials with different, and the trials with the same expressions. Figure 5C shows the distributions of left- and right-wards biases for all participants. Overall, 61% of the participants had a biases towards making leftward eye-movements.

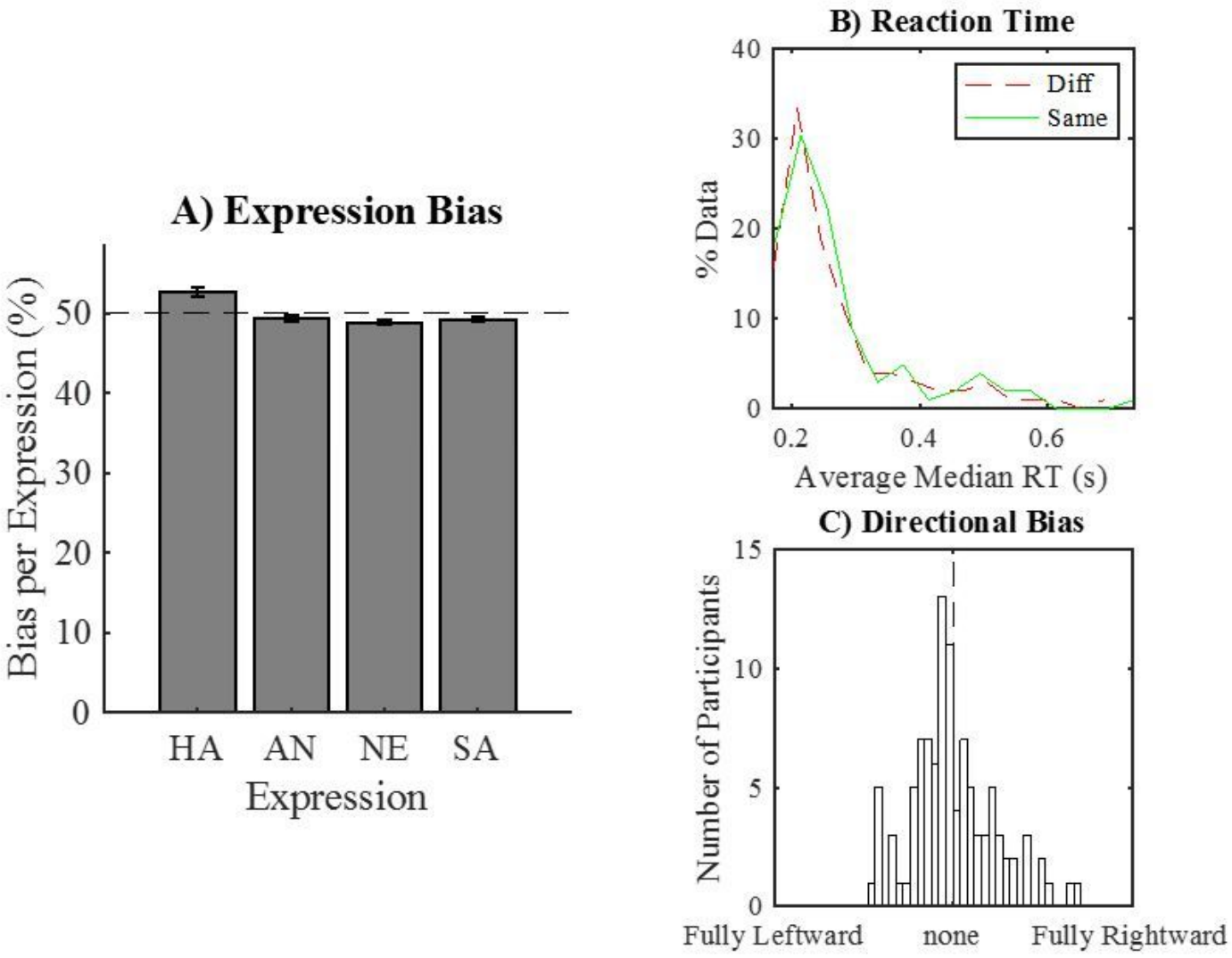

Figure 5

In figure 5A we show the average fraction of selecting, across participants, (y-axis) a particular expression (x-axis; HA: Happy, AN: Angry, NE: Neutral & SA: Sad) for trials in which two different expressions were displayed. Errorbars represent the Standard Error of the Mean. Results show a significant positive bias for happy facial expressions. Figure 5B shows the distributions of the average median reaction times of all participants for the trials with different, and the trials with the same expressions. Figure 5C shows the distributions of left- and right-wards biases for all participants. Overall, 61% of the participants had a biases towards making leftward eye-movements.

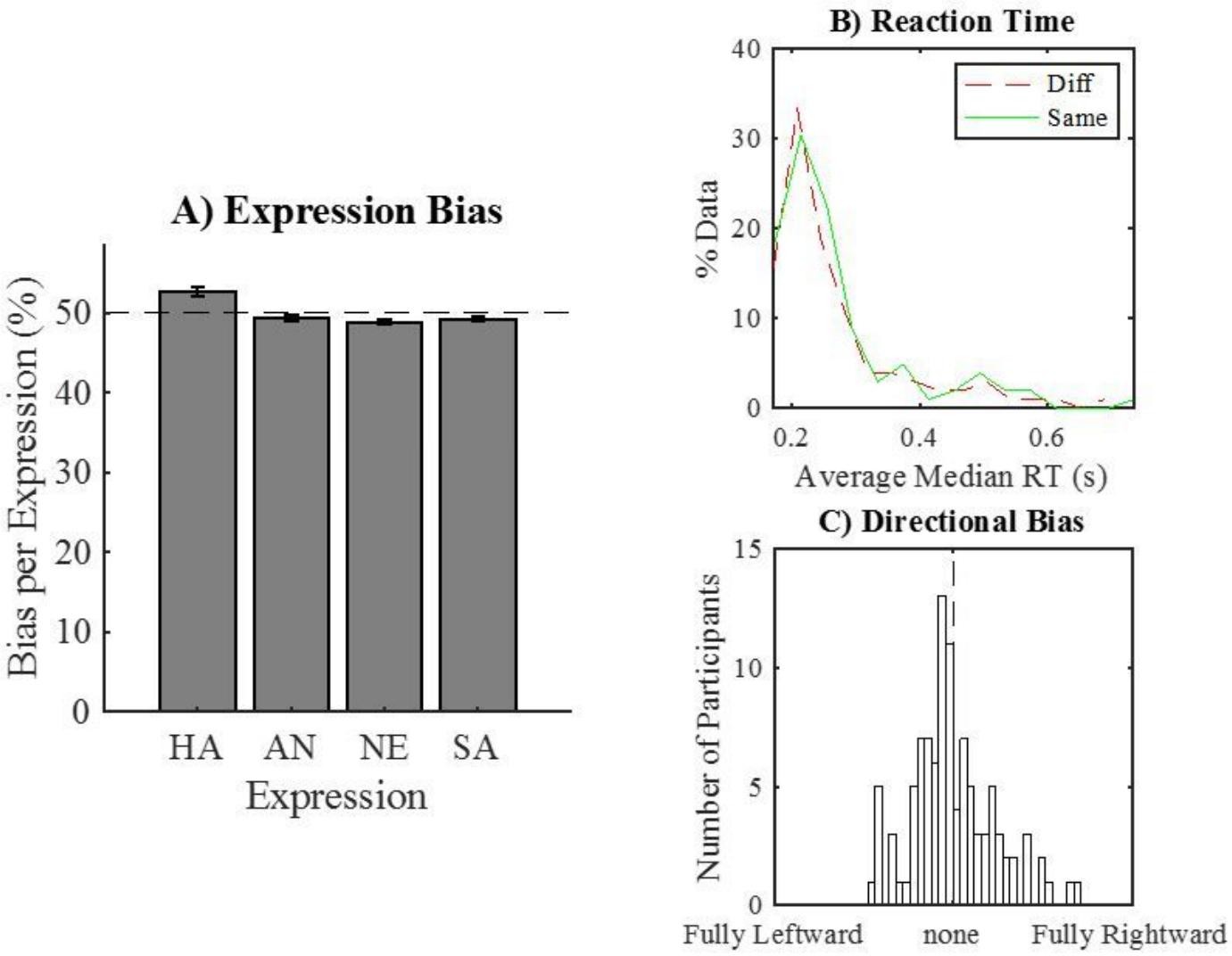

Figure 5

In figure 5A we show the average fraction of selecting, across participants, (y-axis) a particular expression (x-axis; HA: Happy, AN: Angry, NE: Neutral & SA: Sad) for trials in which two different expressions were displayed. Errorbars represent the Standard Error of the Mean. Results show a significant positive bias for happy facial expressions. Figure 5B shows the distributions of the average median reaction times of all participants for the trials with different, and the trials with the same expressions. Figure 5C shows the distributions of left- and right-wards biases for all participants. Overall, 61% of the participants had a biases towards making leftward eye-movements.

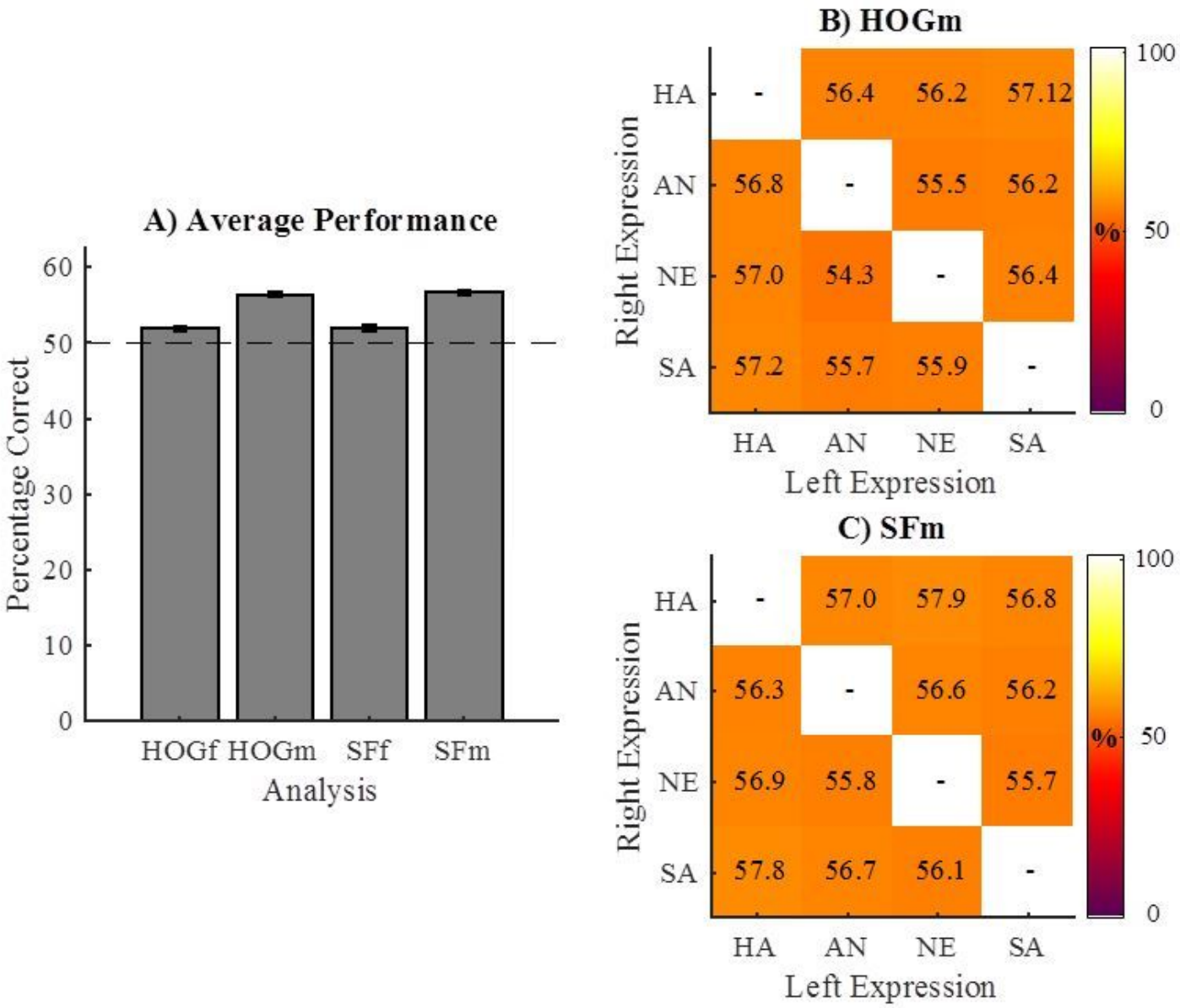

Figure 6

In figure 6A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on the trials where different expressions were presented to the participants. The dotted line represents the overall empirical chance level performance. Errorbars represent the Standard Error of the Mean. 6B-C) Confusion Matrices for the minimal models. For all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that, performance is nearly equal for all combinations of expressions.

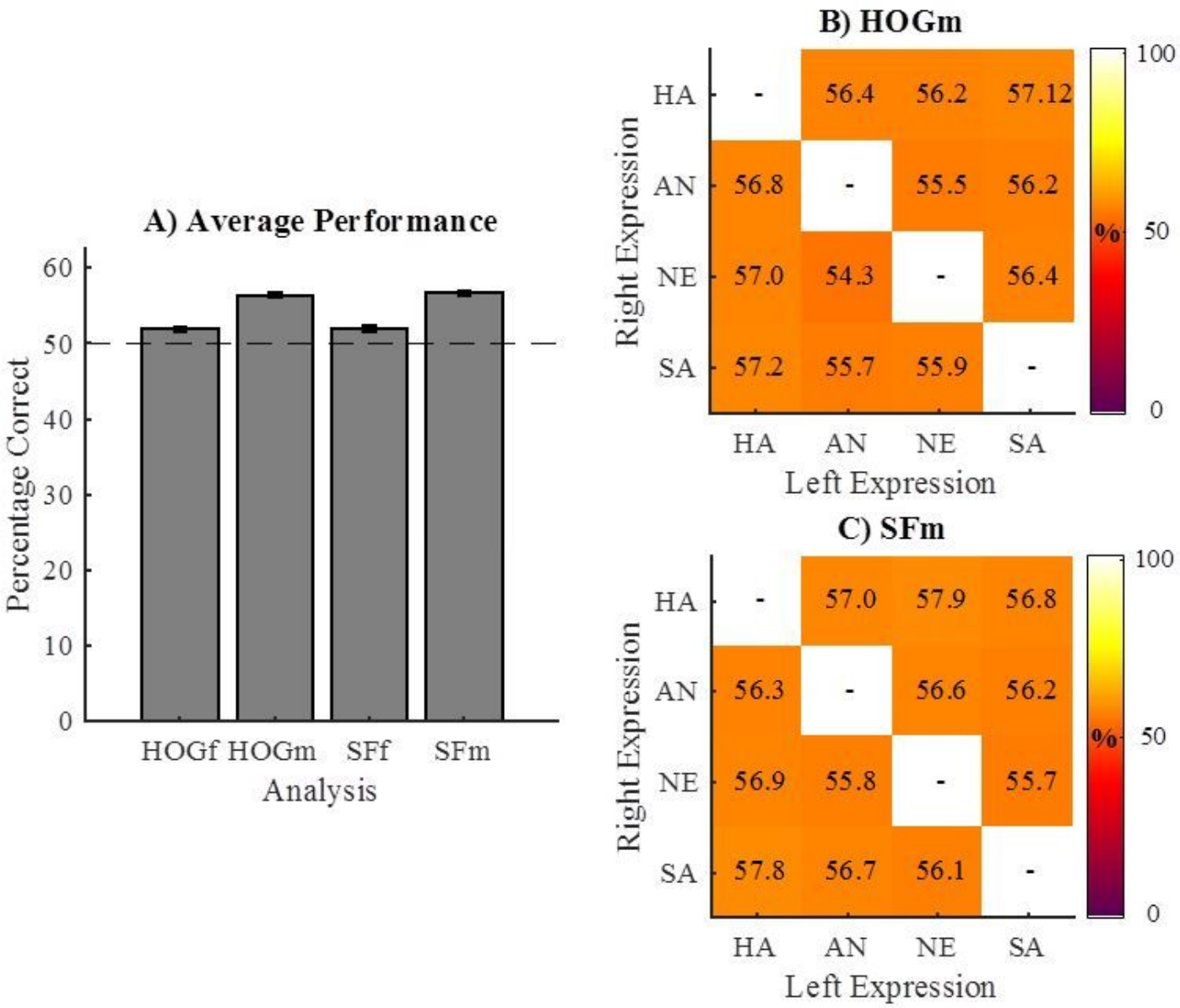

Figure 6

In figure 6A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on the trials where different expressions were presented to the participants. The dotted line represents the overall empirical chance level performance. Errorbars represent the Standard Error of the Mean. 6B-C) Confusion Matrices for the minimal models. For all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that, performance is nearly equal for all combinations of expressions.

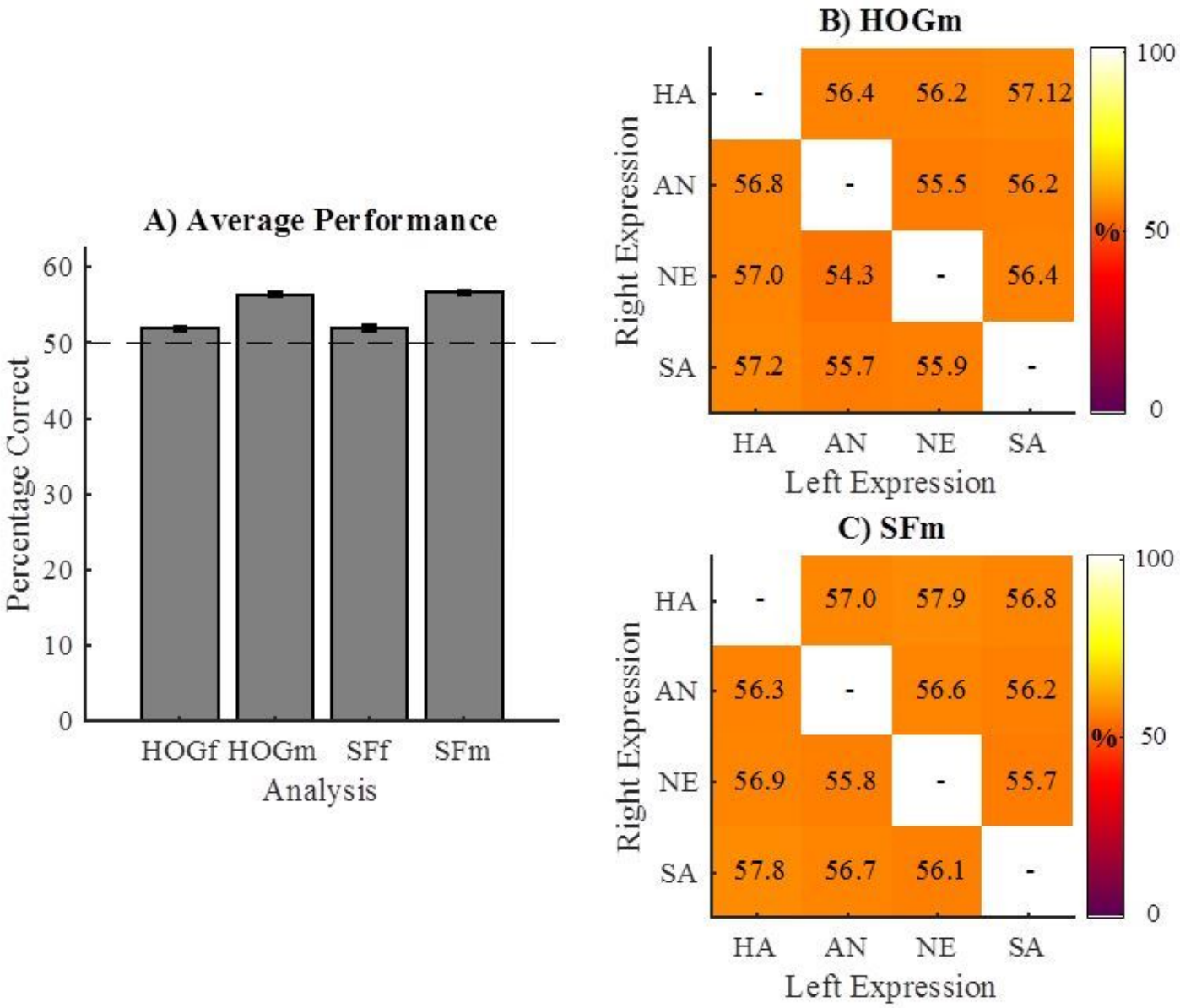

Figure 6

In figure 6A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on the trials where different expressions were presented to the participants. The dotted line represents the overall empirical chance level performance. Errorbars represent the Standard Error of the Mean. 6B-C) Confusion Matrices for the minimal models. For all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that, performance is nearly equal for all combinations of expressions.

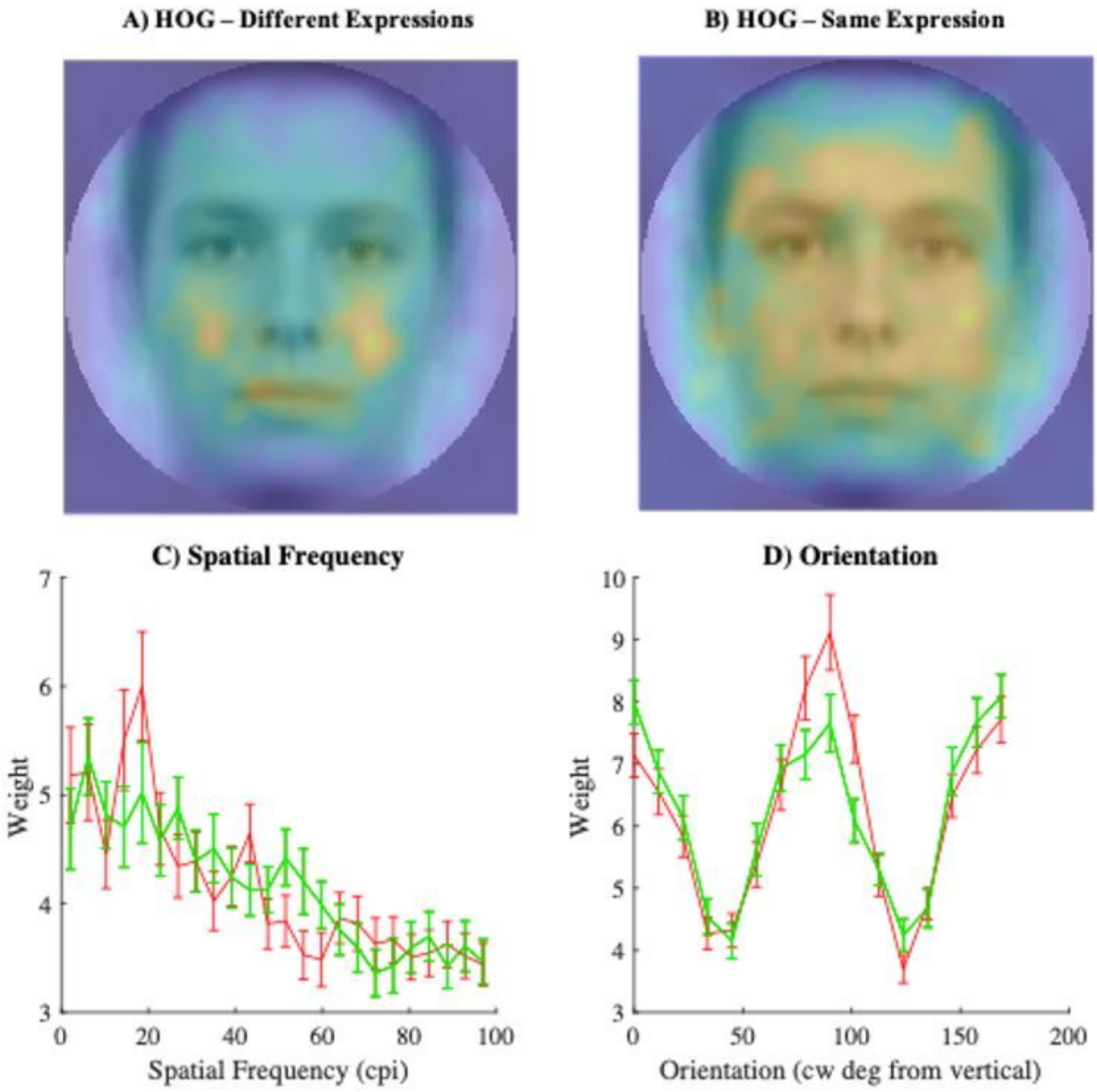

Figure 7

Visual representations of the most relevant features for decoding. 7A-B) Heatmaps reflecting the relevance of spatial locations of the HOG features to decoding either trials with different expressions (A) or the same expression (B) overlaid on the averages of all neutral expressions. As relative importance of a location increases, colour changes from blue through green to yellow. 7C) Here we show the weight, reflecting the percentage of contribution to overall performance, for each band of spatial frequencies used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. 7D) Here we show the weight for each band

of orientations used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. Note that, for both spatial frequency and orientation, the only clear difference is a larger weight for horizontal orientations in trials where the expressions differ.

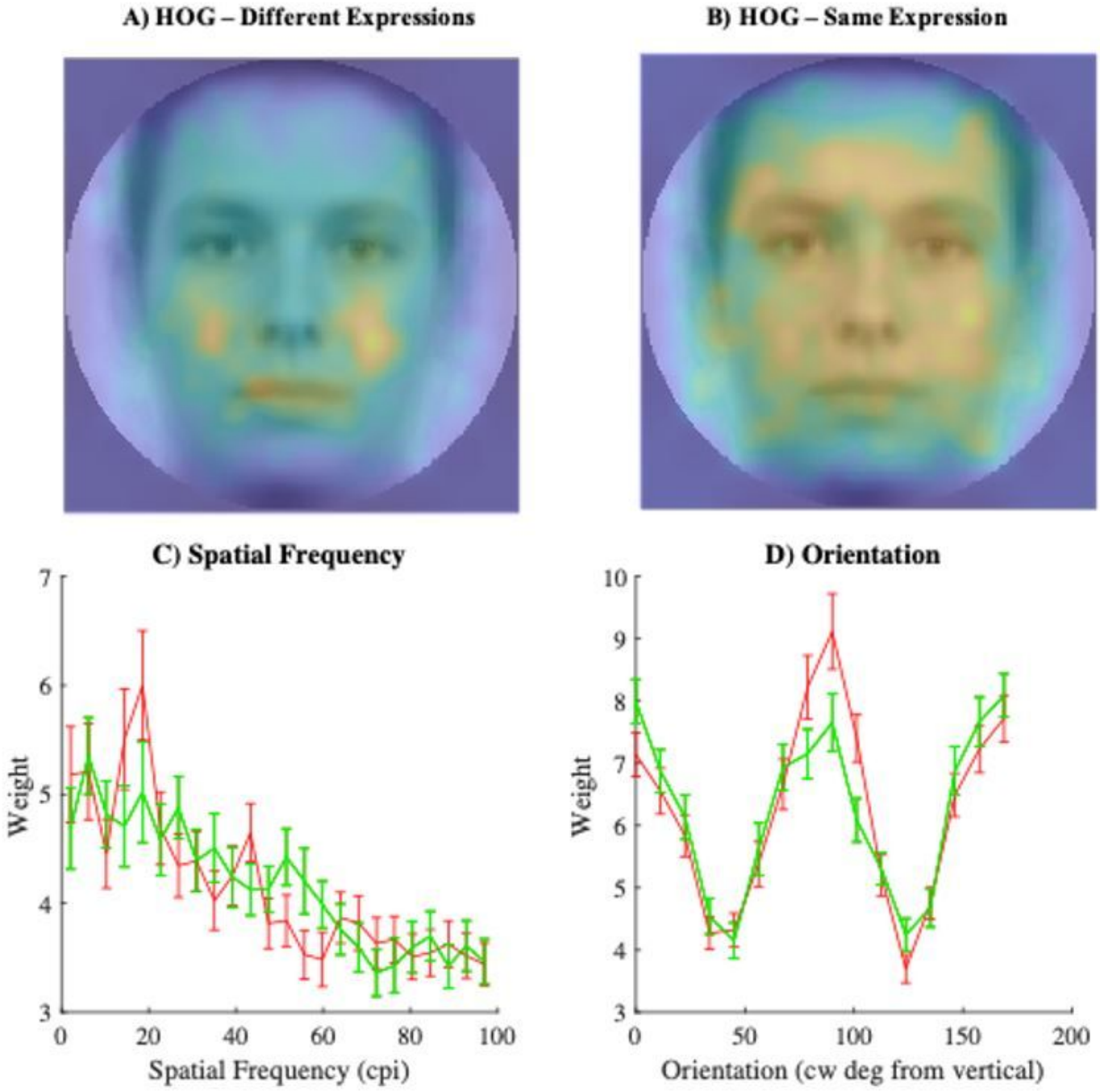

Figure 7

Visual representations of the most relevant features for decoding. 7A-B) Heatmaps reflecting the relevance of spatial locations of the HOG features to decoding either trials with different expressions (A) or the same expression (B) overlaid on the averages of all neutral expressions. As relative importance of a

location increases, colour changes from blue through green to yellow. 7C) Here we show the weight, reflecting the percentage of contribution to overall performance, for each band of spatial frequencies used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. 7D) Here we show the weight for each band of orientations used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. Note that, for both spatial frequency and orientation, the only clear difference is a larger weight for horizontal orientations in trials where the expressions differ.

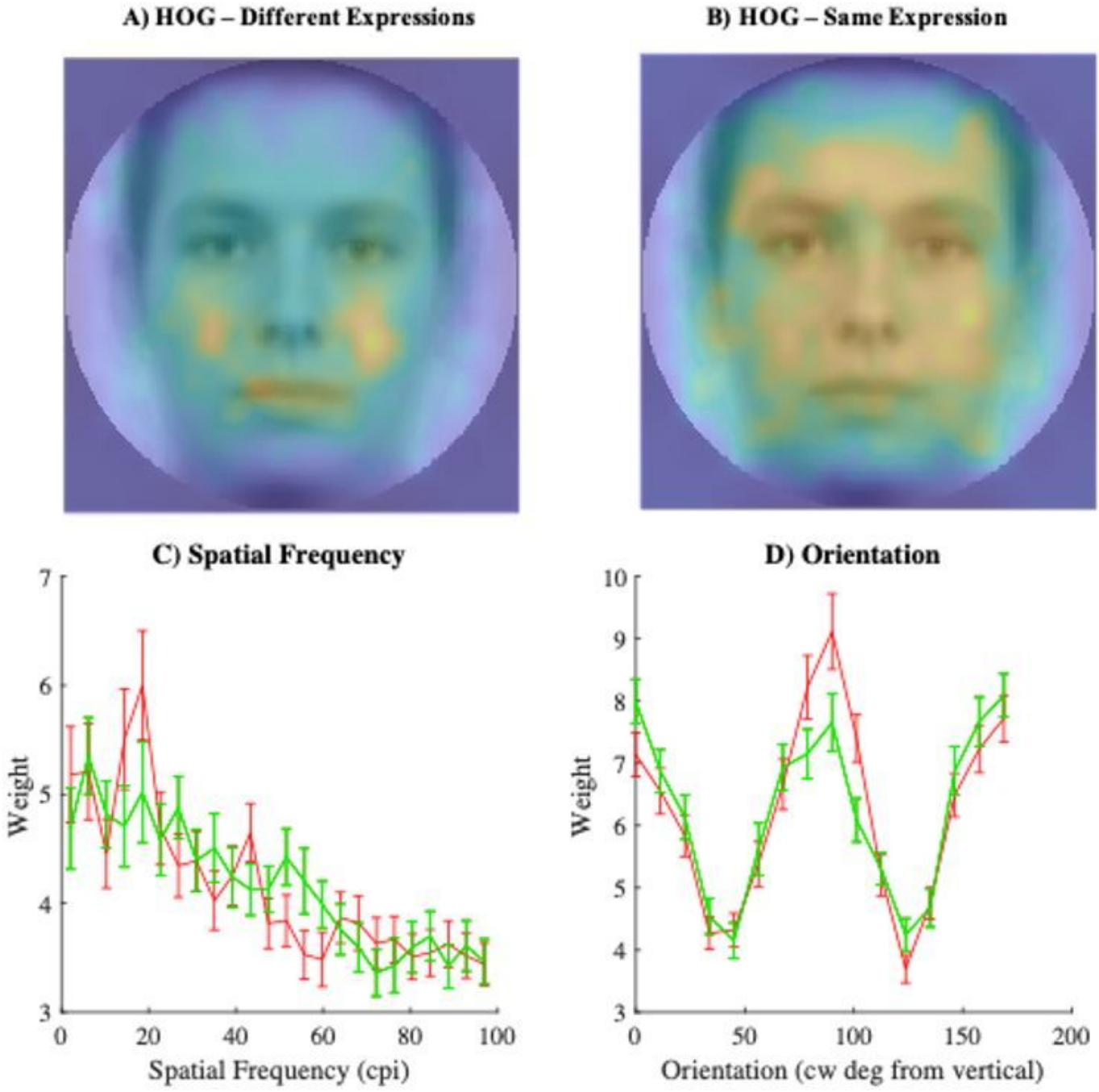

Figure 7

Visual representations of the most relevant features for decoding. 7A-B) Heatmaps reflecting the relevance of spatial locations of the HOG features to decoding either trials with different expressions (A) or the same expression (B) overlaid on the averages of all neutral expressions. As relative importance of a location increases, colour changes from blue through green to yellow. 7C) Here we show the weight, reflecting the percentage of contribution to overall performance, for each band of spatial frequencies used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. 7D) Here we show the weight for each band of orientations used to decode face selection for both trial types (red line, different expressions; green line same expression). Errorbars reflect the standard error of the mean. Note that, for both spatial frequency and orientation, the only clear difference is a larger weight for horizontal orientations in trials where the expressions differ.

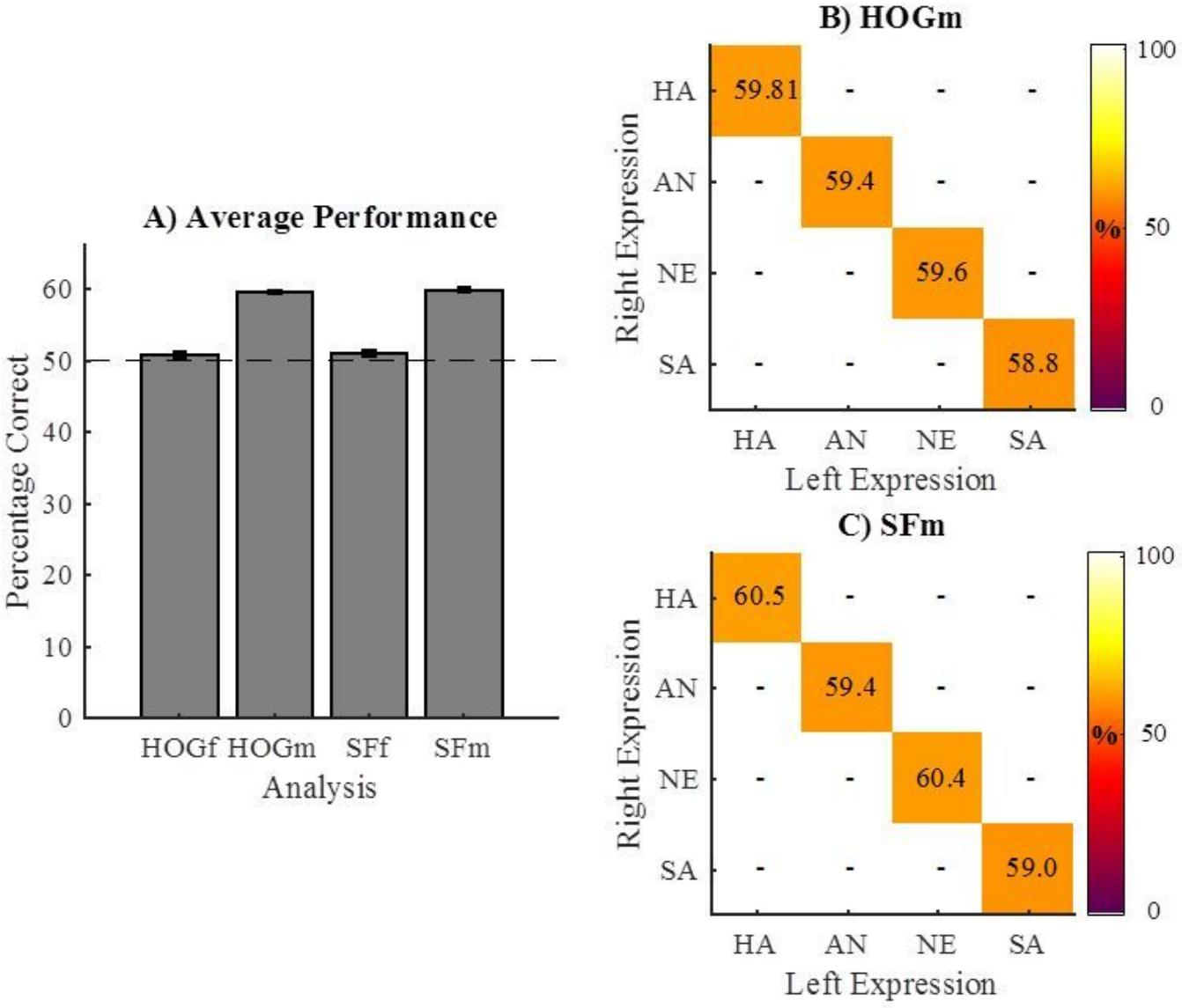

Figure 8

In figure 8A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on the trials where the same expression was present to both the left and the right side of the screen, leaving only a difference in identity. The dotted line represents the overall empirical chance level performance. Errorbars represent the Standard Error of the Mean. 8B-C) Confusion Matrices for the minimal models. For all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that the differences in performance are very small.

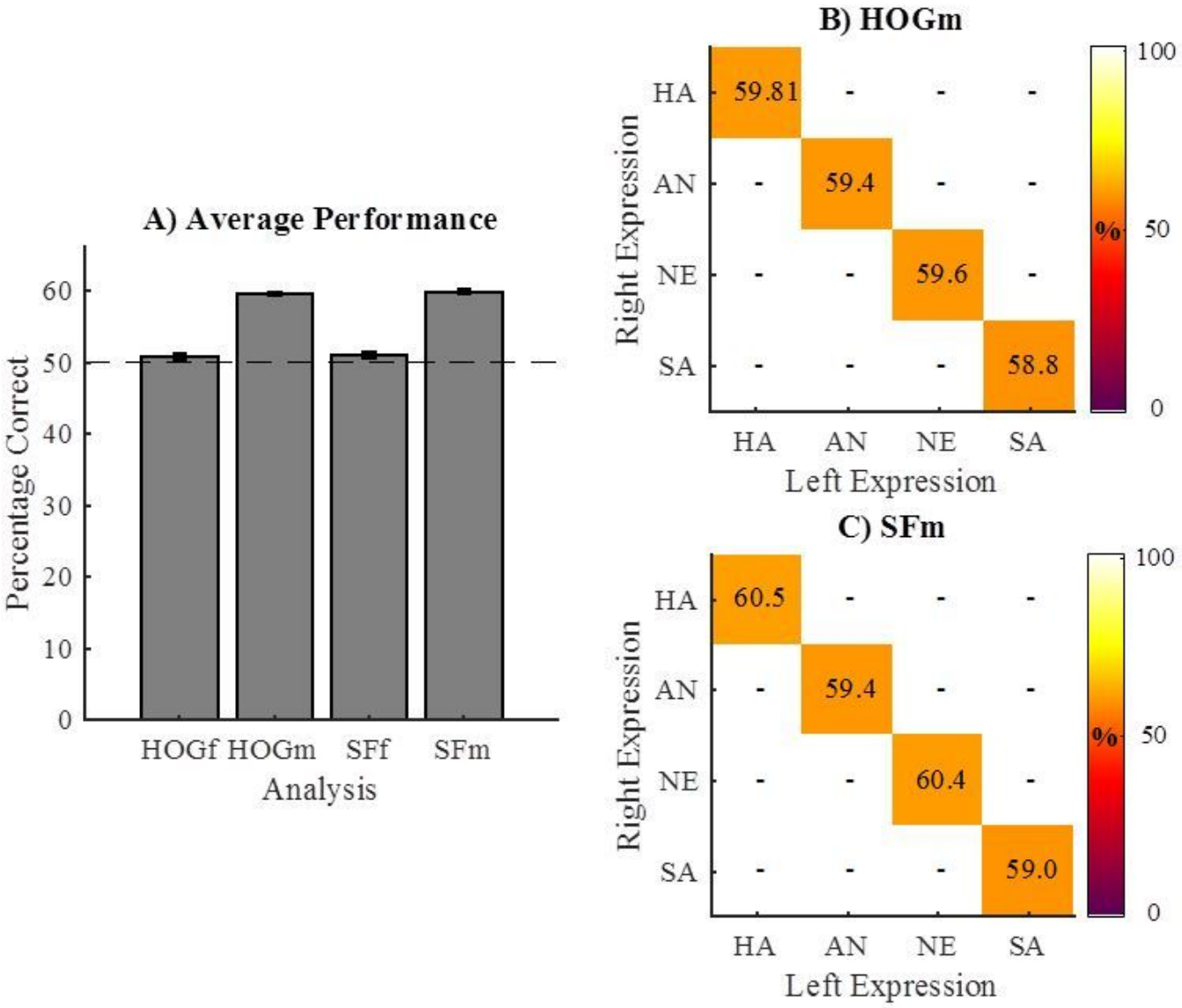

Figure 8

In figure 8A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on

the trials where the same expression was present to both the left and the right side of the screen, leaving only a difference in identity. The dotted line represents the overall empirical chance level performance. Errorbars represent the Standard Error of the Mean. 8B-C) Confusion Matrices for the minimal models. For all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that the differences in performance are very small.

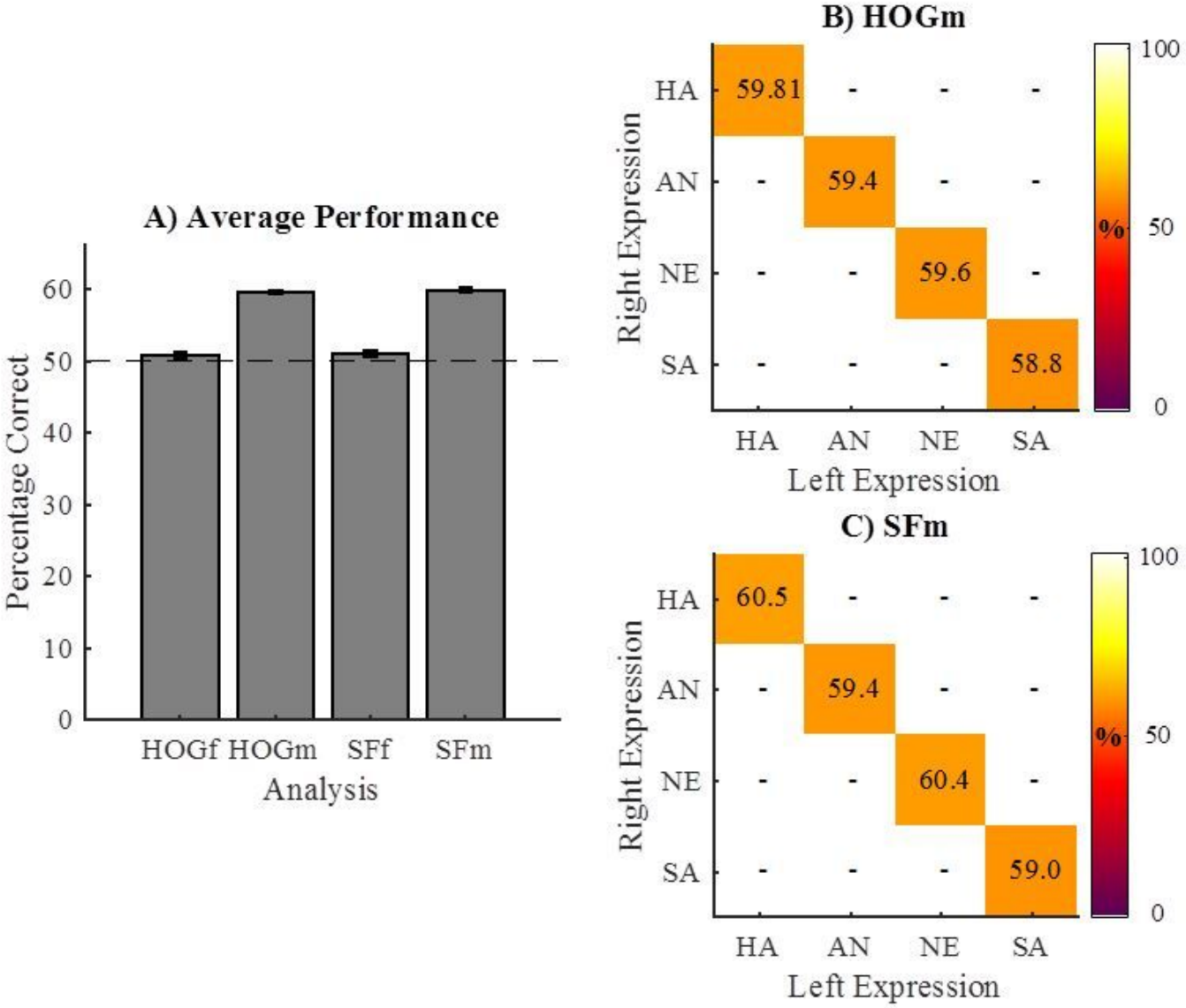

Figure 8

In figure 8A we show the average decoding performance across participants (y-axis) for different modelling procedures (x-axis; HOGf = HOGfull, HOGm = HOGmin; SFf = SFfull; SFm = SFmin) based on the trials where the same expression was present to both the left and the right side of the screen, leaving only a difference in identity. The dotted line represents the overall empirical chance level performance.

Errorbars represent the Standard Error of the Mean. 8B-C) Confusion Matrices for the minimal models. For all trials of each participant, we reorganized the decoding performance to show how well the model performed for different pairs of expressions. Here, performance is represented as a matrix with expression of the left face on the y-axis and right face on the x-axis. Colour intensity reflects the fraction correct for the specific combination of expressions. Note that the differences in performance are very small.

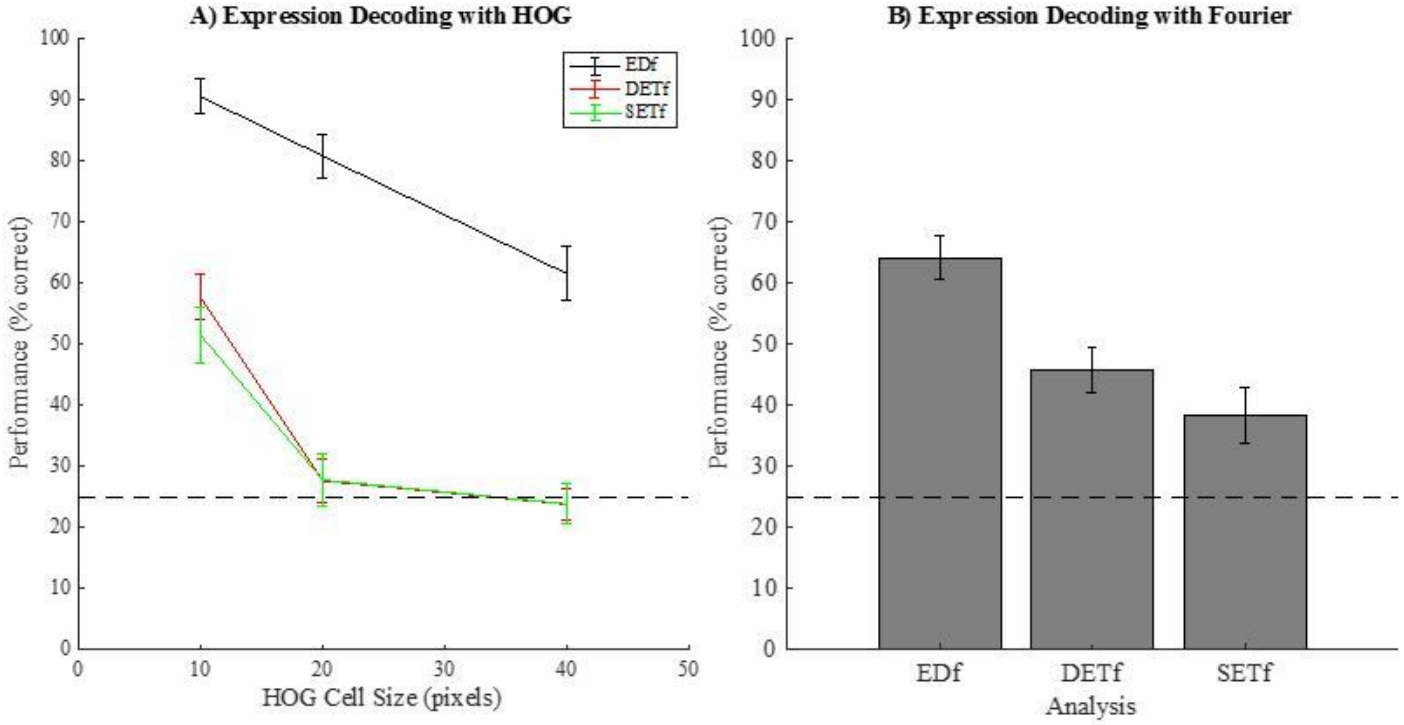

Figure 9

Average decoding performances across all folds (y-axis) based on different sets of HOG and Fourier features. The dotted black lines represent chance level performance. Errorbars represent the Standard Error of the Mean. Figure 9A shows average performance for decoding expressions based on HOG features using three different feature sets at three different spatial resolutions (x-axis). EDf (Expression Decoding features) uses a feature set based on feature selection for expression decoding (i.e. decoding the semantic label of the expression, not selection behaviour), DETf (Different Expressions Trials features) uses the features based on decoding eye-movements towards faces with different expressions and SETf (Same Expressions Trials features) uses the features based on decoding eye-movements towards faces with the same expression. Figure 9B shows average performance for decoding expressions based on Fourier features, again using three different feature sets (x-axis). Note that, for both HOG and Fourier features, the features based on decoding selection behaviour are suboptimal for decoding expressions. Moreover, only the high-resolution HOG features based on decoding eye-movement behaviour are relevant for decoding expressions.

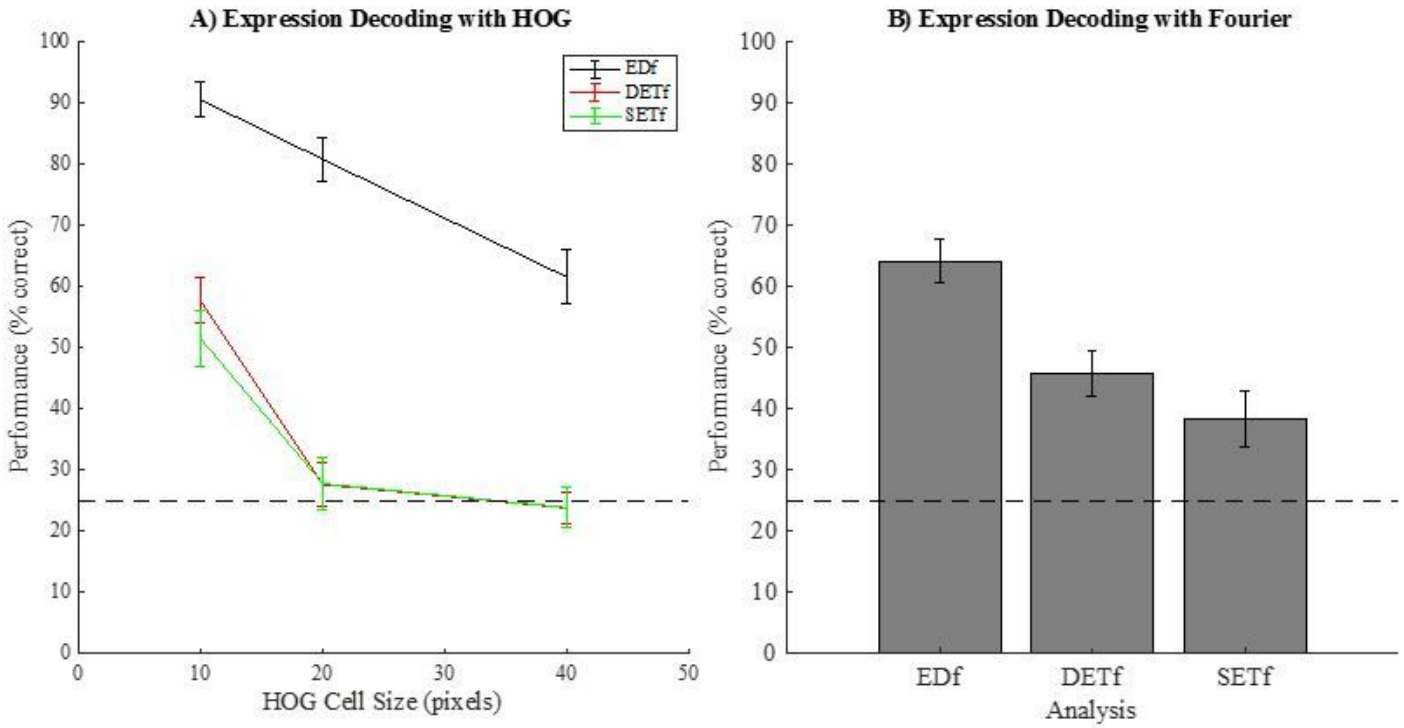

Figure 9

Average decoding performances across all folds (y-axis) based on different sets of HOG and Fourier features. The dotted black lines represent chance level performance. Errorbars represent the Standard Error of the Mean. Figure 9A shows average performance for decoding expressions based on HOG features using three different feature sets at three different spatial resolutions (x-axis). EDf (Expression Decoding features) uses a feature set based on feature selection for expression decoding (i.e. decoding the semantic label of the expression, not selection behaviour), DETf (Different Expressions Trials features) uses the features based on decoding eye-movements towards faces with different expressions and SETf (Same Expressions Trials features) uses the features based on decoding eye-movements towards faces with the same expression. Figure 9B shows average performance for decoding expressions based on Fourier features, again using three different feature sets (x-axis). Note that, for both HOG and Fourier features, the features based on decoding selection behaviour are suboptimal for decoding expressions. Moreover, only the high-resolution HOG features based on decoding eye-movement behaviour are relevant for decoding expressions.

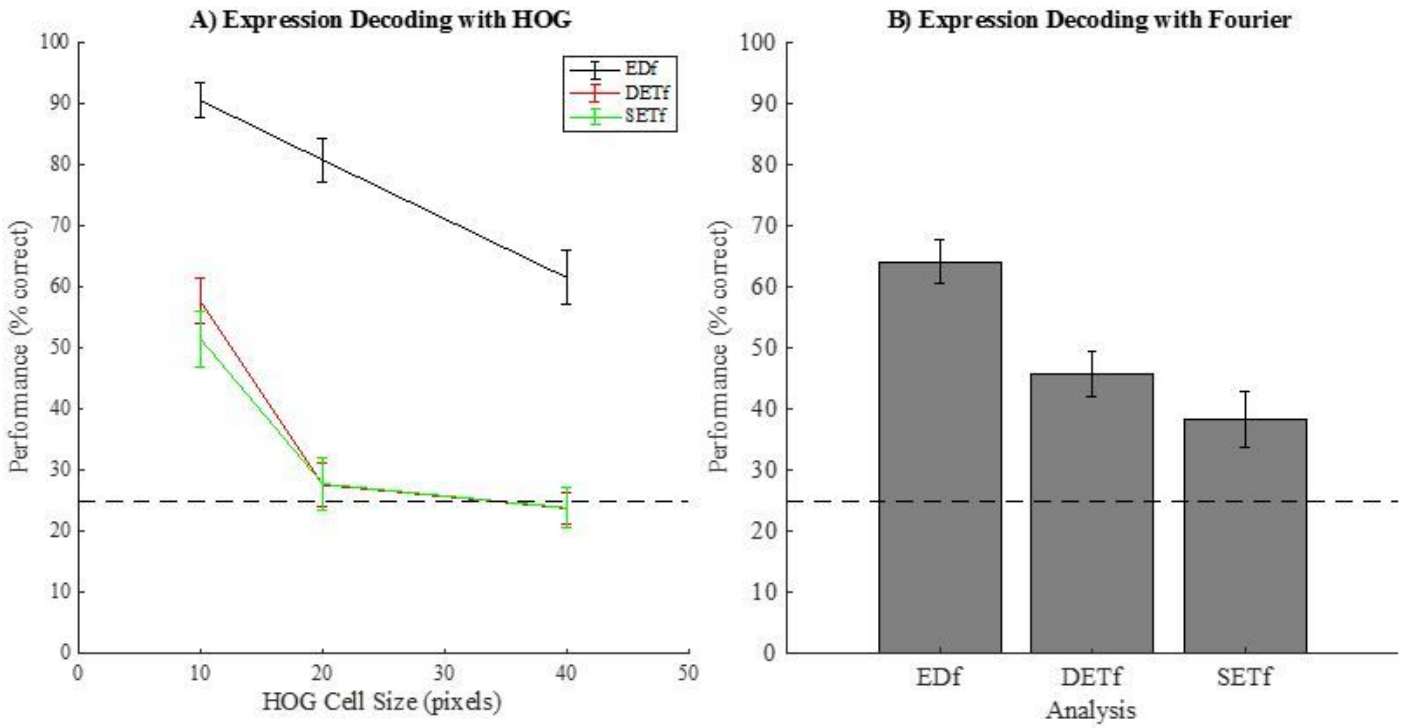

Figure 9

Average decoding performances across all folds (y-axis) based on different sets of HOG and Fourier features. The dotted black lines represent chance level performance. Errorbars represent the Standard Error of the Mean. Figure 9A shows average performance for decoding expressions based on HOG features using three different feature sets at three different spatial resolutions (x-axis). EDf (Expression Decoding features) uses a feature set based on feature selection for expression decoding (i.e. decoding the semantic label of the expression, not selection behaviour), DETf (Different Expressions Trials features) uses the features based on decoding eye-movements towards faces with different expressions and SETf (Same Expressions Trials features) uses the features based on decoding eye-movements towards faces with the same expression. Figure 9B shows average performance for decoding expressions based on Fourier features, again using three different feature sets (x-axis). Note that, for both HOG and Fourier features, the features based on decoding selection behaviour are suboptimal for decoding expressions. Moreover, only the high-resolution HOG features based on decoding eye-movement behaviour are relevant for decoding expressions.

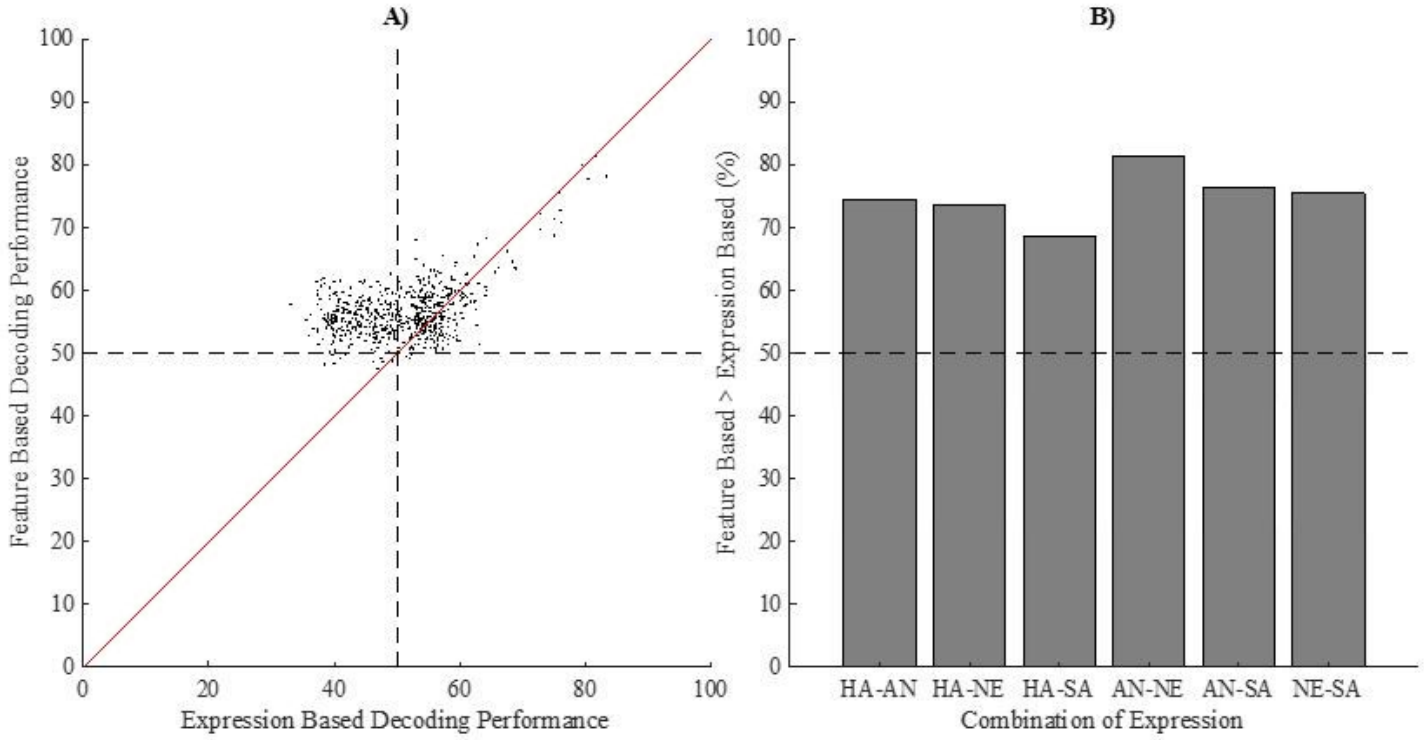

Figure 10

Figure 10A shows the relation between the average percentage correct prediction for each participant and each combination of expressions based on the biases towards expressions (x-axis) and the average corresponding percentage correct prediction based the visual features of the images (y-axis). The dotted vertical line represents chance level performance for expression-based decoding. The dotted horizontal line represents chance level performance for visual feature-based decoding. The solid diagonal line shows where performance would be equal. Figure 10B shows, for each combination of expressions separately, the percentage of the predictions where visual feature-based prediction out-performed prediction based on biases towards expressions. Note that for all expressions this is the case (all values well above 50%).

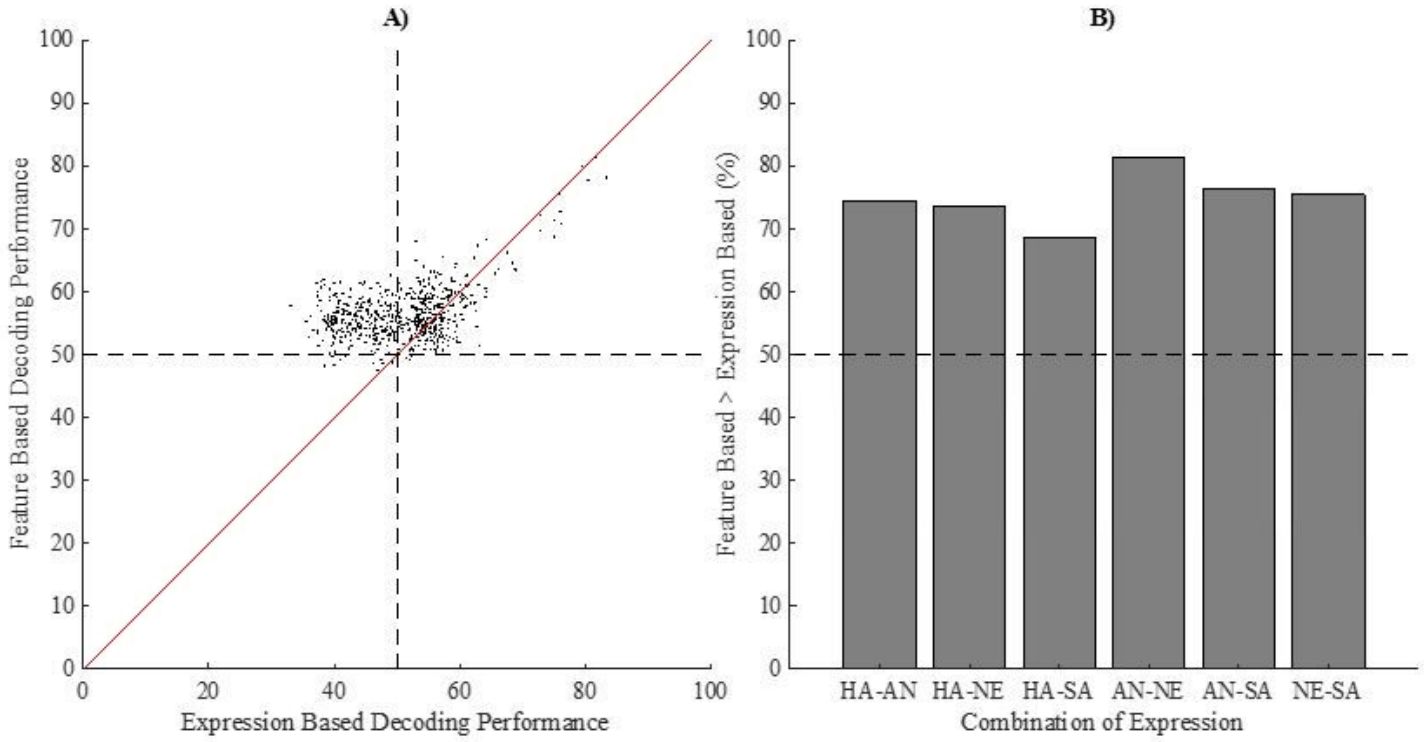

Figure 10

Figure 10A shows the relation between the average percentage correct prediction for each participant and each combination of expressions based on the biases towards expressions (x-axis) and the average corresponding percentage correct prediction based the visual features of the images (y-axis). The dotted vertical line represents chance level performance for expression-based decoding. The dotted horizontal line represents chance level performance for visual feature-based decoding. The solid diagonal line shows where performance would be equal. Figure 10B shows, for each combination of expressions separately, the percentage of the predictions where visual feature-based prediction out-performed prediction based on biases towards expressions. Note that for all expressions this is the case (all values well above 50%).

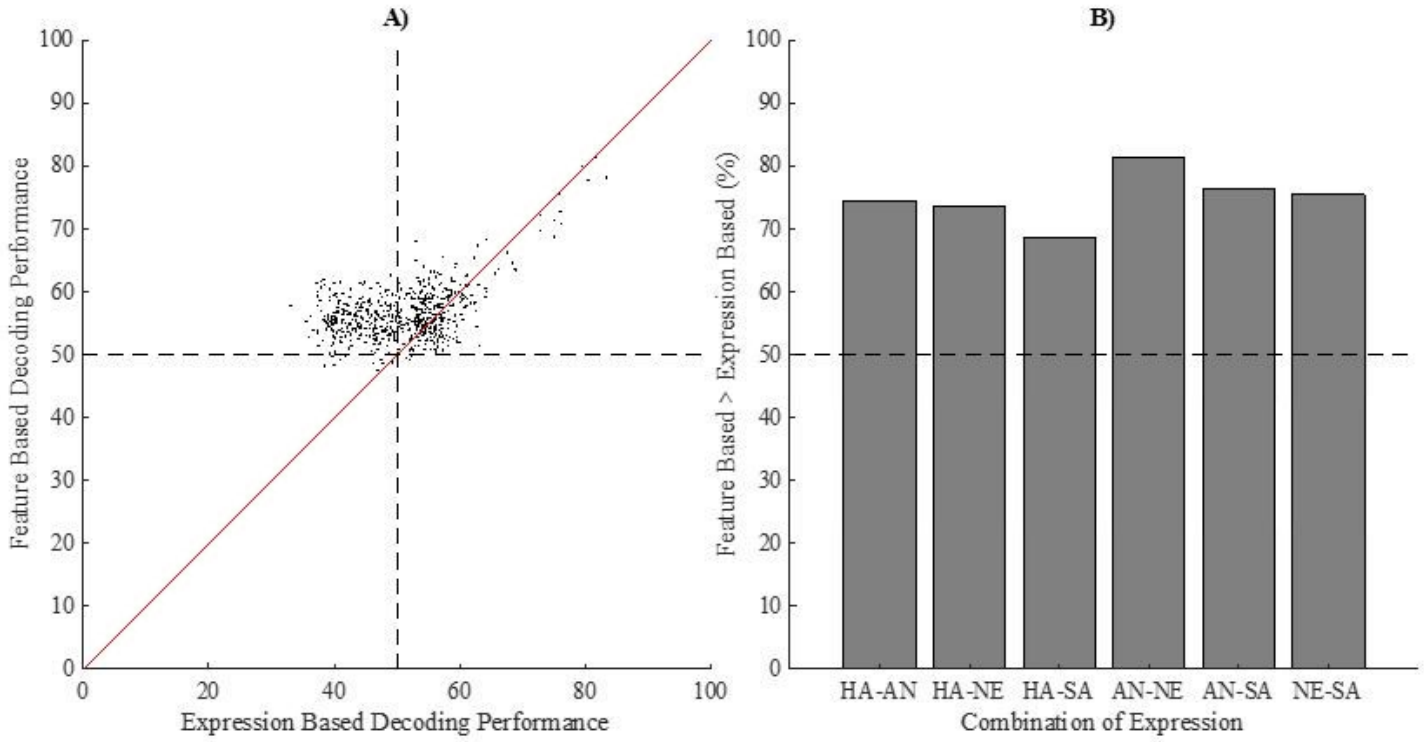

Figure 10

Figure 10A shows the relation between the average percentage correct prediction for each participant and each combination of expressions based on the biases towards expressions (x-axis) and the average corresponding percentage correct prediction based the visual features of the images (y-axis). The dotted vertical line represents chance level performance for expression-based decoding. The dotted horizontal line represents chance level performance for visual feature-based decoding. The solid diagonal line shows where performance would be equal. Figure 10B shows, for each combination of expressions separately, the percentage of the predictions where visual feature-based prediction out-performed prediction based on biases towards expressions. Note that for all expressions this is the case (all values well above 50%).