Article

Sixteen facial expressions occur in similar contexts worldwide

https://doi.org/10.1038/s41586-020-3037-7

Received: 23 January 2020

Accepted: 30 October 2020

Check for updates

Alan S. Cowen[1,2]✉, Dacher Keltner[1], Florian Schroff[3], Brendan Jou[4], Hartwig Adam[3] & Gautam Prasad[3]

Understanding the degree to which human facial expressions co-vary with specific social contexts across cultures is central to the theory that emotions enable adaptive responses to important challenges and opportunities[1–6]. Concrete evidence linking social context to specific facial expressions is sparse and is largely based on survey-based approaches, which are often constrained by language and small sample sizes[7–13]. Here, by applying machine-learning methods to real-world, dynamic behaviour, we ascertain whether naturalistic social contexts (for example, weddings or sporting competitions) are associated with specific facial expressions[14] across different cultures. In two experiments using deep neural networks, we examined the extent to which 16 types of facial expression occurred systematically in thousands of contexts in 6 million videos from 144 countries. We found that each kind of facial expression had distinct associations with a set of contexts that were 70% preserved across 12 world regions. Consistent with these associations, regions varied in how frequently different facial expressions were produced as a function of which contexts were most salient. Our results reveal fine-grained patterns in human facial expressions that are preserved across the modern world.

Emotions arise from appraisals of the challenges and opportunities that we encounter and are associated with patterns of subjective experience, physiology and expression, which are thought to enable adaptive responses to specific social contexts[1–4]. Emotions are theorized to shape relationships from the first moments of life[5], guide judgment, decision-making and memory[6,15], and contribute to our health[16] and well-being[17].

Central to these wide-ranging theories is the idea that emotions arise in specific contexts and manifest in patterns of behaviour that are preserved across cultures. Efforts to document universality in emotion-related behaviours have centred on the recognition of emotional expressions. In these studies, small samples of participants typically label photographs of posed expressions with 5–6 words[7–10]. Because these methods are sensitive to the language[11,12,18,19], norms[12] and values[18,20] of the participants, inconsistent findings[9,10,13] have supported diverging viewpoints regarding the universality of emotion[3,9,10,21].

A more direct approach to universality is to document whether expressive behaviours—in the present study, patterns of facial movement—occur in similar contexts across cultures[22,23]. Evidence of this kind is surprisingly lacking. It is difficult to capture expressive behaviour in naturalistic contexts that trigger strong emotions. The coding of expressive behaviour—especially facial cues—is time-consuming[24]. Moreover, as expressions and contexts are complex, estimating associations between them requires extensive data[25]. As a result, claims that people form specific expressions in similar contexts across cultures—which are central to affective science—have been sparsely investigated[22,26].

The aim of this study is to examine a multitude of context–expression associations across a comprehensive array of cultures. To do so, we used deep neural networks (DNNs) to classify facial expressions and contexts in 6 million naturalistic videos from 144 countries. DNNs apply multiple layers of transformation to inputs (in this case, videos) to predict outputs (for example, perceived facial expression). Recent empirical work guided by computational approaches has documented a wide range of expressions that people associate with distinct emotions, including amusement, awe, contentment, interest, pain and triumph[14,27–30]. Studying these richly varying facial movements allowed us to examine fine-grained associations between context and expression, moving beyond the prevalent focus on 5–6 posed expressions[9,10].

Across two experiments, we uncover theoretically coherent and robust associations between facial expression and social context that are preserved in 12 world regions[2,31]. Specific contexts including fireworks, weddings and sporting competitions are reliably and differentially associated with 16 patterns of dynamic facial expression, such as those often labelled awe, contentment and triumph[14] by English speakers, in a similar manner across world regions. In total, 70% of the variance in the context–expression association was found to be preserved in all 12 world regions that we examined. In revealing universals in expressive behaviour throughout the modern world, our findings directly inform the origins, functions and universality of emotion.

Facial expression and context annotation

To measure facial expressions, we used a DNN that processes videos and annotates 16 patterns of facial movement that, in isolation, tend to be associated with distinct English-language emotion categories[14,32] (Fig. 1a and Extended Data Fig. 1). These 16 categories are not exhaustive

[1]Department of Psychology, University of California Berkeley, Berkeley, CA, USA. [2]Google Research, Mountain View, CA, USA. [3]Google Research, Venice, CA, USA. [4]Google Research, New York, NY, USA. ✉e-mail: alan.cowen@berkeley.edu

of those used to label facial expressions in English, let alone other languages[11,12,14,18]. Some cultures lack perfect translations for these terms[11,12,19] (but do not necessarily lack the facial expressions to which they are applied[12,33]). They are, however, preserved in emotion recognition across several cultures[27,30,32] and account for appraisals such as valence, arousal and avoidance[14,30]. The DNN was trained on human annotations and relies only on pixels from the face[34] (see Methods, 'Facial expression annotation'). As a result, it cannot account for contextual cues and cultural norms that shape judgments of expressive behaviour[7,12,35] nor directly support inferences about the underlying subjective experiences (which are not well understood)[9,10]. Instead, the emotion categories used to refer to the outputs of the DNN should be considered a shorthand for patterns of facial movement that are often labelled with these categories (Fig. 1). The outputs of the DNN are largely invariant to facial demographics, viewpoint and lighting (Extended Data Figs. 1, 2). To protect privacy, annotations were not linked to the identity of any individual within the publicly available videos that we analysed—no facial identification software was used. Instead, annotations were averaged over each video and analysed across thousands of videos at a time.

To annotate contexts in videos, we used separate algorithms in two experiments. In experiment 1, we integrated DNNs that process video content (pixel values) and metadata (titles or descriptions) to classify 3 million videos in terms of 653 contexts, including many with theoretical relevance to emotion (for example, wedding, practical joke or protest). In experiment 2, to rule out confounding visual features, a DNN that relied exclusively on user-generated titles and descriptions was used to classify another 3 million videos in terms of 1,953 contexts (see Methods, 'Context annotation' for details; and Supplementary Figs. 1, 2 for all contexts).

Regional context–expression associations

To aggregate videos for comparison across cultural groups, we divided the 144 countries from which they originated into 12 world regions (Fig. 1b), integrating countries with ethnolinguistic overlap[36–38] into regions with ample videos to estimate context–expression correlations (comprising more than 60,000 videos) (Extended Data Table 2).

We investigated whether different patterns of facial movement occur systematically in different contexts. To characterize associations between facial expressions and numerous contexts, we computed partial correlations between each context annotation and the 16 facial expression annotations across videos from each region. This analysis provided easily interpretable metrics of context–expression associations that can be compared across regions.

In experiment 1, we computed partial correlations between each of the 653 video-based context annotations and the 16 patterns of facial expression in each world region. Many contexts had associations with specific facial expressions that were well-preserved across regions (Fig. 2a). In every region, expressions associated with amusement occurred more often in videos with practical jokes; awe with fireworks; concentration with martial arts; contentment with weddings; doubt with police; pain with weight training; and triumph with sports. These findings are in keeping with theories proposing that facial expressions occur in psychologically relevant contexts[2,4,10,22,31]. Some associations were less intuitive, illustrating that facial expressions can have multiple meanings[39]—for example, expressions of disappointment were associated with music, which probably reflect the sentimental expressions of the performers[40].

In experiment 2, we computed partial correlations between the 1,953 text-based context annotations and the facial expression annotations. Once again, many contexts had associations with specific expressions that were well-preserved across regions (Fig. 2b), even though context annotations based only on titles and descriptions were less accurate (60.7% of annotated contexts were found to be present in the videos, rather than just in the titles and/or descriptions, compared with 91.6% in experiment 1) (Extended Data Table 1). The associations were often similar to those captured in experiment 1, despite potential biases in text-based context annotations (for example, videos titled 'marry me

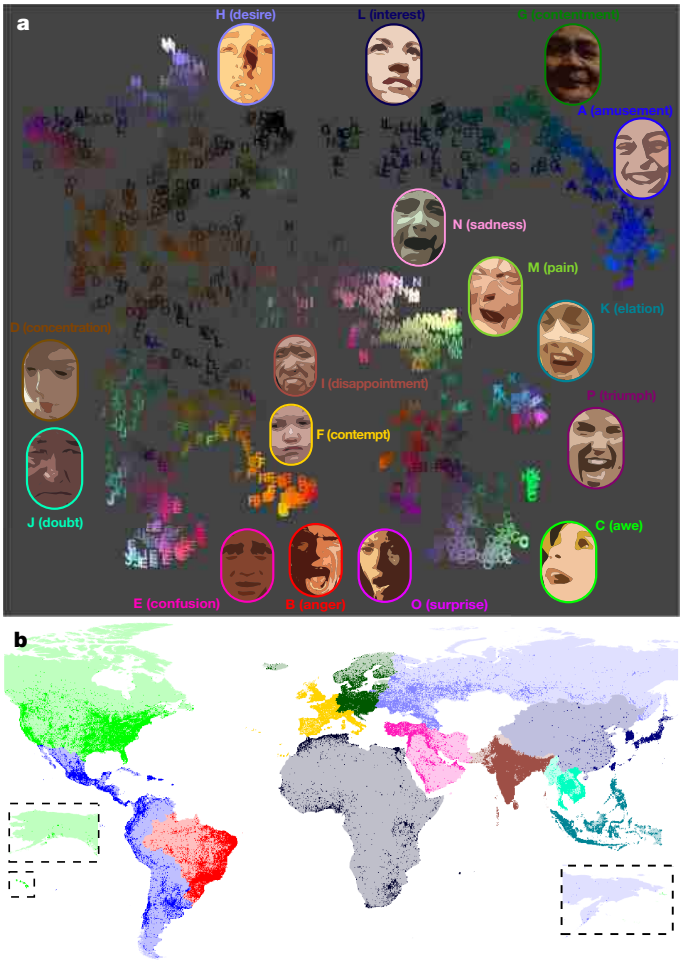


Fig. 1 | Measuring facial expression around the world. a, Facial expression annotations according to the expression DNN. Each of the 16 patterns of facial movement annotated by the expression DNN tends to be associated with a distinct perceived mental state or emotion. These associations also account for broader appraisals attributed to expressions, including valence and arousal[14] (Extended Data Fig. 1d). Here, we present the outputs of the expression DNN applied to 1,456 isolated facial expressions from a previously published study[14]. These outputs have been mapped onto two dimensions using *t*-distributed stochastic neighbour embedding[49], a visualization method that projects similarly annotated data points—here, facial expressions with similar annotations—into similar locations. Each face is assigned a letter that represents its maximal expression DNN score and a colour that represents a weighted average of its two maximal scores, to visualize smooth gradients between expressions. An example of the kind of face that scores highly for each pattern of facial expression is shown (artistically rendered). These examples help to illustrate what is captured by each output of the expression DNN. To delve deeper into what the outputs of the expression DNN represent, explore the annotations of all 1,456 face images in the interactive online map: https://is.gd/PX3u8A. **b**, Division of uploads into 12 world regions. Darker pixels represent 0.25° longitude by 0.25° latitude (<27.9 by 27.9 km) regions in which the videos analysed in experiment 1 were uploaded. Experiment 2 used the same regions but approximate upload coordinates were not available. Some of the artistically rendered faces in **a** are based on photographs originally posted on Flickr by V. Agrawal (https://www.flickr.com/photos/13810514@N07/8790163106), S. Kargaltsev (https://commons.wikimedia.org/wiki/File:Mitt_Jons.jpg), J. Hitchcock (https://www.flickr.com/photos/91281489@N00/90434347/) and J. Smed (https://commons.wikimedia.org/wiki/File:Tobin_Heath_celebration_(42048910344).jpg).

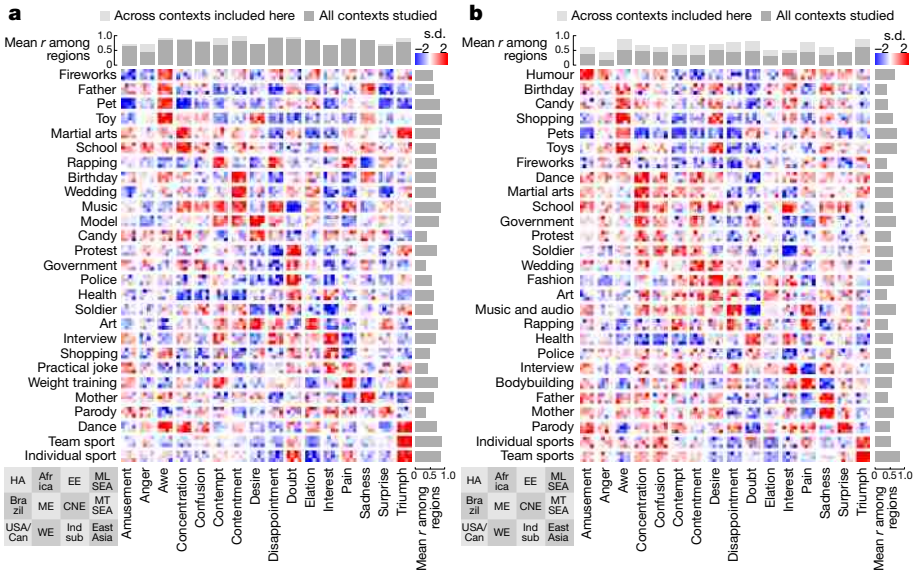


Fig. 2 | Contextual correlates of facial expression in 12 world regions. **a**, Contexts inferred from all video content (experiment 1). Partial correlations computed in experiment 1 between each pattern of facial expression and selected contexts in 12 different world regions. Each rectangle is composed of 12 squares that represent the 12 regions (as indicated in the bottom left). Partial correlations were computed across all videos in each region and controlled for every other pattern of facial expression. Here, correlations are normalized (*z*-scored) within each context (row). For about half (49.6%) of the contexts, including many shown here, we uncovered correlations with specific expressions that were positive in all 12 regions, as represented by the red rectangles (for example, school and 'concentration', candy and 'desire'). Bars at the top of the figure represent mean correlations for each expression across all regions and across the selected contexts (light grey) or all contexts (dark grey). Bars on the right represent mean correlations across all 16 patterns of facial expression and 12 world regions within each context. Contexts for experiment 1 were provided by video topic annotations. See Supplementary Fig. 1 for correlations with every context. **b**, Contexts inferred from only the video title and description (experiment 2). Contexts were inferred from the text rather than visual content. These results therefore rule out confounding factors based on any direct influence of facial expression on context annotation. For about a third (33.5%) of contexts, we uncovered correlations with specific expressions that were positive in all 12 regions (for example, humour and 'amusement', parody and 'surprise'). See Supplementary Fig. 2 for correlations with every context. Anonymized (differentially private[50]) versions of all context–expression correlations are available in the GitHub repository ((https://github.com/google/context-expression-nature-study). CNE, central–northern Europe; EE, eastern Europe; HA, Hispanic America; Ind sub, Indian subcontinent; ME, Middle East; ML SEA, mainland Southeast Asia; MT SEA, maritime Southeast Asia; USA/Can, USA/Canada; WE, western Europe.

Obama' could be misclassified as wedding videos). For instance, in every region, expressions associated with amusement occurred more often in videos with humour; awe with toys; concentration with martial arts; contentment with weddings; pain with bodybuilding; and triumph with sports. These findings, again in keeping with theories proposing that expressions occur in psychologically relevant contexts[2,4,10,22,31], confirm that the results of experiment 1 cannot be explained by artefactual correlations in expression and context annotation.

Cross-regional expression correlations

We next investigated how well associations between context and facial expression are preserved across cultures. The results so far seem to point to robust cultural universals, indicated by the red rectangles in Fig. 2, which represent consistent associations across all 12 regions (for example, martial arts and concentration, police and doubt, team sports and triumph). To characterize these possible universals, we computed second-order correlations between different world regions in context–expression associations. These correlations measure the degree of similarity in associations—of martial arts with concentration, for example—between two regions. In computing between-region correlations, we accounted for the sparsity of some contexts by weighting each context based on its frequency in each region (see Methods, 'Context–expression correlations').

In experiment 1, correlations between world regions in context–expression association (Fig. 3a) ranged from 0.703 (s.e., 0.008) between the Indian subcontinent and East Asia to 0.971 (s.e., 0.005) between USA and Canada (hereafter USA/Canada) and western Europe, with a mean of 0.838. The square root of the minimum correlation between two regions (0.703) approximates their correlation with universal associations from which they separately diverge[41]. This yields a correlation of 0.84, which indicates that all regions share 70% (s.e., 0.8%) of their variance with universal context–expression associations found across the 144 countries represented in this study.

As another approach to estimating universality in context–expression associations, we computed the shared variance between context–expression correlations in each region and the average from the remaining regions (Fig. 3b). A minimum of 70.1% (s.e., 0.6%) of the variance was shared between an individual region (the Indian subcontinent) and the world average, closely corroborating our findings based on pairwise correlations. On average, 82.5% was shared.

The variance that each region shared with the region with which it was most similar was only slightly higher, on average, than the variance each region shared with the world average (84.8% compared with 82.5%), generally reflecting geographical proximity (Fig. 3b). For example, Africa was most similar in its context–expression associations to the Middle East (r^2 = 79.8%, s.e., 1.6%) and the Middle East to the Indian subcontinent (85.0%; s.e., 0.4%). However, Africa and the Indian subcontinent were less similar to each other (67.3%; s.e., 2.1%) than to the world average (82.1% and 70.1%; s.e., 1.3% and 0.8%, respectively). These results suggest that cultural geography and broader universals both influence facial expression.

In experiment 2, correlations between world regions were uniformly positive (Fig. 3a), replicating our findings from experiment 1. However, with an average of 0.596, they were lower in magnitude than in experiment 1 (by about 40%), a difference that could be attributed to the reliance on language-based descriptions to predict contexts. These results suggest that differences in language and norms[11,12,18]

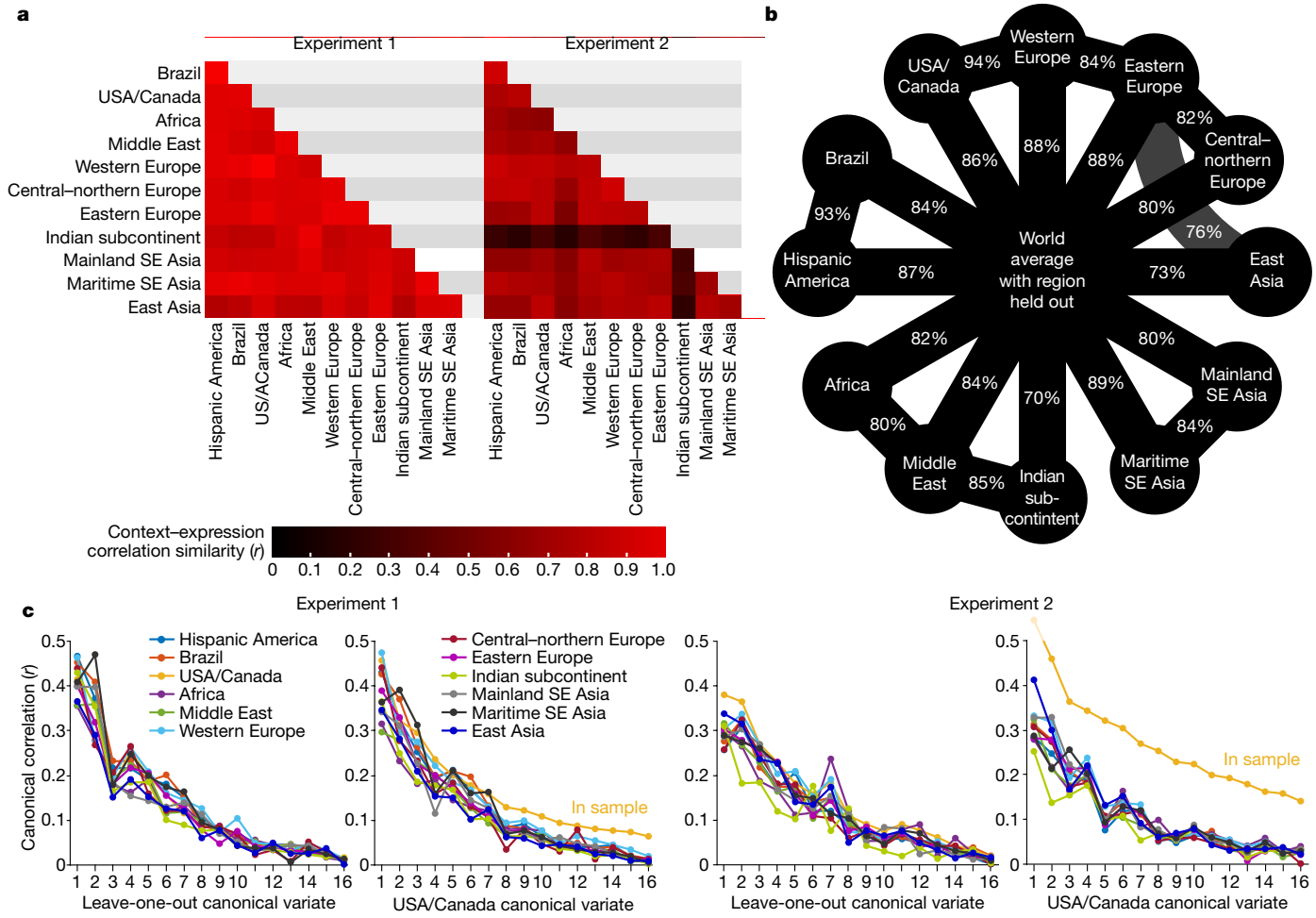


Fig. 3 | Preserved context–expression associations across world regions. **a**, Pairwise correlations between regions. Context–expression associations were well-preserved in experiment 1, with pairwise correlations of at least 0.703 (s.e., 0.008), suggesting that at least 70% of the variance in every region is shared with a universal set of associations (see 'Cross-regional expression correlations'), and with a mean of 0.838. Context–expression associations were moderately well-preserved in experiment 2 ($r \geq 0.159$ (s.e., 0.009) for the Indian subcontinent, an outlier; $r \geq 0.490$ (s.e., 0.014) for other regions; mean, 0.596). These results suggest that context–expression associations are largely universal, at least in videos uploaded online, and differences in language (and context salience)[11,12,18,51] reduce estimates of universality. **b**, Shared variance in context–expression associations between each region and the rest of the world. As another approach to gauging universality, we computed the shared variance (r^2) between context–expression correlations in each region (experiment 1) and the average of the remaining regions. A minimum of 70.1% (s.e., 0.6%) of the variance was shared, corroborating our findings based on pairwise correlations. For comparison, r^2 is also shown between each region and the region with which it shares the most variance. **c**, CCA reveals 16 globally preserved dimensions of context–expression association. In each experiment, CCA was applied in a leave-one-region-out manner across 275,000 videos from the 11 held-out regions (25,000 per region) or for the 333,226 videos from the USA/Canada. The expression and context annotations from the remaining regions were projected onto the extracted canonical variates, and out-of-sample correlations were computed. Correlations were positive for all 16 variates in all regions in both experiments, whether the CCA was trained on geographically balanced videos or on USA/Canada videos alone. Therefore, 16 dimensions were required to account for how expressions occurred in similar contexts across the world ($P < 0.001$, one-sided sign-rank test across regions; false-discovery rate-adjusted $q < 0.001$, Benjamini–Hochberg-corrected).

reduce estimates of cultural universality, and may explain why studies that rely on survey-based measures to capture the meaning of expressions sometimes observe greater variability[9,11,12,18].

Globally preserved facial expressions

The results so far suggest that even nuanced facial expressions have unique contextual associations. In particular, note how the red rectangles in Fig. 2 that represent positive context–expression correlations in all 12 regions are often specific to subtle expressions (for example, toys with awe or school with concentration). To assess how many distinct dimensions of context–expression association were preserved across the 12 world regions, we applied canonical correlations analysis (CCA)[42]. CCA finds independent patterns in two sets of features—here, context and expression—that are maximally correlated with each other (ordered by descending correlation). We applied CCA in two ways: (1) in a leave-one-region-out manner, finding the most-correlated patterns in context and expression within eleven regions and seeing if those patterns held in the twelfth region; and (2) by extracting correlated patterns in the USA/Canada and seeing if they held in every other region. In each case, all 16 canonical correlations were preserved in all 12 regions (Fig. 3c) ($P < 0.001$, one-sided sign-rank test across regions; false-discovery rate-adjusted $q < 0.001$, Benjamini–Hochberg-corrected; see Extended Data Fig. 3 for canonical variate loadings). Given that all canonical correlations can be significant only if each facial expression has unique associations, these results confirm that all 16 expressions have distinct meanings that are preserved across the modern world.

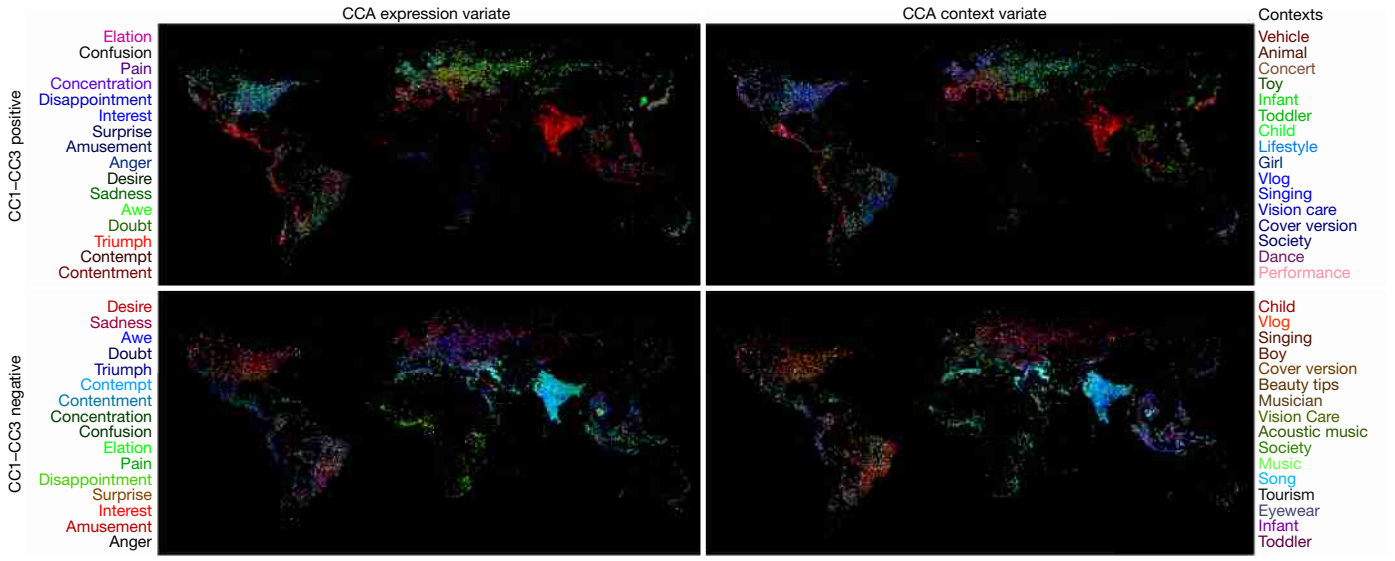


Fig. 4 | Canonical correlations between context and expression are observed at a geographical scale, accounting for differences in the rates at which facial expressions occur in videos from different regions of the world. CCA was applied to expression and context annotations across 300,000 videos balanced across the 12 regions (25,000 videos from each region). Canonical variate scores for each video were then averaged within 0.5° by 0.5° geographical bins (≤55.6 × 55.6 km accounting for approximately 2% of the variance in expression annotations). These averaged scores have been projected as colours onto a flattened globe. Red, first canonical variate; green channel, second canonical variate; blue, third canonical variate. Emotion terms and contextual features adjacent to each map are coloured according to their loadings on the three canonical variates, with positive scores or loadings represented in the top row and negative scores or loadings represented in the bottom row. Broad geographical variations in the facial expressions that occur in videos can be predicted by geographical variations in the contexts that occur in videos. For example, inspection of the positive end of the first canonical variate (top, red) reveals that higher rates of contentment, elation and triumph expressions within the Indian subcontinent are attributable in part to the presence of relatively more videos of concerts and performances in this region. Similarly, inspection of the positive end of the second canonical variate (top, green) reveals that higher rates of awe, doubt and sadness expressions in eastern Europe and South Korea are attributable in part to the presence of relatively more infants and children in videos from these regions. CC, canonical correlation.

Geographical expression–context covariance

We next investigated how facial expression varies across cultures. So far we have identified universal correlations between expression and context. However, because correlations are insensitive to the average rates of each facial expression, it remains possible that there are cultural differences in how frequently different expressions occur[12,43]. A complexity in addressing this possibility is that even if the rates of expressions in online videos differ across world regions, this could be explained by variations in the rates of different contexts.

To address these questions with precision, it is important to trace videos to more fine-grained geographical origins. In experiment 1, the approximate longitude and latitude of upload were available for each video (within several kilometres). To visualize whether geographical variations in average expression were explained by variations in the frequency of different contexts, we applied CCA between the 16 facial expression annotations and 653 contexts across a geographically balanced set of 300,000 videos (25,000 randomly selected per region), then averaged the expression and context scores of videos within 0.5°× 0.5° bins (≤55.6 × 55.6 km).

This analysis revealed geographical variation in the rates of different facial expressions, but these variations closely tracked variations in the contexts found in videos (Fig. 4). This pattern of results is suggestive of how cultural variation observed in survey-based studies of expression may arise partly from differences in the rates of contexts and expressions across cultures and how these contexts and expressions are consequently construed[44].

Discussion

Questions of universality are central to the study of emotion. The most direct approach to understanding universality is to examine whether people in different cultures express emotions similarly in psychologically relevant contexts—weddings, humour, art and sports triumphs. Evidence regarding such associations is surprisingly sparse. Here, drawing on computational approaches to map the meaning of a wide array of emotional expressions, we show that 16 facial expressions are associated with similar contexts in 12 world regions encompassing 144 countries. Nuanced expressions such as awe occur in similar circumstances (fireworks, toys and dance) in every region, pointing to universals in expressive behaviour across the modern world. Overall, 70% of the variation in facial expression across contexts was universal to every world region that we studied (based on the more-accurate context annotations in experiment 1).

Because these findings are based on online videos, they may have been influenced by cultural globalization, particularly through digital platforms. For example, people across the world may have adopted facial expressions from Western media. However, we do not see suggestions of a bias towards similarity with the West. Instead, we see greater similarity between neighbouring regions (Fig. 3b), which is expected if the videos are representative of local cultures as opposed to Western culture. For instance, the Middle East and Indian subcontinent shared 85.0% (s.e., 0.8%) of their variance in context–expression associations, but only 63.9% (s.e., 1.2%) and 53.8% (s.e., 0.9%), respectively, with the USA/Canada. Moreover, if expressions worldwide were strongly influenced by westernization, one would expect the world average to be most strongly correlated with Western regions, but it is most correlated with maritime Southeast Asia (Fig. 3b). Of course, facial expressions may have spread through a more dispersed process of cultural diffusion. Regardless of their origins, we have identified universals throughout the modern world in facial expressions that people produce in diverse contexts. These findings have important implications for methodologies that incorporate measurements of expression, including psychiatric[35] and affective computing[18,35,45]

applications, revealing how they may generalize across modern cultures.

It is worth acknowledging that online videos do not provide an unbiased representation of everyday life. The events that people document will generally have some meaning to them—whether they evoke emotion, commemorate special occasions or convey valued information—and may overrepresent certain emotions (see Extended Data Fig. 4 for a representation of facial expressions in the present study). This may explain less intuitive associations, such as that of mothers with sadness—perhaps in online videos, mothers often accompany crying infants. The factors that motivate people to document events probably influenced the context–expression associations that we uncovered, and their cultural variation may have attenuated estimates of universality.

In training a DNN to classify facial expressions, we used English categories for which not all languages have direct translations[11,12,14,18,19]. Nevertheless, the facial expressions associated with these concepts had psychologically relevant contextual correlates that were preserved across disparate ethnolinguistic regions (for example, Africa and East Asia). To reconcile these findings, one must distinguish the facial expressions that people produce[14] from the categories with which they are parsed[1,7,12,35]. Just as not all languages have a word for purple[46], even though purple hues of light stimulate similar retinal cones in all human populations[47], a universal space of expression–context associations can be partitioned differently by different languages[12]. Indeed, the reliance on language to predict contexts in experiment 2 reduced measurements of universality.

We did not find evidence of demographic biases in the facial expression annotations themselves, after examining annotations in four racial and ethnic and three regional groups (Extended Data Figs. 1, 5). However, it is worth acknowledging that any unexamined demographic biases in expression annotation could potentially have mitigated the degree of universality that we observed, causing identical expressions that appear within the same contexts to be classified slightly differently across regions.

Our findings are consistent with observations that facial expressions can have multiple meanings[39]. For example, pain expressions occurred not only in congruent contexts (for example, weightlifting) but also in contexts related to music, consistent with work documenting the sentimental expressions of musical performers[40]. Research incorporating vocal and bodily expression could potentially disambiguate facial movements, capturing how they can have multiple yet coherent meanings[10,39,48].

Although the contextual correlates of facial expressions were largely consistent around the world, our findings hinted at cultural differences in the rates of expressions (Fig. 4). In keeping with a central claim in cultural approaches to emotion[43,44], differences in the prevalence of facial expressions were explained in part by variations in the rates of different contexts. Future research could examine how variations in facial expression may also be associated with personality, health and cultural dimensions (for example, collectivism).

Nearly 150 years ago, as a central line of justification for his theory of evolution by natural selection, Charles Darwin made the controversial argument that human facial expression is a universal language of social life. This proposal has generated hundreds of empirical studies and considerable debate. Here, using previously undescribed methods and quantitative approaches, we find that 16 kinds of facial expressions systematically co-vary with specific social contexts in 144 countries. Across diverse geographical regions, there is 70% overlap in the associations between context and facial expression—evidence for substantial universality in our cyber-connected world.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-3037-7.

1. Cowen, A. S. & Keltner, D. Clarifying the conceptualization, dimensionality, and structure of emotion: response to Barrett and colleagues. *Trends Cogn. Sci*. **22**, 274–276 (2018).
2. Moors, A., Ellsworth, P. C., Scherer, K. R. & Frijda, N. H. Appraisal theories of emotion: state of the art and future development. *Emot. Rev*. **5**, 119–124 (2013).
3. Ekman, P. & Cordaro, D. What is meant by calling emotions basic. *Emot. Rev*. **3**, 364–370 (2011).
4. Keltner, D. & Haidt, J. Social functions of emotions at four levels of analysis. *Cogn. Emot*. **13**, 505–521 (1999).
5. Niedenthal, P. M. & Ric, F. in *Psychology of Emotion* 98–123 (Routledge, 2017).
6. Keltner, D., Kogan, A., Piff, P. K. & Saturn, S. R. The sociocultural appraisals, values, and emotions (SAVE) framework of prosociality: core processes from gene to meme. *Annu. Rev. Psychol*. **65**, 425–460 (2014).
7. Elfenbein, H. A. & Ambady, N. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychol. Bull*. **128**, 203–235 (2002).
8. Ekman, P. in *The Nature of Emotion* (eds Ekman, P. & Davidson, R.) 15–19 (Oxford Univ. Press, 1994).
9. Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **20**, 1–68 (2019).
10. Cowen, A., Sauter, D., Tracy, J. L. & Keltner, D. Mapping the passions: toward a high-dimensional taxonomy of emotional experience and expression. *Psychol. Sci. Public Interest* **20**, 69–90 (2019).
11. Kollareth, D. & Russell, J. A. The English word disgust has no exact translation in Hindi or Malayalam. *Cogn. Emot*. **31**, 1169–1180 (2017).
12. Mesquita, B. & Frijda, N. H. Cultural variations in emotions: a review. *Psychol. Bull*. **112**, 179–204 (1992).
13. Russell, J. A. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol. Bull*. **115**, 102–141 (1994).
14. Cowen, A. S. & Keltner, D. What the face displays: mapping 28 emotions conveyed by naturalistic expression. *Am. Psychol*. **75**, 349–364 (2020).
15. Holland, A. C. & Kensinger, E. A. Emotion and autobiographical memory. *Phys. Life Rev*. **7**, 88–131 (2010).
16. Kok, B. E. et al. How positive emotions build physical health: perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychol. Sci*. **24**, 1123–1132 (2013).
17. Diener, E., Napa Scollon, C. & Lucas, R. E. The evolving concept of subjective well-being: the multifaceted nature of happiness. *Adv. Cell Aging Gerontol*. **15**, 187–219 (2003).
18. Jou, B. et al. Visual affect around the world: a large-scale multilingual visual sentiment ontology. In *MM'15: Proc. 23rd ACM International Conference on Multimedia* 159–168 (2015).
19. Jackson, J. C. et al. Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019).
20. Tsai, J. L., Knutson, B. & Fung, H. H. Cultural variation in affect valuation. *J. Pers. Soc. Psychol*. **90**, 288–307 (2006).
21. Russell, J. A. Core affect and the psychological construction of emotion. *Psychol. Rev*. **110**, 145–172 (2003).
22. Tracy, J. L. & Matsumoto, D. The spontaneous expression of pride and shame: evidence for biologically innate nonverbal displays. *Proc. Natl Acad. Sci. USA* **105**, 11655–11660 (2008).
23. Martin, R. A. Laughter: a scientific investigation (review). *Perspect. Biol. Med*. **46**, 145–148 (2003).
24. Cohn, J. F., Ambadar, Z. & Ekman, P. in *The Handbook of Emotion Elicitation and Assessment* (eds Coan, J. A. & Allen, J. J. B.) 203–221 (Oxford Univ. Press, 2007).
25. Gatsonis, C. & Sampson, A. R. Multiple correlation: exact power and sample size calculations. *Psychol. Bull*. **106**, 516–524 (1989).
26. Anderson, C. L., Monroy, M. & Keltner, D. Emotion in the wilds of nature: the coherence and contagion of fear during threatening group-based outdoors experiences. *Emotion* **18**, 355–368 (2018).
27. Cordaro, D. T. et al. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion* **18**, 75–93 (2018).
28. Keltner, D. & Cordaro, D. T. in *Oxford Series in Social Cognition and Social Neuroscience. The Science of Facial Expression* (eds Fernández-Dols, J.-M. & Russell, J. A.) 57–75 (Oxford Univ. Press, 2015).
29. Cowen, A. S., Elfenbein, H. A., Laukka, P. & Keltner, D. Mapping 24 emotions conveyed by brief human vocalization. *Am. Psychol*. **74**, 698–712 (2019).
30. Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R. & Keltner, D. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nat. Hum. Behav*. **3**, 369–382 (2019).
31. Keltner, D. & Lerner, J. S. in *Handbook of Social Psychology* (eds Fiske, S. T et al.) (Wiley Online Library, 2010).
32. Cordaro, D. T. et al. The recognition of 18 facial-bodily expressions across nine cultures. *Emotion* **20**, 1292–1300 (2020).
33. Sauter, D. A., LeGuen, O. & Haun, D. B. M. Categorical perception of emotional facial expressions does not require lexical categories. *Emotion* **11**, 1479–1483 (2011).
34. Schroff, F., Kalenichenko, D. & Philbin, J. FaceNet: a unified embedding for face recognition and clustering. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 815–823 (2015).
35. Rudovic, O. et al. CultureNet: a deep learning approach for engagement intensity estimation from face images of children with autism. In *Proc. IEEE International Conference on Intelligent Robots and Systems* 339–346 (2018).

36. Gupta, V., Hanges, P. J. & Dorfman, P. Cultural clusters: methodology and findings. *J. World Bus.* **37**, 11–15 (2002).
37. Rosenberg, N. A. et al. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
38. Eberhard, D. M., Simons, G. F. & Fennig, C. D. (eds) *Ethnologue: Languages of the World* 23rd edn (SIL International, 2020).
39. Aviezer, H., Trope, Y. & Todorov, A. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338**, 1225–1229 (2012).
40. Davidson, J. W. Bodily movement and facial actions in expressive musical performance by solo and duo instrumentalists: two distinctive case studies. *Psychol. Music* **40**, 595–633 (2012).
41. Lord, F. M. & Novick, M. R. in *Statistical Theories of Mental Test Scores* Ch. 2 358–393 (Addison-Wesley, 1968).
42. Hardoon, D. R., Szedmak, S. & Shawe-Taylor, J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**, 2639–2664 (2004).
43. Scherer, K. R., Wallbott, H. G., Matsumoto, D. & Kudoh, T. in *Facets of Emotion: Recent Research* (ed. Scherer, K. R.) 5–30 (Erlbaum, 1988).
44. Markus, H. R. & Kitayama, S. Culture and the self: implications for cognition, emotion, and motivation. *Psychol. Rev.* **98**, 224–253 (1991).
45. Picard, R. W. *Affective Computing* (MIT Press, 1997).
46. Roberson, D., Davies, I. & Davidoff, J. Color categories are not universal: replications and new evidence from a stone-age culture. *J. Exp. Psychol. Gen.* **129**, 369–398 (2000).
47. Gegenfurtner, K. R. & Kiper, D. C. Color vision. *Annu. Rev. Neurosci.* **26**, 181–206 (2003).
48. Emily, M., Sungbok, L., Maja, J. M. & Shrikanth, N. Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions. In *Proc. 2008 IEEE International Conference on Multimedia and Expo* 961–964 (ICME, 2008).
49. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2625 (2008).
50. Dwork, C. in *Theory and Applications of Models of Computation* (eds Agrawal, M. et al.) 1–19 (Springer, 2008).
51. Pappas, N. et al. Multilingual visual sentiment concept matching. In *Proc. 2016 ACM International Conference on Multimedia Retrieval* 151–158 (ICMR, 2016).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Facial expression annotation

To annotate facial expressions, we used a DNN. A DNN is a machine-learning model that is trained to apply multiple layers of transformation to input data—in this case, videos—to predict complex outputs; in this case, judgments of facial expressions. We used a supervised DNN that processed the temporal trajectory of the RGB pixel values that make up a video of a face over the course of up to 1 s, at a rate of 18 frames s^{-1}, to predict the proportions of raters who would attribute each label to each expression. To extract faces from the video, each face was detected within a video frame using a deep convolutional neural network, similar to the method provided by the Google Cloud Face Detection API. Faces were tracked across each video using traditional optical flow methods. The pixels from each face were then read by our facial expression DNN.

Facial expression DNN architecture. Face-based visual features were extracted using layers from the NN2 FaceNet architecture[34]. These layers consisted of an inception (5a) block with a 7 × 7 feature map comprising 1,024 channels, which was fed into a 7 × 7 average pooling layer, generating a 1,024 dimensional vector representing face image features within a given frame of the video. The resulting features were then fed into two long short-term memory layers, each with 64 recurrent cells, to capture the dependence of facial expression recognition on temporally unfolding patterns of facial movement. Finally, the output of the last long short-term memory layer was fed into a mixture of experts model (two experts, plus a dummy expert). A cross entropy loss with a sigmoid activation function was used for the final layer.

Training. The DNN was trained on a total of 273,599 ratings of 186,744 clips of faces on YouTube, and originally evaluated on another 79,597 ratings of 54,406 clips (Extended Data Fig. 5). Raters were English speakers in India. Raters were presented with 1–3-s video clips with bounding boxes placed over individual faces. They were asked to rate the emotion(s) expressed by the face by selecting all that applied from a list of 31 labels (28 emotion categories drawn from a previous study[14], plus boredom, neutral and unsure). Raters were familiarized with the 31 labels in advance. They viewed each video on loop and could pause the videos when needed.

The videos from which the clips were extracted were largely collected manually by raters before any ratings were collected. During this initial video collection process, raters had been instructed to conduct a broad search for videos likely to contain emotional expressions.

We did not see evidence of regional or ethnic biases in DNN prediction performance. Human rating prediction correlations consistently ranged between 0.63 and 0.68 for faces of different races and ethnicities (Asian, Black, Hispanic/Latino and white), with a less than 1% bias by race or ethnicity in any expression dimension (Extended Data Figs. 1, 5). However, it is worth acknowledging that any such biases could potentially have mitigated the degree of universality that we observed in facial expression.

Label selection. From the original 29 emotion labels used to train the DNN, 13 were excluded from the present study due to a combination of modest prediction accuracy (particularly boredom, disgust, embarrassment, pride, realization, relief and sympathy) and high correlation with a better-predicted label (particularly contemplation with concentration, confusion with doubt, distress with pain, and fear with surprise), and/or because uninteresting aspects of facial posture appeared to affect annotations (love annotations were affected by kissing and ecstasy by closed eyes).

Context annotation

To predict features of the contexts portrayed in each video, we relied on proprietary machine-learning models. These models rely on well-known innovations in natural language processing[52] and video content analysis[53]. The model used to annotate topics in experiment 1 used a mixture of experts approach[54] to integrate the outputs of a text DNN applied to the user-generated titles and descriptions of each video with a separate video content analysis DNN applied to the pixels of each video. The model used to annotate topics in experiment 2 comprised the outputs of the text DNN alone. The accuracy of each model was verified by reviewing their outputs for a random sample of videos in experiments 1 and 2 (Extended Data Table 1). Obtained accuracy levels were consistent with the state-of-the-art in both text classification and video content analysis[52,53].

Experiment 1. The model used in experiment 1 integrated the output of DNNs that label text[52] (titles and descriptions provided by video uploaders) and video content (pixel values, similar to tools provided in the Google Cloud Video Intelligence API)[53] to accurately classify thousands of contexts in videos (mean specificity = 91.6%) (Extended Data Table 1). For our purposes, we focused on contexts that occurred at least 30 times across the videos included in our experiment (see 'Video selection' for inclusion criteria), for a total of 653 contexts ranging from practical joke to weight training. The contexts occurred a minimum of 39 times (no contexts occurred 30–38 times) with the vast majority occurring, on average, 11,849 times across the 3,029,812 videos in experiment 1 (see Extended Data Fig. 6 for histogram of occurrence rates). The contexts are intentionally broad. For example, a video labelled 'mother' could include or pertain to mothers in any number of ways, ranging from footage of actual parenting to a man discussing his mother.

A potential concern when analysing these annotations is that, because they rely on the visual content (in pixels) of videos, they could be influenced by visual information from facial expressions. For instance, the classification of a context as a surprise party may be influenced by the presence of facial expressions of surprise. If this occurred, it could result in artefactual correlations with the expression annotations. To address this concern, we measured the extent to which the video topic annotations were in fact influenced by facial expressions using a simulation in which facial expressions were systematically manipulated to be present or absent in controlled contexts (Extended Data Fig. 7). The effect of facial expressions on the video topic annotations was very small ($r^2 < 0.0001$) compared to the variance in video topic annotations explained by facial expressions in our real dataset ($r^2 > 0.02$)—differing by more than two orders of magnitude—mitigating the possibility that artefactual correlations influenced our results. We do not expect the raw r^2 to be high, given that we do not expect a one-to-one mapping between any contextual feature and expression. For instance, a dog can evoke adoration or fear. We examine the r^2 here to characterize the relative magnitude of a potential confounding effect compared to the effects of interest. Nevertheless, in our second experiment, we fully ruled out the influence of such artefactual correlations by deriving context annotations from text alone.

Experiment 2. We derived context annotations from text by applying a natural language processing DNN that predicts topics of text[52] (similar to tools provided in the Google Cloud Natural Language API) in video titles and descriptions, which are provided by video uploaders in their own words. For our analysis, we selected topics that occurred at least 30 times in titles and descriptions of the videos included in our experiment ('Video selection') for further analysis, which resulted in a total of 1,953 contexts that occurred a minimum of 176 times (no

contexts occurred between 30 and 175 times). The average topic occurred 18,570 times across the 3,056,861 videos analysed in experiment 2. A histogram of rates of occurrence of contexts is provided in Extended Data Fig. 6. Because the text topic DNN does not directly analyse video or audio, its context annotations are less accurate in predicting the content of the video than the video topic annotations (mean specificity = 60.7%) (Extended Data Table 1), even though they are highly accurate in predicting whether a phrase relevant to a given topic occurs in the title or description (mean specificity = 92.9%) (Extended Data Table 1).

Video selection

Across two experiments, we selected 6.1 million publicly available videos for analysis using our automated facial expression and context annotation systems. The use of the video data in aggregate form underwent review for alignment with Google's AI Principles (https://ai.google/principles/) and conformed to Google's privacy policy (https://policies.google.com/privacy).

Experiment 1. For our first experiment, we focused on natural footage for which reliable geographical information was available. To find naturalistic footage, we searched for publicly available YouTube videos that were uploaded from mobile phones. YouTube is an ideal source of ecologically valid footage, as it has more than 2 billion monthly active users, is available in more than 80 languages and is used around the world as a social media platform on which most videos are viewed by users within a social network. We restricted our search to videos tagged with a latitude and longitude of upload that matched the country in which the uploader was registered. Furthermore, to focus on naturalistic footage, we filtered out videos predicted by the video topic annotations to include video games and other animated content. This yielded a total of 3,029,812 videos uploaded between 14 July 2009 and 3 May 2018.

To verify that the videos portrayed their culture of origin, we ensured that they largely depicted people who, when filmed speaking, spoke languages widely spoken in the country of origin (see Extended Data Tables 2, 3 for proportions of languages detected via automated methods and manual inspection). On the basis of a manual inspection of a balanced sample of 300 videos from across the 12 regions, we also verified that the videos depicted people whose appearance on human inspection was consistent with their being of native origin, to the limited extent to which geographical origin can be gauged qualitatively on the basis of accent, dress and physical features (96.7% of videos) (Extended Data Table 3). However, this does not rule out the influence of globalization and online media, considered further in the 'Discussion'.

Experiment 2. The videos from experiment 1, as with many videos on YouTube, typically lacked detailed descriptions, making them poor candidates for annotation by the text topic DNN. We therefore collected an entirely new set of videos for experiment 2. To ensure that we would have the power to investigate correlations between contexts and facial expressions, we includes publicly available videos that had titles and descriptions pertaining either to the contexts that we explored in experiment 1 or to emotions. To do so, we first searched for videos with a wide range of context- and emotion-related substrings within their English-translated titles and descriptions (Supplementary Data 1; to the extent that translations were inaccurate, representation of corresponding contexts could be reduced in non-English-speaking cultures, augmenting cultural differences). We then retrieved the full native-language titles and descriptions for those videos and computed text topic annotations. Finally, to avoid synthetic faces, we filtered out videos predicted by the text topic DNN to include video games and animated content. This yielded a total of 3,056,861 videos uploaded between 27 December 2005 and 15 April 2019. Again, when filmed speaking, people in these videos spoke languages widely spoken in the region of origin (Extended Data Table 2), and on the basis of manual inspection of a balanced sample of 300 videos from across the 12 regions, these videos largely depicted people who appeared to be of native origin (88.3% of videos) (Extended Data Table 3).

Context–expression correlations

To capture context–expression associations, we computed partial correlations between the context annotations and the expression annotations across videos. Partial correlations for each expression annotation were controlled for the other expression annotations. To measure the degree to which world regions were similar in their context–expression associations, we computed second-order correlations between different world regions across the context–expression partial correlations (Fig. 3a). To do so, we flattened the matrices that represent the context–expression partial correlations in each region (resulting in vectors of length 643 × 16 in experiment 1 and 1,953 × 16 in experiment 2 for each region) and correlated the resulting vectors between regions. It was important to account for sampling error, particularly for contexts that occurred infrequently (see Extended Data Fig. 6 for a histogram of rates of occurrence of contexts). Correlations were thus weighted on the basis of the frequency of each context in each region. More specifically, they were weighted by $\sqrt{p_1(1-p_1)p_2(1-p_2)}$ in which p_i is the proportion of times that the context occurred in each region, which is the product of the standard deviations of the two proportions and approximates the signal variance available to estimate correlations with expression (for fixed sample size of videos).

As another approach to gauging universality, we computed the shared variance (r^2) between context–expression correlations in each region (experiment 1) and the average context–expression correlations from the remaining regions (Fig. 3b). When we computed the average correlations across regions, we computed the Fisher transformation[55], averaged, then computed the inverse Fisher transformation. When we computed the correlation between each region and the average, contexts were weighted as before, and weights were averaged across regions.

To compute standard errors for correlation or shared variance estimates, we repeated all analyses after bootstrap resampling of the videos. Resampling was performed 50 times[56].

These analyses were performed using custom code in MATLAB. See also Extended Data Table 4 for a summary of how we addressed potential confounding factors in the correlation between expression and context annotations.

CCA

To assess how many distinct dimensions of context–expression association were preserved across the 12 world regions, we applied CCA between expression annotations and context annotations across videos from all but one region and evaluated each canonical correlation in the held-out region. For each extracted dimension, we then applied a one-sided sign-rank test to the 12 canonical correlations from each held-out region. We used a one-sided test because we are interested only in preserved (positive) correlations. These analyses were performed using custom code in MATLAB.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Anonymized (differentially private) versions of the context–expression correlations in each country are available from the GitHub repository (https://github.com/google/context-expression-nature-study). Owing to privacy concerns, video identifiers and annotations cannot be provided.

Code availability

Code to read and visualize the anonymized context–expression correlations within each world region is available from the GitHub repository (https://github.com/google/context-expression-nature-study). Code to generate interactive online maps (Fig. 1a) is available in the GitHub repository (https://github.com/krsna6/interactive-embedding-space)[57]. The trained video-processing algorithms used in this study are proprietary, but similar tools to annotate contexts in video and text, detect faces and classify the language of speech are available via the Google Cloud Video Intelligence API, Natural Language API and Speech-to-Text API, respectively (see https://cloud.google.com/apis).

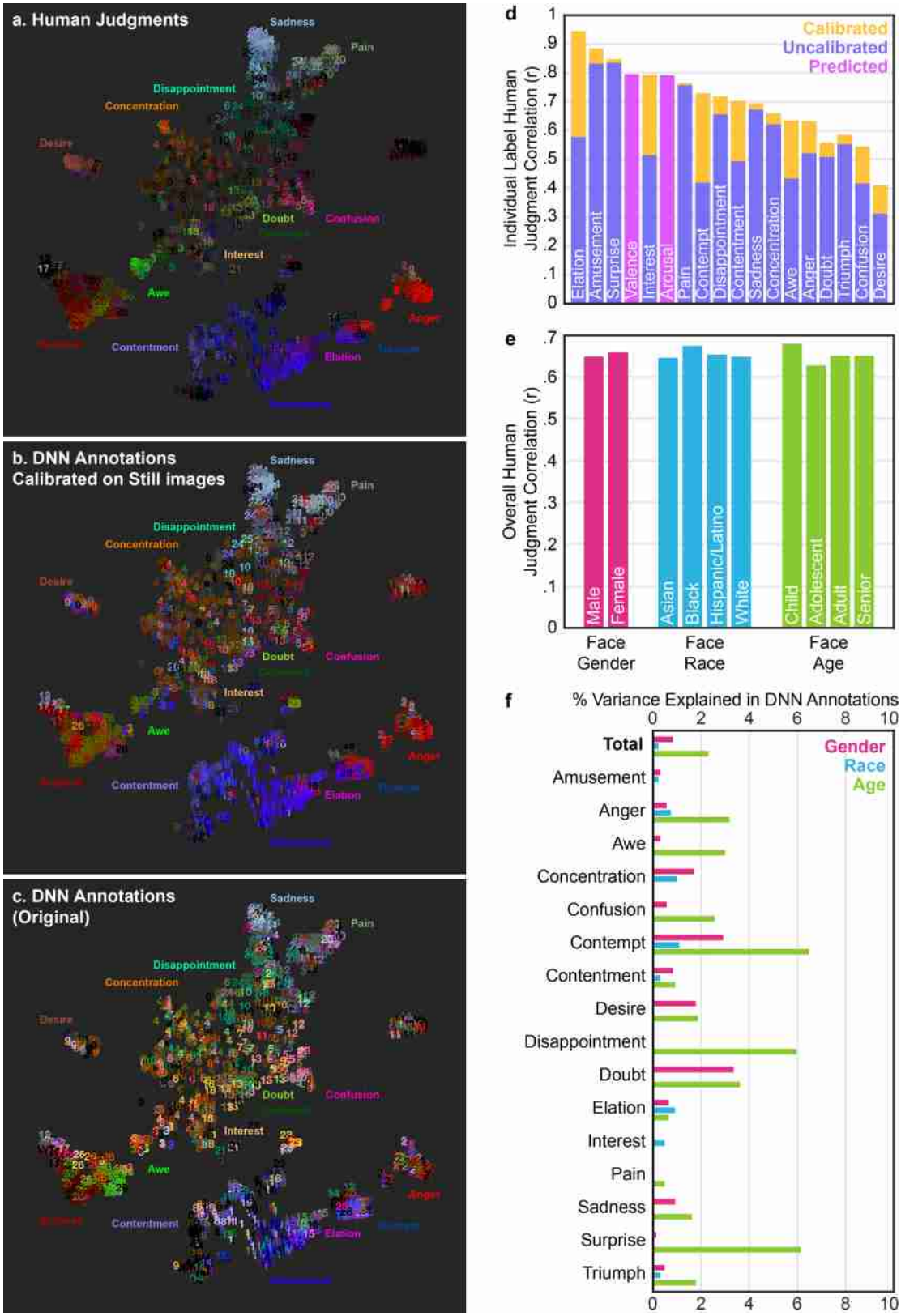
52. Kowsari, K. et al. Text classification algorithms: a survey. *Information (Switzerland)* **10**, (2019).
53. Lee, J., Natsev, A. P., Reade, W., Sukthankar, R. & Toderici, G. in *Lecture Notes in Computer Science* 193–205 (Springer, 2019).
54. Yuksel, S. E., Wilson, J. N. & Gader, P. D. Twenty years of mixture of experts. *IEEE Trans. Neural Netw. Learn. Syst.* **23**, 1177–1193 (2012).
55. Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507–521 (1915).
56. Efron, B., Rogosa, D. & Tibshirani, R. in *International Encyclopedia of the Social & Behavioral Sciences* 2nd edn (eds Smelser, N. J. & Baltes, P. B.) 492–495 (Elsevier, 2015).
57. Somandepalli, K. & Cowen, A. S. A simple Python wrapper to generate embedding spaces with interactive media using HTML and JS. https://doi.org/10.5281/zenodo.4048602 (2020).
58. Weyrauch, B., Heisele, B., Huang, J. & Blanz, V. Component-based face recognition with 3D morphable models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2004).
59. Bartlett, M. S. The general canonical correlation distribution. *Ann. Math. Stat.* **18**, 1–17 (1947).
60. John, O. P. & Soto, C. J. in *Handbook of Research Methods in Personality Psychology* (eds Robins, R. W. et al.) Ch. 27, 461 (Guilford, 2007).
61. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
62. Powers, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2**, 37–63 (2011).
63. Benjamini, Y. & Yu, B. The shuffle estimator for explainable variance in FMRI experiments. *Ann. Appl. Stat.* **7**, 2007–2033 (2013).
64. Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.* **11**, e1005206 (2015).
65. Mordvintsev, A., Tyka, M. & Olah, C. Inceptionism: going deeper into neural networks. *Google Research Blog* 1–8 (2015).
66. Brendel, W. & Bethge, M. Approximating CNNs with bag-of-local-features models works surprisingly well on Imagenet. In *Proc. 7th International Conference on Learning Representations* (ICLR, 2019).

Acknowledgements This work was supported by Google Research in the effort to advance emotion research using machine-learning methods. Some of the artistically rendered faces in Fig. 1 are based on photographs originally posted on Flickr by V. Agrawal (https://www.flickr.com/photos/13810514@N07/8790163106), S. Kargaltsev (https://commons.wikimedia.org/wiki/File:Mitt_Jons.jpg), J. Hitchcock (https://www.flickr.com/photos/91281489@N00/90434347/) and J. Smed (https://commons.wikimedia.org/wiki/File:Tobin_Heath_celebration_(42048910344).jpg). We acknowledge the Massachusetts Institute of Technology and to the Center for Biological and Computational Learning for providing the database of facial images used for the analysis shown in Extended Data Fig. 2.

Author contributions A.S.C. conceived and designed the experiment with input from all other authors. A.S.C. and G.P. collected the data. A.S.C., F.S., B.J., H.A. and G.P. contributed analytic tools. A.S.C. analysed the data. A.S.C., D.K. and G.P. wrote the paper with input from all other authors.

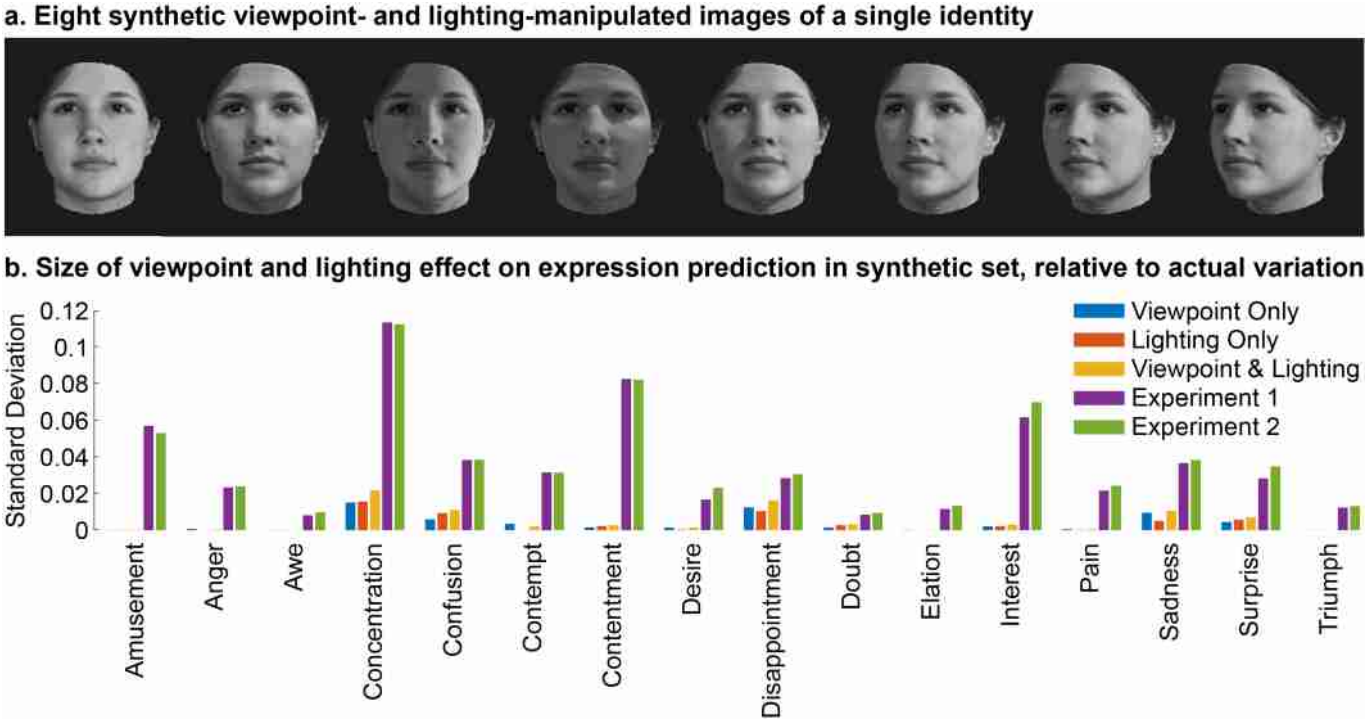
Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-3037-7.
Correspondence and requests for materials should be addressed to A.S.C.
Peer review information *Nature* thanks Jeffrey Cohn, Ursula Hess and Alexander Todorov their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

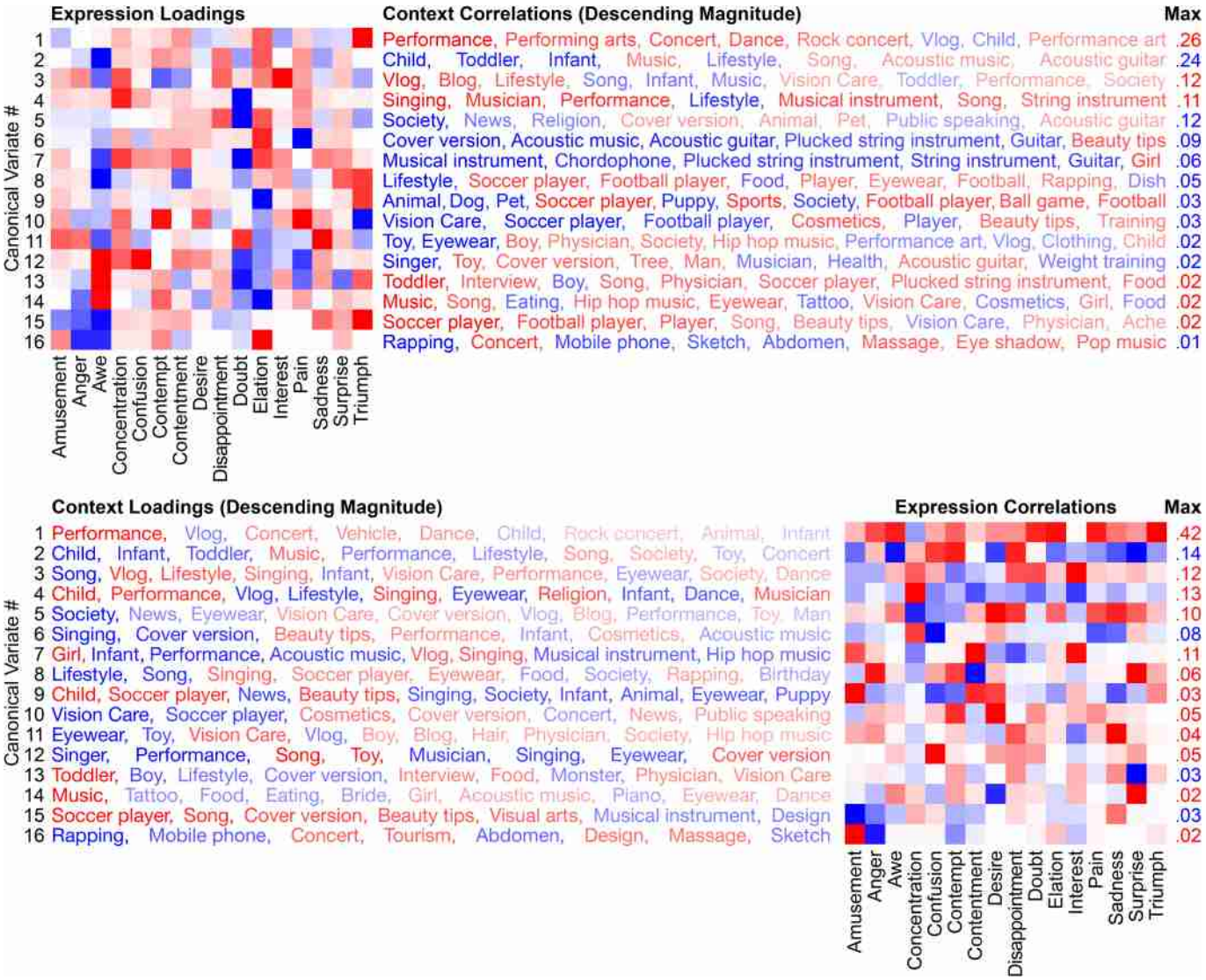


Extended Data Fig. 1 | See next page for caption.

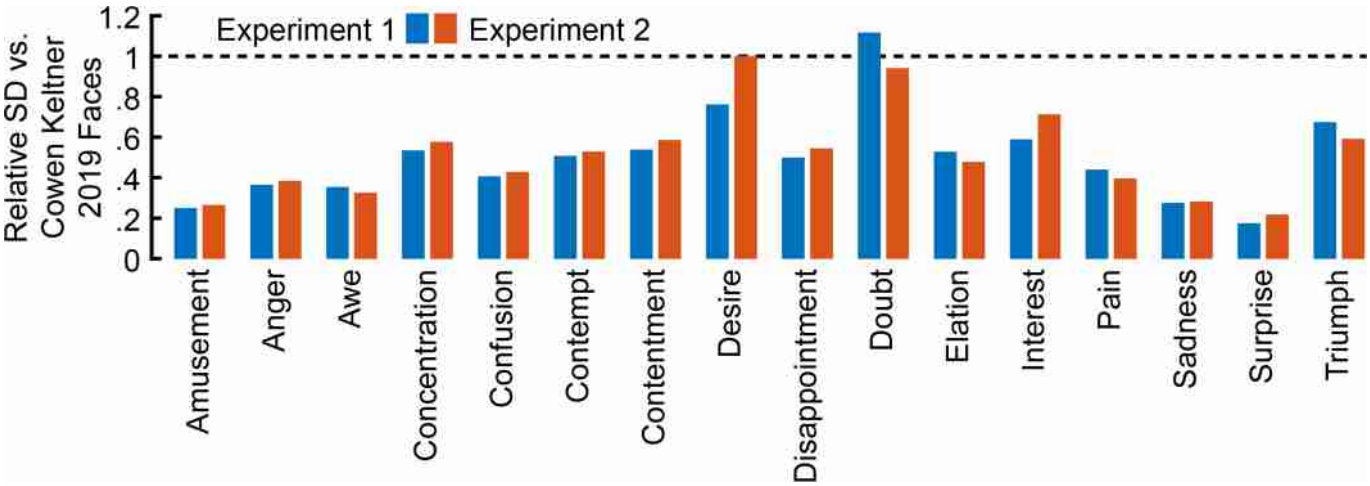
Extended Data Fig. 1 | The expression DNN predicts human judgments and is largely invariant to demographics. a–**c**, Accuracy of the expression DNN in emulating human judgments. Human judgments (**a**) and the annotations of the expression DNN (**b**, **c**) have been projected onto a map of 1,456 facial expressions adapted from a previously published study[14]. Human judgments and the annotations of the expression DNN are represented using colours, according to the colour scheme used previously[14]. Probably because the expression DNN was trained on dynamic faces, it can in some cases make systematic errors in predicting the judgments of static faces (**c**). For example, a number of static faces of surprise were more strongly annotated by the expression DNN as awe (**c**, bottom left), probably because dynamic faces that convey surprise are distinguished in part by dynamic movement. This problem is mitigated when the expression DNN is calibrated for still images (**b**). To calibrate the DNN, multiple linear regression is applied in a leave-one-out manner to predict human judgments of the still images from the DNN annotations. After calibration on still images, we can see that the annotations of the expression DNN are fairly accurate in emulating human judgments (overall $r = 0.69$ between calibrated annotations of the expression DNN and human judgments after adjusting for explainable variance in human judgments[33]). **d**, The expression DNN can emulate human judgments of individual emotions and valence and arousal with moderate to high accuracy. Individual expression DNN predictions are correlated with human judgments across the 1,456 faces. Valence and arousal judgments (also from the previously published study[14]) were predicted using multiple linear regression in a leave-one-out manner from the 16 facial expression DNN annotations. **e**, The expression DNN is reliable for different demographic groups. By correlating the predictions of the expression DNN (calibrated for static images) across subsets of the 1,456 faces from the previous study[14], we can see that the expression DNN is accurate for faces from different demographic groups (adjusted for explainable variance in human judgments). **f**, The expression DNN has little bias across demographic groups. To assess demographic bias, the annotations of each face of the expression DNN were predicted by averaging the annotations of the expression DNN across all other faces from the same demographic group. The variance in the annotations of the expression DNN explained by demographic group in this dataset was low, even though no effort was originally made to balance expressions in this dataset across demographic groups. Gender explained 0.88% of the total variance, race explained 0.28% and age explained 2.4%. Results for individual expressions were generally negligible, although age did explain more than 4% of the variance for three expressions—contempt, disappointment and surprise (maximum, 6.2% for surprise). Note that these numbers only provide a ceiling for the demographic bias, given that explained variance may also derive from systematic associations between expression and demographics in this naturalistic dataset—for example, because older people are less often pictured playing sports, they are less likely to be pictured with expressions that occur during sports. We can conclude that the expression DNN is largely unbiased by race and gender, with age possibly having at most a minor influence on certain annotations.



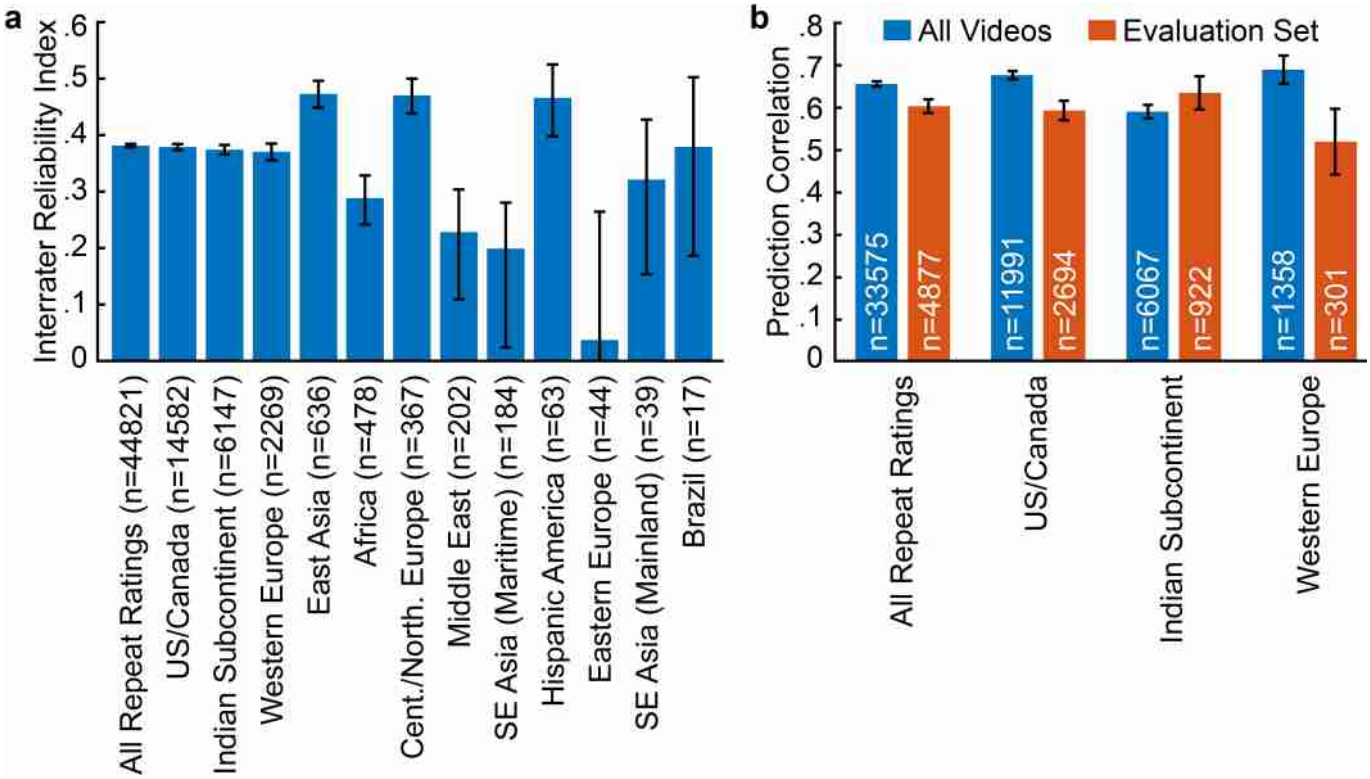
Extended Data Fig. 2 | Annotations of the expression DNN are largely invariant to viewpoint and lighting. a, To account for possible artefactual correlations owing to the effect of viewpoint and lighting on facial expression predictions, an in silico experiment was conducted. Predictions of facial expressions were applied to 3,240 synthetic images from the MIT-CBCL database, in which three-dimensional models of 10 neutral faces were rendered with 9 viewpoint conditions and 36 lighting conditions[58]. The variance explained in each facial expression annotation by viewpoint condition, lighting condition and their interaction was then computed. **b**, The explained standard deviation by viewpoint, lighting and their interaction is plotted alongside the actual standard deviation of each expression annotation in experiments 1 and 2. We note that the effects of viewpoint and lighting are small, except perhaps in the case of disappointment. This is unsurprising, given that faces are centred and normalized before the prediction of the expression.



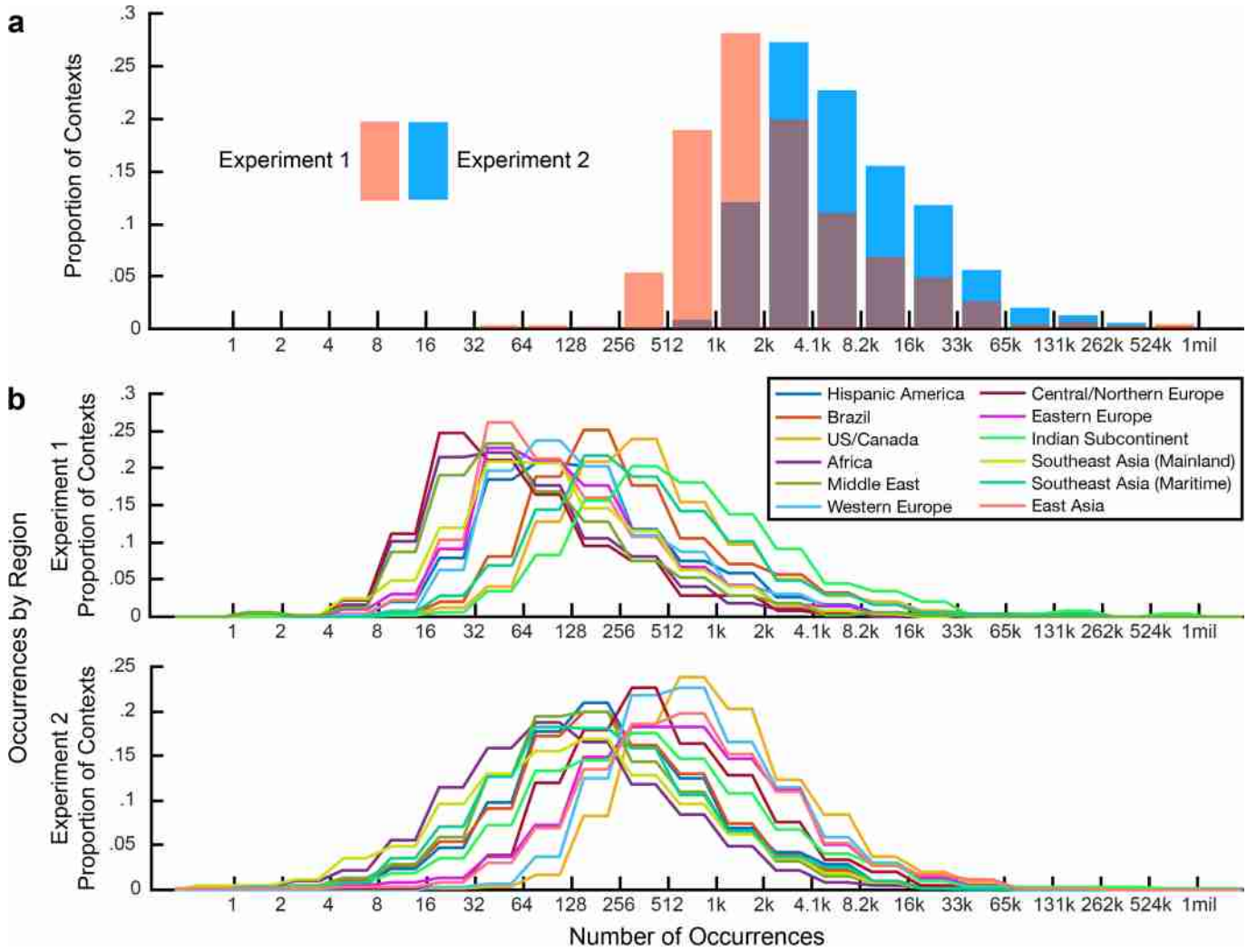
Extended Data Fig. 3 | Loadings of 16 global dimensions of context–expression associations revealed CCA. Given that all 16 possible canonical correlations were discovered to be preserved across regions (using cross-validation; Fig. 3c), we sought to interpret the 16 underlying canonical variates. To do so, we applied CCA between facial expression and context to a balanced sample of 300,000 videos from across all 12 regions (25,000 randomly selected videos per region). Left, loadings of each canonical variate on the 16 facial expression annotations and the maximally loading context annotations. Right, correlations of the resulting canonical variates with individual contexts and expressions, revealing how each variate captures context–expression associations. Unsurprisingly, a traditional parametric test for the significance of each canonical correlation revealed that all 16 dimensions were highly statistically significant ($\chi^2 = 860.4$ for 16 variates, $P < 10^{-8}$, Bartlett's χ^2 test[59]), a necessary precondition for their significant generalizability across all cultures (Fig. 3c).



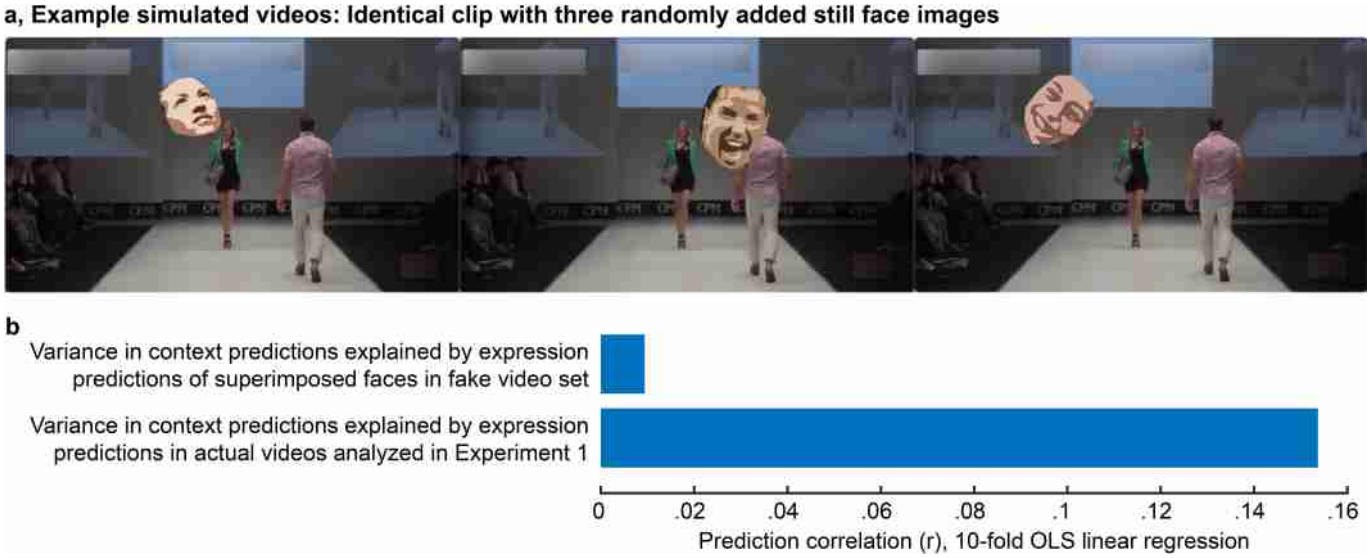
Extended Data Fig. 4 | Relative representation of each facial expression in the present study compared to the previous study. In Fig. 1, we provide an interface for exploring how 1,456 faces[14] are annotated by our facial expression DNN. Here, we analyse what the relative representation of these different kinds of facial expression within the present study was compared to within these 1,456 images. For each kind of facial expression, we plot the ratio of the standard deviation of our facial expression DNN annotations in the present study, averaged over each video, to the standard deviation of the annotations over the 1,456 faces. Given that the standard deviation in the present study was computed over averaged expressions within videos, it was expected to be smaller than the standard deviation over the 1,456 isolated expressions, generally yielding a ratio of less than 1. Nevertheless, it is still valid to compare the relative representation of different kinds of expression. We find that in both experiments within the present study, expressions labelled amusement, awe, sadness and surprise were particularly infrequent compared to those labelled concentration, desire, doubt, interest and triumph by the expression DNN. However, our findings still revealed culturally universal patterns of context–expression association for the less-frequent kinds of facial expression. Still, it is important to note that our measurements of the extent of universality may be differentially influenced by the expressions that occurred more often. Note that given limitations in the accuracy of our DNN, we were unable to examine 12 other kinds of facial expression that had been documented previously[14] (for example, disgust, fear), and are unable to address the extent to which they are universal.



Extended Data Fig. 5 | Neither interrater agreement levels nor prediction correlations in the data originally used to train and evaluate the expression DNN reveal evidence of regional bias. a, Interrater agreement levels by upload region among raters originally used to train the expression DNN. For 44,821 video clips used in the initial training or evaluation of the DNN, multiple ratings were obtained to compute interrater reliability. For a subset of 25,028 of these clips, we were able to ascertain the region of upload of the video. Here, across all clips and each of the 12 regions, we compare the interrater reliability index (the square root of the interrater Pearson correlation[41,60], the Pearson correlation being equivalent to the Matthew's correlation coefficient for binary judgments[61,62]), which reflects the correlation of a single rater with the population average[41,60]. We can see that the interrater reliability index converges on a similar value of around 0.38 in regions with a large number of clips (the USA/Canada, Indian subcontinent and western Europe). This is an acceptable level of agreement, given the wide array of options in the rating task (29 emotion labels, plus neutral and unsure). We do not see a significant difference in interrater reliability between ratings of videos from India (the country of origin and residence of the raters) and the two other regions from which a large diversity of clips were drawn (the USA/Canada and western Europe). Error bars represent the standard error; *n* denotes the number of clips. To compute interrater reliability, we selected two individual ratings of each video clip, subtracted the mean from each rating, and correlated the flattened matrices of ratings of the 16 emotion categories selected for the present study across all clips. We repeated this process across 100 iterations of bootstrap resampling to compute standard error. **b**, Human judgment prediction correlations by region. We applied the trained expression DNN to the video clips from each region. To compute unbiased prediction correlations, it is necessary to control for interrater agreement levels by dividing by the interrater reliability index, which is the maximum raw prediction correlation that can be obtained given the sampling error in individual human ratings[41,60,63,64]. Given that the interrater reliability index could be precisely estimated for three of the regions (**a**), we computed prediction correlations for those regions. We did so across all video clips used in the training or evaluation in the DNN (blue), which may be subject to overfitting, and for a subset of video clips used only in the evaluation of the DNN (red). In both cases, prediction correlations are similar across regions, exceed human levels of interrater agreement (**a**) and are consistent with prediction correlations derived from a separate set of images rated by US English speakers, which also showed no evidence of bias to ethnicity or race (Extended Data Fig. 1e). Error bars represent the standard error; *n* denotes the number of clips.



Extended Data Fig. 6 | Rates of the context occurrence. a, Proportion of contexts by number of occurrences (out of around 3 million videos) for experiments 1 and 2. The minimum number of occurrences of any given context was 39 for experiment 1 and 176 for experiment 2, but most contexts occurred much more often. Note that the number of occurrences is plotted on a logarithmic scale. **b**, Proportion of contexts by number of occurrences in each region. Certain contexts were rarer in particular regions, especially regions with fewer videos overall. Still, the vast majority of contexts occurred at least dozens of times in every region.



Extended Data Fig. 7 | Video-based context annotations are insensitive to facial expression. To account for possible artefactual correlations owing to the direct influence of facial expression on the video topic predictions (and vice versa), an in silico experiment was conducted. **a**, First, 34,945 simulated videos were created by placing 1,456 tightly cropped facial expressions[14] on each of 24 3-s clips from YouTube videos at random sizes and locations. Note that given the convolutional architecture of the DNN that we use, which largely overlooks strangeness in the configuration of objects within a video[65,66], randomly superimposed faces should have a similar effect on annotations to real faces. Examples shown here are artistically rendered. **b**, The video topic and the expression DNN were then applied to these videos. The variance in the video topic annotations explained by the expression predictions was then computed using ordinary least squares linear regression. The amount of variance in context predictions explained by randomly superimposed expressions (prediction correlation $r = 0.0094$) was negligible compared to the amount of variance explained in context predictions by the expression predictions in actual videos from experiment 1 ($r = 0.154$, tenfold ordinary least squares linear regression applied to a subset of 60,000 videos from experiment 1; 5,000 from each of the 12 regions). Therefore, any direct influence of facial expression on the video topic annotations had a negligible effect on the context–expression correlations that we uncovered in experiment 1.

Extended Data Table 1 | Manual inspection of videos annotated with selected contexts

Video & Text Context	Present in Video	Text-Only Context	Present in Video	Pertinent Phrase in Title/Description
Art	3/3	Art	2/3	3/3
Birthday	3/3	Birthday	0/3	3/3
Candy	2/3	Candy	2/3	3/3
Dance	3/3	Dance	3/3	3/3
Father	3/3	Father	2/3	3/3
Fireworks	2/3	Fireworks	2/3	3/3
Government	2/3	Government	3/3	3/3
Health	3/3	Health	2/3	2/3
Individual sport	3/3	Individual sports	3/3	3/3
Interview	2/3	Interview	1/3	3/3
Martial arts	3/3	Martial arts	2/3	3/3
Model	3/3	Fashion	2/3	2/3
Mother	2/3	Mother	1/3	3/3
Music	3/3	Music & audio	2/3	3/3
Parody	3/3	Parody	1/3	2/3
Pet	3/3	Pets	2/3	2/3
Police	1/3	Police	2/3	3/3
Practical joke	3/3	Humor	3/3	3/3
Protest	3/3	Protest	2/3	3/3
Rapping	3/3	Rapping	1/3	3/3
School	3/3	School	1/3	3/3
Shopping	3/3	Shopping	0/3	3/3
Soldier	3/3	Soldier	2/3	3/3
Team sport	3/3	Team sports	2/3	2/3
Toy	3/3	Toys	2/3	2/3
Wedding	3/3	Wedding	0/3	3/3
Weight training	3/3	Bodybuilding	3/3	3/3
Avg. specificity	**91.6%**	**Avg. specificity**	**60.7%**	**92.9%**

Three videos annotated with each of 27 contexts by the video-topic DNN and the text-topic DNN were hand inspected by A.S.C. on the basis of whether they contained natural images or videos consistent with that context. For the text-topic DNN, the videos were also inspected on the basis of whether a phrase relevant to the context was present in the user-generated title or description of the video.

Extended Data Table 2 | Countries included and languages spoken in the 12 geographical regions

Region	Countries/Territories from Which Videos Were Uploaded	# of Videos, Exp. 1	Spoken Languages Detected >10% Experiment 1	# of Videos, Exp. 2	Written Languages Detected >10% Experiment 2
Hispanic America	Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, Guadalupe, Guatemala, Honduras, Mexico, Nicaragua, Panama, Peru, Puerto Rico, El Salvador, Uruguay, and Venezuela	158,446	Spanish (88%)	192,671	Spanish (88%), English (11%)
Brazil	Brazil	286,202	Portuguese (90%)	181,854	Portuguese (89%)
US/Canada	United States and Canada	333,226	English (94%)	593,064	English (70%)
Africa	49 countries spanning the African continent	78,691	Arabic (49%) English (34%), French (14%)	77,086	English (48%), Arabic (45%)
Middle East	Iran, Iraq, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, Yemen	138,830	Arabic (44%), Turkish (33%), English (17%)	139,823	Turkish (65%), Arabic (18%), English (14%)
Western Europe	Andorra, Belgium, France, Great Britain, Ireland, Italy, Luxembourg, Monaco, Netherlands, Portugal, Spain	124,693	English (43%), French (17%), Spanish (16%), Italian (12%)	504,625	English (32%), Spanish (25%), French (19%), Italian (14%)
Central and Northern Europe	Austria, Croatia, Czech Republic, Denmark, Estonia, Finland, Germany, Hungary, Iceland, Latvia, Liechtenstein, Lithuania, Norway, Poland, Slovakia, Slovenia, Sweden, Switzerland	61,453	English (34%), German (20%), Polish (12%)	313,944	German (35%), English (31%), Polish (13%)
Eastern Europe	Armenia, Azerbaijan, Belarus, Georgia, Moldova, Russia, Ukraine	129,123	Russian (88%)	440,321	Russian (86%), English (10%)
Indian Subcontinent	Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, Sri Lanka	1166151	English (87%)	480,697	English (81%)
Mainland Southeast Asia	Cambodia, Laos, Myanmar, Thailand, Vietnam	110,856	Vietnamese (49%), Thai (35%)	114,694	Vietnamese (62%), Thai (20%), English (14%)
Maritime Southeast Asia	Brunei, Indonesia, Philippines, Singapore, Timor-Leste	363,160	Indonesian (57%), English (37%)	148,624	Indonesian (75%), English (19%)
East Asia	China, Hong Kong, Japan, Macau, Mongolia, North Korea, South Korea, Taiwan	69,456	Korean (39%), Japanese (28%), Mandarin (17%)	648,113	Korean (40%), Japanese (32%), Chinese (14%), English (12%)

Language was detected using a version of the methods provided in the Google Cloud Speech-to-Text API and Natural Language API.

Extended Data Table 3 | Manual inspection of videos from each region

Region	Experiment 1	Experiment 2
Hispanic America	Natural faces: 24/25 Native origin: 25/25	Natural faces: 21/25 Native origin: 22/25 Native comments: 25/25
Brazil	Natural faces: 25/25 Native origin: 24/25	Natural faces: 20/25 Native origin: 24/25 Native comments: 24/25
US/Canada	Natural faces: 25/25 Native origin: 24/25	Natural faces: 19/25 Native origin: 21/25 Native comments: 20/25
Africa	Natural faces: 24/25 Native origin: 23/25	Natural faces: 23/25 Native origin: 19/25 Native comments: 22/25
Middle East	Natural faces: 23/25 Native origin: 24/25	Natural faces: 21/25 Native origin: 23/25 Native comments: 25/25
Western Europe	Natural faces: 25/25 Native origin: 23/25	Natural faces: 24/25 Native origin: 23/25 Native comments: 23/25
Central and Northern Europe	Natural faces: 25/25 Native origin: 24/25	Natural faces: 19/25 Native origin: 20/25 Native comments: 21/25
Eastern Europe	Natural faces: 24/25 Native origin: 24/25	Natural faces: 20/25 Native origin: 19/25 Native comments: 23/25
Indian Subcontinent	Natural faces: 24/25 Native origin: 24/25	Natural faces: 24/25 Native origin: 24/25 Native comments: 25/25
Mainland Southeast Asia	Natural faces: 25/25 Native origin: 25/25	Natural faces: 18/25 Native origin: 22/25 Native comments: 25/25
Maritime Southeast Asia	Natural faces: 24/25 Native origin: 25/25	Natural faces: 21/25 Native origin: 23/25 Native comments: 25/25
East Asia	Natural faces: 25/25 Native origin: 25/25	Natural faces: 22/25 Native origin: 21/25 Native comments: 24/25
Overall	Natural faces: 97.7% Native origin: 96.7%	Natural faces: 84.0% Native origin: 87.0% Native comments: 94.0%

Videos from each region were manually inspected (by A.S.C. and G.P.) for whether they could be verified as containing natural images or videos of faces (out of a random sample of 25 videos), if so whether they primarily included people who appeared native to the limited extent to which geographical origin can be gauged based on spoken languages, accent, dress and physical features (out of 25 videos containing faces) and, for experiment 2, whether commenters on the video also appeared to be native based on language or username (out of 25 that had comments; comments were rare on videos used in experiment 1). Note that for the assessment of native origin, A.S.C. inspected 20 videos per region per experiment and G.P. inspected 5 videos per region per experiment, with similar results: 96.3% versus 98.3% of videos in experiment 1; 85.8% versus 91.7% in experiment 2, respectively.)

Extended Data Table 4 | Measures taken to address possible confounding factors

Possible Confound	Measures Taken
The video topic annotations may be influenced by facial expression, resulting in artifactual correlations between the video topic annotations and facial expression annotations in Experiment 1.	1. We created an artificial dataset in which 1456 tightly cropped faces were randomly resized and pasted into a set of 24 3s YouTube clips. We then applied the video topic DNN to these artificial videos. The variance explained in context predictions by expression predictions in this fake dataset was negligible compared to the variance explained in our real dataset (Extended Data Fig. 8), indicating that our results were unlikely to be driven by artifactual correlations. 2. In Experiment 2 we derive context annotations using only user-generated video titles and descriptions, which are not subject to this potential confound.
The expression annotations may be influenced by surrounding context, resulting in artifactual correlations between the expression annotations and the context annotations.	The expression DNN only receives pixels only from the face.
The expression DNN annotations may be influenced by demographics, lighting and viewpoint, which may result in correlations with contexts that tend to involve specific demographics or kinds of lighting and viewpoint.	The outputs of the DNN are largely invariant to facial demographics (Extended Data Fig. 1). Faces are frontalized and luminance normalized prior to expression prediction. We also utilized a database of 3D faces that have identical expressions but are digitally manipulated to produce a range of distinct lighting and viewpoint conditions. Lighting and viewpoint had a negligible impact on extracted facial expression annotations (Extended Data Fig. 2).
Content may be reuploaded in different countries, driving artifactual similarities across countries in contextual correlates of expression.	1. We took measures to ensure that as many as possible of the videos used in each experiment originated in the countries from which they were uploaded, including (a) verifying that the coordinates of upload were based within the country (Experiment 1), (b) verifying that the languages spoken in the videos were languages native to each country (Experiment 1, Extended Data Table 1), and (c) verifying that the titles and metadata were written in a language native to each country (Experiment 2, Extended Data Table 1). 2. We reviewed 25 videos per experiment from each country to verify that the people in the videos appeared to be native of the countries from which the videos were uploaded (Extended Data Table 2).

nature research

Corresponding author(s): Alan S. Cowen

Last updated by author(s): Oct 2, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Custom and proprietary code was used to search for videos and process the video content and metadata using machine learning algorithms. These data collection and processing steps can partially be replicated using the Google Cloud Video Intelligence and Natural Language APIs, as referenced in the manuscript. Anonymized (differentially private) versions of the context-expression correlations in each country will be made available via Github under the repository github.com/alanscowen/contextexpression.
Data analysis	Analysis of the processed data was performed using custom code in Matlab version R2018B. Code to read and visualize the anonymized context-expression correlations in each country will be made available via Github under the repository github.com/alanscowen/contextexpression.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data presented in the manuscript was publicly available at the time of data analysis and largely remains publicly available, although the data is owned by the original YouTube contributors who are free to remove their videos from the Internet any time. However, we are unable to release identifiers of the specific videos we analyzed. Anonymized (differentially private) versions of the context-expression correlations in each country for each experiment are available in github.com/alanscowen/contextexpression. The MIT CBCL Database (used in Extended Data Figure 2) is available upon request at http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A quantitative observational study correlating facial expressions with other video content in YouTube videos.
Research sample	A total of approximately 6 million videos uploaded to YouTube. Experiment 1: We sought to focus on natural footage for which reliable geographic information was available. To find naturalistic footage, we searched for publicly available YouTube videos that were uploaded from mobile phones. The search was restricted to YouTube videos tagged with a latitude and longitude of upload that matched the country in which the uploader was registered. Furthermore, to focus on naturalistic footage, we filtered out videos predicted by the video topic annotations to include video games and other animated content. This yielded a total of 3,029,812 videos. Experiment 2: The videos from Experiment 1, like many videos on YouTube, typically lacked detailed descriptions, making them poor candidates for annotation by the text topic DNN. Thus, we collected a new set of videos for Experiment 2. To ensure that we would have the power to investigate correlations between contexts and facial expressions, we sought to include publicly available videos that had titles and descriptions pertaining either to the contexts we explored in Experiment 1 or to emotions. To do so, we first searched for videos with a wide range of context- and emotion-related substrings within their English-translated titles and descriptions (Dataset S3; note that to the extent that translations were inaccurate, representation of corresponding contexts could be reduced in non-English-speaking cultures, exacerbating cultural differences). We then retrieved the full native-language titles and descriptions for those videos and computed text topic annotations. Finally, to avoid synthetic faces, we filtered out videos predicted by the text topic DNN to include video games and animated content. This yielded a total of 3,056,861 videos.
Sampling strategy	N/A -- Full population of videos meeting the criteria specified above were included in the study.
Data collection	Processing of YouTube videos was performed on temporary cloud computing system instances without permanently downloading any video data or metadata.
Timing	Videos included in Experiment 1 were uploaded between July 14, 2009 and May 3, 2018. Videos included in Experiment 2 were uploaded between December 27, 2005 and April 15, 2019. Facial expression annotations were generated between May 3, 2018 and May 1, 2019. Context annotations were generated between the time of upload of each video and May 1, 2019. All statistical analyses were performed between May 3, 2018 and May 1, 2019.
Data exclusions	All data meeting the criteria specified above were included in the study.
Non-participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	N/A
Ethics oversight	The use of the video data in aggregate form underwent review for alignment with Google's AI Principles (see https://ai.google/principles/) and conformed to Google's privacy policy (see https://policies.google.com/privacy).

Note that full information on the approval of the study protocol must also be provided in the manuscript.