Attention, Perception, & Psychophysics (2021) 83:2159–2173
https://doi.org/10.3758/s13414-021-02281-6

# Facial expressions can be categorized along the upper-lower facial axis, from a perceptual perspective

Chao Ma[1] · Nianxin Guo[1] · Faraday Davies[2] · Yantian Hou[1] · Suyan Guo[1] · Xun Zhu[1]

Accepted: 9 February 2021 / Published online: 23 March 2021
© The Psychonomic Society, Inc. 2021

## Abstract

A critical question, fundamental for building models of emotion, is how to categorize emotions. Previous studies have typically taken one of two approaches: (a) they focused on the pre-perceptual visual cues, how salient facial features or configurations were displayed; or (b) they focused on the post-perceptual affective experiences, how emotions affected behavior. In this study, we attempted to group emotions at a peri-perceptual processing level: it is well known that humans perceive different facial expressions differently, therefore, can we classify facial expressions into distinct categories in terms of their perceptual similarities? Here, using a novel non-lexical paradigm, we assessed the perceptual dissimilarities between 20 facial expressions using reaction times. Multidimensional-scaling analysis revealed that facial expressions were organized predominantly along the upper-lower face axis. Cluster analysis of behavioral data delineated three superordinate categories, and eye-tracking measurements validated these clustering results. Interestingly, these superordinate categories can be conceptualized according to how facial displays interact with acoustic communications: One group comprises expressions that have salient mouth features. They likely link to species-specific vocalization, for example, crying, laughing. The second group comprises visual displays with diagnosing features in both the mouth and the eye regions. They are not directly articulable but can be expressed prosodically, for example, sad, angry. Expressions in the third group are also whole-face expressions but are completely independent of vocalization, and likely being blends of two or more elementary expressions. We propose a theoretical framework to interpret the tripartite division in which distinct expression subsets are interpreted as successive phases in an evolutionary chain.

**Keywords** Emotion · Face perception · Categorization · Facial expression

## Introduction

Researchers generally agree that facial expressions have evolved to signal a specific emotional status of the signaler and served a critical communicative function between conspecifics (for review, see Dezecache, Mercier, & Scott-Phillips, 2013). According to this point of view, if the function of facial expressions is to convey emotional information to conspecifics, the evolution of the capacity to visually express emotions had to be accompanied by the ability to visually interpret them in order to maximize transmission through a noisy medium. That is, if our face evolved in part as a device to optimize the transmission of the emotional signals, then our visual system probably co-evolved as an efficient decoder of these signals (see Schyns, Petro, & Smith, 2009). This suggests an isomorphism between the visual signal (e.g., happy) and its visual decoding (i.e., analyzing the happiness). This isomorphism brings up a potentially interesting question, can we classify emotions in terms of their perceptual characters? The idea is that according to the above-mentioned isomorphism, if several individual emotional facial expressions are visually processed similarly, they likely belong to the same superordinate emotion category.

It is well known that different facial expressions are detected or recognized at different latencies and accuracies (see Nummenmaa & Calvo, 2015). However, to our knowledge, no study has classified facial expressions in terms of their perceptual characters. Previous investigations targeting the categorization of emotion have typically taken one of two approaches. The first approach describes a discrete,

✉ Xun Zhu
zhuxun@z-and-z.com; zhuxun@shzu.edu.cn

[1] Department of Psychology, Normal College, Shihezi University, Xinjiang, China

[2] Department of Psychiatry and Behavioral Sciences, College of Medicine, Medical University of South Carolina, Charleston, SC, USA

categorical model of emotion, which propounds that emotional facial expressions can be categorized into a few canonical prototypes, for example, happiness, anger, sadness, fear, disgust, and surprise (Ekman, 1992). Each expression prototype is defined by a specific combination of facial muscle movements, called action units (AUs). The AU activations are consistent within and different between emotional categories (Du, Tao, & Martinez, 2014). That is, this approach categorizes emotions according to their visual cues. The second approach describes a dimensional model of emotion, which assumes that emotions are centered on subjective experiences. Since language is the primary access to affective experiences, the mutual relation between the semantic fields of emotion prototypes defines the structure of the emotion space, also called the *subjective emotion space* (Sokolov & Boucsein, 2000). Efforts to understand this space have been focused on its dimensionality, for example, valence and arousal (Kuppens, Tuerlinckx, Russell, & Barrett, 2013).

In short, to categorize emotions, previous studies focused either on its pre-perception visual cues or on its post-perceptual conscious experiences. However, there are still fundamental limitations to both approaches. The pre-perception approach assumes that the fundamental differences between facial expressions are rooted in the pictorial level – either at a pixel level (Calder et al., 2001) or at an AU level (Martinez, 2017). A major problem with this approach is that the search for the brain's region of interest responsible for individual emotion categories or AUs has come up empty handed (see Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012). The post-perceptual approach posits that we code emotions along continuous affective dimensions, for example, valence and arousal. The problem with this approach is that the dimensional structure reflects how people parse emotions (converting the non-verbal emotional information into verbal-linguistic concepts) rather than how people recognize emotion. A developmental study showed that the rise of such a multidimensional representation of emotion is mediated by the increase in verbal knowledge and is associated with the general ability to represent non-emotional stimuli dimensionally (Nook, Sasse, Lambert, McLaughlin, & Somerville, 2017). It is possible that the structure of emotion may not be the same as the structure of emotional language. Together, current categorization approaches are inconclusive. Further research is needed to better understand the structure of emotion.

In this contribution, unlike previous studies that focused either on the pre-perception visual cues or the post-perceptual conscious experiences, we attempt to categorize emotions at a peri-perceptual level, namely based on how humans perceive different facial expressions differentially. We argue that this approach is feasible and necessary, for the following reasons:

First, previous research has shown that different expressions were processed differentially during visual perception. It is well known that different facial expressions were detected or recognized at different latencies and accuracies (Nummenmaa & Calvo, 2015). Previous research also suggested that the brain has multiple emotion-processing mechanisms. For example (see Gainotti, 2020), the valence-specific hypothesis posits that the left hemisphere is specialized for processing positive affect while the right hemisphere is specialized for negative affect; and the approach/withdrawal hypothesis posits that emotion is associated with left or right lateralization according to the extent to which it is accompanied by approach or avoidance motivation. Moreover, besides cortical pathways, a subcortical pathway might also exist (for review, see Liebenthal, Silbersweig, & Stern, 2016). The strongest evidence supporting the existence of a subcortical pathway came from case studies of “affective blindsight,” in which V1 damaged patients could correctly “guess” whether a face was depicting certain expressions, especially happiness, anger, or fear, despite their insistence that they saw nothing and were “just guessing.” Evidence converges on the role of subcortical structures of old evolutionary origin such as the pulvinar and amygdala in mediating affective blindsight and nonconscious perception of emotions (for review, see Celeghin, de Gelder, & Tamietto, 2015). Overall, previous studies suggested that different facial expressions might be processed differently. Consequently, classifying facial expressions according to their perception signatures is likely feasible.

Second, we argue that the perception of facial expressions – i.e., the interpretation of the facial display – may be at least as important, if not more important, than the emotional facial cues. Several lines of evidence have suggested that displacements of facial features (i.e., AU activations) do not necessarily convey signals of affect; instead, the facial display is needed to be interpreted by the brain to generate the psychological sense of that display. For example, by morphing all facial features of emotional expressions in the opposite direction from the neutral expressions by an amount equivalent to the difference between the emotional and neutral expressions, we can create artificial “anti-expressions,” for example, anti-happy (Sato & Yoshikawa, 2009). Anti- and normal expressions are equivalently different from neutral expressions in the displacement of facial features (but in opposite direction). However, anti-expressions do not express the emotional messages opposite to the normal expressions, instead, they express almost no emotional messages and are perceived as emotionally neutral (Sato & Yoshikawa, 2009, 2010). The "anti-expressions" demonstrated that there is no linear relationship between AU activations and the emotional messages. In addition, the AUs are also unlikely to have universal affective meaning. As an example, although facial musculature is almost identical for chimpanzees and humans (Burrows, 2008), and there is a potential homology between several chimpanzee expressions and human expressions (Preuschoft

& van Hooff, 1995), chimpanzees use somewhat different combinations of AUs to express their prototypical expressions (Parr, Waller, Vick, & Bard, 2007). This is not surprising, as chimpanzees lack the key upper-face visual contrasts that can be seen in humans. Chimpanzees have larger brows than humans, but the detection of brow movement is enhanced in humans due to our hairless forehead. Furthermore, the lack of white sclera in chimpanzees apparently impedes the detection of facial motion near the eyes. In comparison, the human eye is especially visible relative to all other primates (Kobayashi & Kohshima, 2001). The absence of contrast in the chimpanzee upper face makes discriminating upper-face facial movements extremely challenging, therefore, unlike humans, chimpanzee emotions are conveyed mainly using lower-face AUs. Together, it seems that the AUs have no innate signal function, so the brain needs to interpret the face display to decipher its affective contents (but see Martinez, 2017, for a different view). The structure of emotion in the perception stage will, therefore, be of particular interest to explore.

To our knowledge, there are few if any studies that have investigated the structure of emotion in the realm of its visual perception stage. In this study, we aim to address this missing aspect. We first measure the perceptual dissimilarities between 20 prototypical facial expressions, then use multidimensional scaling (MDS) to summarize the structure of emotional expressions. We therefore look for the most meaningful set of axes or clusters in the resulting MDS solution and validate our interpretations using an independent set of data. We present our results below.

## Materials and methods

### Participants

Sixty-nine participants were enrolled. Eight were excluded due to low response rates (< 80%) or bad eye fixations (< 300 ms per trial averaged). Therefore, 61 participants (mean age 21.4 years, SD 2.5, 32 females (29 Chinese and 32 Turkic)) were included in the analysis and presentation of findings. The sample size was determined as, when assuming a medium effect size (0.5), to achieve a power of 0.95, requiring 52 participants (significance threshold 0.05; two-tailed repeated-measure t-test). The medium effect size was assumed according to Lench, Flores, and Bench's meta-analysis (2011), which reported that the overall effect size associated with comparisons between discrete emotions was 0.51 across 687 studies.

We recruited our participants from two distinct ethnic groups for the reasons explained below. Emotions are deciphered from faces, but faces are not simply blank "canvases" upon which facial expressions manifest their emotional message. The "canvas" might affect how emotions are perceived (Hess, Adams, & Kleck, 2009; Wang, 2018). Given that we focus on the visual perceptual stage, this factor should be constrained. Accordingly, ethnicity is incorporated into our task paradigm: facial expressions were posed in two different "canvases": same-race faces and other-race faces. This enabled us to investigate whether the perceptual patterns are stable.

In the Xinjiang province where our university is located, the demographic balance is 52% Turkic (including 46% Uyghur and 6% Kazakh) and 45% Chinese (including 40% Han and 5% Hui). Accordingly, our participants have considerable experience with other-race faces. Turkic people speak a Turkic language written with an Arabic script and are as distinct in appearance from the Chinese as Native Americans are from Caucasians. They are genetic descendants of western and eastern Eurasian populations, with the western Eurasian versus eastern Eurasian genetic ratio for Uyghur and Kazakh being 54:46 and 34:66, respectively (Lou et al., 2015). To better illustrate the difference, average-faces of these ethnic groups are presented in Online Supplemental Material (OSM) S1.

### Stimuli

We tested 20 facial expressions, based on the emoji set (https://findicons.com/pack/1039/manto) we choose to use. Reasons for choosing this emoji set included: it is three-dimensional (3D) and has high resolution, is composed of moderate numbers of emotion, and most emojis are used highly frequently online. Sixteen paid adults (half male, eight of each race) were recruited to pose these 20 expressions. Facial occlusions were minimized with no eyeglass or jewelry. They were also asked to uncover their forehead to fully show the eyebrows. The pictures were taken against a blue background without flash using a Nikon D200 digital camera in our lab's studio. The experimenter showed them the emoji and suggested a possible situation that might cause this emotion. Photos were taken at the apex of the expression. For each emotion, four photographs were taken. The two that better depicted that emotion were chosen by a three-member panel. The final image inventory consisted of 640 images (320 pairs) , and is available publicly at our lab website http://lab.z-and-z.com/shares/Emotion20.zip sized 0.8G.

### Apparatus and procedure

The participant was seated comfortably in a dimly lit, acoustically shielded EEG room. Stimuli were presented on a 23-in. LED monitor at its native resolution of 1,680 × 1,050, guided by Eprime 2.0 software. At a viewing distance of 60–70 cm (no chinrest was used), faces subtended 6–7° of the visual angle ear to ear. Full-color images were used to produce a more realistic representation of the human face.

This 960-trial study (20 expressions × 2 races × 24 repetitions each) was organized into two sessions with a filler task (reported elsewhere) between sessions. Each trial consisted of an emoji in the center of the screen at its native resolution (128 × 128 pixels) and two pictures of that expression (posed by the same person) on its left and right (resized to 588 × 682 pixels) for 3 s or until a response was made. Participants completed a two-alternative forced-choice task in which they were required to indicate which of the two simultaneously presented pictures better portrayed the target emotion using appropriate keys in a keyboard (labeled "left" and "right"). A 400-ms fixation period interleaved between trials. Participants were given a self-timed rest period in the middle of each session. Trial sequences were randomized for each participant (Fig. 1).

The key challenge of this experiment is properly measuring the perceptual similarities between facial expressions. Here, we measure the perceptual dissimilarities by recognition time. The recognition time could serve as an indirect measurement reflecting the perceptual processing when the intrusion of non-perceptual processes was minimized. On the constructivist view, the recognition of emotion depends on the perception of the emotional expressions and on the ability to make sense of such expressions. We aimed to focus on the first stage and control for the effects of the ensuing affective encoding stage. To meet this end, we employed a non-lexical paradigm, as the post-perceptual affective encoding stage requires conceptual knowledge whereas language scaffolds concept knowledge in humans. Most prior research into this topic have involved lexical processing, for example, Schlosberg's (1952) well-known two-dimensional circular model was based on the interchangeability of lexical labels for emotions. Previous studies that purposely avoided explicit judgment of the emotions have used either a multiple sorting paradigm (Lindquist, Gendron, Barrett, & Dickerson, 2014) or an odd-one-out paradigm (Nishimura, Maurer, & Gao, 2009), or required participants to report similarity ratings of a heterogeneous pair of expressions (Shah & Lewis, 2003). However, a potential methodological issue with these paradigms is that they still couldn't prevent the application of linguistic labels that may occur during the deliberation of similarity decisions.

Our paradigm was designed to minimize lexical and conceptual engagement. It required the participant to rate which of the two simultaneously presented photos was a better representation of the target emotion (which was also presented in a non-verbal, visual way, i.e., via an emoji). In such a discrimination task, the decision can be made either perceptually (e.g., by visual inspection) or conceptually (i.e., involves the application of linguistic labels to the facial expression and the comparison of the labels). The idea was that if both photos depicted the target emotion and are quite similar (since they were posed by the same person), the decision couldn't be made conceptually: we didn't have proper linguistic labels to describe such subtle differences. Consequently, behavioral differences between different expressions could be primarily attributed to the difference in the visual perception of the facial expressions, and the reaction time (RT) differences could be used to indirectly characterize perceptual dissimilarities between pairs of expressions.

Remote, contact-free acquisition of eye-tracking data was carried out simultaneously using SMI RED 500 systems (SensoMotoric Instruments Inc. USA) with a sample rate of 500 Hz. The eye-tracker was placed in front of the subject, just below the monitor where the stimuli were presented. Standard nine-point calibration and validation procedures were carried out. Default criteria for fixations (minimum fixation duration 100 ms), saccades (0.6° 45°/s) and blinking as implemented in the system were used. Areas of interests were defined as the face photo regions combined (i.e., include both photos, but

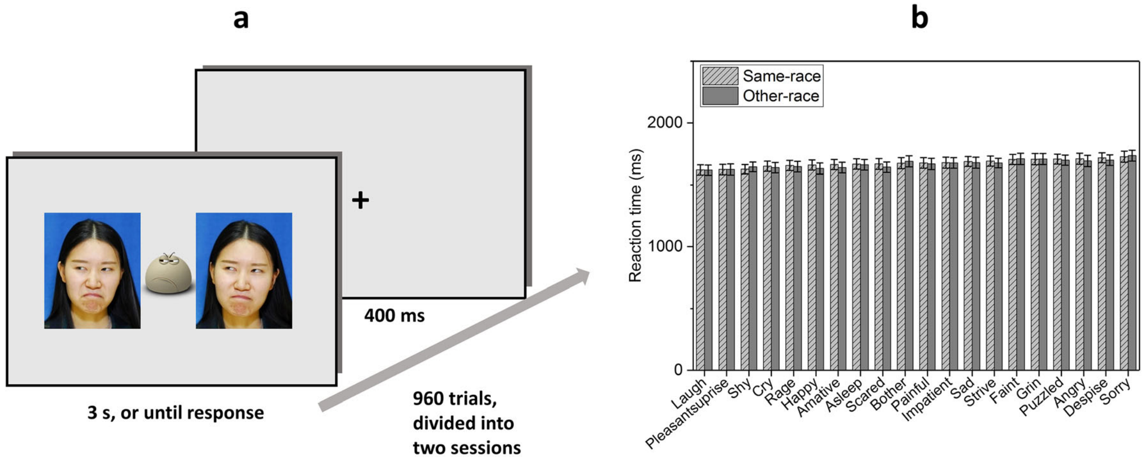

**Fig. 1** Task paradigm **a)** Participants were required to indicate which of the two simultaneously presented photos better portrayed the target emotion (depicted by an emoji showed centrally). **b)** Reaction times (RTs) for the 20 emotional expressions (overall average RT = 1,673 ms, SD = 317 ms). The main effect of expression type is significant, $F(19) = 10.05$, $p < 0.001$, $\eta_p^2 = 0.143$. Error bars depict the standard error of mean

exclude the emoji region). Various eye-tracking measures, i.e., gaze location, number of fixations, gaze duration, and pupil size, were extracted. EEG signals were also collected simultaneously but are not reported here.

### MDS analysis

Individual response times less than 500 ms or exceeding 2,500 ms were excluded (average RT = 1,673 ms, SD = 317 ms), accounting for 5.6% of the data. No responses accounted for 3.3% of the data. Separate analyses was run for same-race and other-race data.

The first stage of MDS is the construction of distance matrices showing the pair-wise distances between facial expressions. We define the perceptual distance between two facial expressions *i* and *j* as the absolute RT difference between the pair of facial expressions, that is, |RT*i*-RT*j*|. In this way, we generated one distance matrix for each participant. These distance matrices were submitted to a weighted multidimensional scaling function in SPSS 23 (also known as individual differences scaling or INDSCAL, which is a procedure that enables the model to account for individual differences in cognitive processes or perceptions; Carroll & Chang, 1970). This MDS analysis aims to represent the facial expressions in a low-dimensional space, so it is easier to identify the key features of the data.

A bootstrap analysis (see Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005) was performed to assess the instability of the location of facial expressions in the MDS space. A new MDS space was created by randomly selecting 50 subjects, with replacement (i.e., one subject might be used multiple times). This technique was repeated 100 times. The generated 100 MDS spaces were then aligned by in-plane rotation and were superimposed on a single representation. The position of each facial expression in these 100 analyses defined a "cloud" of points on this single representation. The idea is that the greater the variability between subjects, the more each space must be different, thus, the size of the "cloud" expresses the stability of each facial expression in the MDS space.

### Verifying the interpretation of MDS dimensions

Interpreting the feature dimensions of the generated MDS space is highly subjective. However, it is possible to indirectly verify our interpretations by a follow-up analysis (Hout et al., 2016; Hout, Papesh, & Goldinger, 2013). In this analysis, a new group of participants (n = 30, aged 24.3 ± 1.2 years, 11 males; none took part in the previous experiment) were asked to rate each stimulus with regard to how much it represents the dimensions we proposed. For each facial expression (one picture at a time, all same-race pictures), participants completed a four-alternative forced-choice task in which they indicated which of the facial features (mouth, eyes, eyebrows, others) were most salient for defining this expression. Also, participants were asked to rate valence, arousal, and ambiguity using a 5-point Likert-scale. These ratings were then regressed over the coordinates derived from each dimension in the resulting same-race bootstrap MDS analysis. A high regression weight could be taken as evidence that a particular dimension reflects the hypothesized construct.

### Clustering analysis and the validation of clusters

We note that the identification of interpretable axes for the MDS solution is not always the best way to discern interesting patterns. Perhaps we can also identify clusters of facial expressions that have a practical significance. Therefore, we conducted a hierarchical cluster analysis on the MDS solutions derived from same-race RT data, and the resulting dendrogram was then used to identify the number of potential clusters. Cluster memberships were then determined by subsequent *k*-mean cluster analysis.

Since RT data is used to define the clusters, to avoid circularity we need independent data to validate the goodness of clustering. Therefore, eye-tracking data, which is acquired simultaneously with the task, were used to validate the clustering results. In this procedure, we averaged each participant's eye-tracking measurements (i.e., number of fixations, gaze duration, gaze location, and pupil size) across emotions for each cluster. Then, we ran a repeated-measures ANOVA with the clusters being the independent variable and various eye-tracking measurements being the dependent variable. The results should verify whether there are statistically significant differences between the clusters. Since behavioral results revealed only a weak overall effect of race, same-race and other-race data were combined in these analyses.

## Results

### MDS results

Given that people encounter same-race faces much more often than other-race faces, we first examined the same-race data. We limited the analyses to two-dimensional (2D) solutions, according to the scree plots (see OSM S2), and based on the fact that a low-dimensional solution promptly allowed us to take advantage of the eye's ability to spot patterns in the plots. The solution (Fig. 2a) delivered a roughly circular arrangement of the facial expressions with no facial expression in the center of the space, commensurate with previous two-dimensional solutions (for lists, see Posner, Russell, & Peterson, 2005, and Shah & Lewis, 2003). The major contrast between our results and the previous circumplex results is that the structure of our space is not dominated by changes in

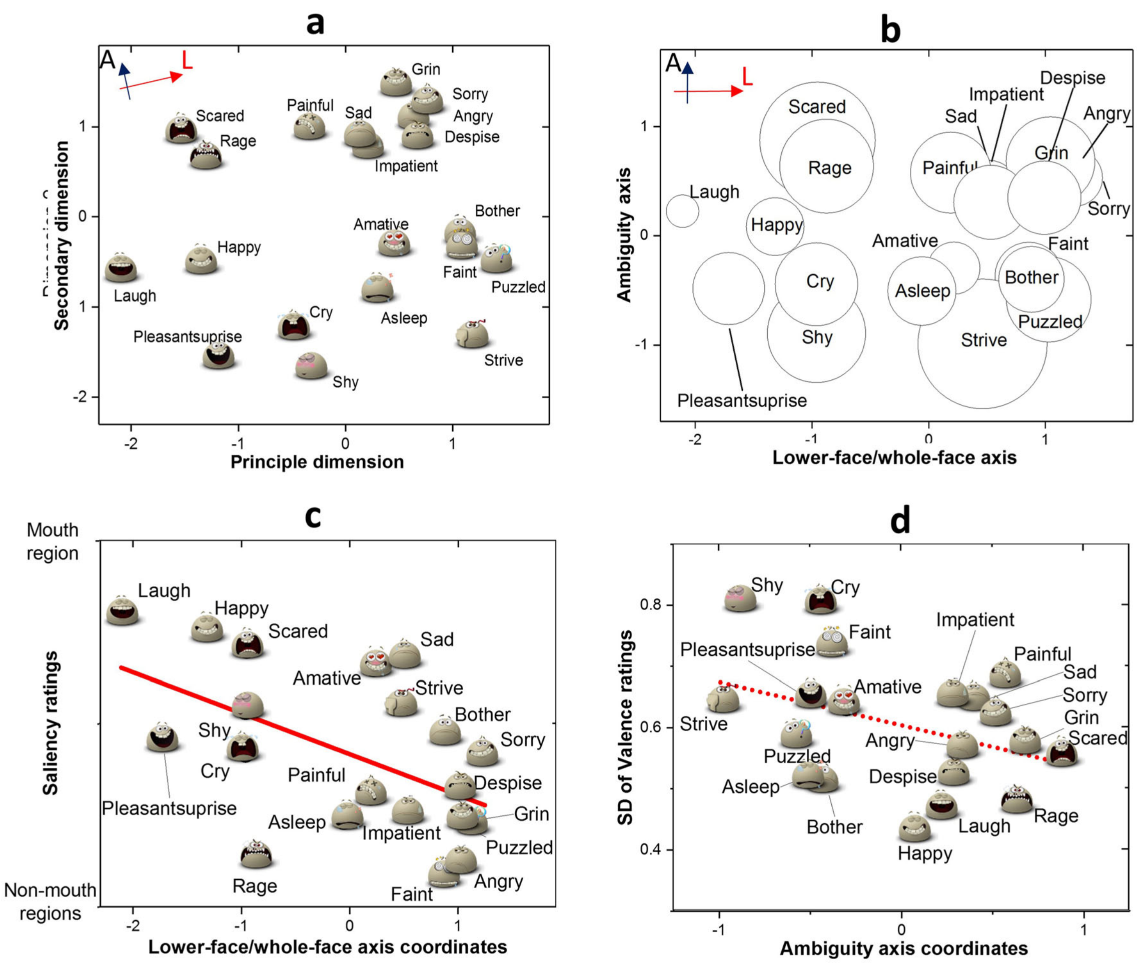

**Fig. 2** MDS solution, based on same-race reaction times (RTs) **a)** Locations of the 20 facial expressions are indicated together with the label given to the poser. The proposed dimensions are depicted in the upper-left corner. **L**: lower-face/ whole-face axis; **A**: ambiguity axis. Please note that Euclid space is rotation invariant, and perhaps some optimal rotation would provide a better fit and allow meaningful axes to emerge. **b)** Bootstrap MDS solution, which expresses the stability of each facial expression in the 2D space. The radius of the sphere depicts the mean root-squared distance, which characterized the amount of 2D dispersion around the centroid of each emotion. **c)** For the principal MDS dimension (lower-face/upper-faces axis), the MDS coordinates are significantly correlated to individual ratings (p = 0.017). **d)** For the secondary MDS dimension (ambiguity axis), the coordinates are *marginally* related to the standard deviation of valence rating (p=0.096)

emotional valence. Previous studies bore a remarkable resemblance to each other, with emotions appearing to lie on a circular manifold embedded within the Euclid space, and valence consistently being the predominant dimension. In contrast, each quadrant of our facial expression space is populated by both positively and negatively valenced emotions.

The MDS solution's instability was assessed by a follow-up bootstrap analysis (Fig. 2b), in which the size of the "cloud" expresses the stability of each facial expression in the space. As illustrated in Fig. 2b, the surfaces covered by the "cloud" were fairly small, providing evidence for the reliability of the geometric representation.

It is worth noting that, as we expected, the space we derived is more likely to be a perceptual space rather than a semantic space. If the space is semantic, one would expect that facial expressions that fell within the same meaning would be indistinguishable. Our results, however, showed that faces depicting semantically close emotions, for example, happy and grin or cry and sad, are perceived as being different. In comparison, sad and angry, which are semantically far from each other, are perceived as being similar. Interestingly, during visual recognition, people do often mistake anger for sadness (Du & Martinez, 2011). This demonstrated that, as expected, the lexical engagement was not significant in our task paradigm.

## Interpret the dimensions of the MDS space

A low-dimensional MDS solution permits visual inspection. It seems that the principal dimension, that is, a dimension running from bottom left to top right, appears to characterize something about the mouth (Fig. 2a). The individual facial expressions at the low end of this dimension tend to have salient mouth features, that is, a larger mouth (for the superstimuli such as emoji, the salient features are usually highly exaggerated). Facial expressions at the high end of this dimension are mostly whole-face emotions with diagnosing features in both the mouth and the eye regions.

For the axis orthogonal to the principal axis, the visual inspection approach proved to be very inefficient: there is no obvious interpretation of this axis visually. Nevertheless, if interpreted conceptually, prominent differences could be noted. Expressions at the higher end of this axis (upper part of the space) transmit more a precisely specified illocutionary force. In contrast, expressions at the lower end of this axis are either difficult to interpret (e.g., facial expressions such as striving, asleep, and faint are less distinct in their visual cues), or context-dependent (e.g., for cry there are sad tears, happy tears, and angry tears; for shy, the illocutionary force also varies widely). Therefore, the second dimension presumably characterizes ambiguity.

Because the order of the dimensions reflects their relative importance (i.e., the degree to which a particular dimension explains variance), our results showed that facial expressions are organized primarily along the upper-lower face axis and secondarily along the ambiguity axis.

## Validating the interpretation of dimensions

The interpretations of the MDS dimensions were validated by data from a new group of participants. For the principal dimension (upper-lower face axis), regression analysis revealed that the MDS coordinates agree with the individual "mouth region" ratings (beta = -0.526, p = 0.017, $R^2 = 0.277$; Fig. 2c). The coordinates are also counter-correlated with the "eye region" ratings (beta = 0.447, $p = 0.048$, $R^2 = 0.200$), but were uncorrelated with both the "eyebrow region" and "other region" ratings ($ps > 0.1$). The results confirmed the validity of our claims that the upper-lower axis is the predominant dimension underlying the perceptual facial expression space.

For the secondary dimension (ambiguity axis), we found that the coordinates are not related to explicit ambiguity ratings ($p > 0.1$). However, it is *marginally* correlated with an alternative type of ambiguity: the standard deviation of valence rating (beta = -0.382, $p = 0.096$, $R^2 = 0.146$; Fig. 2d). That is, this dimension might not be best interpreted as explicit ambiguity, but might reflect the ambiguity of perceived valence. However, the effect is only marginally significant. Further investigation is necessary before any definite conclusion as to the true nature of the second axis can be arrived at.

## Cluster analysis of the MDS solution

Crucially, some features of the MDS solution show that the present findings are not entirely in support of a dimensional perspective. As can be appreciated from Fig. 2, the distribution of expressions in the MDS space is markedly uneven: facial expressions are likely forming clusters. Moreover, no expression is located near the center of our MDS space. In the classic valence-arousal space, there is an instinctive assumption that neutral or ambiguous expressions should represent the center of the space. However, it is hard to describe what kinds of expressions should populate the center of our MDS space. Hence, the lower-face/upper-face axis may superficially represent the structure of emotion, but some of the assumptions involved in accepting a strict finite-dimensional model are not satisfied. A discrete, multicompartment construct may be more appropriate.

Therefore, cluster analysis is used with the purpose of identifying homogeneous expression subsets. Based on the dendrogram of the hierarchical cluster analysis (Fig. 3a), the number of clusters was determined as three. Cluster memberships were determined by subsequent k-mean cluster analysis and were superimposed on the MDS solution (Fig. 3b).

"Mouth" expressions: Happy, Crying, Pleasant surprise, Laughing, Scared, Rage, Shy. These facial expressions have salient mouth features.

"Whole-face" expressions: Sad, Angry, Painful, Grinning, Despise, Sorry, Impatient. For these expressions, both the eye and the mouth regions are diagnostic.

"Blended" expressions: Amative, Bothered, Puzzled, Asleep, Striving, Faint. These are also whole-face expressions, but their visual cues are less distinct or are difficult to pose. These are likely emotion-blends that involve several simultaneous superimposed/masked elementary expressions.

## Same race versus other race

MDS solutions for same-race and other-race data were remarkably similar (Fig. 3b and c). To quantitatively assess the convergence of same-race and other-race MDS solutions, we calculated a multiple correlation between the two dimensions. The multiple correlation coefficient was 0.78 with x-dimension and 0.72 with y-dimension. Multiple rather than bivariate correlations were used because of the slight rotational difference between the two solutions. We conclude that during the emotion perception, the platform (i.e., racial) difference, if it exists at all, is small. Since people encounter same-race faces much more often than other-race faces, the cluster memberships derived from same-race data (Fig. 3b) were used throughout this study.

## Validating the clustering solution

Given that we define the clusters using RT data, to avoid circularity we need independent data to validate the goodness of clustering. Eye-tracking data, which was acquired simultaneously with task performance (i.e., the RT), was therefore used to assess the goodness of clustering result.

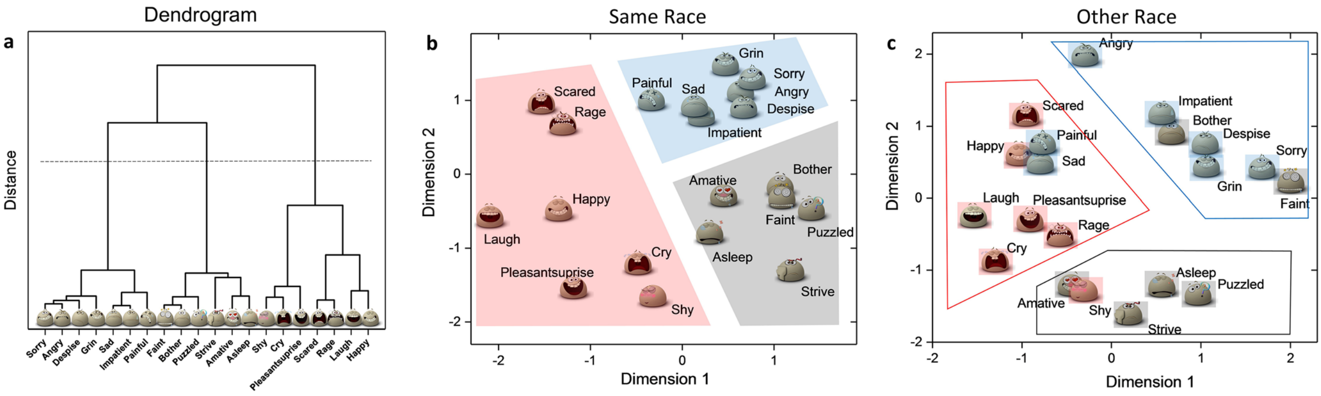

**Fig. 3** Cluster analysis **a)** Dendrogram visualizing the hierarchical clustering of emotions. The number of clusters was determined as three (the dash line), as the distances between the clusters (the vertical segments of the dendrogram) were highest for three clusters. **b)** The same-race data. Cluster analysis suggests a tripartite division within the 2D space. Red, "mouth" expressions; Blue, "whole-face" expressions; Black, "blended" expressions. **c)** The other-race data. The colored outlines depict the cluster analysis results on the solution and the background color of each expression depicts their cluster membership in the same-race scenario. The other-race result closely resembles the same-race result, except that it seems to be rotated approximately 45°clockwise relative to the configuration of the same-race solution. However, since rotation is arbitrary for the Euclidian space, this difference is unimportant

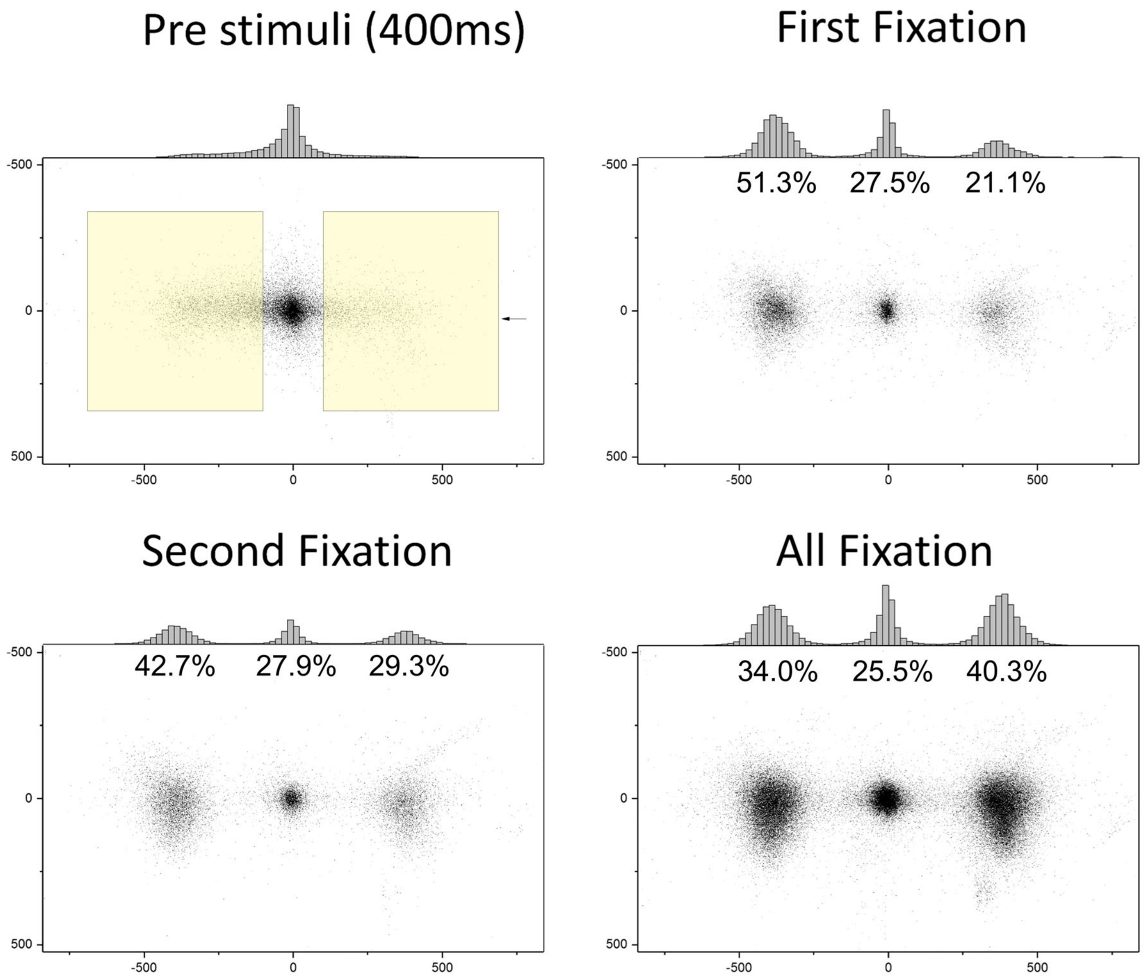

**Fig. 4** Visualization of gaze data. Gaze distribution of the eye on the 2D screen across all subjects, collapsed across all trials. Dots represent fixation locations of the eye. The upper panes showed the histograms of fixation locations along the horizontal axis. Please note that the histograms are not to the same scale **a)** Pre-trial fixation pattern. The two rectangles represent the area of interest (AOI) used for the ensuing cluster validation analysis. The arrow indicates the cut-off y-axis coordinate used to define upper face and lower face. **b)** Locations of the first fixation. There is a strong tendency to fixate the left face (51.3% of all fixations). **c)** Participants still tend to fixate the left face for the second fixation, but the tendency declines (now only 42.7%). **d)** When pooling all data together, the ratio of left face fixations to right face fixations is 34%: 40%

The gaze distribution patterns (Fig. 4) provided useful information about how participants complete the task. We have expected that the participant will fixate on the emoji initially and then vacillate back and forth between the two face images until a decision is reached. However, eye-tracking data showed that participants tend to fixate the left face image first (about 51% of the first fixations and 43% of the second fixations were directed to the left face image), then they check the right face image. Overall, the participants fixated slightly more frequently on the right image (left : right = 34% : 40%), although behaviorally they didn't prefer the right image to the left image (participants choose the right image in 51% of the trials). The percentage of fixations directly on the emoji is relatively constant across time (about 25% of all fixations), but the duration of fixations (231 ms) was much longer than two real faces (averaged 180 ms). Accordingly, for eye-tracking analysis, the area of interest (AOI) was defined as the two real face regions (see Fig. 4a), and all fixations that were directed outside the AOI were excluded, including those directed toward the emoji region. The gaze distribution also indirectly suggested that the participant relied on perceptual information to complete our task. If they relied on conceptual labels, they should have fixated first on the emoji to generate the emotion label. However, this is not the case: the three images in the screen were actively compared throughout the time course.

To verify whether there were statistically significant differences between the clusters, we ran repeated-measures ANOVAs with the clusters being the independent variable and various eye-tracking measurements, i.e. number of fixation, gaze duration, gaze location, and pupil size, being the dependent variable. Means and standard deviations of all measures used in this study are presented by cluster in Table 1 and Fig. 5. Analysis of the proportion of fixations directed to lower faces (Fig. 5a) yielded a significant main effect of clusters, $F(2) = 43.30$, $p < 0.001$, $\eta_p^2 = 0.419$. The proportion of lower face fixations was defined as the ratio of the number of fixation points directed to the lower face to the total number of fixations, with the lower face being defined as below the middle of the nose (which is roughly the lower two-thirds of the face, y-axis coordinates > 25; see Fig. 4a). Post hoc pairwise comparisons indicated that this ratio was smallest for "blended" expressions and greatest for "mouth" expressions (*ps* < 0.05). The main effect of clusters on pupil size was also significant ($F(2) = 5.9$, $p = 0.007$, $\eta_p^2 = 0.09$, Greenhouse-Geisser corrected values). Pupil size was largest for "mouth" expressions: mouth > whole-face, $p = 0.086$; mouth > blended, $p < 0.001$. For duration of fixations (Fig. 5b), the main effect of clusters is also significant, $F(2) = 12.6$, $p < 0.001$, $\eta_p^2 = 0.174$. Post hoc analysis indicates that participants fixated longer on "mouth" expressions. Analysis of the number of fixations also yields a significant main effect of emotion clusters ($F(2) = 9.5$, $p = 0.001$, $\eta_p^2 = 0.137$), with post hoc pairwise analysis revealing that "blended" expressions receive more fixations than others.

In sum, eye-tracking data gave external validation to the goodness of clusters derived from RT data by showing that different looking/gaze patterns were associated with each cluster. Additionally, the finding that the proportion of fixations directed to lower faces is different across clusters provides additional support to the existence of the low-face/whole-face axis.

## Discussion

The main results can be summarized as follows. We found that in the perceptual perspective, facial expressions are organized predominately along the upper-lower face axis and grouped into three superordinate categories: the "mouth" expressions, which comprise facial expressions that have salient mouth features; the "whole-face" expressions, which are visual displays with diagnosing features in both the mouth and the eye regions; and the "blended" expressions, which are blends of two or more elementary expressions. The specific nature of our findings is discussed below.

### Structure of facial expression in the perception perspective

The idea that facial expressions are organized along the vertical face axis is not new (e.g., Ross, Prodan, & Monnot, 2007), but it is especially intriguing as humans have separate brainstem nuclei and cortical regions that control upper and

**Table 1** Clustering evaluation results. The mean values and standard deviations of each behavioral and eye-tracking measurement were averaged across each of the clusters. Data for individual expressions are reported in Online Supplemental Material S3

| Expression groups | RT (ms) | | Duration of fixations (ms) | No. of fixations (per image) | Pupil size change (%) | Proportion of lower-face fixations (%) |
|---|---|---|---|---|---|---|
| | Same-race | Other-race | | | | |
| Mouth | 1,643 ± 307 | 1,635 ± 317 | 178.6 ± 35.9 | 2.97 ± 1.06 | 6.7 ± 2.7 | 58.6 ± 15 |
| Whole-face | 1,701 ± 328 | 1,694 ± 335 | 181.6 ± 37.3 | 2.88 ± 1.06 | 6.5 ± 2.7 | 56.6 ± 15 |
| Blended | 1,685 ± 315 | 1,679 ± 319 | 182.1 ± 36.6 | 3.08 ± 1.13 | 6.4 ± 2.7 | 53.1 ± 16 |

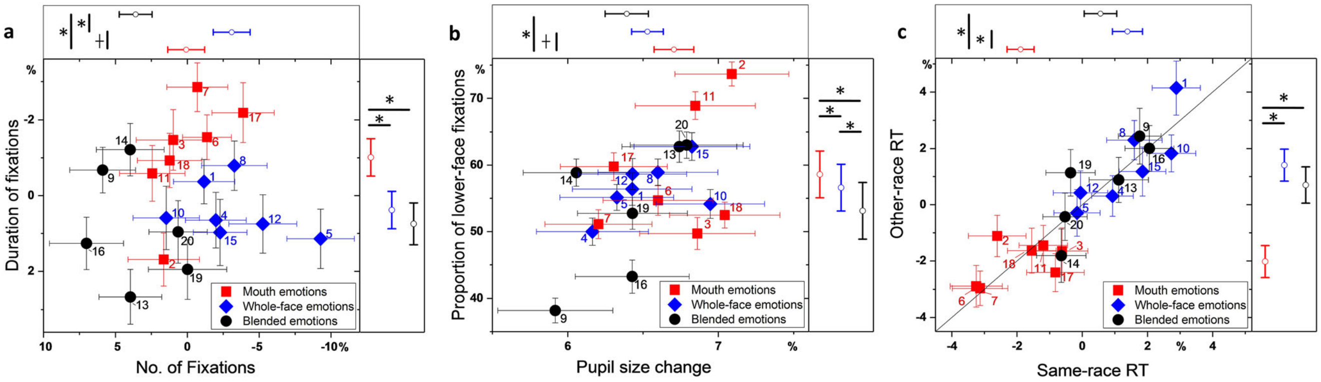

**Fig. 5** Cluster validation. To quantitatively evaluate the effect of clustering, we ran repeated-measures ANOVAs to test whether there are statistically significant differences between the clusters (as illustrated in the upper and right panes). The cluster memberships were derived from same-race data (see Fig. 3b) and are highlighted in colors: Red, "mouth" expressions; Blue, "whole-face" expressions; Black, "blended" expressions. **a)** Main effects of clusters were significant for both pupil size and the proportion of lower-face fixations. **b)** Main effects of clusters were significant for both durations of fixation and the number of fixations. Both variables were converted into percent changes from average, to facilitate the comparison of effect size across variables. **c)** For both same-race and other-race reaction times, the main effects of clusters were also significant. However, since analyses on reaction time were post hoc, they couldn't be used to validate the goodness of clustering results. Error bars depict standard error. * denotes p<0.05; ┼: marginally significant, p<0.1; Labels: 1 = Sorry; 2 = Shy; 3 = Scared; 4 = Sad; 5 = Painful; 6 = Pleasant surprise; 7 = Laugh; 8 = Grin; 9 = Faint; 10 = Despise; 11 = Rage; 12 = Impatient; 13 = Striving; 14 = Amative; 15 = Angry; 16 = Puzzled; 17 = Happy; 18 = Crying; 19 = Bothered; 20 = Asleep

lower face musculature: the upper third of the face receives input from both the ipsilateral and the contralateral cerebral control, whereas the lower two-thirds of the face are controlled contralaterally (Müri, 2016). Different innervation patterns in the upper versus the lower face suggest that the upper-lower face axis may play an important regulatory role in the production of facial expression. Ross et al. (2007) further argued that forebrain control of facial expressions is also more powerfully organized across the upper-lower axis.

We note that this dimensionality is sharply in contrast with many previous studies in which valence was consistently the principal axis (for lists, see Posner et al., 2005, and Shah & Lewis, 2003). However, as mentioned in the *Introduction*, previous studies cannot be accepted as completely removing the artifacts of conceptual knowledge from the structure of emotion. There is compelling evidence that conceptual knowledge, especially emotional concepts, is involved in the formation of the representational structure of emotion (Brooks & Freeman, 2018) and in the recognition of emotional valence (Lindquist, 2013). As an example, semantic dementia patients (who have impaired access to conceptual knowledge) showed deficits in the recognition of valence, while they still are able to recognize several individual emotional expressions (happy, surprise, and, to a lesser degree, fear; Macoir, Hudon, Tremblay, Laforce, & Wilson, 2019). On the constructivist view, emotion recognition depends on the perception of emotional expressions, and on the ability to make sense of such expressions. This latter stage requires conceptual knowledge to allow the emergence of the psychological sense of emotion, such as valence. In our study, we purposely controlled the conceptual involvement and focused on the perceptual stage. It is therefore not surprising that our results showed that facial expressions are not organized along the valence axis. In fact, it is not unexpected that the structure of emotional affect differs from the structure of emotional signals (i.e., the facial expression), as the main purpose of emotional signal, i.e., accurately delivering a variety of information to conspecifics, is different from that of emotional affect (i.e., to help us act with minimal thinking; Tooby & Cosmides, 2008).

Further inspection of the MDS solution showed that its structure is more complicated than a continuous 2D space. A discrete, multicompartment construct may be more appropriate. Cluster analysis identified three superordinate categories: i.e., "mouth," "whole-face," and "blended" expressions. But, just because statistical tests declare our tripartite classification as "adequate," it does not mean that it is meaningful. On the methodological ground, the classification should also be conceptual and biological. From a conceptual perspective, categories should be defined by logical-formal criteria, i.e., within a category the elements share certain common properties that are sufficient to determine whether a single element belongs to that category. From a biological perspective, by claiming that emotions could be partitioned into subsets, it is assumed that there is an innate, biological mechanism that links to each. These are discussed below.

## Conceptualization of the expression subsets

Here, we propose that the tripartite division derived empirically can be conceptualized according to how the facial displays interact with acoustic communications. Emotional communication is achieved predominantly visually, but voices are also natural carriers of emotion, and parallel the face in that they also convey a person's identity and emotional status.

The "mouth" expressions could be conceptualized as vocal emotions. These expressions usually link with species-specific vocal calls. Human (and also great apes') vocal calls are associated with especially urgent functions, for example, to deter predators, warning or attracting partners, keeping in-group contact, mating. These visual-vocal combos are naturally selected (e.g., via sensory adaption; Lee, Mirza, Flanagan, & Anderson, 2014; Susskind et al., 2008) to express a particular affective state, implying transmission of a corresponding illocutionary force along with predictable perlocutionary reaction from receivers.

The "whole-face" expressions could be conceptualized as prosodic emotion. They are visual displays that do not link to specific vocalization. They still closely associate with acoustic communications, but as paralinguistic cues, i.e., affective prosody.

In contrast to the above two subsets, in which the changes in affect correspond to predictable changes both visually and vocally, "blended" expressions are pure-visual emotions. They likely form by mixing two or more expressions, are typically more complex and nuanced, and usually involve greater theory-of-mind decision making. Their meanings might be more adequately understood from a linguistic perspective.

## Multiple facial expression subsets as an evolutionary adaptation

Another important issue to address is whether these facial expression subsets are pure conceptual or are biological realities. In this section, we propose that distinct expression subsets could be interpreted as successive phases in an evolutionary chain.

"Mouth" expressions are likely related to our primate heritage. For instance, chimpanzee (which offer some intriguing parallels to our ancestors 4–8 million years ago; Wood & Harrison, 2011) facial expressions include the bared-teeth display, pant-hoot, play face, scream, alert face, pout, and whimper (Parr, Waller, & Heintz, 2008), all accompanied by characteristic vocalization (Parr, Cohen, & de Waal, 2005). In some sense, their expressions are byproducts of their vocal calls: lips act as an articulator in vocalization, so the shape of the lip is physically connected to the vocal sound. Disruption of the acoustic-visual congruence reduces speech intelligibility, known as the McGurk Effect (McGurk & MacDonald, 1976). These vocal-visual combos are thought to be innate and present early in life regardless of hearing or visual status (Valente, Theurel, & Gentaz, 2018). The visual-vocal bonding assures that the information transmitted by the lower face is identical to the information transmitted by the vocal call. Hence, lower-face features are most salient for the visual perception of these emotions. In fact, as mentioned in the *Introduction*, the absence of contrast in the chimpanzee upper-face makes discriminating their upper-face movements extremely challenging. Due to this limitation, the chimpanzee's (and probably our ancestor's) expression inventory likely relies mostly on lower-face movements, as the contrast between white teeth and dark lips allows movements in the lower-face to be more readily detectable.

Evolving a high contrast upper face might be a turning point in our evolutionary course. Interestingly, humans not only evolved a high-contrast upper face but also changed from avoiding direct eye contact into routinely using direct eye contact during social activities (which allows us to make use of the upper-face visual information). Non-human animals typically avoid direct eye contact as it is an especially aggressive and threatening signal. In contrast, eye contact elicits pro-social behavior in humans. In fact, Perea-García and others (2019) speculated that human scleral depigmentation arises from processes of selection against aggression. This key evolutionary step could have two effects: First, enhanced upper face contrasts, i.e., hairless forehead and bright sclera, together with the acquirement of the ability to perceive them, enable us to display more varieties of facial movements using the upper face. This greatly expands our expression inventory. Second, and more importantly, the involvement of the upper face in the production of expression decouples the above-described visual-vocal bonding. Since the upper face is not a part of the vocal tract, the affective states transmitted via the upper face are no longer able to physically map to vocal sounds (and vice versa). This de-bonding allows humans to display an expression independent of what information the acoustic channel is transmitting, and to produce a sound that has no species-universe function or meaning. The resulting functional flexibility, i.e., the ability to decouple vocalizations from accompanying motivational states and using vocalizations in a goal-directed manner, is the *sine qua non* language and develops in humans around 4 months old (Oller et al., 2013). In comparison, although chimpanzees can also use a particular call in varying contexts (Laporte & Zuberbuhler, 2011; Notman & Rendall, 2005) that may suggest variable functions, such variations have never extended to the degree of functional flexibility as in humans.

Notably, at this evolutionary stage only the suprularyngeal portion of the vocal tract, i.e., the lip/mouth, becomes independent from the emotional contents. In the other parts of the vocal tract, i.e., laryngeal activity and respiratory movements, the same set of vocal tract muscles still simultaneously convey the speech and emotional contents: Changes in breathing patterns, loudness fluctuations, and the rhythmic structure represent the most salient acoustic correlates of affective prosody. At some point in time, the emotional expressions became completely independent from acoustic communication. This disassociation allows humans to freely mix expressions to meet growing contextual social needs. Blending cannot occur if facial displays are still linked to vocalization, because the production of voice (and the prosody of voice) can't be blended. Developmental studies indicated a greater frequency of blended display as infants become toddlers (Hyson & Izard, 1985). Chimpanzees do have the ability to blend expressions. Their facial blends, however, are characterized by one full facial expression that, over a brief period of time, turns into a different full expression, both morphologically and acoustically (Parr et al., 2005).

By being able to blend expressions, the emotional content of face display can be systematically and continuously constructed along various dimensions, such as valence and arousal, filling all portions of affective space. Hence, the "blended" expressions are more closely associated with verbal-linguistic operations, and their visual cues are less distinct. For example, on viewing a scowling expression, participants might offer responses like "angry," "sad," "confused," "hungry," or even "wanting to avoid a social interaction" when they are provided with stories about those emotions (Carroll & Russell, 1996). This suggests that unlike "mouth" and "whole face" expressions, "blend" expressions are more likely to be visual cues rather than visual signals (for a discussion of the difference between a visual cue and a visual signal, see Dezecache et al., 2013).

In summary, we proposed that during evolution, our expression inventory was greatly expanded to meet growing contextual social needs, and different subsets of expressions were evolved at different evolutional stages. Our theory is supported in part by neuroscientific evidence. It seems that both vocal and visual components of "mouth" emotions are governed more subcortically, while others are processed cortically. A dual-pathway processing system has been proposed for affective acoustic communication. Studies showed that affective non-speech vocalizations (e.g., scream, cry, shriek) are distinguished from their neutral counterparts as early as 150 ms after sound onset (Sauter & Eimer, 2010), whereas the perception of affective prosody is significantly slower, diverging from that of neutral speech 200 ms after word onset (Paulmann & Kotz, 2008). Notably, non-speech vocalizations are likely being processed subcortically, and the amygdala is particularly responsive to non-verbal emotional vocalizations (see Fruhholz, Trost, & Grandjean, 2014). The processing of affective prosody is more cortical and has been only inconsistently associated with the amygdala (Adolphs, Tranel, & Buchanan, 2005; Bach, Hurlemann, & Dolan, 2013). Congruously, the visual processing of affective facial expression is also comprised of a slower cortical pathway and a faster subcortical pathway (see *Introduction*). The fundamental prioritization of the "mouth" expressions in our brain (i.e., via vantage, reflex-like subcortical processing) is consistent with the idea that primate vocal calls are usually associated with especially urgent functions. In comparison, the cortical emotion processing may have evolved to engage in more deliberated responses and greater theory-of-mind decision makings (see Said, Haxby, & Todorov, 2011).

## The triune model of emotion

To summarize, we turn to a hypothetical model depicted schematically in Fig. 6. Inspired by MacLean's "Triune Brain" model, we describe the structure of facial expressions in terms of three distinct compartments that have been assembled and emerged along the evolutionary pathway. Unlike MacLean's triune brain model, in our model the compartments are not hierarchical, rather, they are better imagined as being in parallel: To deliver more variety of emotional signals, evolution favors building expansions rather than rebuilding from the bottom up. Older expressions have proven their effectiveness for meeting specific social needs, there is no reason for them to disappear, and they function alongside newly emerged ones. Newly evolved expressions are no less basic than older ones, they are just equally fundamental in meeting our social needs.

The triune model might help to reconcile the discrete-dimensional debate in emotion research. One common point of debate in emotion research is whether the basic, irreducible elements of emotion are discrete "basic emotions," or continuous dimensions such as valence, arousal, etc. Our triune model, instead, proposes a fusion between discrete categorical and continuous dimensional models: the "mouth" and "whole-face" emotions are discrete while "blended" emotions are dimensional. This probably accounts for why, although the discrete theory and the dimensional theory are mutually exclusive, data exist to support both.

## The upper-lower axis and ambiguity axis of facial expression

Lastly, although we take a discrete perspective to generate our theoretical premise (i.e., the triune model of emotion), we believe that the dimensional perspective also has something to offer.

The lower-face/upper-face axis is already deeply embedded in the triune model of emotion. What would be potentially interesting is that previous literature also suggested a left-right difference in terms of neural and behavioral patterns. It is well established that emotions are lateralized to the right brain hemisphere (see Lindell, 2018). This cortical asymmetry leads to an expressional asymmetry (the left side of the face is more emotionally expressive, mobilizing earlier and moving more) and also manifests in asymmetries when perceiving emotion. This right lateralization represents an old evolutionary lineage (Lindell, 2013). Nevertheless, it seems that the left-hemiface's superiority is more predominant in the lower face. Losin, Russell, Freeman, Meguerditchian, and Hopkins

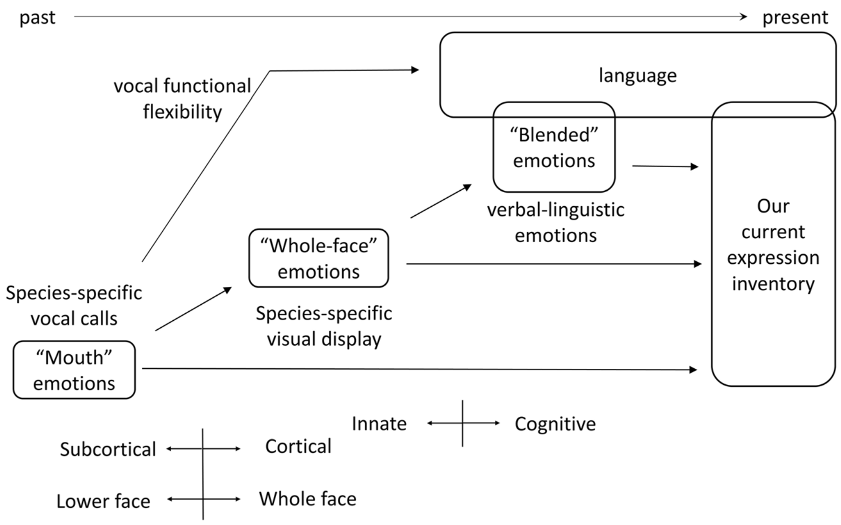
past
present
vocal functional flexibility
language
"Blended" emotions
verbal-linguistic emotions
Our current expression inventory
"Whole-face" emotions
Species-specific visual display
Species-specific vocal calls
"Mouth" emotions
Innate
Cognitive
Subcortical
Cortical
Lower face
Whole face

**Fig. 6** The proposed triune model of emotion

(2008) found that chimpanzees' expressions (pant-hoot, food-bark) were expressed more intensely on the left side of the face – but only in the mouth region and not in eye regions. Asthana and Mandal (1997) reported a similar effect in humans. These observations are consistent with the different neuroanatomic connections for the upper versus the lower face, i.e., a unilateral, contralateral innervation of the lower face and a bilateral innervation of the upper face from both cortical regions and the face nucleus (Müri, 2016). Together, it appears that facial expressions are organized predominantly across the horizontal facial axis and secondarily across the vertical axis.

For the ambiguity axis, we showed that this axis does not reflect the explicit ambiguity rating, instead, it correlates with discordant valence ratings across participants. That is, people recognize the expressions located at the high end of the ambiguity axis effortlessly, but the effect of these expressions varies. This is similar to the ambiguity in language (ambiguity is a central problem in language comprehension; MacDonald, Pearlmutter, & Seidenberg, 1994). According to Piantadosi, Tily, and Gibson (2012), ambiguity allows for greater communicative efficiency, and any efficient communication system will necessarily be ambiguous when context is informative about meaning. From this perspective, the evolutionary constraint (whatever it is) that conserves ambiguity in communication should work on both verbal and non-verbal signals (i.e., facial expression). Therefore, it could be speculated that the same neural-cognitive mechanism underlies the ambiguity in facial expressions and in languages. This postulation is consistent with Neta et al.' (2013) finding that showed that core cortical processes engaged in ambiguity resolution are domain-general. The domain-generality is also predicted by our triune model of emotion, in which we posit that the newly evolved "blended" expressions are closely associated with verbal-linguistic functions. However, it is unclear whether the ambiguity is the result of the varied illocutionary force such expressions can deliver, or, a perhaps more likely alternative is that it reflects individual differences in interpreting the expressions (e.g. Petro, Tong, Henley, & Neta, 2018). Further investigations are clearly needed to clarify this.

## Conclusion

To conclude, the current study used a novel task paradigm with minimal dependence on conceptual thinking to investigate the structure of emotion during the visual perception stage. The most remarkable finding is that facial expressions are organized along the upper-lower face axis and can be clustered into three superordinate categories. We propose a triune model to consolidate these results. The basic underlying assumptions of this triune model might be of potential value in understanding the neural circuits and evolutionary trajectory of the emotional-charged stimuli.

We would like to mention several methodological issues surrounding this preliminary study. First, in this study facial expressions are posed rather than spontaneous. This raises issues regarding validity. Unfortunately, before the experiment, we didn't explicitly verify whether these photos accurately depicted the target emotion. An artificially inflated RT might be observed if the photos are not in agreement with the target emotion: in this case, the participant might just vacillate back and forth with their decision because neither is a good representation of the target. Second, the task requires the participants to compare two expressions. This is not what we usually do in the natural circumstances of emotion recognition, as shown by the long RT and the complex gaze pattern. Lastly, we did not test the "neutral" emotion, which could be a crucial reference in emotion researches. These could serve as research topics for future studies.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13414-021-02281-6.

**Acknowledgements** We sincerely thank our editor and reviewers for their careful reading of the earlier draft of this manuscript and suggesting substantial improvements. We also thank those who participated in this study. This work was sponsored in part by the National Science Foundation of China (NFSC) [grant 31660280] awarded to XZ.

**Open Practices Statements** The raw data have been made public at our lab website http://lab.z-and-z.com/shares/Emotion20_data.tar.gz (sized 60M). The experiment was not preregistered.

## Declarations of interest

None declared.

## References

Adolphs, R., Tranel, D., & Buchanan, T. W. (2005). Amygdala damage impairs emotional memory for gist but not details of complex stimuli. Nat Neurosci, 8(4), 512-518. https://doi.org/10.1038/nn1413

Asthana, H. S., & Mandal, M. K. (1997). Hemiregional variations in facial expression of emotions. Br J Psychol, 88 ( Pt 3), 519-525. https://doi.org/10.1111/j.2044-8295.1997.tb02654.x

Bach, D. R., Hurlemann, R., & Dolan, R. J. (2013). Unimpaired discrimination of fearful prosody after amygdala lesion. Neuropsychologia, 51(11), 2070-2074. https://doi.org/10.1016/j.neuropsychologia.2013.07.005

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition and Emotion, 19(8), 1113-1139. https://doi.org/10.1080/02699930500204250

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature Human Behaviour, 2*(8), 581-591. https://doi.org/10.1038/s41562-018-0376-6

Burrows, A. M. (2008). The facial expression musculature in primates and its evolutionary significance. Bioessays, 30(3), 212-225. https://doi.org/10.1002/bies.20719

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. Vision Research, 41(9), 1179-1208. https://doi.org/10.1016/S0042-6989(01)00002-5

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika, 35(3), 283-319. https://doi.org/10.1007/BF02310791

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. J Pers Soc Psychol, 70(2), 205-218. https://doi.org/10.1037//0022-3514.70.2.205

Celeghin, A., de Gelder, B., & Tamietto, M. (2015). From affective blindsight to emotional consciousness. Consciousness and Cognition, 36, 414-425. https://doi.org/10.1016/j.concog.2015.05.007

Dezecache, G., Mercier, H., & Scott-Phillips, T. C. (2013). An evolutionary approach to emotional communication. Journal of Pragmatics, 59, 221-233. https://doi.org/10.1016/j.pragma.2013.06.007

Du, S., & Martinez, A. M. (2011). The resolution of facial expressions of emotion. Journal of Vision, 11(13), 24-24. https://doi.org/10.1167/11.13.24

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. Proc Natl Acad Sci U S A, 111(15), E1454-1462. https://doi.org/10.1073/pnas.1322355111

Ekman, P. (1992). Are there basic emotions? Psychol Rev, 99(3), 550-553. https://doi.org/10.1037/0033-295x.99.3.550

Fruhholz, S., Trost, W., & Grandjean, D. (2014). The role of the medial temporal limbic system in processing emotions in voice and music. Prog Neurobiol, 123, 1-17. https://doi.org/10.1016/j.pneurobio.2014.09.003

Gainotti G. (2020) The History of Research on Emotional Laterality. In: Emotions and the Right Side of the Brain. Springer, Cham. https://doi.org/10.1007/978-3-030-34090-2_4

Hess, U., Adams, R. B., & Kleck, R. E. (2009). The face is not an empty canvas: how facial expressions interact with facial appearance. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), 3497-3504. https://doi.org/10.1098/rstb.2009.0165

Hout, M. C., Godwin, H. J., Fitzsimmons, G., Robbins, A., Menneer, T., & Goldinger, S. D. (2016). Using multidimensional scaling to quantify similarity in visual search and beyond. Atten Percept Psychophys, 78(1), 3-20. https://doi.org/10.3758/s13414-015-1010-6

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. Wiley Interdiscip Rev Cogn Sci, 4(1), 93-103. https://doi.org/10.1002/wcs.1203

Hyson, M. C., & Izard, C. E. (1985). Continuities and changes in emotion expressions during brief separation at 13 and 18 months. Developmental Psychology, 21(6), 1165-1170. https://doi.org/10.1037/0012-1649.21.6.1165

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. J Hum Evol, 40(5), 419-435. https://doi.org/10.1006/jhev.2001.0468

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. Psychol Bull, 139(4), 917-940. https://doi.org/10.1037/a0030811

Laporte, M. N., & Zuberbuhler, K. (2011). The development of a greeting signal in wild chimpanzees. Dev Sci, 14(5), 1220-1234. https://doi.org/10.1111/j.1467-7687.2011.01069.x

Lee, D. H., Mirza, R., Flanagan, J. G., & Anderson, A. K. (2014). Optical origins of opposing facial expression actions. Psychol Sci, 25(3), 745-752. https://doi.org/10.1177/0956797613514451

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin, 137*(5), 834-855. https://doi.org/10.1037/a0024244

Liebenthal, E., Silbersweig, D. A., & Stern, E. (2016). The Language, Tone and Prosody of Emotions: Neural Substrates and Dynamics of Spoken-Word Emotion Perception. Front Neurosci, 10, 506. https://doi.org/10.3389/fnins.2016.00506

Lindell, A. (2013). Continuities in Emotion Lateralization in Human and Non-Human Primates. Frontiers in Human Neuroscience, 7(464). https://doi.org/10.3389/fnhum.2013.00464

Lindell, A. (2018). Chapter 9 - Lateralization of the expression of facial emotion in humans. In G. S. Forrester, W. D. Hopkins, K. Hudry, & A. Lindell (Eds.), Progress in Brain Research (Vol. 238, pp. 249-270): Elsevier. https://doi.org/10.1016/bs.pbr.2018.06.005

Lindquist, K. A. (2013). Emotions Emerge from More Basic Psychological Ingredients: A Modern Psychological Constructionist Model. Emotion Review, 5(4), 356-368. https://doi.org/10.1177/1754073913489750

Lindquist, K. A., Gendron, M., Barrett, L. F., & Dickerson, B. C. (2014). Emotion perception, but not affect perception, is impaired with semantic memory loss. Emotion, 14(2), 375-387. https://doi.org/10.1037/a0035293

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. Behavioral and Brain Sciences, 35(3), 121-143. https://doi.org/10.1017/S0140525X11000446

Losin, E. A., Russell, J. L., Freeman, H., Meguerditchian, A., & Hopkins, W. D. (2008). Left hemisphere specialization for oro-facial movements of learned vocal signals by captive chimpanzees. *PLoS One, 3*(6), e2529. https://doi.org/10.1371/journal.pone.0002529

Lou, H., Li, S., Jin, W., Fu, R., Lu, D., Pan, X., … Xu, S. (2015). Copy number variations and genetic admixtures in three Xinjiang ethnic minority groups. Eur J Hum Genet, 23(4), 536-542. https://doi.org/10.1038/ejhg.2014.134

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, 101(4), 676-703. https://doi.org/10.1037/0033-295X.101.4.676

Macoir, J., Hudon, C., Tremblay, M. P., Laforce, R. J., & Wilson, M. A. (2019). The contribution of semantic memory to the recognition of basic emotions and emotional valence: Evidence from the semantic variant of primary progressive aphasia. Soc Neurosci, 14(6), 705-716. https://doi.org/10.1080/17470919.2019.1577295

Martinez, A. M. (2017). Visual perception of facial expressions of emotion. Curr Opin Psychol, 17, 27-33. https://doi.org/10.1016/j.copsyc.2017.06.009

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746-748. https://doi.org/10.1038/264746a0

Müri, R. M. (2016). Cortical control of facial expression. Journal of Comparative Neurology, 524(8), 1578-1585. https://doi.org/10.1002/cne.23908

Neta, M., Kelley, W. M., & Whalen, P. J. (2013). Neural responses to ambiguity involve domain-general and domain-specific emotion processing systems. Journal of Cognitive Neuroscience, 25(4), 547-557. https://doi.org/10.1162/jocn_a_00363

Nishimura, M., Maurer, D., & Gao, X. (2009). Exploring children's face-space: a multidimensional scaling analysis of the mental representation of facial identity. J Exp Child Psychol, 103(3), 355-375. https://doi.org/10.1016/j.jecp.2009.02.005

Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2017). Increasing verbal knowledge mediates development of multidimensional emotion representations. Nat Hum Behav, 1, 881-889. https://doi.org/10.1038/s41562-017-0238-7

Notman, H., & Rendall, D. (2005). Contextual variation in chimpanzee pant hoots and its implications for referential communication. Animal Behaviour, 70(1), 177-190. https://doi.org/10.1016/j.anbehav.2004.08.024

Nummenmaa, L., & Calvo, M. G. (2015). Dissociation between recognition and detection advantage for facial expressions: a meta-analysis. Emotion, 15(2), 243-256. https://doi.org/10.1037/emo0000042

Oller, D. K., Buder, E. H., Ramsdell, H. L., Warlaumont, AS., Chorna, L., & Bakeman, R. (2013). Functional flexibility of infant vocalization and the emergence of language. Proc Natl Acad Sci U S A, 110(16), 6318-6323. https://doi.org/10.1073/pnas.1300337110

Parr, L. A., Cohen, M., & de Waal, F. (2005). Influence of social context on the use of blended and graded facial displays in chimpanzees. International Journal of Primatology, 26(1), 73-103. https://doi.org/10.1007/s10764-005-0724-z

Parr, L. A., Waller, B. M., & Heintz, M. (2008). Facial expression categorization by chimpanzees using standardized stimuli. Emotion, 8(2), 216-231. https://doi.org/10.1037/1528-3542.8.2.216

Parr, L. A., Waller, B. M., Vick, S. J., & Bard, K. A. (2007). Classifying chimpanzee facial expressions using muscle action. Emotion, 7(1), 172-181. https://doi.org/10.1037/1528-3542.7.1.172

Paulmann, S., & Kotz, S. A. (2008). Early emotional prosody perception based on different speaker voices. Neuroreport, 19(2), 209-213. https://doi.org/10.1097/WNR.0b013e3282f454db

Perea-García, J. O., Kret, M. E., Monteiro, A., & Hobaiter, C. (2019). Scleral pigmentation leads to conspicuous, not cryptic, eye morphology in chimpanzees. Proceedings of the National Academy of Sciences, 116(39), 19248-19250. https://doi.org/10.1073/pnas.1911410116

Petro, N. M., Tong, T. T., Henley, D. J., & Neta, M. (2018). Individual differences in valence bias: fMRI evidence of the initial negativity hypothesis. Social Cognitive and Affective Neuroscience, 13(7), 687-698. https://doi.org/10.1093/scan/nsy049

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122(3), 280-291. https://doi.org/10.1016/j.cognition.2011.10.004

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol, 17(3), 715-734. https://doi.org/10.1017/S0954579405050340

Preuschoft, S., & van Hooff, J. A. (1995). Homologizing primate facial displays: a critical review of methods. Folia Primatol (Basel), 65(3), 121-137. https://doi.org/10.1159/000156878

Ross, E. D., Prodan, C. I., & Monnot, M. (2007). Human facial expressions are organized functionally across the upper-lower facial axis. Neuroscientist, 13(5), 433-446. https://doi.org/10.1177/1073858407305618

Said, C. P., Haxby, J. V., & Todorov, A. (2011). Brain systems for assessing the affective value of faces. Philos Trans R Soc Lond B Biol Sci, 366(1571), 1660-1670. https://doi.org/10.1098/rstb.2010.0351

Sato, W., & Yoshikawa, S. (2009). Anti-expressions: Artificial control stimuli for the visual properties of emotional facial expressions. Social Behavior and Personality: An International Journal, 37(4), 491-502. https://doi.org/10.2224/sbp.2009.37.4.491

Sato, W., & Yoshikawa, S. (2010). Detection of emotional facial expressions and anti-expressions. Visual Cognition, 18(3), 369-388. https://doi.org/10.1080/13506280902767763

Sauter, D. A., & Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience, 22*(3), 474-481. https://doi.org/10.1162/jocn.2009.21215

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. J Exp Psychol, 44(4), 229-237. https://doi.org/10.1037/h0055778

Schyns, P. G., Petro, L. S., & Smith, M. L. (2009). Transmission of facial expressions of emotion co-evolved with their efficient decoding in the brain: behavioral and brain evidence. PLoS One, 4(5), e5625. https://doi.org/10.1371/journal.pone.0005625

Shah, R., & Lewis, M. (2003). Locating the neutral expression in the facial-emotion space. Visual Cognition, 10(5), 549-566. https://doi.org/10.1080/13506280244000203a

Sokolov, E. N., & Boucsein, W. (2000). A psychophysiological model of emotion space. Integr Physiol Behav Sci, 35(2), 81-119. https://doi.org/10.1007/bf02688770

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W., & Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. Nat Neurosci, 11(7), 843-850. https://doi.org/10.1038/nn.2138

Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In M. Lewis, J. M. Haviland-Jones, & L. Feldman Barrett (Eds.), Handbook of Emotions (3rd ed., pp. 114–137). New York, NY: The Guilford Press.

Valente, D., Theurel, A., & Gentaz, E. (2018). The role of visual experience in the production of emotional facial expressions by blind people: a review. Psychon Bull Rev, 25(2), 483-497. https://doi.org/10.3758/s13423-017-1338-0

Wang, S. (2018). Face size biases emotion judgment through eye movement. Sci Rep, 8(1), 317. https://doi.org/10.1038/s41598-017-18741-9

Wood, B., & Harrison, T. (2011). The evolutionary context of the first hominins. Nature, 470(7334), 347-352. https://doi.org/10.1038/nature09709

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.